Creating More Personas Improves Representation of Demographically Diverse Populations: Implications Towards Interactive Persona Systems

Author(s): Salminen, Joni; Jung, Soon-Gyo; Nielsen, Lene; Jansen, Bernard

Title: Creating More Personas Improves Representation of Demographically Diverse Populations: Implications Towards Interactive Persona Systems

Year: 2022

Version: Accepted manuscript

Please cite the original version:

Salminen, J., Jung, S-G., Nielsen, L. & Jansen, B. (2022). Creating More Personas Improves Representation of Demographically Diverse Populations: Implications Towards Interactive Persona Systems. In: *Participative Computing for Sustainable Futures: Proceedings of the 12th Nordic Conference on Human-Computer Interaction (NordiCHI'22)*, 1-11. New York: Association for Computing Machinery. https://doi.org/10.1145/3546155.3546654

Creating More Personas Improves Representation of Demographically Diverse Populations: Implications Towards Interactive Persona Systems

JONI SALMINEN, University of Vaasa, Finland
SOON-GYO JUNG, Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
LENE NIELSEN, IT University of Copenhagen, Denmark
BERNARD J. JANSEN, Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar

Personas represent distinct user types. However, while online user data can be demographically and behaviorally heterogeneous, most studies generate less than ten personas, regardless of how heterogeneous the data is. Because all persona creation efforts need to assign a number of personas to create, assigning this number evokes a fundamental question, How many personas to create?. To address this question, we apply data-driven persona creation in a dataset with 250 million YouTube views from a global news and media organization. We focus on a statistically optimal number of personas, namely, how the distribution of demographic persona attributes deviates from the baseline user data. Altering the number of generated personas, ranging from 5 to 160 personas per set, we find that more personas cover more age groups and countries, thus improving the statistical correspondence with the raw user data, and increasing the representation of demographic diversity by including more fringe user segments. While the user representation continuously improved with more personas, the relative diversity gain was maximal with 40 personas, implying that, using our data, one ought to create more than 4 times more personas than generally advocated. The results imply that organizations with heterogeneous online audiences benefit from many personas in terms of more inclusive user representation. We further demonstrate how an interactive persona system can help stakeholders navigate many personas with possibly smaller cognitive effort.

Additional Key Words and Phrases: Data-driven personas, user segmentation, number of personas

1 INTRODUCTION

A persona (see Figure 1 for an example) is a *fictitious person representing an underlying customer or user group*, often the core users of a product or system, although personas can also represent the potential or desired users of a system [9], website audience segments [46], or social media followers [33]. Personas are used for enhancing stakeholders' perceived empathy towards users [24]. In contrast, a *user segment* is defined as a non-personified representation of users. Whereas personas typically include personified information, such as a name and picture [25], such details are not conventionally included when presenting user segments.

Personas are an important method for user understanding in human-computer interaction (HCI) [9]. Yet, nobody seems to know how many to exactly create. The current body of knowledge proposes no definitive and empirically strong guidelines based on scientific evidence. Most typically, a relatively small number of personas is created, citing presumed concerns over the manageability and cognitive load for users to handle many personas [24]. However, as data-driven personas (DDPs) and interactive persona systems become more common [1, 15], it is worthwhile to question this *small personas hypothesis* by asking if more personas should be created than conventionally is done, and if interactive systems can provide features that afford persona users to overcome the manageability issue of 'too many personas' [24].

Indeed, a fundamental question in designing personas (and more broadly any form of user segmentation) is: *How many personas (or segments) should one create?* This question is particularly important for data-driven persona development [21] where the number of personas is a hyperparameter that persona creators can easily change [14] – for example, creating 100 personas relative to 5 personas has little to no additional cost. DDPs are often created from social media and/or web

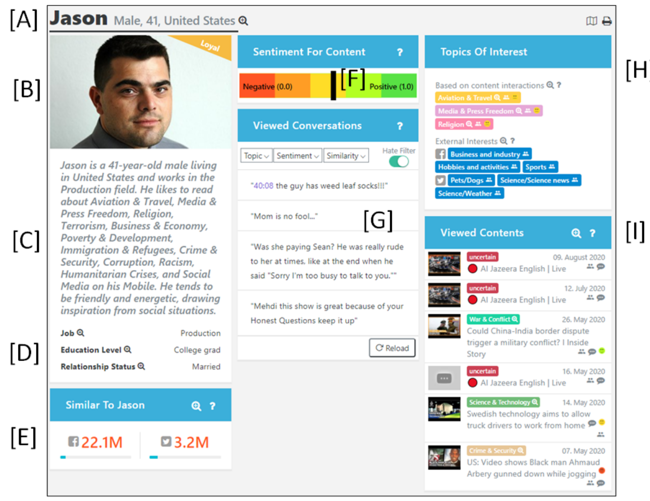


Fig. 1. Example of a DDP, with [A] name, age, gender, and country; [B] picture, [C] text description, [D] sociographics, [E] audience size, [F] sentiment, [G] quotes, [H] topics of interest, and [I] the most viewed content.

analytics user data, referred to as 'personified big data' [40], which is done using data science algorithms and completed by leveraging web interaction techniques to provide the personas to their destined users [1, 2, 5, 15, 16, 22, 34, 39, 46].

Traditionally, choosing the number of personas is based on either exploratory data analysis, heuristic rules, or a combination of the two. Research suggests that the number of personas created tends to be small, within the range of 3-10 [24], which is in line with other forms of user segments [10]. Despite the small number of personas being more or less the norm, it is not self-evident that creating only a handful of personas would result in better design outcomes and more user-centric decisions than creating more personas. In fact, instead of arguing for better design outcomes for end-users, the decision for a small number of personas is often made at least partially due to *presumed* idea that stakeholders (i.e., designers, marketers, managers, public health professionals, analysts, and others engaged with personas or user segments in their daily jobs) are unable to cope with many segments. Despite this widespread belief, we could locate no study that empirically shows it as true. In fact, the *small personas hypothesis* appears to be a truism, one of many in the persona field [30], with the adverse side-effect of discouraging explorations with a higher number of personas and interaction techniques that would support using a larger number of personas for user-centered design.

This question matters because, even when the data might allow for increasing the number of personas, the existence of the *small personas hypothesis*—defined as a tendency of scholars suggesting that creating a small number of personas is more desirable than creating a high number of personas—may pose a mental blockade for researchers to pursue the development of innovative user interface (UI) and user experience (UX) techniques for presenting more personas, such as providing persona users with more flexibility to browse the available personas. In other words, the focus of research might miss a relevant topic (i.e., 'How to develop techniques that support serving more personas to stakeholders?') while accepting the default premise that a handful of personas would be adequate.

Moreover, the norm of generating only a handful of personas may unnecessarily compress large and heterogeneous online audiences into a handful of personas, a result that can enhance stereotyping and biased user-centered design instead of taking personas away from these problems [44].

In this research, we address the question of 'optimal number of personas' by statistical analysis, leaving the user experience aspects for future work. We define this *numeric optimality* as an optimal number of user segments (e.g., personas) that best approximates the underlying data about users. (See the methodology section for further details.) We specifically focus on the following research questions (RQs):

- RQ1: Does the demographic diversity of personas increase with the number of personas?
- RQ2: Is there an effect of decreasing returns, where the marginal gain of demographic diversity stales?
- RQ3: How can web interaction techniques support a flexible number of personas?

The motivation of RQ2 is to investigate the nature of how the demographic diversity of the personas changes along with their number. In turn, addressing RQ3 with design concepts demonstrates how generating more personas can make "missing" user segments more prominent and visible for stakeholders in these organizations, and how interactive system features can support stakeholders' process of discovering specific user segments [31], even when these segments exceed a handful.

The results have implications to persona design, specifically, how many personas to generate from a given dataset. They also contribute to the broader field of user segmentation, as segmentation faces the same challenge of determining segment size. Consequently, investigating the question of optimal persona (segment) number is a worthwhile research objective when using algorithms for persona generation or user segmentation in general.

2 RELATED STUDIES

2.1 How Many Personas Have Been Created?

To investigate the number of personas in previously created persona sets, we analyzed 51 research articles from the ACM Digital Library that reported from specific cases of persona creation. These articles were identified among full-text articles that mentioned "personas" in their abstract. Looking at the number of personas in the sets (see Figure 2) it becomes clear that most have 3 personas, but also 4 in the set seems to be quite typical. Most commonly, the articles developed three personas (n = 15, 29.4%). Altogether, close to three out of four articles (n = 37, 72.5%) developed between 1 to 5 personas. How it became three and whether three is better than 10 or more is not clear from the literature. It is also common that the specific numbers are mentioned without any reference to research or reasoning as to why this specific number is used for persona creation. This indicates that persona research predominantly applies a small number of personas, without necessarily justifying this choice in any clear terms. For example, a study mentions creating 5-15 personas [2]—despite there being many more observed patterns in the data—simply because a relatively small number is suggested in the literature. This example underlines how the heuristic rules carry over from one study to another, without anyone questioning them in the process.

We also investigated recommendations given in prominent HCI textbooks and persona research articles. Most of these recommendations for the number of personas to create are below 10 personas [9, 23, 24, 27] because this is a "manageable" number [27]. However, there is a scarcity of empirical work addressing this question using quantitative data and data-driven (algorithmic) methods. The relationship between the source data and the varying number of personas generated from it is not well known, and the field of HCI lacks research efforts in this regard. The next section reviews the scarce work that we could locate.

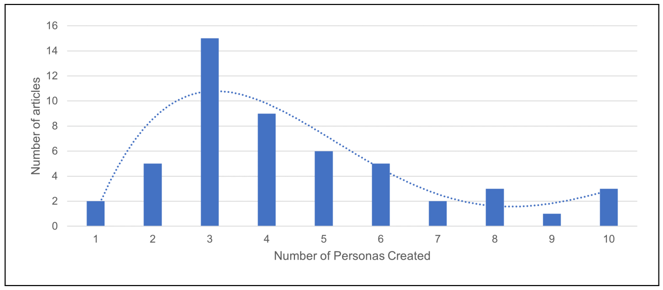


Fig. 2. Number of personas in research articles (N = 51). Researchers tend to generate a small number of personas. None of the articles created more than ten personas.

2.2 Empirical Work on Optimal Number of Personas

In one of the rare empirical studies, Chapman et al. [7] evaluate personas as quantitative information. They find that more granular personas represent smaller audiences, thus associating the question of numeric optimality with the number of attributes assigned to the user segment. Nevertheless, no detailed analysis was done on the optimal number of personas. Since computational methods have been adopted in the persona creation by some [1, 2, 21, 46], algorithmically generated user personas address some of the challenges of dealing with online user data by (a) being based on large volume of quantitative user data, (b) describing a wide range of behaviors and demographics across the digital user population, and (c) affording updatability via the use of application programming interfaces (APIs) and automatic data analysis pipelines. It is also generally established that the number of DDPs is more malleable than those created using manual methods, as this number can be arbitrarily changed as a hyperparameter of an algorithm [14].

Using algorithmic persona creation, Jansen and colleagues [14] focused on reducing the number of DDPs to a "manageable" number that efficiently represents the heterogeneous user population, while maintaining the "meaning" (i.e., demographic similarity) of the personas. They applied an algorithmic cost function to collapse the set to the minimum needed to represent the whole population, collapsing 1593 personas to 493, and thus decreasing the number of DDPs by 69%. Algorithmic cost function is a procedure of assessing the performance of an algorithm or a model. It is based on calculating the difference between the generated outputs (e.g., personas) and the actual data (e.g., the raw user data). The difference will be high if the outputs are far from the actual data, and vice versa. Results of Jansen and colleagues [14] suggest that using algorithms can result in not a handful but hundreds of DDPs for organization with large worldwide user populations.

Canossa and Drachen [6] use subsets to expand the number of player personas from a smaller number of "main" personas, with fewer primary personas that describe more significant user segments or are otherwise more impactful for a given design task. In contrast, the secondary set contains personas describing smaller user segments. Looking at the industrial use of personas, Nielsen and Hansen [25] found that the number of personas varies between 3-12, with most companies having sets between 3-6, but two persona sets had 10 and 12 personas.

2.3 The Optimal Number of User Segments

There is similar lack of work addressing the optimal number of user segments. In their literature review, Fu et al. [10] found that most user segments (clusters) are in the range of 3-8, with an average of 5.5 segments. The highest number found by the researchers was 8 segments (4 studies), and the smallest number was 3 segments (4 studies) [10].

Overall, the optimal number of segments or clusters is considered as an open research question in computer science, with several metrics and approaches suggested [35, 36, 45]. Computer scientists generally tend to favor data-driven determination of segments, where a threshold value is set either by visualizing the data—the so-called elbow method [41]—or by examining outcome metrics, such as residual error [3] or variance [45]. In contrast, social media analysts are often more concerned with presenting a number of segments that stakeholders can cope with and that can be feasibly presented to them using media such as presentation slides, PDFs, or paper prints. Nonetheless, like in computer science, the number of optimal segments has been debated over the years also in marketing [4, 19, 42, 43], with few definitive answers. Researchers seem to coalesce on a small number of segments, though, with the—often implicit—assumption that stakeholders cannot cope with a large number.

Overall, the expansion of digital end-user data, including social media and online analytics user data about demographics, behaviors, engagement, sentiment and other variables of interest, has led to exciting opportunities for user segmentation to understand users and social media audiences in digital environments, applications, and systems [40]. However, extracting value from digital user data remains challenging for people involved in analyzing user behavior and making user-centric decisions. Therefore, despite the availability of online analytics data—and perhaps partly because of it—converting this data into practically useful insights remains challenging [13]. This has inspired researchers to suggest automatic tools for user analytics and user behavior pattern detection [37, 47]. Obtaining full value from online user data would seem to direct us away from the small personas hypothesis, simply due to the fact that online user populations are heterogeneous. For example, a YouTube channel can have viewers from more than 100 countries, old and young, male and female, interested in sports while others in gaming. So, it is unlikely that such heterogeneous populations could be effectively represented by a handful of personas.

3 RESEARCH GAP AND HYPOTHESES

The "optimal" number of segments is one of the fundamental questions in persona design and other fields depending on user segmentation, such as Web analytics, health informatics, social media marketing, and so on. This is because analysts working with user data need to determine the number of segments they create in one way or another. This number can either "emerge" from the data (induction) or be set *a priori* (deduction). Previous research has largely overlooked the aspect of how well user segments representations, such as personas, represent real data distributions. The criticism of persona creation being based on heuristics rather than empirical evidence goes back to at least 2006 [8]. Still, the rationale for assigning the number is typically based on the assumption that users lack the capacity to process many personas. However, we could not locate any study that empirically substantiates this assumption.

Previous studies imply, although not directly show, suggest that the number of personas results in more diverse user representation. Therefore, we expect that the more personas are created, the larger the representation of a range of demographics and behaviors in the user populations (e.g., website users, social media followers, online news audiences). We formulate an explicit hypothesis to test demographic traits as variants of this hypothesis: **H1: More personas results in a better representation of users' (a) gender, (b) age, and (c) country.** Contrary to the main-stream belief of creating a handful personas, it is reasonable to expect that more personas may improve the *demographic diversity* of user representation [11]—i.e., the inclusion of more demographic groups in the persona set—although this remains to be empirically shown, hence the need for the current hypothesis.

Often, segmentation follows the law of decreasing returns (i.e., marginal benefits become smaller over time) and algorithms tend to convergence to some specific values, which human data scientists may obtain via visualization and heuristic rules. It is important to test whether these conventions also hold in the context of data-driven personas, which is why we propose the following hypothesis: **H2: There is an effect of decreasing returns, where the marginal gain of personas' demographic diversity decreases after a peak.**

4 METHODOLOGY

We present an empirical analysis of what happens when increasing the number of algorithmically generated personas, particularly in relation to baseline data about social media users. We increase the number of DDPs using a geometrical series with a multiplier of two—that is, 5, 10, 20, 40, 80, and 160. Geometrical series is chosen because this enables us to generate personas from a quite broad overall range with reasonable interval values. We start from five personas, as this falls in the range of often-used numbers for personas in the literature, and provides already some opportunity for demographic variation to emerge (with two personas, for example, the opportunity of getting different demographic values is very small because there simply is too few values to include). The results are then analyzed to address how well the DDPs in each set correspond with the baseline data's actual demographic properties.

4.1 Data Collection

Our dataset consists of 246,082,804 engagements (Views) on YouTube Channel of a large, international news and media organization. The view counts and demographic groups form the bases of DDPs, for which we apply a previously validated method called Automatic Persona Generation (APG) [1, 2] [1]. The view counts originate from 11, 077 videos published between January 2016 and February 2019 on the organization's YouTube Channel. The users of these channels locate in 172 countries and regions and the age range varies from 13 to 65+, with 25-34 being the most common age category – see Figure 3 that shows the distribution of demographic groups in the baseline data. As such, this dataset exemplifies a large and heterogeneous user base in an online setting.

Using this source data, APG generated sets of [5, 10, 20, 40, 80, 160] personas. While the APG methodology is validated and explained more thoroughly in separate research articles [1, 2], the following section gives a brief overview for the reader's convenience.

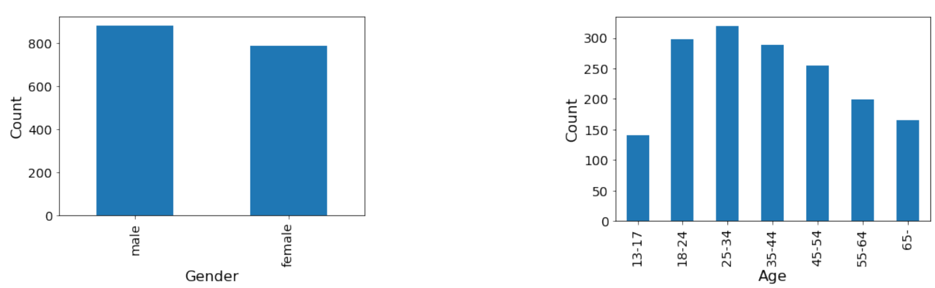


Fig. 3. Gender and age distributions of the baseline data from YouTube. Count indicates the number of demographic groups.

4.2 Persona Generation

4.2.1 Data Collection from Online Platforms. Online analytics platforms (e.g., Google Analytics, YouTube Analytics, Facebook Insights) enable the collection of user data automatically via application programming interfaces (APIs) [2].

[1]https://persona.qcri.org
[2]Note that accessing the analytics data requires authorization from the owner of the analytics property.

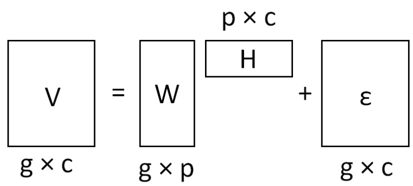


Fig. 4. NMF used in APG [2]. The interaction matrix V is decomposed to two matrices, W and H, by p latent patterns indicating the number of segments. ϵ indicates the reproduction error.

Typically, online platforms aggregate this data to protect the privacy of individual users. An example of an aggregated user segment is Female, 44-55, Denmark, containing no personally identifiable information. The segments given by the online analytics platforms tend to contain information on the gender, age, and country of the users, typically collected from the users upon registration. Various interaction metrics can be retrieved for each group (e.g., clicking and viewing content). For example, the user segment [Female, 45-54, Denmark] can have [2, 333] views for [Video X]. Using the APIs of online analytics platforms, APG collects this aggregated data for products and engagement metrics. For example, from YouTube Analytics, APG collects videos and their view counts, whereas, from Google Analytics, APG collects pages and the number of sessions.

4.2.2 Algorithmic Processing of the Data. After collecting the data from an online analytics platform (with the channel owner's permission), APG uses an automated computational approach to transform data into an *interaction matrix* that captures the engagement between users and digital content (e.g., online videos, webpages, etc.) [1, 2]. This matrix, $\mathbf{V}$, is a $\mathbf{g} \times \mathbf{c}$ matrix of $\mathbf{g}$ user segments and $\mathbf{c}$ digital contents. The elements of $\mathbf{V}$, V_{ij}, is any count metric that reflects the engagement of user segment G_i with digital content C_j. For example, in YouTube Analytics, V_{ij} is a view count for a particular video, C_j from the user segment G_i. The user segment is characterized by gender, age, and country (e.g., Male, $35 - 44$, Finland). Using $\mathbf{V}$ as the basis, non-negative matrix factorization (NMF) [17] is applied to detect $\mathbf{p}$ latent patterns that represent the user segment's digital content preferences (see Figure 4). APG then chooses a representative demographic group for each underlying pattern and enriches this demographic group with personified information (picture, name, topics of interest, quotes, etc.) to generate a complete persona profile (an example provided in Figure 1).

In Figure 1, the *Persona Profile* section [A] contains basic information of the DDP, including name, gender, age, and country. These correspond to the content of a typical persona profile [25] The *About persona* section [B] includes a text description of the DDP, generated using a dynamic template. The *Topics of interest* section [C] describes the DDP's interests based on the classification of the content the DDP has interacted with. The *Most Viewed Videos* section [E] lists the content the DDP has viewed, while the *Quotes* section [D] includes comments pulled from this content. Finally, the *Potential reach* section [F] shows the number of people similar to the DDP, calculated by querying the Facebook Marketing API[3] with the location, gender, age, and interests corresponding to the DDP's characteristics.

4.3 Statistical Distance Metrics

In information science, statistics, and probability theory, statistical distance deals with quantifying the similarity or difference between two variables or probability distributions. A probability distribution expresses a series of probabilities for a set of possible outcomes. In this study, we equate the proportions of demographic attribute values to a probability

[3]https://developers.facebook.com/docs/marketing-apis/

distribution. For example, say there are 3 personas, of which 2 are Indian and 1 is American; then the probability distribution would be [2/3, 1/3] for the persona set's nationalities to be Indian and American, respectively. If the baseline dataset from which the personas were created contained 66% Indian users and 34% American users, we would say that the statistical distance between these two objects – the persona distribution and the baseline dataset – would be very close. The farther the ratios of persona nationalities go from the "ideal" ratio indicated by the baseline data, the higher the statistical distance is.

Following this intuition, we calculate the distribution for each persona attribute for the baseline data and for the generated DDPs sets. The age groups, genders, and countries used as persona attributes originate from YouTube Analytics that uses binary gender categorization and seven age groups (see Figure 3). Then, we compare these distributions using three statistical distance metrics: Kullback-Leibler Divergence (KL) [12], Hellinger Distance (HD) [38], and Kolmogorov-Smirnov statistic (KS) [18]. The lower the values these metrics give, the closer the attributes in the set of DDPs are to the baseline data. The statistical metrics deployed quantify the similarity of the generated personas and the raw data. Each metric uses a different formula for that, but they all compare proportions of persona features against the corresponding proportions of features in the baseline data. In basic terms, the metrics indicate a difference between two probability distributions, a and b, which here are *personas' demographic attributes* and *demographic attributes in the raw data*. The supplementary material provides mathematical formulation of these metrics for interested readers.

We chose these three metrics as they are among the standard metrics for measuring statistical distance of two distributions and, thus, correspond to our research problem. In addition, three metrics (as opposed to using only one) help establish consistency of the results, avoiding the results being dependent on the mathematical nature of any one metric. The chosen three metrics enable the comparison of data distributions between the baseline data and the generated DDPs, including the share of males and females, age groups, and countries.

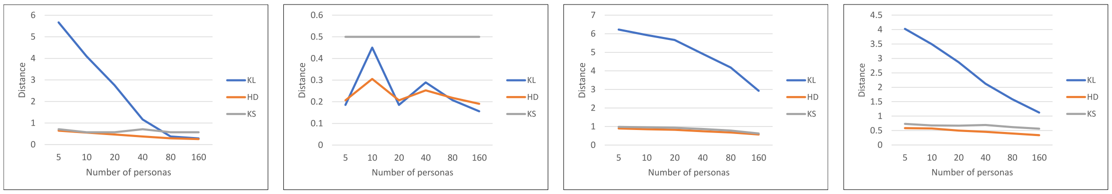
5 RESULTS

Figure 5 summarizes the results of statistical distance comparisons. Generally, KL shows that persona distance to baseline data drastically improves when increasing the number of personas, especially for age and country. The KS metric struggles with gender because of the binary values of this variable. Apart from that, KS and HD provide very similar results, except for age, in which KS produces a stable pattern rather than a clear decrease. Overall, the values indicate that accuracy of the personas increases with their number, which is further supported by Table 1.

From the distance metrics, we observe no effect of diminishing returns in the sense that the distance between DDPs and the baseline data distributions would stabilize. This suggests that even more than 160 DDPs are needed if one wants to achieve numeric optimality with this dataset. It is likely that at some point, the information gains would stabilize. However, our results suggest that this possibly requires hundreds of DDPs, and certainly more than a handful of user personas typically created. Therefore, **H1 is supported overall: More personas result in a better representation of user data.** More specifically, when the number of personas doubles, the three metrics' values decrease by 12.5% on average, indicating a better representation of the baseline data.

5.1 Gender

The *KL* and *HD* metrics give almost identical results for all categories (see Figure 5), whereas *KS* gives distinctly different results for age and gender. This is likely because the mathematical properties of *KS* are more appropriate for a higher number of potential values in a category. The relatively low values of the metrics for gender, especially in the 5 and 20 DDP sets (see Figure 5b) indicate that *numeric optimality for gender can be achieved even with a low number*



(a) Age (b) Gender (c) Country (d) Average

Fig. 5. Results of comparing the persona sets with the underlying data distributions. Closer to zero indicates better correspondence with the baseline data.

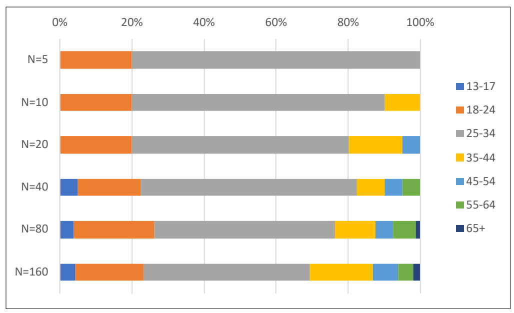


Fig. 6. Age diversification effect. The color coding shows how the age of the DDPs diversifies over the number of DDPs (Red = fewer DDPs, Green = more DDPs).

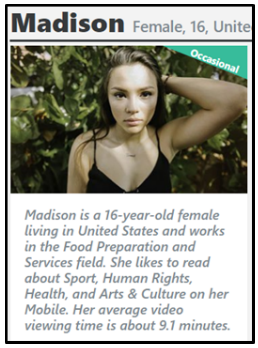
of DDPs. Therefore, **H1a is <u>not</u> supported: There is no evidence that more personas would result in a better representation of user gender.** This can be explained as a natural consequence of gender having only two values for the algorithm to choose from; in other words, it is relatively easier for the algorithm to generate representative personas than it is when the attribute has a lot of values to choose from.

5.2 Age

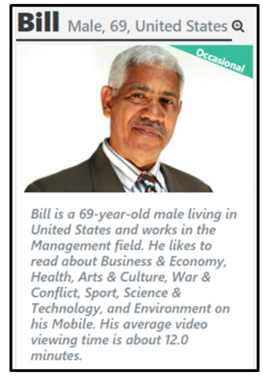
Interestingly, even though age also has a relatively small number of possible values (seven age groups), the approximation of age seems to stabilize only after 80 DDPs (see Figure 5a). Especially the elderly age groups are underrepresented with smaller numbers of DDPs (see Figure 6). Among the "new" ages appearing (relative to 5 DDPs) are 35-44 (at 10 DDPs), 45-54 (at 20 DDPs), 13-17 and 55-64 (at 40 DDPs), and 65+ (at 80 DDPs). Overall, creating fewer personas tends to prioritize age groups with larger representation in the baseline data, so that many personas are required in order for the algorithm to find a persona for minority groups. Only at 80 DDPs there is at least one DDP for each age group.

The fact that the youngest age group of 13-17 also grows at a pattern of $2 \longrightarrow 3 \longrightarrow 7$ (see the first row in Figure 6) indicates that the effect on persona age is symmetrical: both the youngest and oldest age groups become more prevalent with more personas. Moreover, even though all ages are covered at 80 DDPs, age groups become more balanced with 160 DDPs, as shown by the decreasing values in Figure 5a. This diversification effect can be quantified using the Entropy metric (H) that considers the probability P of an age group i across all the age groups M:

$$H = -\sum_{i=1}^{M} P_i \, log_2 \, P_i \tag{1}$$



(a)



(b)

Fig. 7. **(a) Madison**, a young segment only visible after 40 DDPs, and **(b) Bill**, an elderly segment only visible after 80 DDPs. Personas from these age groups appear only when increasing the number of DDPs, making demographic outlier user segments more concrete for decision makers.

The obtained values indicate that DDPs become more diversified when going from 80 (H = 1.43) to 160 (H = 1.51) personas. To corroborate this, we also calculated Simpson Dominance Index (S):

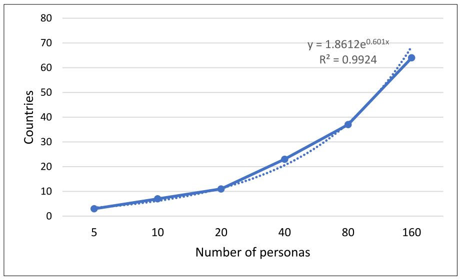
$$D = 1 - \sum_{i=1}^{M} n(n-1)/N(N-1), \tag{2}$$

where n is number of personas within a given age group, and N is the number of personas in the set. The values of this metric are to be interpreted in reverse to H, so that a higher number indicates a particular category value is dominating. The results of S are consistent with H, so that 80 DDPs have more age dominance (S = 32.1%) than 160 DDPs (S = 28.9%). This indicates that even though all age groups are already represented among 80 personas, increasing the number beyond this makes rarer age groups appear more often. In aggregate, the results indicate that **H1b is supported: Creating more personas results in a better representation of user age.** The implication is that personas from the youngest and eldest age group would not appear at all before 80 personas (see Figure 7).

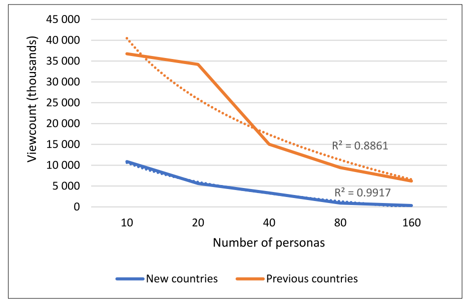
5.3 Country

Because country contains 172 possible values, personas with a "new" country emerge even within the set of 160 DDPs (e.g., France, Norway, Denmark). Percent-wise, 40% of the persona countries are new in the set of 10 DDPs, 35% in the 20 DDPs, 27.5% in the 40 DDPs, 17.5% in the 80 DDPs, and 16.9% in the 160 DDPs. For example, in the 40 persona set, there are personas from 11 new countries that did not emerge in the smaller persona sets: Australia, United Arab Emirates, Saudi Arabia, Germany, Singapore, Kenya, South-Africa, Netherlands, Brazil, Morocco, and Taiwan.

In the baseline data, the average view counts across all the persona sets is higher for the "old" countries that had emerged in the smaller DDP sets (M = 20.3 million) than for the new countries (M = 4.2 million). Welch's t-test shows that this difference between the groups is significant, t(86.9) = -7.6, p < .0001. Therefore, **H1c is supported: More personas results in a better representation of user country.** Figure 8 illustrates the found differences.



(a)



(b)

Fig. 8. (a) Unique countries in a persona set. (b) The viewcount differences for new and previously emerging countries in a given persona set. Diverging lines indicate that, as the number of personas increases, smaller audiences are included in the persona sets.

Table 1. Average decrease of the distance metrics (KL, HD, KS). The numbers indicate that increasing the number of personas improves the representativeness of the personas relative to the baseline user data.

	Age	Gender	Country	Average
from 5 to 10	-20 %	63 %	-4 %	-8 %
from 10 to 20	-16 %	-30 %	-4 %	-11 %
from 20 to 40	-18 %	26 %	-10 %	-10 %
from 40 to 80	-37 %	-14 %	-11 %	-16 %
from 80 to 160	-12 %	-13 %	-22 %	-17 %

5.4 RQ2: Is There an Effect of Decreasing Returns?

The results from the distance metrics (see Table 1) show that more personas are always better for capturing demographic attributes from the baseline data. Doubling the number of DDPs yields a general trend of negative growth (R^2 = 0.897), which indicates that the increase in the number of DDPs improves the approximation of the baseline data about the users in a reasonably consistent manner. This implies that, within the tested range of personas, there is no clear effect of diminishing returns when it comes to measuring the distance between the personas and the underlying dataset.

However, are the effects consistent when considering the personas' demographic diversity? To investigate this, we define a metric called *Diversity* (D), which equals the number of unique attributes in the persona set.

We compute D separately for age, gender, and country, and take their sum that defines how many unique values a persona set has. We then compare D of each set to the previous set to obtain a growth rate in unique demographic values, obtaining what we call the *Relative Diversity Gain* (RDG) between two successive persona sets i and j:

$$RDG = \frac{D_j - D_i}{D_i} \tag{3}$$

The RDG indicates the growth in new attribute values that were *not* included in the previous set of personas. Figure 9 shows that the set of 40 personas has the largest gain of new demographics relative to the previous sets. This implies that, although gains may continue after 40 personas, the proportional gains will not be as great in terms of increasing persona diversity. As such, **H2 is supported: There is an effect of decreasing returns, where the marginal gain of personas' diversity decreases after a peak.** In our experimental range of 5 to 160 personas, this peak was achieved with 40 personas.

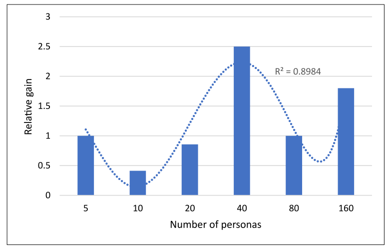


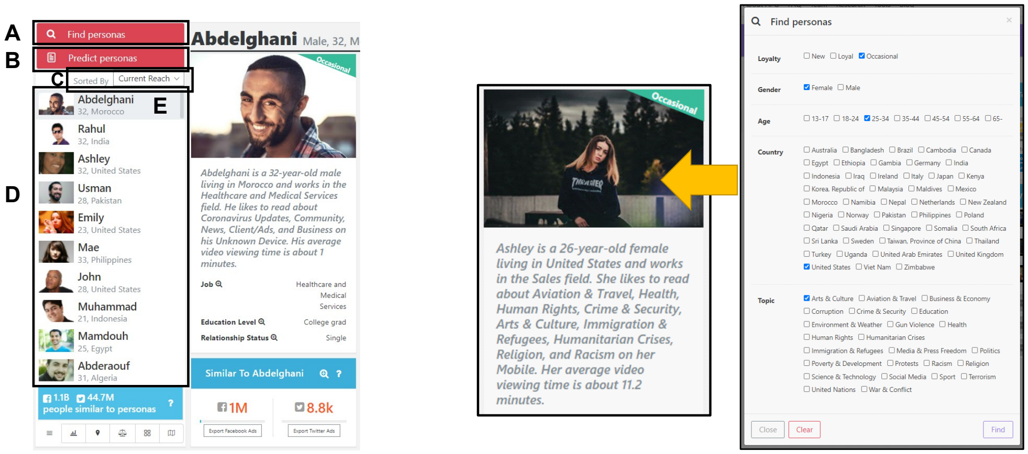
Fig. 9. RDG indicates a point where the number of unseen demographics peaks. Y-axis indicates the proportional increase in unique demographic attributes over the *previous* persona set. RDG peaks at 40 personas.

5.5 RQ3: How Can System Features be Developed to Support a Flexible Number of Personas?

In practice, we suggest developing techniques that support *user choice* for the number of DDPs for a given task. Our experiences from the field (observational user studies with a real persona system) imply that it is challenging to get stakeholders spontaneously to change the number of DDPs, even when a persona system has this functionality. It appears that the stakeholders simply lack the context to decide why a specific number (e.g., 5 or 15) would be better and predominantly accept the system's default number. However, a useful way for getting the same outcome is giving the stakeholders *navigational filters* that enable them to narrow down the DDPs by age, gender, country, persona's interests, and so on. Figure 10 illustrates interactive web techniques for a flexible number of personas. In Figure 10**(a)**, the several functionalities demonstrated include an option for [A)] finding personas based on their demographics, interests, or other information (Figure 10**(b)** further illustrates a use case for this), [B] predicting what personas among the total number would be interested in a given social media content, [C] sorting the personas based on their representativeness of the baseline user data or other criteria. The personas are shown in a scrollable sidebar [D] and selecting a given persona [E] will load the persona's full profile.

Figure 10**(b)** illustrates a user segment identification task [31], "Finding Ashley". Using the search functionality of an interactive system, stakeholders can narrow down the DDPs according to task-specific requirements. *Scenario:* a stakeholder wants to find a target segment for her new campaign. She wants to target occasional audience members that are Female, 25-34 years of age, from the United States, and interested in Arts & Culture. Out of the 100 available DDPs, the system finds one matching DDP. Screenshots from a real persona system (https://persona.qcri.org). We leave the user evaluation of this approach for future work.

Interactive features such as this can signify a *paradigm shift* for personas: rather than the persona developers deciding the right number of personas on stakeholders' behalf, the stakeholders will make this choice themselves, possibly without even thinking of a specific number. Logically, "persona search" will enable the users to narrow down the list of candidate personas for a given user-centric task. Trends supporting such a shift include the availability of online user data, automation of data analysis, and the rise of interactive Web systems for DDPs [29]. We believe the idea of "persona search" to be a great step in the right direction by better supporting human teams in finding the main personas from a larger group of automatically generated personas. The idea is that interactive features in an interactive persona system make the user's cognitive cost a constant rather than a variable. This removes the cognitive cost constraint, and the number of personas can be set either by data-driven way , or by task fit, e.g., by searching for personas that match specific attributes that the designer seeks.



(a)

(b)

Fig. 10. Interactive features for managing with a large number of personas (explained in text).

6 DISCUSSION

6.1 Implications for Persona Creation

The results indicate that creating more personas is beneficial for the better representation of the user base's demographic features, which is particularly important when dealing with large and heterogeneous online audiences. We also illustrate some key interaction features to deal with a larger of personas. The current study therefore presents one of the few – if not the only – studies exclusively focusing on the question of how many personas to create. Specifically, our study provides primary empirical analysis towards determining how many personas might be needed in international contexts with highly diverse user populations. Our findings support the anecdotal observations from previous research that one *needs* more personas to describe diverse and international audiences [26, 32].

More specifically, the results illustrate two specific diversification effects: age and country diversification effects. *Age diversification* means that the algorithm selects more age groups when increasing the number of personas, thus increasing the age diversity of the persona sets. *Country diversification* means that smaller geographic regions (in terms of their engagement with online content) become included in the persona sets when the number of personas increases.

We also address the calls in persona research for more empirical studies of representative and accurate personas [7, 8, 39] by showing that the increase in the number of DDPs reduces the statistical distance (error) of the personas relative to the baseline data. This also has a positive effect on the inclusivity of the personas [11], as the demographic diversity and representation of demographically marginalized groups increases with more personas.

Thus, the advantages for increasing the number of personas are two-fold: (a) representing the data more accurately (which is a *statistical* benefit evidenced by the decreasing trend in the statistical metrics when increasing the number of personas); and (b) representing fringe audience segments (which is a *diversity* benefit evidence by the increasing diversity metric when increasing the number of personas).

Because of the observed consistent pattern of better approximating the baseline data, simple techniques such as the elbow method are ill-equipped for determining a single "optimal" number of DDPs using our data (after which adding more DDPs would not add much information). Instead, it appears even more than 160 DDPs are needed for the marginal error to stabilize. The numerically optimal number of personas appears to be "as many as possible". We devise a metric—the Relative gain of diversity—to partially address this challenge. Applying the RDG on our dataset yields 40 as the number that most increases the diversity in the persona set. The advantage of using this metric is that it can provide a useful way of determining a cut-off point for the number of personas in cases where the marginal error seems to continuously decrease with the number of personas.

The current study has important implications for creation of personas to portray online audiences, and other audiences comprising a wide range of behaviors and demographics. It sheds light on the ability of "more" personas to cover more demographic attributes of the underlying data about users, and embodies a methodology to create effective international and global personas [26]. This may provide tangible design advantages realized in the course of making decisions for a broader mental model that accounts for demographic diversity and can possibly help mitigate the designer's bias of focusing on majority groups available in the data [44]. Since the presenting demographically diverse user segments is considered instrumental for mitigating the fundamental risk of stereotyping [11], our results involve implications to *fair* persona creation (i.e., one that increases the diversity of different demographic groups being represented by the personas), which is seen as a crucial dimension in ethical use of personas and user datasets [28].

Overall, the research challenges the established "small number of personas" hypothesis. This is an important contribution as existing studies are often based on heuristically determined numbers without any in-depth reasoning, except that too many personas can be too complicated for users to handle. Consequently, this premise can hinder and restrict further development of data-driven user segmentation. In contrast, our findings suggest that data-driven approaches can more actively be applied to investigate the optimal number of personas than done so far in the literature. Given that that there is no clear evidence of how many user segments or personas stakeholders can cope with, the question of how many personas to generate ought to be be explored with an open mind.

6.2 Implications for Organizations Using Personas

The applicability of our findings concerns organizations with large and heterogeneous, often international or global, audiences. These include, among others, news and media companies, international e-commerce sites, social media channels and influences, public health broadcasters, and others.

Our research has three key takeaways for persona developers that create personas for such organizations.

- **Data-driven approaches create pressure for going beyond a handful of personas.** The computational cost for increasing the number of data-driven personas is negligible, as one can generate either 5 or 500 segments with a click of a button, which drastically differs from manual persona creation. With large digital user populations, assuming each segment is unique and needs a persona, 3-10 is simply not enough. In lack of other analyses, our findings point to 40 as a reasonable benchmark for a global user population.
- **More DDPs better portray marginal audiences from digital user data.** The results show no point of diminishing returns: from a numeric point of view, more DDPs are always better, as they cover more unique user characteristics, specifically characteristics with multiple levels, such as age and country.
- **Persona traits with fewer classes (e.g., gender) can be approximated more easily than traits with more classes (e.g., country)**. Therefore, persona creators should examine their data distributions carefully prior to deciding the number of personas.

Using the premise that each user segment needs a persona, the traditional small number of personas cannot adequately work for organizations with large and diverse user populations, especially in digital environments. Coincidentally, this may be a factor in the criticism that personas have been argued to be of little value in the design process [20]. When fewer personas are created, personas that could be relevant for design could be ignored, as these personas never appear for the stakeholders. Figure 7 shows two examples of DDPs "missing" from the sets with a smaller number of DDPs. Hence, with more personas, one can concretely observe and clarify what type of people are missing when using a smaller segmentation number. Some of these personas can be important when designing social media content for different markets. The implications impact organizations with broad international audiences, as such audiences include many countries and age groups that the personas should represent. It is crucial that these organizations set the number of personas high enough to cover at least the central audience segments in the data. According to our results, 40 personas provide the highest diversity gain in terms of demographic features.

6.3 Limitations and Future Work

One can criticize the approach of increasing the number of personas based on the fact that fringe personas still remain rare even in the bigger persona sets. For example, even if the chance of picking an older persona increases with more personas, the overall chance will still be very small (assume two personas in the 40-persona set would be above 65 years old; this would mean 2 / 40 = 5% chance of a random persona being above 65 years). This argument is, on one hand, correct and, on the other hand, incorrect. It is correct if one presumes that the designer is randomly assigned a persona among the generated ones. But it is incorrect if one presumes that the designer can choose the persona (in which case the choice is not random, but influenced by the designer's preference). With more personas, the designer will be exposed to the age group 65+, an age group that did not even exist among the smaller set of personas and that, therefore, could not previously be part of her consideration set. So, given that there is a system that serves more personas efficiently, having more personas increases the designer's freedom of choice.

Our results do not provide a definitive answer to the question of how many personas should one create. The answer to this fundamental question might be Pareto optimal—a balance between the human need for fewer personas (i.e., cognitive cost) and the representational quality (i.e., diversity) of many personas. In this study, we did not address the element of cognitive cost, which means that finding an appropriate heuristic for this two-sided optimization problem remains for future work.

In our results, p=40 produced the biggest relative gain of demographic variety. This implies that the traditional 'less than ten personas' is most likely inadequate for representing large and heterogeneous online populations (LHOPs). From this, we can conceive two alternative research directions to pursue: either (1) to look into increasing the number of personas as a problem that applies universally to all persona types, or (2) to separate LHOPs as a special case of persona creation, for which a higher number of personas than ten is generally required (whereas the other persona types might not be so much affected). A corollary is that the heterogeneity of the data should contribute to deciding the number, most likely in a decisive way.

Finally, intelligent features regarding how to recommend decision makers specific personas based on their tasks remains an interesting and impactful area for future research (especially user studies) that is currently missing any meaningful contributions.

7 CONCLUSION

The study illustrates that increasing the number of data-driven personas from digital user data increases the accuracy with which these personas represent the data that was used to generate them. We also show that the higher number of personas means the personas are more inclusive, providing a better representation of fringe user groups. The gains of increasing personas' demographic diversity were relatively highest when the number of personas was forty. The results also imply that the more demographic variation the user attributes have in the baseline data, the more personas should be generated to cover this variation adequately. Hence, persona creators should examine their data distributions carefully prior to deciding the number of personas. In conclusion, more personas should be generated and considered when studying, understanding, and making decisions about a user base.

REFERENCES

[1] Jisun An, Haewoon Kwak, Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen. 2018. Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining* 8, 1 (2018). https://doi.org/10.1007/s13278-018-0531-0

[2] Jisun An, Haewoon Kwak, Joni Salminen, Soon-gyo Jung, and Bernard J. Jansen. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Transactions on the Web (TWEB)* 12, 3 (2018).

[3] Ivan E. Auger and Charles E. Lawrence. 1989. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology* 51, 1 (1989), 39–54. Publisher: Springer.

[4] Derrick S. Boone and Michelle Roehm. 2002. Evaluating the appropriateness of market segmentation solutions using artificial neural networks and the membership clustering criterion. *Marketing Letters* 13, 4 (2002), 317–333.

[5] J. Brickey, S. Walczak, and T. Burgess. 2012. Comparing Semi-Automated Clustering Methods for Persona Development. *IEEE Transactions on Software Engineering* 38, 3 (May 2012), 537–546. https://doi.org/10.1109/TSE.2011.60

[6] Alessandro Canossa and Anders Drachen. 2009. Play-personas: behaviours and belief systems in user-centred game design. In *IFIP Conference on Human-Computer Interaction* (2009). Springer, 510–523.

[7] C.N. Chapman, E. Love, R.P. Milham, P. ElRif, and J.L. Alford. 2008. Quantitative evaluation of personas as information. In *Human Factors and Ergonomics Society 52nd Annual Meeting* (2008). 1107–1111.

[8] Christopher N. Chapman and Russell P. Milham. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. In *Human Factors and Ergonomics Society Annual Meeting* (2006), Vol. 50. 634–636.

[9] Alan Cooper. 1999. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (1st Edition).* Pearson Higher Education.

[10] Xin Fu, Xi Chen, Yu-Tong Shi, Indranil Bose, and Shun Cai. 2017. User segmentation for retention management in online social games. *Decision Support Systems* 101 (Sept. 2017), 51–68. https://doi.org/10.1016/j.dss.2017.05.015

[11] Joy Goodman-Deane, Sam Waller, Dana Demin, Arantxa González-de Heredia, Mike Bradley, and John P. Clarkson. 2018. Evaluating Inclusivity using Quantitative Personas. https://doi.org/10.21606/drs.2018.400

[12] Peter Hall. 1987. On Kullback-Leibler loss and density estimation. *The Annals of Statistics* (1987), 1491–1519. Publisher: JSTOR.

[13] Hosagrahar Visvesva Jagadish. 2015. Big data and science: Myths and reality. *Big Data Research* 2, 2 (2015), 49–52. Publisher: Elsevier.

[14] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. 2019. Creating Manageable Persona Sets from Large User Populations. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). ACM, Glasgow, United Kingdom, 1–6. https://doi.org/10.1145/3290607.3313006

[15] Soon-gyo Jung, Joni Salminen, Haewoon Kwak, Jisun An, and Bernard J. Jansen. 2018. Automatic Persona Generation (APG): A Rationale and Demonstration. In *CHIIR '18: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Jersey, USA, 2018-03). ACM, 321–324. https://doi.org/10.1145/3176349.3176893

[16] Dannie Korsgaard, Thomas Bjørner, Pernille Krog Sørensen, and Paolo Burelli. 2020. Creating user stereotypes for persona development from qualitative data through semi-automatic subspace clustering. 30, 1 (2020), 81–125. https://doi.org/10.1007/s11257-019-09252-5

[17] Daniel D. Lee and Sebastian H. Seung. 1999. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* 401, 6755 (1999), 788–791.

[18] Hubert W. Lilliefors. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association* 62, 318 (1967), 399–402. Publisher: Taylor & Francis Group.

[19] Claudio Marcus. 1998. A practical yet meaningful approach to customer segmentation. *Journal of Consumer Marketing* 15, 5 (Oct. 1998), 494–504. https://doi.org/10.1108/07363769810235974

[20] Nicola Marsden and Maren Haag. 2016. Stereotypes and Politics: Reflections on Personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016) *(CHI '16)*. ACM, San Jose, USA, 4017–4031. https://doi.org/10.1145/2858036.2858151

[21] Jennifer Jen McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Florence, Italy, 1521–1524. https://doi.org/10.1145/1357054.1357292

[22] Joanna Misztal-Radecka and Bipin Indurkhya. 2020. Persona Prototypes for Improving the Qualitative Evaluation of Recommendation Systems. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA, 2020-07-14) *(UMAP '20 Adjunct)*. Association for Computing Machinery, 206–212. https://doi.org/10.1145/3386392.3399297

[23] Steve Mulder and Ziv Yaar. 2006. *The User is Always Right: A Practical Guide to Creating and Using Personas for the Web.* New Riders. Google-Books-ID: gLjPMUjVvs0C.

[24] Lene Nielsen. 2019. *Personas - User Focused Design* (2nd ed. 2019 edition ed.). Springer, New York, NY, USA.

[25] Lene Nielsen, Kira Storgaard Hansen, Jan Stage, and Jane Billestrup. 2015. A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *International Journal of Sociotechnology and Knowledge Development* 7, 1 (Jan. 2015), 45–61. https://doi.org/10.4018/ijskd.2015010104

[26] Lene Nielsen, Kira Storgaard Nielsen, Jan Stage, and Jane Billestrup. 2013. Going Global with Personas. In *Proceedings of the INTERACT 2013 conference* (2013). Springer, Berlin, Heidelberg, Cape Town, South Africa, 350–357. https://doi.org/10.1007/978-3-642-40498-6_27

[27] John Pruitt and Jonathan Grudin. 2003. Personas: Practice and Theory *(DUX '03)*. ACM, San Francisco, California, USA, 1–15. https://doi.org/10.1145/997078.997089

[28] Joni Salminen, Willemien Froneman, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. [n.d.]. The Ethics of Data-Driven Personas. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts* (Honolulu, HI, USA, 2020-04-25) *(CHI '20)*. Association for Computing Machinery, 1–9. https://doi.org/10.1145/3334480.3382790

[29] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, Shammur Absar Chowdhury, and Bernard J. Jansen. 2020. A Literature Review of Quantitative Persona Creation. In *Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI'20)* (2020). ACM, Honolulu, Hawaii, USA.

[30] Joni Salminen, Soon-Gyo Jung, Kamal Chhirang, and Bernard Jansen. 2021. Instilling Knowledge Claims of Personas from 346 Research Articles. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 450. Association for Computing Machinery, 1–9. https://doi.org/10.1145/3411763.3451619

[31] 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)* (Honolulu, Hawaii, USA, 2020). ACM. https://doi.org/10.1145/3313831.3376770

[32] Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2019. Detecting Demographic Bias in Automatically Generated Personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019) *(CHI EA '19)*. ACM, New York, NY, USA, LBW0122:1–LBW0122:6. https://doi.org/10.1145/3290607.3313034 event-place: Glasgow, Scotland Uk.

[33] Joni Salminen, Rohan Gurunandan Rao, Soon-gyo Jung, Shammur A. Chowdhury, and Bernard J. Jansen. 2020. Enriching Social Media Personas with Personality Traits: A Deep Learning Approach Using the Big Five Classes. In *Artificial Intelligence in HCI (Lecture Notes in Computer Science)*, Helmut Degen and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 101–120. https://doi.org/10.1007/978-3-030-50334-5_7

[34] Joni Salminen, Sercan Şengün, Haewoon Kwak, Bernard J. Jansen, Jisun An, Soon-gyo Jung, Sarah Vieweg, and Fox Harrell. 2018. From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *First Monday* 23, 6 (June 2018). https://doi.org/10.5210/fm.v23i6.8415

[35] Stan Salvador and Philip Chan. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE international conference on tools with artificial intelligence*. IEEE, 576–584.

[36] Marko Sarstedt, Jan-Michael Becker, Christian M. Ringle, and Manfred Schwaiger. 2011. Uncovering and treating unobserved heterogeneity with FIMIX-PLS: which model selection criterion provides an appropriate number of segments? *Schmalenbach Business Review* 63, 1 (2011), 34–62. Publisher: Springer.

[37] Drew Schmidt, Wei-Chen Chen, Michael A. Matheson, and George Ostrouchov. 2017. Programming with BIG data in R: Scaling analytics from one to thousands of nodes. *Big Data Research* 8 (2017), 1–11. Publisher: Elsevier.

[38] Douglas G. Simpson. 1987. Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American statistical Association* 82, 399 (1987), 802–807. Publisher: Taylor & Francis Group.

[39] Dimitris Spiliotopoulos, Dionisis Margaris, and Costas Vassilakis. 2020. Data-Assisted Persona Construction Using Social Media Data. *Big Data and Cognitive Computing* 4, 3 (Sept. 2020), 21. https://doi.org/10.3390/bdcc4030021 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[40] Phillip Douglas Stevenson and Christopher Andrew Mattson. 2019. The Personification of Big Data. *Proceedings of the Design Society: International Conference on Engineering Design* 1, 1 (July 2019), 4019–4028. https://doi.org/10.1017/dsi.2019.409

[41] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering*, Vol. 336. IOP Publishing, 012017. Issue: 1.

[42] Thorsten Teichert, Edlira Shehu, and Iwan von Wartburg. 2008. Customer segmentation revisited: The case of the airline industry. *Transportation Research Part A: Policy and Practice* 42, 1 (Jan. 2008), 227–242. https://doi.org/10.1016/j.tra.2007.08.003

[43] Aaron Tkaczynski, Sharyn R Rundle-Thiele, and Nina Katrine Prebensen. 2018. To segment or not? That is the question. *Journal of Vacation Marketing* 24, 1 (Jan. 2018), 16–28. https://doi.org/10.1177/1356766716679482

[44] Phil Turner and Susan Turner. 2011. Is stereotyping inevitable when designing with personas? *Design studies* 32, 1 (2011), 30–44.

[45] Kari T. Vasko and Hannu TT Toivonen. 2002. Estimating the number of segments in time series data using permutation tests. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 466–473.

[46] Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. 2016. Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016) *(CHI '16)*. ACM, San Jose, California, USA, 5350–5359.

[47] Hongbo Zou, Yongen Yu, Wei Tang, and Hsuan-Wei Michelle Chen. 2014. FlexAnalytics: a flexible data analytics framework for big data applications with I/O performance improvement. *Big Data Research* 1 (2014), 4–13. Publisher: Elsevier.