Survey2Persona: Rendering Survey Responses as Personas

Joni Salminen
University of Vaasa, Vaasa, Finland
jonisalm@uwasa.fi

Soon-Gyo Jung
Qatar Computing Research Institute,
Hamad Bin Khalifa University
sjung@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute,
Hamad Bin Khalifa University
bjansen@hbku.edu.qa

ABSTRACT

Data-driven persona generation can benefit from stakeholder inputs while offloading the complexities of high-dimensional datasets. To this end, we present Survey2Persona (S2P), an interactive web interface for real-time persona generation from survey data. The users of the web interface—the designers—can upload survey data and have the interface automatically generate personas. Researchers and practitioners can use S2P to explore different respondent types in their survey datasets in a privacy-preserving manner, which is akin to making the analytical journey more productive, enjoyable, and human-centered. We make the system publicly available and provide argumentation about its novelty and value for user modeling and human-computer interaction communities.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI).

KEYWORDS

Personas, surveys, data-driven personas, human-centered analytics

ACM Reference Format:
Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2022. Survey2Persona: Rendering Survey Responses as Personas. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct), July 04–07, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3511047.3536403

1 INTRODUCTION

Personas are fictitious people that represent real user groups [10], with the purpose of helping designers put themselves in the position of end-users of software systems, applications, and other tangible or intangible offerings provided to society at large [33]. Personas are, thus, a tool of user-centered design and a research topic within human-computer interaction (HCI) and the broader field of user modeling. In competitive markets, it is important that products are human-centric, that is, that they solve real end-user needs and, therefore, deliver concrete value to end-users [3, 16].

The gap between 'what is possible' in design and 'what solves user needs' is what personas, as one design instrument among many others [7], aim at mitigating. The strength of personas is seen in their ability to enhance empathy [9, 12], which is rooted in the idea that designers are better able to empathize with the needs realistic people types than nameless and faceless 'user segments' or 'target groups' [14]. Even though personas have traditionally been created using qualitative data [32, 34, 35] and manual data analysis, as data science technologies are rapidly becoming more broadly available [30], *automation* of persona creation has increased its feasibility. To this end, researchers have proposed methodologies [1, 2] and systems [19, 20, 22] for automatic persona generation. While these systems have contributed to data-driven personas [29]—defined as personas created using quantitative techniques, usually with the help of algorithms and evaluation metrics—most of these automated persona systems rely on a massive amount of social media or system log data, rather than survey data [18, 51].

This shortcoming matters for three main reasons. First, **(1)** survey data is the most popular source of data for persona generation, with roughly half (47%) of the articles reviewed in previous work [39] using surveys for persona generation. Second, **(2)** even though the use of survey data for persona generation is highly prevalent among researchers (e.g., [6, 25, 52]), the personas generated from surveys tend to be static and not *interactive*, which hinders the way designers interact with them for design purposes. Third, **(3)** designers typically cannot participate in the process of quantitatively creating personas from survey data, a process typically governed by opaque data science algorithms [44], hindering empathy. These limitations restrict the applicability of personas when deploying survey data for persona creation.

Our motivating question in this work is: "*How can we generate personas automatically using survey data?*" While previous persona systems have focused on demographic and behavioral online analytics [51] or social media analytics [2], there is a lack of replicable, scalable, and interactive systems for designers to work with personas using survey data. Additionally, purely computational techniques have their advantages and drawbacks—at worst, quantitative persona generation can be riddled with complex and opaque generation methods that are both difficult to understand by designers and whose decisions are hard to explain and justify [44]; e.g., *Why this number of personas? Why this persona has this age?* And so on. These explainability concerns can be alleviated when designers are provided a self-help tool for persona creation, making them effectively take part in the process of persona generation instead of simply consuming the "ready" personas created by data analysts and algorithms. To this end, we propose *Survey2Persona (S2P)*, an interactive web system that enables the designer to upload survey data about users and generate personas from this data without making any complex technical choices (the system is available online at https://s2p.qcri.org.. So, to summarize, the problem we address is: despite the considerable progress in data science algorithms and associated technologies, as well as in data-driven persona generation, there are no comparable tools for creating personas from survey

UMAP '22 Adjunct, July 04–07, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9232-7/22/07.
https://doi.org/10.1145/3511047.3536403

data, despite the broad popularity of surveys as a data collection method in HCI and social sciences.

2 CONCEPTUAL UNDERPINNINGS AND RELATED WORK

In terms of concepts, here we use 'designer' to refer to persona users [32–34], but in reality, persona users can be any stakeholders in whatever profession that deals with decision making about end-users (e.g., content creators, marketers, advertisers, HR managers, analysts, videogame developers, public health professionals, etc.). This is because personas are a cross-sectional decision-making support tool [17]; they can be broadly applied to any range of user-related decision making (e.g., targeting, copywriting, and so on). In contrast, 'end-user' refers to the people that personas represent. These can be, for example, social media audiences, customers, product users, patients, and so on.

Personas have many benefits, including that they enable a higher degree of user-centricity [9, 12] which, in turn, enables empathetic design outcomes, manifesting in products that solve real user needs in an effective manner [35]. Because of this tenet, understanding how designers interact with personas is vital. Despite the benefits, the potential of personas for enhancing user-centered design is curbed by some challenges, for example, (a) leading designers to stereotypical assumptions about the end-users [48, 49]; (b) personas being perceived as abstract, misleading, or useless [28]; and (c) unclear value of personas for design that may lead to (d) personas being questioned in practice [13, 37].

In separated study lines, researchers have made progress, on the one hand, with *interactive persona systems* [1, 19, 22] and, on the other hand, with *persona creation from survey data* (e.g., [5, 15, 23, 26, 45]). However, no previous attempt, to our knowledge, has combined these two attempts – therefore, the current state-of-the-art is lacking an easy-to-use and practical interactive data-driven persona system that would enable the designer to upload survey data and, with low effort, generate personas using a quantitatively principled approach in the back-end.

Research has also observed other issues relating to persona generation. Salminen et al. [38] conducted a literature review of 49 quantitative persona generation papers. They found that methodologically, the most common persona generation technique was clustering (N=17, 34.6% of the reviewed papers), with other techniques including principal component analysis (PCA) (n=5, 10.2%), latent semantic analysis (LSA) (n=5, 10.2%), and non-negative matrix factorization (NMF) (n=4, 8.2%). Why is there such a strong dominance of using clustering or dimensionality reduction for persona generation? We believe this stems from three main drivers: first, (a) quantitative persona generation has traditionally been understood as a user segmentation or customer segmentation problem. This means, methods addressing that type of problem have been adopted, with the consequence of almost exclusively focusing on data dimensionality and clustering. Second, (b) there is a lack of precedent or alternatives *not* using these techniques for persona generation – instead, it is customary to use a complex statistical algorithm in the absence of other solutions.

Moreover, (c) there is a phenomenon called 'mystique of numbers' [47], meaning that complexity is seen as a virtue. We believe that this tendency takes place both among persona creators that prefer technically complex methods because these are easier to publish than simpler solutions that are seen as uninteresting generally in academia [4]. This tendency is also among designers that may think that a complex algorithm has superhuman accuracy and capabilities, making the generated personas more valuable than those created, for example, using simple qualitative analysis.

Our approach is based on the idea that, even though persona generation is typically addressed using data dimensionality and clustering algorithms, this need not be. Instead, personas can be generated using a much simpler statistical heuristic approach that still maintains the benefits of being data-driven and also enables more flexibility using an interactive generation process that considers designers' information needs as a part of real-time persona generation. In the following section, we discuss these aspects in greater detail.

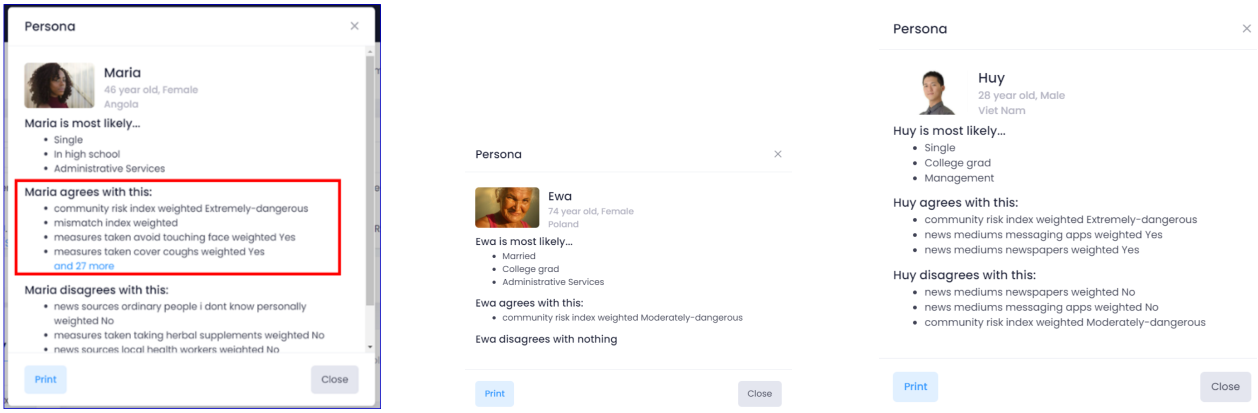
3 APPROACH

A crucial intuition behind the system is the analogy between survey items and personas' quotes. What we mean by this is that survey items typically contain a statement (e.g., "I trust Google"), and the respondent chooses an option from a Likert scale (typically, between "Strongly agree" and "Strongly disagree"). When constructing personas, we can use these statements as quotes (quotes are an essential part of a persona profile as they convey the attitudes of the persona). Therefore, the group of respondents that is likely to agree with "I trust Google" will have this statement shown as their quote in the persona profile under "This persona agrees with" and the group that is likely to disagree with the statement will have it under their "This persona disagrees with" section (see Figure 1 for illustration). The important intuition here is that surveys are a way to organize information about people's attitudes (the same applies to personas!). Survey creation is an informative process that reveals the organization's / researchers' information needs about users [23]. Both of these premises are essential for generating personas with high practical value.

3.1 Algorithm Design

We apply a simple intuition to construct the personas from survey data in real-time: *outlier detection*. This means that whatever demographic or item a designer is interested in, we investigate the mean scores of each demographic group in the sample and select those that considerably deviate from the mean score of all demographic groups. The idea is that if a demographic group's mean score is higher than s standard deviations (SD) from the mean of the whole sample, the demographic group has a tendency of agreeing with the statement more than other demographic groups. If the group's mean score for the statement is lower than s standard deviations from the mean of the whole sample, then the group has a tendency of disagreeing with the statement (relative to other groups).

The parameter s is the number of standard deviations that the systems test, both for positive (+) and negative (-) respective to the mean. The possible values of s stem for common values in outlier detection literature (e.g., [11, 24, 36]), ranging from 1 to 3, with 0.5 increments (i.e., 1.0, 1.5, 2.0, 2.5, 3.0). For example, 99.87% of normally distributed data are situated within the range of mean ±3



Maria
46 year old, Female
Angola

Maria is most likely...
- Single
- In high school
- Administrative Services

Maria agrees with this:
- community risk index weighted Extremely-dangerous
- mismatch index weighted
- measures taken avoid touching face weighted Yes
- measures taken cover coughs weighted Yes

and 27 more

Maria disagrees with this:
- news sources ordinary people i dont know personally weighted No
- measures taken taking herbal supplements weighted No
- news sources local health workers weighted No

(a) Maria agrees with 31 statements

Ewa
74 year old, Female
Poland

Ewa is most likely...
- Married
- College grad
- Administrative Services

Ewa agrees with this:
- community risk index weighted Moderately-dangerous

Ewa disagrees with nothing

(b) Ewa agrees with 1 statement and disagrees with 0

Huy
28 year old, Male
Viet Nam

Huy is most likely...
- Single
- College grad
- Management

Huy agrees with this:
- community risk index weighted Extremely-dangerous
- news mediums messaging apps weighted Yes
- news mediums newspapers weighted Yes

Huy disagrees with this:
- news mediums newspapers weighted No
- news mediums messaging apps weighted No
- community risk index weighted Moderately-dangerous

(c) Huy agrees with 3 statements and disagrees with 3

Figure 1: A persona with too much information (a) and too little information (b). S2P's approach can is a trade-off of satisfying information needs (without information overload) and finding notable distance in a group's answer to the general tendency of the whole sample's answers, which produces personas that have adequate but not too much information (c).

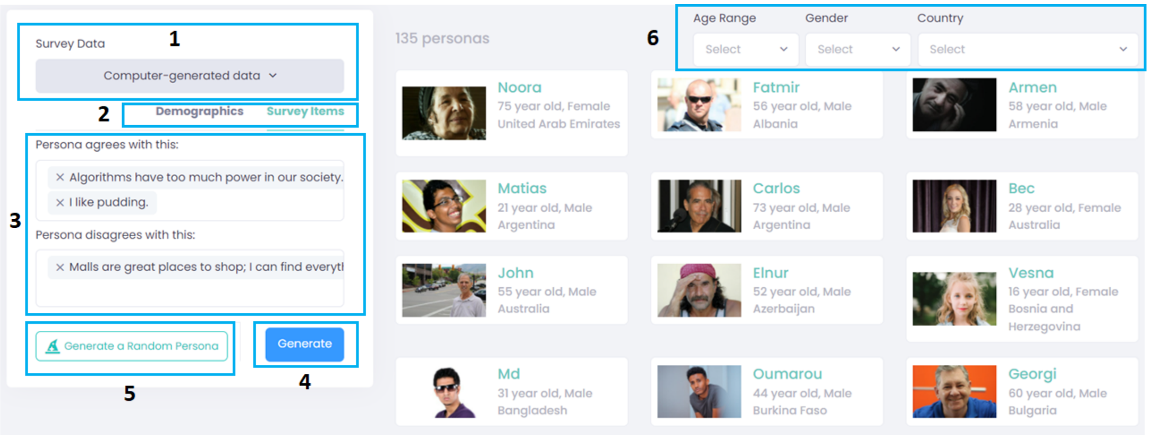
SD [27]. The system computes the deviations of each demographic for each statement using these five values separately, and then counts the number of deviations (either positive or negative) that each demographic group has. In our experiments, we found that setting a universal value for s that would always be used for all datasets is not appropriate, because this results in a high number of scenarios where the generated personas either have too many statements they agree and disagree with (typically when the SD value is on the low end), or there are no personas at all (typically when the SD value is on the high end). Similarly, picking a value from the midpoint also does not always produce personas that would have adequate information. By 'adequate,' we mean that a persona should, on the one hand, include *enough* information to be useful but, on the other hand, it cannot have too much information. These two extremes, as well as the balanced case, are illustrated in Figure 1

A persona with too much information is not unique enough and has too many statements for the designer to effectively make sense of. In contrast, a persona with too little information presents the designer with nothing to learn about. We solve this problem by defining a range of statements (3-7) that the personas need to have. In other words, each persona has at least three and at maximum seven agree and disagree statements. In our experiments with the test datasets, we found that this range both is able to generate personas in the overwhelming majority of cases and also generates personas that are not cognitively overloading to the system user (see Figure 1c). Therefore, apart from this boundary heuristic that we justify by the desire to keep the information in the personas 'not too scarce, not too exhausting', the persona generation process in S2P is completely data-driven (i.e., we do not impose any manual rules or decisions regarding the selection of information). We chose 3-7 as the range for three primary reasons. First, (a) from a visual glance, it appears to generate personas that 'not too little, not too much' information. Second, (b) it corresponds to typical persona information content, as persona profiles typically have ~5 quotes. Third, (c) this range is close to the "magical number of seven plus minus two", introduced famously by Miller in 1956 [31]. In brief, this magical number postulates that, on average, people can hold 7 ± 2 items in their short-term memory.

3.2 System Design

Personas in S2P can be generated either by using demographic characteristics or using survey items. This provides versatility in the process. For example, a scenario where demographic generation is appropriate could be a United Nations cross-country survey about social matters. An analyst might want to know what kind of answers young women in developing countries gave. To investigate this *Scenario 1*, the analyst selects an age range and countries in S2P, and the system generates personas representing distinct response types from the corresponding demographics. Similarly, an analyst might want to know what people in different countries think about a particular issue. In this *Scenario 2*, the analyst uses the item-based persona generation (see Figure 2) to select the specific item of interest, and the system generates personas that either agree or disagree with the item based on the analyst's choice. Moreover, an analyst might have a question, "I wonder what people in Finland think about this matter?". In this case, they can first select the item(s) of interest, generate the personas, and then use filters to narrow down the persona country to Finland. In our previous user studies [41–43], we have discovered that these types of scenarios are typical among stakeholders, as they often want to know about a specific aspect/attitude or target group. If personas are static and created without consideration to the varying information needs, they risk falling short in practical use.

The actual persona profiles in S2P are simple. We wanted to keep the personas simple purposefully, because a clean interface helps designers focus on the essential information. As can be seen from Figure 3, the included information corresponds to typical basic information in persona templates [34, 40]. The name is obtained



Persona agrees with this:

× Algorithms have too much power in our society.

× I like pudding.

Persona disagrees with this:

× Malls are great places to shop; I can find everytl

Noora
75 year old, Female
United Arab Emirates

Fatmir
56 year old, Male
Albania

Armen
58 year old, Male
Armenia

Matias
21 year old, Male
Argentina

Carlos
73 year old, Male
Argentina

Bec
28 year old, Female
Australia

John
55 year old, Male
Australia

Elnur
52 year old, Male
Azerbaijan

Vesna
16 year old, Female
Bosnia and Herzegovina

Md
31 year old, Male
Bangladesh

Oumarou
44 year old, Male
Burkina Faso

Georgi
60 year old, Male
Bulgaria

Figure 2: UI elements of S2P. User can select the dataset (1), then whether to generate personas based on demographic attributes or survey items (2). When generating based on survey items, the user can select statements that persona agrees or disagrees with (3). The algorithm then creates the personas in real-time based on these choices (4). The user also can generate a random persona (5), in which case the system will automatically choose random values for the statements or demographics. The generated personas appear on the right-hand side, and the user has the option to further narrow down the personas based on demographic filters (6).

using GAN2Name [21], a social media -based algorithm that outputs a person's likely based on demographic attributes. The picture is obtained from a proprietary database that contains thousands of demographically tagged facial images [17]. Even though the current persona layout is clean and simple, it is not out of the question that its information content can be expanded in future work – this depends on designers' needs and wishes that we will investigate in future user studies.

4 EVALUATION

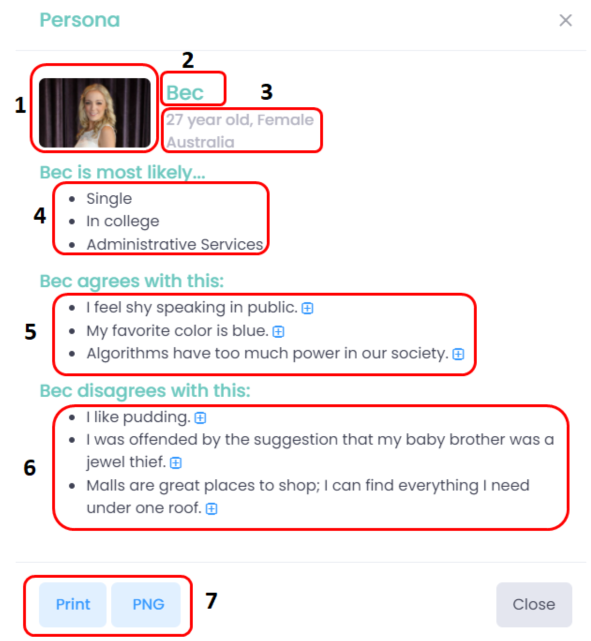
We tested S2P with three datasets to ensure that the system could actually generate personas out of different demographics and survey items. The datasets are: (1) the "COVID-19 Beliefs, Behaviors & Norms Survey" (*COVID-19*) published by MIT in collaboration with Facebook and the World Health Organization (retrieved using the API on the project's website, https://covidsurvey.mit.edu/), (2) the "American Trends Panel Data" (*AMTRENDS*) from Pew Internet Research (available at https://www.pewresearch.org/social-trends/dataset/american-trends-panel-wave-77/), (3) an artificial dataset (*ARTIFICIAL*) that we created by randomly generating statements using a text generator and then randomly assigning values from the Likert range of 1-5 to 1708 unique demographic groups (i.e., each group was assigned a random value for each statement). The latter dataset enables us to test S2P with a high number of demographic groups, whereas the two real datasets enable us to see how the system works with real datasets in the wild. Table 1 shows the number of personas generated using different demographic and statement combinations, whereas Table 2 summarizes the dataset properties and average numbers of personas generated.

When there is a medium or high number of demographic groups, many personas are generated. In contrast, with few demographics, only a few personas are generated. The number of demographic groups is more important than the number of statements and responses; with only ten statements, with a high number of demographic groups, S2P still generates a high number of personas. Symmetrically, a decent sample size of 10,169 is not adequate to generate a high number of personas when there are only a few statements and demographic groups. Overall, S2P's ability to generate personas does not appear great when there is a small number of statements or demographic groups. Future research is needed to define the lower boundary conditions for meaningful persona generation using S2P, but at this point, the data suggests that S2P is the most useful for large-scale survey datasets with a variety of statements (>10), demographic groups (>100), whereas the number of responses seems less sensitive and could be assigned based on general sampling rules (e.g., >250 [46]). These recommendations match the general notion of data-driven personas being useful for large datasets [18, 29], whereas creating personas from small datasets is often more meaningful using manual means [33].

5 DISCUSSION

5.1 Strengths

The work-in-progress system we demonstrate here represents a step towards a higher degree of designer-participation, while still attempting to maintain the benefits of quantitative methods. S2P is the first interactive persona system to our knowledge that generates personas truly in real-time. Previous approaches have either generated the whole persona set before displaying it to designers or only partially generated them in real-time by retrieving only some persona information (e.g., quotes) in real-time [22]) based on designers' information needs. Also, previous systems rely on clustering to generate a set of static personas that cannot be modified by the designer once they have been created, or the only way to do so is the repeat the whole generation process [2]). S2P provides a fast and responsive interface for designers to generate the personas



Bec

27 year old, Female
Australia

Bec is most likely...

- Single
- In college
- Administrative Services

Bec agrees with this:

- I feel shy speaking in public. ⊞
- My favorite color is blue. ⊞
- Algorithms have too much power in our society. ⊞

Bec disagrees with this:

- I like pudding. ⊞
- I was offended by the suggestion that my baby brother was a jewel thief. ⊞
- Malls are great places to shop; I can find everything I need under one roof. ⊞

Explanations:

(1) Manually curated and labeled picture database (copyrights acquired)
(2) GAN2Name algorithm (outputting a person's first name based on their demographics
(3) Demographic group in the survey data used for the persona generation (age is selected randomly from a range; (e.g., 28 years from the range of 25-34)
(4) Most common sociographic information for this demographic group in Facebook user population
(5) Statements that the persona agrees with more than an average demographic group
(6) Statements that the persona disagrees with more than an average demographic group
(7) Option to print the persona in a format that can be attached to presentations.

Figure 3: Bec, generated using an artificial test dataset. Upon clicking a persona in the generation view (see Figure 2), the full persona profile is shown, with the persona's picture, name, demographic and sociographic information, and the statements persona agrees and disagrees with. There is also a possibility to print the persona profile in a picture file.

Table 1: Number of personas generated using the test datasets. Age groups in parentheses are for the COVID-19 dataset, age groups not in parentheses are for the two other datasets.

	COVID-19	AMTRENDS	ARTIFICIAL
Age group: 18-24 (13-19)	12	1	12
Age group: 25-34 (20-30)	49	0	16
Age group: 35-44 (31-40)	39	0	16
Age group: 45-54 (41-50)	30	2	16
Age group: 55-64 (51-60)	30	N/A	17
Age group: 65+ (61-70)	24	N/A	14
Gender: Male	106	0	58
Gender: Female	90	1	45
Gender: Non-binary	2	2	N/A
Country: random among available	6 (Venezuela)	3 (United States)	1 (Lebanon)
1 random agree statement (over 5 iterations)	M=110.2, SD= 15.9	M=1.8, SD=0.8	M=337.6, SD= 16.5
1 random disagree statement (over 5 iterations)	M=102.4, SD=29.0	M=1.6, SD=0.5	M=352.4, SD= 20.1
2 random agree statements and 2 random disagree statements	M=57.4, SD=9.7	M=1.4, SD=0.9	M=224, SD=25.2

Table 2: Test dataset properties. Colors indicate high level (green), medium level (yellow), and low level (red).

	COVID-19	AMTRENDS	ARTIFICIAL
Number of statements	🟩 140	🟥 5	🟥 10
Number of responses	🟩 ~2,000,000	🟨 10,169	🟩 51,240
Number of dem. groups with responses	🟨 756	🟥 12	🟩 1,708
Avg. # of personas generated	**50.6**	**1.3**	**92.4**

based on specific demographics or survey items in the dataset. This also addresses one of the perennial issues of persona generation – how many personas to generate [8]. Because the personas generated using S2P correspond directly to the designer's information needs, the number always falls in between a user-friendly range [33] while still being based on the personas' deviations from the average response behaviors in the data.

5.2 Limitations and Future Work

Concerning limitations and future research, despite justification we provided, our 'three to seven statements' condition is a heuristic decision. This range could be 2-5, or 3-10, with varying ramifications. Thus, a good research topic is: *what should be the optimal range for information in persona profiles?* We could vary this range to a reasonable extent and observe effects on designers' use of the personas. Also, there are issues with using mean-based SD for outlier detection, including poor performance with imbalanced (non-normal) data and small sample sizes [50]. We are aware of these issues and are working on addressing them, with one potential solution being MAD (mean absolute deviation) [27]. Finally, there are other challenges we have observed across different survey datasets: (a) surveys that are split by some answer choice, (b) binary or other categorical answers not following the Likert scale, (c) open-ended answers, (d) items that do not include statements or are formulated as questions, (e) latent constructs, i.e., multiple items belong to the same measure. Addressing these challenges will require further research and development. User studies are also required for understanding how designers use interactive persona systems such as S2P for data exploration and more user-centered decision making. The current iteration of the system is available at https://s2p.qcri.org.

6 CONCLUSION

This work presents the first step of automatically creating personas from survey datasets and presenting these personas to designers using an interactive online system. The core idea is when the demographic group's mean score in the characteristics of interest is higher than the global mean, it is considered an outlier. In this on-going work, we found that our approach for this goal works as intended, generating personas from multiple test datasets. At the same time, we also observed multiple further development points that this work communicates to the computing community. More development can increase the robustness of S2P for effective persona creation from survey data.

REFERENCES

[1] [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. *Soc. Netw. Anal. Min.* 8, 1 (2018), 54. DOI:https://doi.org/10.1007/s13278-018-0531-0

[2] [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 Trans. Web TWEB* 12, 4 (2018), 27. DOI:https://doi.org/10.1145/3265986

[3] [3] Eevelyn S. Arkush and Steven A. Stanton. 1988. Measuring the Value of End-User Computing. *J. Inf. Syst. Manag.* 5, 4 (January 1988), 62–63. DOI:https://doi.org/10.1080/07399018808962942

[4] [4] Michael S. Bernstein, Mark S. Ackerman, Ed H. Chi, and Robert C. Miller. 2011. The trouble with social computing systems research. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems.* 389–398.

[5] [5] Richard D. Boyce, Isabelle Ragueneau-Majlessi, Jingjing Yu, Jessica Tay-Sontheimer, Chris Kinsella, Eric Chou, Mathias Brochhausen, John Judkins, Brandon T. Gufford, Bruce E. Pinkleton, Rebecca Cooney, Mary F. Paine, and Jeannine S. McCune. 2018. Developing User Personas to Aid in the Design of a User-Centered Natural Product-Drug Interaction Information Resource for Researchers. *AMIA. Annu. Symp. Proc.* 2018, (December 2018), 279–287.

[6] [6] J. Brickey, S. Walczak, and T. Burgess. 2012. Comparing Semi-Automated Clustering Methods for Persona Development. *IEEE Trans. Softw. Eng.* 38, 3 (May 2012), 537–546. DOI:https://doi.org/10.1109/TSE.2011.60

[7] [7] John M. Carroll. 1997. Scenario-based design. In *Handbook of human-computer interaction.* Elsevier, 383–406.

[8] [8] Chris Chapman, Edwin Love, Russell P. Milham, Paul ElRif, and James L. Alford. 2008. Quantitative Evaluation of Personas as Information. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1107–1111. DOI:https://doi.org/10.1177/154193120805201602

[9] [9] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), Association for Computing Machinery, Honolulu, HI, USA, 1–13. DOI:https://doi.org/10.1145/3313831.3376461

[10] [10] Alan Cooper. 1999. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity* (1 edition ed.). Sams - Pearson Education, Indianapolis, IN.

[11] [11] Carl Ernst, Alexandre Bureau, and Gustavo Turecki. 2008. Application of microarray outlier detection methodology to psychiatric research. *BMC Psychiatry* 8, 1 (2008), 1–8.

[12] [12] Bruna Moraes Ferreira, Simone DJ Barbosa, and Tayana Conte. 2016. Pathy: Using empathy with personas to design applications that meet the users' needs. In *International Conference on Human-Computer Interaction*, Springer, 153–165.

[13] [13] Erin Friess. 2012. Personas and decision making in the design process: an ethnographic case study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1209–1218. DOI:https://doi.org/10.1145/2207676.2208572

[14] [14] Jonathan Grudin. 2006. Why Personas Work: The Psychological Evidence. In *The Persona Lifecycle*, John Pruitt and Tamara Adlin (eds.). Elsevier, 642–663. DOI:https://doi.org/10.1016/B978-012566251-2/50013-7

[15] [15] Richard J. Holden, Anand Kulanthaivel, Saptarshi Purkayastha, Kathryn M. Goggins, and Sunil Kripalani. 2017. Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *Int. J. Med. Inf.* 108, December (December 2017), 158–167. DOI:https://doi.org/10.1016/j.ijmedinf.2017.10.006

[16] [16] Prateek Jain, Soussan Djamasbi, and John Wyatt. 2019. Creating value with proto-research persona development. In *International Conference on Human-Computer Interaction*, Springer, 72–82.

[17] [17] Bernard J. Jansen, Joni Salminen, and Soon-gyo Jung. 2020. Data-Driven Personas for Enhanced User Understanding: Combining Empathy with Rationality for Better Insights to Analytics. *Data Inf. Manag.* 4, 1 (2020), 1–17. DOI:https://doi.org/10.2478/dim-2020-0005

[18] [18] Bernard Jansen, Joni Salminen, Soon-gyo Jung, and Kathleen Guan. 2021. *Data-Driven Personas* (1st ed.). Morgan & Claypool Publishers.

[19] [19] Soon-gyo Jung, Jisun An, Haewoon Kwak, Moeed Ahmad, Lene Nielsen, and Bernard J. Jansen. 2017. Persona Generation from Aggregated Social Media Data. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17), ACM, Denver, Colorado, USA, 1748–1755.

[20] [20] Soon-Gyo Jung, Joni Salminen, and Bernard J. Jansen. 2020. Giving Faces to Data: Creating Data-Driven Personas from Personified Big Data. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (IUI '20), Association for Computing Machinery, Cagliari, Italy, 132–133. DOI:https://doi.org/10.1145/3379336.3381465

[21] [21] Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen. 2021. All About the Name: Assigning Demographically Appropriate Names to Data-Driven Entities. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, Virtual conference. Retrieved from http://hdl.handle.net/10125/71108

[22] [22] 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*, ACM, New Jersey, USA, 321–324. DOI:https://doi.org/10.1145/3176349.3176893

[23] [23] Hae Min Kim and John Wiggins. 2016. A Factor Analysis Approach to Persona Development using Survey Data. In *Proceedings of the 2016 Library Assessment Conference*, 11.

[24] [24] Judith Koppehele-Gossel, Lisa Hoffmann, Rainer Banse, and Bertram Gawronski. 2020. Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *J. Exp. Soc. Psychol.* 87, (2020), 103905.

[25] [25] 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. *User Model. User-Adapt. Interact.* 30, 1 (March 2020), 81–125. DOI:https://doi.org/10.1007/s11257-019-09252-5

[26] [26] Mingyu Lee, Jiyoung Kwahk, Sung H. Han, Dawoon Jeong, Kyudong Park, Seokmin Oh, and Gunho Chae. 2020. Developing personas & use cases with user survey data: A study on the millennials' media usage. *J. Retail. Consum. Serv.* 54, (May 2020), 102051. DOI:https://doi.org/10.1016/j.jretconser.2020.102051

[27] [27] Christophe Leys, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 49, 4 (July 2013), 764–766. DOI:https://doi.org/10.1016/j.jesp.2013.03.013

[28] [28] 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*, ACM Press, Austin, Texas, USA, 1219. DOI:https://doi.org/10.1145/2207676.2208573

[29] [29] 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. DOI:https://doi.org/10.1145/1357054.1357292

[30] [30] T. Mijač, M. Jadrić, and M. Ćukušić. 2018. The potential and issues in data-driven development of web personas. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1237–1242. DOI:https://doi.org/10.23919/MIPRO.2018.8400224

[31] [31] G. A. Miller. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 2 (March 1956), 81–97.

[32] [32] Lene Nielsen. 2002. From User to Character: An Investigation into User-descriptions in Scenarios. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (DIS '02), ACM, London, England, 99–104. DOI:https://doi.org/10.1145/778712.778729

[33] [33] Lene Nielsen. 2019. *Personas - User Focused Design* (2nd ed. 2019 edition ed.). Springer, New York, NY, USA.

[34] [34] 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. *Int. J. Sociotechnology Knowl. Dev.* 7, 1 (2015), 45–61. DOI:https://doi.org/10.4018/ijskd.2015010104

[35] [35] Lene Nielsen and Kira Storgaard Hansen. 2014. Personas is applicable: a study on the use of personas in Denmark. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, Toronto, Ontario, Canada, 1665–1674.

[36] [36] Thomas V. Pollet and Leander van der Meij. 2017. To remove or not to remove: the impact of outlier handling on significance testing in testosterone data. *Adapt. Hum. Behav. Physiol.* 3, 1 (2017), 43–60.

[37] [37] Kari Rönkkö, Mats Hellman, Britta Kilander, and Yvonne Dittrich. 2004. Personas is Not Applicable: Local Remedies Interpreted in a Wider Context. In *Proceedings of the Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices - Volume 1* (PDC 04), ACM, Toronto, Ontario, Canada, 112–120. DOI:https://doi.org/10.1145/1011870.1011884

[38] [38] Joni Salminen, Kathleen Guan, Soon-gyo Jung, Shammur Absar Chowdhury, and Bernard J. Jansen. 2020. A Literature Review of Quantitative Persona Creation. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, Honolulu, Hawaii, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376502

[39] [39] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2021. A Survey of 15 Years of Data-Driven Persona Development. *Int. J. Human–Computer Interact.* 0, 0 (April 2021), 1–24. DOI:https://doi.org/10.1080/10447318.2021.1908670

[40] [40] Joni Salminen, Kathleen Guan, Lene Nielsen, Soon-gyo Jung, Shammur Absar Chowdhury, and Bernard J. Jansen. 2020. A Template for Data-Driven Personas: Analyzing 31 Quantitatively Oriented Persona Profiles. In *Human Interface and the Management of Information. Designing Information. HCII 2020.*, S. Yamamoto and H. Mori (eds.). Springer, Copenhagen, Denmark, 125–144.

[41] [41] Joni Salminen, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Findings of a User Study of Automatically Generated Personas. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, ACM Press, Montreal QC, Canada, 1–6. DOI:https://doi.org/10.1145/3170427.3188470

[42] [42] Joni Salminen, Soon-gyo Jung, Jisun An, Haewoon Kwak, Lene Nielsen, and Bernard J. Jansen. 2019. Confusion and information triggered by photos in persona profiles. *Int. J. Hum.-Comput. Stud.* 129, (September 2019), 1–14. DOI:https://doi.org/10.1016/j.ijhcs.2019.03.005

[43] [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)*, ACM, Honolulu, Hawaii, USA. DOI:https://doi.org/10.1145/3313831.3376770

[44] [44] Joni Salminen, Joao M. Santos, Soon-gyo Jung, Motahhare Eslami, and Bernard J. Jansen. 2019. Persona Transparency: Analyzing the Impact of Explanations on Perceptions of Data-Driven Personas. *Int. J. Human–Computer Interact.* 0, 0 (November 2019), 1–13. DOI:https://doi.org/10.1080/10447318.2019.1688946

[45] [45] Katharina Schäfer, Peter Rasche, Christina Bröhl, Sabine Theis, Laura Barton, Christopher Brandl, Matthias Wille, Verena Nitsch, and Alexander Mertens. 2019. Survey-based personas for a target-group-specific consideration of elderly end users of information and communication systems in the German health-care sector. *Int. J. Med. Inf.* 132, (December 2019), 103924. DOI:https://doi.org/10.1016/j.ijmedinf.2019.07.003

[46] [46] Felix D. Schönbrodt and Marco Perugini. 2013. At what sample size do correlations stabilize? *J. Res. Personal.* 47, 5 (2013), 609–612.

[47] [47] 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* (CHI EA '10), ACM, New York, NY, USA, 4721–4732. DOI:https://doi.org/10.1145/1753846.1754221

[48] [48] Phil Turner and Susan Turner. 2011. Is stereotyping inevitable when designing with personas? *Des. Stud.* 32, 1 (2011), 30–44.

[49] [49] Tammara Turner, Danah Boyd, Gary Burnett, Karen E. Fisher, and Tamara Adlin. 2007. Social types and personas: Typologies of persons on the web and designing for predictable behaviors. *Proc. Am. Soc. Inf. Sci. Technol.* 44, 1 (2007), 1–6.

[50] [50] Jiawei Yang, Susanto Rahardja, and Pasi Fränti. 2019. Outlier detection: how to threshold outlier scores? In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, ACM Press, Sanya, China, 1–6. DOI:https://doi.org/10.1145/3371425.3371427

[51] [51] 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* (CHI '16), ACM, San Jose, California, USA, 5350–5359.

[52] [52] Haining Zhu, Hongjian Wang, and John M. Carroll. 2019. Creating Persona Skeletons from Imbalanced Datasets - A Case Study using U.S. Older Adults' Health Data. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (DIS '19), ACM, New York, NY, USA, 61–70. DOI:https://doi.org/10.1145/3322276.3322285