A Literature Review of Quantitative Persona Creation

Joni Salminen[*†], **Kathleen Guan**[§], **Soon-gyo Jung**[*], **Shammur A. Chowdhury**[*], **Bernard J. Jansen**[*]
[*]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar [†]University of Turku, Turku, Finland, [§]Georgetown University, Washington, D.C., United States
{jsalminen, sjung, shchowhdury, bjansen}@hbku.edu.qa, kwg6@georgetown.edu

ABSTRACT

Quantitative persona creation (QPC) has tremendous potential, as HCI researchers and practitioners can leverage user data from online analytics and digital media platforms to better understand their users and customers. However, there is a lack of a systematic overview of the QPC methods and progress made, with no standard methodology or known best practices. To address this gap, we review 49 QPC research articles from 2005 to 2019. Results indicate three stages of QPC research: *Emergence*, *Diversification*, and *Sophistication*. Sharing resources, such as datasets, code, and algorithms, is crucial to achieving the next stage (*Maturity*). For practitioners, we provide guiding questions for assessing QPC readiness in organizations.

Author Keywords
Personas; literature review; quantitative persona creation

CSS Concepts
• **Human-centered computing~Human computer interaction (HCI)**

INTRODUCTION

Impact of personas

Defined as imaginary people describing real user segments [2], personas are consistently garnering interest from HCI researchers and practitioners in software development, design, marketing, healthcare, gaming, and other domains [47]. Personas are needed to go “beyond segmentation” [38] (p. 60) and to “give faces to data” [66] (p. 135). Personas provide “shared mental models” [6] (p. 63), facilitate team members’ communication about users, and help empathize with those using the outputs the organization creates [61]. In the era of “personified big data” [74] (p. 4019), personas are useful for segmenting large and diverse online audiences [65] and can bring about productivity benefits in organizations employing them [23].

Criticism of qualitative persona creation

Even though using mixed methods for persona creation is advocated in the HCI literature [62], the lack of time or resources often results in researchers and practitioners choosing either a purely quantitative or purely qualitative method [77]. Among the available choices, personas are most often created using qualitative data approaches. Brickey et al. [8] found that 81% of persona creation efforts reported in academic literature use qualitative techniques (e.g., interviews, field studies, usability tests, ethnography).

Despite this, purely qualitative persona generation has been widely criticized, with the main criticism being [71]:

- **High Cost:** Qualitative persona development typically requires several months of effort from start to finish and costs tens of thousands of dollars [23], which leaves the persona technique inaccessible for organizations with limited financial resources.
- **Lack of Objectivity and Rigor:** Because qualitative methods are flexible, the real-world applications of constructing personas may differ [90]. As mentioned by Jansen et al. [36], “Quantitative personas are seen as a way to overcome subjectivity both in interpretation and segmentation of available data.” (p. 2128).
- **Lack of Scaling:** Persona development that requires lots of manual labor adapts poorly to large datasets (“Big Data” [74]) that are increasingly common in organizations analyzing online user behavior [1]. The logic of scalability in QPC methods is adopted from machine learning; persona creators can annotate a small portion of the data (e.g., 10%) and let the algorithm classify the rest.
- **Non-Representative Data:** Qualitative personas often use small data not representing the entire user base [10].
- **Expiry:** Personas risk expiration whenever users or user behavior changes. Frequently changing behavior is typical for many online contexts, e.g., purchase behavior [44], search behavior [37], and content consumption [16, 42].

While qualitative and quantitative methods for personas have *each* been subject to criticism in their own right, some criticism is shared. These include (a) the risk of personas being abstract and inaccurate [10, 51], (b) personas simplifying complex human behaviors into simple archetypes that may be useful only to a degree [49,

CHI '20, April 25–30, 2020, Honolulu, HI, USA

ACM ISBN 978-1-4503-6708-0/20/04…$15.00
https://doi.org/10.1145/3313831.3376502

80], and (c) personas being "just" one method of user-centric design while other methods can be better in some use cases [55].

The promise of quantitative methods

Due to these shortcomings, quantitative persona creation (QPC) has gathered increasing interest from HCI scholars and practitioners alike [7, 43, 52, 54]. In the HCI research, QPC is contributing to the larger goal of creating more accurate and more compelling user archetypes from data.

We define QPC as follows: *using algorithmic methods to create accurate, representative, and up-to-date personas from numerical and textual data*. Besides addressing the shortcomings of qualitative methods, QPC can increase the scientific verifiability of personas, as well as their credibility for stakeholders, as QPC has the clout of using "real data" [73]. Ideally, quantitative personas are statistically representative, replicable, and verifiable – i.e., there is a metric that tells how well the specific method works [10, 73].

Researchers and practitioners are being pulled toward QPC also by the availability of online user data [1]. When personas were first introduced in the late 1990s, the Internet was still a nascent technology, and there were few tools to collect and process large amounts of user or customer data. The methodologies and platforms for collecting user data and automatically processing them have vastly developed. This development has dramatically increased the feasibility of QPC in online settings where personified big data about users or customers [15] can be collected through social media platforms and online analytics tools (e.g., Google / YouTube / Twitter Analytics and their APIs).

Simultaneously, data science tools have greatly evolved, including programming languages (e.g., R, Python) and libraries (e.g., *scikit-learn*), making a variety of statistical techniques and computational approaches accessible for persona creation. For textual data, natural language processing (NLP) provides a wide array of techniques, while numerical data can be analyzed using clustering, factor analysis, principal component analysis, and so on.

These developments have resulted in a "*shift from using qualitative data towards using quantitative data for persona development*" [57] (p. 1427). At the same time, there is considerable fragmentation of possible approaches to QPC, resulting in a need for providing an overview of the field. However, we could not locate a comprehensive review of QPC methods within the HCI literature. This gap makes it difficult for researchers to position their work or identify pivotal opportunities in the field. As reviewers, we have observed this problem first-hand, as Ph.D. students and other researchers getting familiarized with quantitative personas, struggle to submit research that would position their QPC work within the existing body of literature and the "methodological continuum".

Scope of the research

This systematic review seeks to examine research on QPC to (1) methodically collect, analyze, and synthesize all related literature within the QPC domain, (2) provide an overview of main QPC methods and their strengths and weaknesses, (3) understand the current status and evolution of the field, as well as (4) derive implications for future research and practice, including research agenda and guidelines. To this end, we formulate the following research questions:

- How has QPC research and methods developed over time?
- What are the key trends and gaps in QCP research?

Following the approach of previous literature reviews in HCI and computer science, we used Association of Computing Machinery's (ACM) Digital Library (DL) and Google Scholar databases to collect and analyzed 49 research articles that developed personas using quantitative methods published between January 2005 and August 2019.

RELATED LITERATURE

Methods of persona development

Mulder and Yaar [60] refer to three main ways of creating personas: (1) qualitative personas, (2) qualitative personas with quantitative validation, and (3) quantitative personas (i.e., QPC). Other HCI researchers refer to hybrid personas that use mixed methods (e.g., [62, 68]). Essentially, all persona creation methods are based on four main steps: (a) data collection, (b) segmentation and grouping, (c) analysis of the qualitative and/or quantitative data, and (d) creating/writing persona profiles to present the user segments and their attributes as user archetypes [87, 90].

A short history of QCP

The concept of "data-driven persona" is first mentioned by Williams in 2006 [86] and popularized by McGinn and Kotamraju [52] in 2008. The purpose of being "data-driven" goes further back in the HCI literature – in fact, personas were always intended to use real data about the user. As Gaiser et al. [25] (p. 521) note, "*In order to fulfill standards of a scientific method, personas can't be created arbitrarily. Personas have to be grounded in data, at best, both qualitative and quantitative data of surveys with the target audience.*" Similarly, Pruitt and Grudin [62] (p. 1) argue that *"[personas] provide a conduit for conveying a broad range of qualitative and quantitative data, and focus attention on aspects of design and use that other method do not*."

Placing QPC into the historical context, its major drivers are (a) the availability and abundance of user data and (b) the rapid development of data analysis algorithms that have changed since the early days of personas.

Previous literature reviews

The methodological diversity within QPC has been widely noted in the literature. For example, Zhu et al. [90] cite several methods: *affinity diagrams*, *decision trees*, *exploratory factor analysis* (EFA), *hierarchical clustering*, *k-means clustering*, *latent semantic analysis* (LSA), *multidimensional scaling analysis* (MSA), and *weighted graphs*. Minichiello et al. [59] provide a similar list of semi-automated methods: cluster analysis (including both hierarchical and k-means), factor analysis, principal component analysis (PCA), and LSA. These reviews, however, are superficial, as they typically only list the methods and do not discuss them further.

In the few literature reviews that assess QPC methods [7, 8, 79], the focus is on clustering, thereby ignoring the methodological diversity under the QPC umbrella. In addition, there are some conceptual articles that discuss the role of personas in the era of online analytics [65], list methodological arguments against qualitative personas [10] or quantitative ones [73] (the former typically questioning the rigor of qualitative methods and the latter warning about the "mystique of numbers" in disguising flawed quantitative segmentation), or provide guidelines for successful persona creation [62]. However, the articles of this type neither focus on quantitative person creation nor provide a systematic methodology to review the work in this domain.

Establishing the research gap

We could locate no prior systematic literature review focused on QPC, given the diversity of its methods. Previous articles that provide a more thorough literature review focus on the use of clustering [7, 8, 79]. This is not ideal, as the diversity of the methods mentioned in other articles clearly indicates that there are several other quantitative techniques that have been applied for QPC and, therefore, a degree of fragmentation that should be investigated more thoroughly.

However, the articles pointing out the methodological plurality only superficially cite the methods, without presenting a detailed overview of them, their popularity, or their strengths and weaknesses. Thus, there is a need to systematically map these attempts to provide useful insights for both persona researchers and practitioners. As noted by Dillahunt et al. [18], "*literature reviews have proved useful and influential by identifying trends and gaps in the literature of interest and by providing key directions for short- and long-term future work.*" (p. 1).

METHODOLOGY

To find the articles, two databases were chosen based on their coverage (Google Scholar) and relevance for the topic of personas/HCI (ACM DL). Identical literature searches were carried out for each database on June 2019. For ACM DL, we used the actual website[1]. For Google Scholar, we used the Publish or Perish software[2] previously employed in systematic literature reviews [48]. Snowball sampling was also used to detect additional articles, as suggested for systematic literature reviews [63]. *Supplementary Material* includes a detailed description of the literature searches, with this section giving an overview of the process.

The search phrases were devised based on the authors' previous knowledge of the field and included references to QPC ("quantitative personas", "data-driven personas", "procedural personas") as well as to specific methodologies ("automatic persona generation", personas + cluster analysis | clustering | conjoint analysis | factor analysis | latent semantic analysis | matrix factorization | principal component analysis). Both plural and singular of the word "persona" were used. To focus on English-speaking articles, we included negative search words in Spanish ('y', 'con', 'de'), as "persona" is the Spanish word for person.

The initial search yielded 149 articles, of which 116 (78%) from ACM DL and 33 (22%) from Google Scholar. The results were combined and manually de-duplicated by a research assistant. The deduplicated articles (N=138) were manually screened by reading the abstracts. The articles passing the screening were read in full and further assessed to ensure the inclusion criteria were met (see Figure 1).

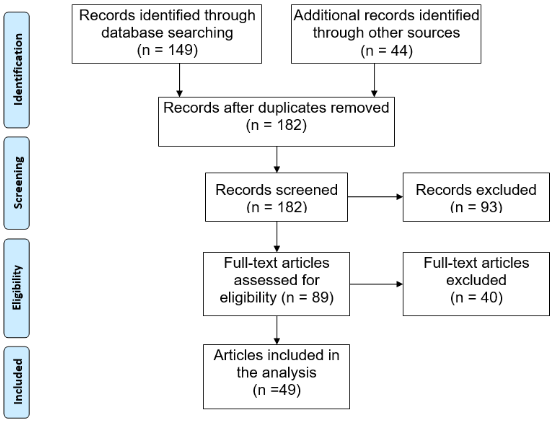

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 149)
Additional records identified through other sources (n = 44)
Records after duplicates removed (n = 182)
Records screened (n = 182)
Records excluded (n = 93)
Full-text articles assessed for eligibility (n = 89)
Full-text articles excluded (n = 40)
Articles included in the analysis (n =49)

Figure 1: PRISMA flow chart [82] of the literature collection

The assessment took place by reading the full articles. More articles were retrieved by applying snowball sampling based on reading the screened articles and identifying other research articles that applied QPC. All the articles retrieved via snowball sampling (N=44) were assessed by reading the full article. The inclusion criteria included:

- **full research article** (no short articles, books or theses) [*screening stage*]

[1] https://dl.acm.org

[2] https://harzing.com/resources/publish-or-perish

- **published in peer-reviewed journal or conference** [*screening*]
- **written in English language** [*screening*]
- **empirical paper that develops personas using quantitative data** [*screening/assessment*]

The total number is 149 search-retrieved + 44 snowball-retrieved = 193 considered records. These contain 182 unique articles; thus N=11 (5.7%) were duplicates. In addition, we discarded non-English articles (N=11, 6.0% of the unique articles), non-peer-reviewed articles (N=40, 22.0%), non-full articles (N=24, 13.2%), and articles not actually developing quantitative personas (N=80, 44.0%). In total, 133 articles (73.1%) were excluded (note that summing up the class percentages does not match this number because a paper can have many exclusion criteria.).

The final collection includes 49 articles, of which 23 (46.9%) were retrieved via searches and 26 (53.1%) via snowball sampling. Of the search articles, 15.4% were kept and of the snowball articles, 59.1%. The following information was extracted from the articles using a standardized data extraction form [78]:

- article information: title, year, keywords, publication venue and type (conference/journal)
- authors' institution locations (countries)
- use of quantitative methodology and mixed methods
- data source (e.g., survey, social media)
- data size (number of analysis units, e.g., participants)
- validation metrics and methods
- authors' suggestions for future work

The extracted data was analyzed, and the findings are presented in the following sections.

RESEARCH INTEREST IN QPC

Research interest over time

The earliest paper applying QPC that we could locate was published in 2005 by Aoyama [3]. The researcher applied conjoint analysis to create personas for software embedded in digital consumer products. Figure 2 shows a stagnating number of QPC articles per year at first, and then a steep increase since 2014. In 2018, publication count reached its peak at N=11 articles (2017 saw only eight articles). Note that the 2019 articles (N=1) are omitted from this figure because, at the time of writing, the full year has not passed.

Publication types

Conference articles are more common (N=36, 73%) than journal articles (N=13, 27%), perhaps reflecting the strong position of conference venues in computer science research. The ACM CHI Conference on Human Factors in Computing Systems (N=4) is the only publication venue with more than two articles. The early development of QPC is characterized by a lack of journal publications. The first journal publication took place at IEEE Transactions on Software Engineering in 2012 [8]. In 2019, there were more journal than conference publications for the first time in the history of QPC.

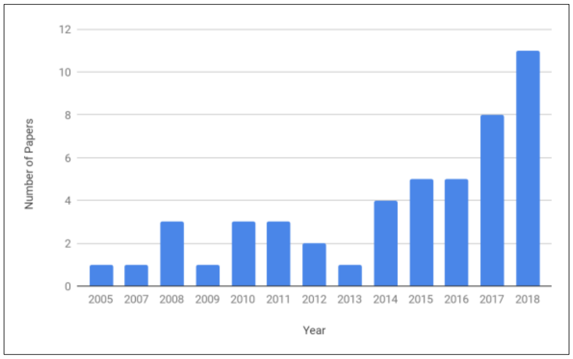


Figure 2: Number of QPC research articles over time

Prominent work

The citation numbers for the most prominent work (see Table 1) were retrieved from Google Scholar in July 2019 (max = 330, min = 0, mean = 26 citations). The mode is 0 citations; 8 articles (16.3%) of the articles have no citations at all. There is a weak positive correlation between the years-of-age of the paper and the number of citations (r = 0.26). The most cited paper [45] uses dialogues from online data to develop persona-based conversation models. McGinn and Kotamraju's [52] paper represents a seminal work of persona generation based on users of big data analytics software.

Title and Year	Authors	Citations
A Persona-Based Neural Conversation Model (2016)	Li et al. [45]	330
Data-Driven Persona Development (2008)	McGinn and Kotamraju [52]	114
Learning Latent Personas of Film Characters (2013)	Bamman et al. [5]	108
Defining Personas in Games Using Metrics (2008)	Tychsen and Canossa [81]	108
Persona-and-Scenario Based Requirements Engineering for Software Embedded in Digital Consumer Products (2005)	Aoyama [3]	85
Persona-Scenario-Goal Methodology for User-Centered Requirements Engineering (2007)	Aoyama [4]	63
A Latent Semantic Analysis Methodology for the Identification and Creation of Personas (2008)	Miaskiewicz et al. [54]	54
Evolving personas for player decision modeling (2014)	Holmgard et al. [33]	45
Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry (2016)	Zhang et al. [89]	30
Invoking the User from Data to Design (2014)	Tempelman-Kluit and Pearce [76]	27

Table 1: Most cited articles of QPC research (Top 10)

METHODS FOR QPC

Data sources

The most typical source for data collection is using surveys, with 55% of the articles reporting the use of surveys. The second most popular data source is the use of web and social media data (27% of the articles). This category includes sourcing data from social media platforms (e.g., YouTube [2]), discussion forums [34], as well as user clicklogs [77] and telemetry [89]. Interestingly, two articles used device-collected data, including GPS signal [27] and comfort levels [72]. Even though the use of device-collected data is currently marginal, "personal big data" provides interesting information about users, e.g., health and wellness.

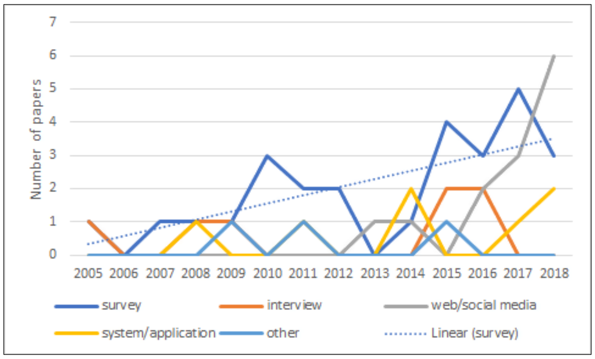


Figure 3: Popularity of data techniques for QPC over time

Survey data (blue line in Figure 3) has been consistently popular format of data; however, the focus is shifting from survey to web data (grey line in Figure 3), with web/social media data seeing a rise since 2015 (virtually non-existent before) and 2018 marks the first year that web data exceeds survey data in QPC.

Also, behavioral data describing actual user interactions is becoming more common [58]. Seven articles (14.3%) use more than one data source. The most common data source association is surveys and interviews (N=5, 71.4% of the multiple data sources). Multiple data sources are used for creating hybrid personas by combining qualitative and quantitative data [68]. This is seen to enhance both the breadth (through quantitative data) and depth (through qualitative data) of the personas.

Popularity of methods

To extract the frequency of the methods, we manually tallied the individual methods mentioned in each article. The most popular method (see Table 2) was k-means clustering (N=11, 22.4%), followed by hierarchical clustering (N=6, 12.2%). In total, clustering methods were used in more than a third of the articles (N=17, 34.6%) [4, 5, 79]. Nonetheless, there was a great variety in methods used, with many articles introducing new models such as the neural speaker model developed [45], the Dirichlet persona model [5], the ego-splitting algorithm [20], and more.

In addition, 19 articles (38.8%) combined quantitative and qualitative methods, and 27 (55.1%) combined multiple quantitative methods (e.g., k-means clustering with principal component analysis). While no specific combination of quantitative methods dominated, combinations often included at least one type of clustering analysis.

Method	Description	Frequency
K-means Clustering (KMC)	Machine learning algorithm that classifies a dataset using a predetermined prime number (k) of clusters.	N=11 (22.4%)
Hierarchical Clustering (HC)	Machine learning algorithm that computes distances between different elements to produce clusters in a hierarchical order based on similarity.	N=6 (12.2%)
Principal Component Analysis (PCA)	Linear dimension-reduction algorithm used to extract information by removing non-essential elements with relatively fewer variations.	N=5 (10.2%)
Latent Semantic Analysis (LSA)	Machine learning algorithm that uses singular value decomposition to detect hidden semantic relationships between words.	N=5 (10.2%)
Non-negative Matrix Factorization (NMF)	Matrix factorization method in which matrices are constrained as non-negative. A matrix is decomposed into two matrices to extract sparse and meaningful features.	N=4 (8.2%)

Table 2: Most popular QPC analytics methods

Quantitative evaluation of QPC methods

Validation of quantitative personas varies by the method applied. **KMC** was validated by calculating the Euclidean distance between the different variables [75, 84] or by conducting Chi-squared tests [75]. A few articles [83, 89, 90] qualitatively validated clusters by engaging subject matter experts as well as users in reviewing the clustering results.

For **HC**, Miaskiewicz et al. [54] and Mesgari et al. [53] both validated their results by looking at the relations between variables within clusters; the former calculated cosine similarity of the angles between pairs of non-zero vectors, while the latter calculated Pearson correlation (the extent of linear relationship between two variables). Holden et al. [32] validated their results with Kruskal-Wallis test and Welch's ANOVA to determine statistical significance between different variables as well as a test for variance, respectively.

PCA was often used in combination with others; in fact, all the articles that applied PCA complemented it with at least another quantitative method. As a result, validation metrics also varied, including Cohen's kappa (a statistical measure of interrater agreement of generated and expert-created clusters) [7, 8], Euclidean distances of different variables [84], Spearman's correlation [13], and even qualitative review with survey participants [79].

Similar to PCA, **LSA** is often combined with other methods, especially hierarchical cluster analysis [7, 8, 54].

Researchers validated their results through cosine similarity tests. Cosine similarity was also used by An et al. [2] to validate the results of **NMF** by calculating it for pairs of personas until the closest pairs were determined. In another study employing NMF [1], researchers used the Kendall rank correlation coefficient to compare the ranking of personas' demographic groups (DGs) with that of DGs in the raw data.

Qualitative evaluation of QPC methods

Use of qualitative validation is common. Of the 19 articles that used mixed methods, 8 (42.1%) incorporated qualitative methods to the validation stage only, while 7 (36.8%) incorporated qualitative methods to both initial data collection and validation. As seen in Figure 4, mixed quantitative-qualitative methodologies have consistently been incorporated, with peaks in 2010 and 2015 (in proportion to the total number of articles published per year). These peaks may be attributed to rises in popularity of incorporating qualitative aspects to validation, such as expert or user consultations after data analyses are complete.

In total, 15 articles (30.6%) incorporated qualitative feedback to their persona validation stages. These generally involved re-gathering members of the initially surveyed population to evaluate the quantitatively generated personas in a focus group setting. An exception is Dupree et al. [19] who recruited an additional population group familiar with the paper's context to anecdotally evaluate the relevance and representativeness of the generated personas. The individuals were tasked with self-identifying with one of the final five personas and rating how realistic they are.

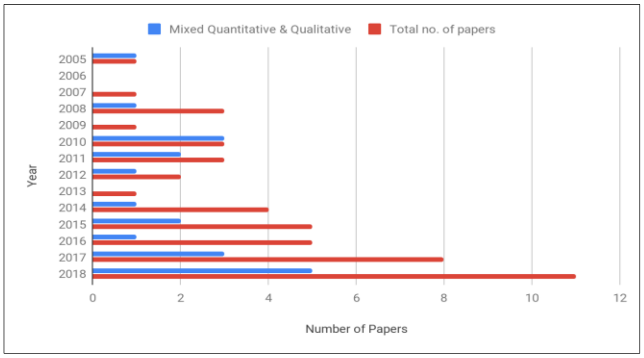


Figure 4: Articles using mixed methods over time

Generally speaking, the validation tends to be informal and not thoroughly described in the articles. Only a couple of authors had a formal process in place. Out of these, Hirskyj-Douglas et al. [30] included open-ended questions in their data collection survey so dog owners could elaborate on their pets' backgrounds and daily activities. Miaskiewicz and Luxmoore [56] systematically identified specific surveyed users to represent the personas and further interview based on k-means distance measures; afterward, they quantitatively compared these individuals' characteristics with the generated personas.

	Emergence (2005–2008)	Diversification (2009–2014)	Sophistication (2015–present)
Conceptual emphasis	• focus on purely quantitative approaches as "the new thing"	• innovations in hybrid approaches (both for mixed qual-quant and quant-quant) • reaching "self-awareness" (via literature reviews)	• fragmentation • expansion to non-humans (animal personas, robot personas) • interaction between persona users and persona models/systems
Method emphasis	• experiments with multiple methods (no "dominant method") • applying well-established quantitative methods: factor analysis, descriptive statistics, cluster analysis	• mixed methods rising: combining qualitative and quantitative approaches • quantitative side dominated by clustering • beginning of deploying NLP techniques	• mixed method: multiple quantitative methods • introduction of matrix factorization and deep learning; standardized hybrid approaches (quant. and qual. *and* mixed quant.) • expanding behavioral personas
Data emphasis	• surveys, interviews, statistics	• surveys and qualitative enrichment • behavioral data	• web and social media data; online analytics platforms and APIs • combining textual and numerical data; census data (large-scale surveys); personal big data
Context emphasis	• focus on software development and engineering • personas for games and e-commerce introduced	• diverse contexts, e.g., gaming, knowledge management, emergency preparedness	• expansion to new domains: health informatics, privacy, social media, journalism, and fashion
Venue emphasis	• focus on conferences • research volume low	• focus on conferences • research volume low	• focus conferences and journals • research volume increasing

Table 3: Development of QPC research

Salminen et al. [68] consulted qualitative data of social media users in the geographical region in the forms of Instagram public profiles and semi-structured interviews. These were used to enrich further and improve the automatically generated personas.

Some articles also consulted subject-matter experts during the validation stages [19, 52]; these evaluations varied in informality and ranged from brief discussions to quantitative coding of interrater agreement levels. However, the extent to which observations from the expert evaluations led to modifications in the personas was unspecified in all the articles mentioning this form of validation.

EVOLUTION OF QPC RESEARCH

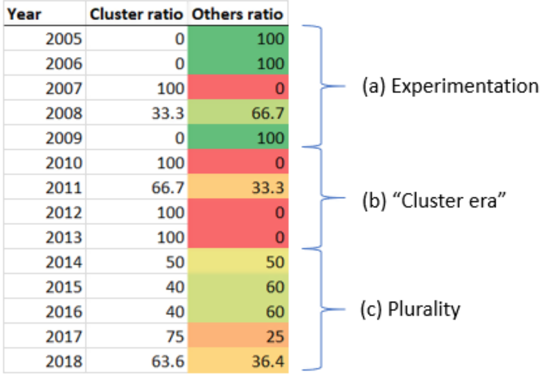
Synthesizing our results (see Table 3), QPC research is divided into three periods: (1) Emergence (2005–2008) that consists of early development and trials, (2) Diversification (2009–2014) that can be seen as a transition period that saw the beginning of some transformations that would be more established in the current third stage, and (3) Sophistication (2015–present) that marks the revitalized interest in QPC.

Emergence: The first stage is marked by a focus on the basics: establishing the need for quantitative methodologies in persona domain [52] and experimenting with different methods, especially those well-known in quantitative research tradition. The contextual focus is on software development, especially requirements engineering [3, 4]. There is also experimentation with using clickstreams and statistics from gaming software [81], even though the main focus is on the use of survey data.

Diversification: In the second stage, contexts expand, but the methods stale. Clustering becomes the dominant method for a few years (see Figure 5). However, there are first experiments with NLP techniques [5]. The field reaches a degree self-awareness, marked by literature reviews focused on different clustering methods [8]. Introduction of behavioral data takes place [50], and simulation is first attempted with personas [40]. From the second stage, QPC personas have gradually been used for analyzing different demographic segments, such as Vietnamese youth [13] and European senior citizens [87]. In such research, personas are merely a means to an end (i.e., understanding the data), not the focus of the research.

Sophistication: In the third stage, researchers expand the notion of behavior; not only for behavioral data [1] but also for using behavioral theories for interpreting quantitative personas [36]. Deep learning is applied to make personas interactive [45] using sophisticated neural networks [11] and new data sources, most notably web and social media data, emerge. Research starts to pay attention to the longitudinal aspect of personas evolving over time [33]. Health context is introduced [32, 83], along with other new domains.

The goal of fully automated persona generation emerges [2] with an associated system development that enables persona users to interact with the personas [1]. In the third stage, clustering remains popular but is no longer dominant; rather, researchers apply multiple quantitative methods simultaneously (see Figure 5).



Year	Cluster ratio	Others ratio
2005	0	100
2006	0	100
2007	100	0
2008	33.3	66.7
2009	0	100
2010	100	0
2011	66.7	33.3
2012	100	0
2013	100	0
2014	50	50
2015	40	60
2016	40	60
2017	75	25
2018	63.6	36.4

(a) Experimentation
(b) “Cluster era”
(c) Plurality

Figure 5: Clustering vs. other methods – the figure shows the percentage of articles using clustering vs. other methods per year. These periods roughly match with the three stages.

While clustering methods are consistently popular, a rise in other methods has been observed since 2014 (see Figure 5). The year 2018 saw the least proportion of articles conducting cluster analyses since 2015. This can be attributed to new models and methodologies, such as the Dirichlet Persona Model [5] and non-negative matrix factorization [1, 2].

Dataset sizes (means and medians) are increasing (see Table 4). The standard deviation also increases, indicating that, in the third era, researchers still use small datasets, but they are now also using larger datasets.

	Emergence (2005–2008)	Diversification (2009–2014)	Sophistication (2015–present)
Mean	343	2,034 (493%)	**14,447 (610%)**
Max	1,300	12,496 (861%)	**170,704 (1266%)**
Median	31	100 (223%)	**435 (335%)**
SD	638	4,003 (527%)	**39,141 (878%)**

Table 4: Survey sample sizes of QPC studies over time. Highest values bolded, growth to the previous period in parentheses.

Interestingly, more data does not necessarily result in more rounded personas, since some of the rich narrative-like personas were generated as early as 2005 (see Figure 8).

In the third stage, there is also an increase in publication numbers relative to earlier years. The first and second stages are characterized by relatively low level of research (see Figure 2), but comparatively, the research volume is increasing in the third stage, with the Sophistication stage averaging 7.25 publications per year, a 211% increase over Diversification (mean = 2.33) and 480% increase over Emergence stage (mean = 1.25). Moreover, the first and second stages are marked by the popularity of survey data; while survey data remains popular in the third stage, online data sources are gaining momentum.

The self-awareness beginning in the second stage has reached maturity, with researchers acknowledging the challenges of QPC [57, 71]. Accumulated experience over the use of methods has helped paint a broader picture of the field. These include at least (1) *data quality* (as in: "garbage in, garbage out"), (2) *data availability* (meaning that information to create useful personas is not always available and platforms constantly change their rules about what data they share), (3) *method-specific weaknesses* (e.g., clustering not reflecting multiple demographics per behavior type [1], and (4) *fallacy of perfection* (i.e., high expectancy of automation and objectivity, whereas the methods require judgment calls like setting the right number of clusters).

As knowledge on QPC has increased, the weaknesses and shortcomings of the methods are also becoming more known. Therefore, the third stage is characterized both by promise and trust in the potential, as well as eagerness to address the outstanding challenges.

RESEARCH TRENDS

RT1: Higher degrees of automation

Overall, 5 articles (10.2%) use application programming interfaces (APIs) to collect data for persona creation. However, social media is more widely used. The sources include WeChat user data [84], YouTube Analytics [1, 2, 68], Google Analytics [57], Twitter FireHose [45], and Wikipedia [5]. The most common social media platform was YouTube (N=4). The advantages of APIs are aligned with the benefits of QPC, including speed, updatability, volume, and cost [14]. In addition, data structures of online platforms regarding user attributes are similar, meaning the same methods can be applied across data sources [2]. The API usage is an increasing trend, as 3 out of 5 articles using APIs are from 2018. We expect API-based data collection for personas to become more common in the future. Using pre-existing data is highly lucrative for persona developers due to time and cost benefits [90].

Several authors [20, 35, 56] express plans to further refine and automate their methods, even to fully automated persona generation [1, 2]. However, these attempts are still on-going – as noted by Mijač et al. [57]: "*Examples of an automatic update of personas are scarce and even those are not fully implemented but are rather on the level of proof-of-concept.*" (p. 1431). Salminen et al. [71] provide a roadmap for automatic persona generation, and complete QPC systems appear achievable near term.

Attempting the goal of full automation means that there is increasing complexity in methods and system architectures, as more and more computational techniques are needed to discern specific nuances of online audiences. For example, there may be a need for one algorithm to detect demographic attributes, another one for behaviors, and a third one for persona's pain points. This shows in an increase of articles that apply multiple quantitative methods (see Figure 6).

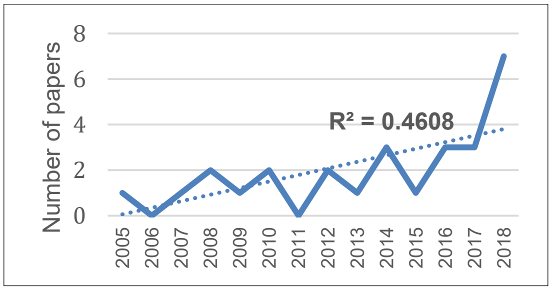
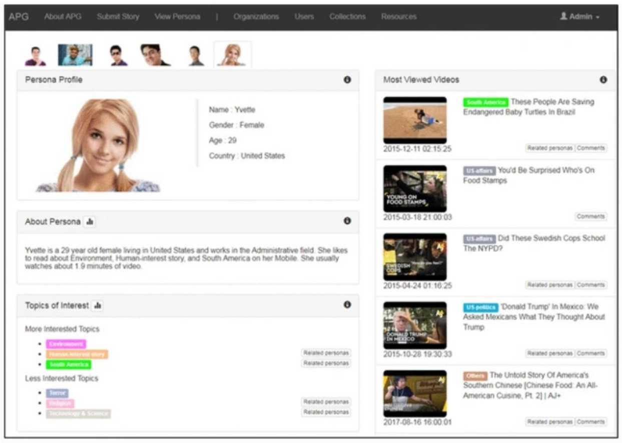


Figure 6: Use of multiple quantitative methods over time

RT2: Interactive persona systems

Interactivity – i.e., persona users interacting with personas – marks another important research trend. This trend is reflected by the development of systems toward real-time creation of personas where users can choose the data from which the personas are generated and how personas to generate [2, 57, 65]. This line of work can result in "customizable" or "tailored" personas based on the specific needs of the persona users (see Figure 7).



Name : Yvette
Gender : Female
Age : 29
Country : United States

Yvette is a 29 year old female living in United States and works in the Administrative field. She likes to read about Environment, Human-interest story, and South America on her Mobile. She usually watches about 1.9 minutes of video.

More Interested Topics
- Environment
- Human-interest story
- South America

Less Interested Topics
- Terror
- Religion
- Technology & Science

South America These People Are Saving Endangered Baby Turtles In Brazil 2015-12-11 02:15:25

US affairs You'd Be Surprised Who's On Food Stamps 2015-03-18 21:00:03

US affairs Did These Swedish Cops School The NYPD? 2015-04-24 01:16:25

US politics 'Donald Trump' In Mexico: We Asked Mexicans What They Thought About Trump 2015-10-28 19:30:33

Others The Untold Story Of America's Southern Chinese [Chinese Food: An All-American Cuisine, Pt. 2] | AJ+ 2017-08-16 16:00:01

Figure 7: Persona system with interactive elements (2018) [68]

Interactivity can also be of help in addressing some of the QCP challenges. For example, Brickey et al. [8], Bamman et al. [5], and Holden et al. [32] highlight the limitations in contextualizing personas when it comes to unexpected outliers and deciding which traits are actually applicable. To alleviate this challenge, Zhang et al. [89], Tychsen and Canossa [81], and Miaskiewicz et al. [54] have suggested incorporating user evaluations of the personas to the validation stages in order to capture the relevant yet also comprehensive traits. Giving users agency and "power" over persona information design can solve some of the black box problems of separating the persona creation to algorithms and applications to humans.

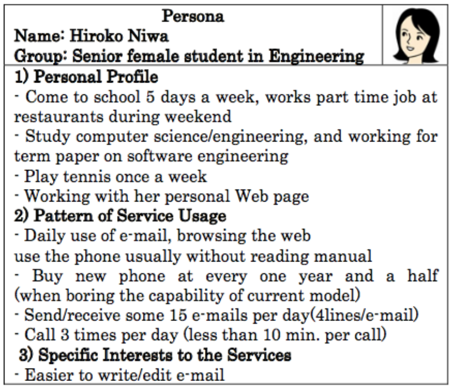
Another line of work is personas for chatbots or dialogue systems emerging in the field of NLP [11, 45]. These "chatbot personas" can enable interactive discussions for querying the sentiments and opinions of the fictional (but data-driven) user segments they reflect. The goal is to capture distinct conversational styles that reflect the personality of the persona. Systems can then be built to allow users to "speak" to personas.

Yet another line of work in interactivity is the "procedural persona" in gaming – these personas are virtual agents that are able to make real-time decisions based on environmental stimuli, reflecting a specific game-playing type of users (e.g., "monster hunter") [33]. Procedural personas enable game developers to test how personas (i.e., archetypical player types) react to changes in the game world.

Finally, some studies aim at *predicting* persona preferences, e.g., how they would respond to a simulated or real news story [2] or lifestyle articles such as fashion [17]. Conceptually, the studies pave the way for merging personas with recommendation system research and also involve notable commercial opportunities.

RT3: Interplay between automatic and manual

One of the consistent trends is the combination of quantitative and qualitative persona creation methods. These include incorporating qualitative components to the data collection [30, 52, 77] and validation stages [3, 4, 19, 56]. A common approach is to use quantitative data to explore user behavior and enhance these behavioral archetypes ("skeletons", "templates") with qualitative insights to create and finalize more rounded personas [58, 68] (see Figure 8).



Persona
Name: Hiroko Niwa
Group: Senior female student in Engineering

1) Personal Profile
- Come to school 5 days a week, works part time job at restaurants during weekend
- Study computer science/engineering, and working for term paper on software engineering
- Play tennis once a week
- Working with her personal Web page

2) Pattern of Service Usage
- Daily use of e-mail, browsing the web
use the phone usually without reading manual
- Buy new phone at every one year and a half (when boring the capability of current model)
- Send/receive some 15 e-mails per day(4lines/e-mail)
- Call 3 times per day (less than 10 min. per call)

3) Specific Interests to the Services
- Easier to write/edit e-mail

Figure 8: Narrative persona profile (2005) [3]

In general, manual work steps for QPC include at least (1) algorithmic choices (e.g., choosing the "right" number of personas for clustering/NMF); (2) writing the persona descriptions/narratives (i.e., "transferring data into narrative persona descriptions" [87]); and (3) evaluating personas' usefulness, credibility, and other persona user perceptions.

Our analysis of the challenges of QPC points out that (a) the challenges of qualitative persona creation do not "go away" with QPC, and (b) qualitative methods can be used for addressing the QPC challenges (as is done *vice versa*). For example, the lack of in-depth can be addressed using qualitative methods to collect and analyze information about users' pain points and motivations [32].

Finally, one re-emerging theme is usefulness/validation of QPC. Articles throughout Emergence [77], Diversification [9], and Sophistication [56] deal with the aspect of generating real value for organizations and individuals with QPC, as well as struggles with organizational adoption.

NOTABLE RESEARCH GAPS

Standards and best practices. Due to the divergence of the methods, there is no unified metric for measuring the quality of quantitative personas, apart from preliminary attempts to create a standardized questionnaire for measuring user perceptions of the personas [69]. In the absence of quality standards, researchers struggle to benchmark their results. The lack of standardization of QPC methods (i.e., there not being one standard methodology but instead many) makes the quality of different methods difficult to compare. For example, clustering is evaluated with a different metric than matrix factorization. Having a unified way to compare different methods would enable benchmarking of results and clear demarcation of scientific progress.

Ethics of QPC. Few articles mention ethical considerations such as data privacy, algorithmic transparency, and risk of creating personas that represent averages or majority groups rather than diversity (resulting from the way the applied statistical methods tend to work). Data privacy is mentioned in one article stating that online platform datasets are typically aggregated, preserving the privacy of individual users [87]. However, using social media data "in the wild" might have issues of informed consent [21]. Particularly, using social media data presents confidentiality risks for participants, as users can be directly identified through profile characteristics or quotes. Persona creators should be aware that harm from online research can occur for classes of people and communities [31].

Moreover, QPC articles are exceedingly focusing on "core users", "representative segments", or other forms of majority users. What is clear from our findings is that the authors of QPC articles tend to consider inclusivity from the perspective of statistics, not from the perspective of fairness. Interestingly, this goes against the "mainstream" persona research, with inclusivity, stereotypes, and "fringe personas" being recognized as increasingly important [29, 49].

These issues matter for both the HCI community and the organizations analyzing user behavior, as useful insights on usability/user experience can often be found in outliers and minority segments. Toward this end, new QPC approaches (e.g., outlier detection for personas) are needed. While many of the articles did pose inclusivity in their future work sections, this was more in terms of improving statistical representativeness, i.e., what characteristics are being mistaken as "fringe" but are nonetheless highly relevant.

Loss of user immersion? The current body of research does not answer if something is lost in QPC relative to qualitative persona creation. The reason why this might take place is that the iterative, inductive process of qualitative persona creation is seen to increase user immersion by itself [12, 46, 61]. Often, end users of personas *co-create* the personas with HCI professionals, which can enhance the shared mental models among team members [12, 61]. Because QPC techniques typically differ drastically from this workshop-driven, collaborative process, it is worthwhile to ask if these positive aspects are lost and, if so, how they could be retained without losing the other positive aspects of QPC.

TAKEAWAYS (ESPECIALLY) FOR RESEARCHERS

We have separated implications to researchers and practitioners, with the former focusing on development of research practices of QPC and the latter on applicability.

Reaching maturity. To reach the next step of QPC research – *Maturity* – the following is needed from HCI researchers:

- *show progress* – e.g., conduct replication studies applying the method to different datasets or different methods to the same dataset. This implies sharing datasets, code, and algorithms.
- *conduct comparative studies* to assess different methods by their technical merits and the overlaps/deviations of the resulting personas
- *conduct formal evaluation studies* to assess both accuracy (internal validation) and impact (external validation) of the created personas.

Building a research community. Many of the gaps in the current body of research could be addressed by building a stronger research community around QPC. This could take place by organizing workshops, networking/meetups, or even via establishing a *special interest group* (SIG) of QPC.

Setting baseline methods. Based on its popularity, k-means clustering could be used as a baseline method for QPC, meaning that results would be compared to k-means output. Replication of previous results and showing progress – in terms of both technical accuracy and practical usefulness – are critical aspects for the progress in QPC research.

Target real use cases and measure. The authors in the QPC articles suggested ways of going beyond mere persona creation to testing the usefulness of the personas in meeting stakeholder goals [26, 56, 64, 75, 85]. These can take the form of longitudinal studies on how the personas are adapted, used, and implemented by stakeholders, and to what effect. For example, in healthcare, such initiatives would involve designing tailored medical interventions to subpopulations represented by the personas and evaluating how health outcomes develop over time [75, 83]. Some studies show promise using longitudinal data and standardized algorithms to compare persona sets over time [39] and organizational units [88]. Evaluation of QPC can be inspired by qualitative persona evaluation studies [24, 51].

TAKEAWAYS (ESPECIALLY) FOR PRACTITIONERS

If in doubt, cluster. Clustering is the most common method of choice. These techniques, including k-means clustering, hierarchical clustering, and others, are well-established and can be combined with other methods such as EFA or PCA in the data exploration stage or qualitative methods in the persona writing stage. However, clustering does include some limitations discussed earlier in this manuscript. Other methods, such as NMF, can partially address these concerns, but each method involves some degree of subjectivity.

Avoid "mystique of numbers". One should not blindly believe the outputs of statistical methods. Additional steps, such as ensuring data quality and triangulating the results with other methods, such as qualitative interviews, are necessary. Therefore, practitioners with limited knowledge about quantitative methods should "ask stupid questions" to avoid the "mystique of numbers" [73], including asking clarification about how the personas were created, what manual choices the creation process involved, and how the results were validated. Being critical pays off.

Consider human bias. Surveys are the most popular data format for QPC. However, even when analyzed quantitatively, survey data may include several issues of validity (e.g., social desirability bias [22]), especially relative to behavioral data. In a similar vein, setting the number of personas, applying hyperparameters for algorithms and other steps that involve human judgment are subject to human bias. Therefore, "quantitative" does not automatically mean "objective" or truthful, which is critical to acknowledge.

Consider "algorithmic bias". The community is becoming increasingly aware of algorithmic biases, meaning that data and algorithms may introduce undesired generalizations into the personas [28, 67]. Relying solely on quantitative data might lead to ignoring minority groups and inclusivity [49], as statistical methods tend to "favor" majority groups and obscure the outliers and deviations within user groups. Sometimes these outliers would be interesting, like the most loyal users that comprise only a small portion of the whole but have a decisive impact. To counter this, QPC applications can, e.g., split the dataset into "majority" and "minority" and generate personas separately for each.

Quantity of data does not automatically mean better quality. Any biases and errors in the data are inherited to personas. For example, when generating personas from online analytics data, the measurement error is unknown. QPC represents "best efforts" to make use of available data; however, the data sources should not be blindly trusted. To increase trust in quantitative personas, creators can (a) apply triangulation by independent samples to corroborate personas and (b) increase "persona transparency" [70] including clear statements of where the data originates, how it was collected, and what were the analysis steps that resulted in the visible personas.

Validate for both accuracy and usefulness. Personas, both qualitative and quantitative, should ideally be validated for accuracy (i.e., truthfulness vis-a-vis real user base) and usefulness (i.e., do they serve decision makers' goals). Specific persona validation methods mentioned by Minichiello et al. [58] include on-site visits, dissemination, and feedback from persona users, log file verification, and persona user and usage observations.

Is QPC for you? Organizations are encouraged to consider the following questions before initiating QPC projects:

- **Do you offer products/services in online environments?** (e.g., e-commerce, social media)
- **Do you have a large and diverse user/customer base?** (e.g., international audience, patient population)
- **Have you collected digital information on your users/customers?** (e.g., CRM system, Web log files, electronic health records, etc.)
- **Are the user attributes you are interested in easily quantifiable?** (e.g., engagement with online content)

If the answers to these questions are mostly positive, QPC techniques can be beneficial for enhanced user insights.

These guiding questions are important for mapping the *QPC readiness* of an organization and for avoiding conflated expectations about the applicability of QPC. In some cases, especially when deep understanding about the goals and motivations of the users are needed, qualitative persona creation may be more applicable than pure-form QPC. Naturally, mixed methods can also be applied to enhance quantitative personas with qualitative insights.

Canonical sources of QPC methods are as follows: factor analysis [41], clustering [8], and matrix decomposition [2]. The role personas amidst online analytics has also been discussed in adjunct work [65, 71] and a research roadmap for automatic persona generation has been proposed [66].

CONCLUSION

Quantitative persona generation lacks shared resources and benchmarks. Technical trends highlight automation and interactivity, while human aspects highlight the need for ethical considerations. Qualitative approaches remain important as a source of supporting information and evaluation. The suggestions we give to researchers (about sharing resources) and practitioners (about considering QPC readiness) can help people navigate this space.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comments that significantly enhanced the quality of this manuscript.

REFERENCES

[1] An, J. et al. 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). DOI:https://doi.org/10.1007/s13278-018-0531-0.

[2] An, J. et al. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Transactions on the Web (TWEB)*. 12, 4 (2018), Article No. 27. DOI:https://doi.org/10.1145/3265986.

[3] Aoyama, M. 2005. Persona-and-scenario based requirements engineering for software embedded in digital consumer products. *Proceedings of the 13th IEEE International Conference on Requirements Engineering (RE'05)* (Washington, DC, USA, Aug. 2005), 85–94.

[4] Aoyama, M. 2007. Persona-Scenario-Goal Methodology for User-Centered Requirements Engineering. *Proceedings of the 15th IEEE International Requirements Engineering Conference (RE 2007)* (Delhi, India, Oct. 2007), 185–194.

[5] Bamman, D. et al. 2013. Learning Latent Personas of Film Characters. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Sofia, Bulgaria, 2013), 10.

[6] Blanco, E. et al. 2014. Role of personas and scenarios in creating shared understanding of functional requirements: an empirical study. *Design Computing and Cognition'12*. Springer. 61–78.

[7] Brickey, J. et al. 2010. A Comparative Analysis of Persona Clustering Methods. *AMCIS 2010 Proceedings* (2010).

[8] Brickey, J. et al. 2012. Comparing Semi-Automated Clustering Methods for Persona Development. *IEEE Transactions on Software Engineering*. 38, 3 (May 2012), 537–546. DOI:https://doi.org/10.1109/TSE.2011.60.

[9] Chapman, C. et al. 2015. Profile CBC: Using Conjoint Analysis for Consumer Profiles. *Sawtooth Software Conference Proceedings* (2015).

[10] Chapman, C.N. and Milham, R.P. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Oct. 2006), 634–636.

[11] Chu, E. et al. 2018. Learning Personas from Dialogue with Attentive Memory Networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct. 2018), 2638–2646.

[12] Cooper, A. 1999. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams - Pearson Education.

[13] Dang-Pham, D. et al. 2015. Demystifying online personas of Vietnamese young adults on Facebook: A Q-methodology approach. *Australasian Journal of Information Systems*. 19, 0 (Nov. 2015). DOI:https://doi.org/10.3127/ajis.v19i0.1204.

[14] De Souza, C.R. et al. 2004. How a good software practice thwarts collaboration: the multiple roles of APIs in software development. *ACM SIGSOFT Software Engineering Notes*. 29, 6 (2004), 221–230.

[15] Del Vecchio, P. et al. 2017. Creating value from Social Big Data: Implications for Smart Tourism Destinations. *Information Processing & Management*. (2017).

[16] Del Vicario, M. et al. 2017. News consumption during the Italian referendum: A cross-platform analysis on facebook and twitter. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2017), 648–657.

[17] Dhakad, L. et al. 2017. SOPER: Discovering the Influence of Fashion and the Many Faces of User from Session Logs using Stick Breaking Process. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17* (Singapore, Singapore, 2017), 1609–1618.

[18] Dillahunt, T.R. et al. 2017. The sharing economy in computing: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction*. 1, CSCW (2017), 38.

[19] Dupree, J.L. et al. 2016. Privacy Personas: Clustering Users via Attitudes and Behaviors Toward Security Practices. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), 5228–5239.

[20] Epasto, A. et al. 2017. Ego-Splitting Framework: From Non-Overlapping to Overlapping Clusters. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2017), 145–154.

[21] Fiesler, C. and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media+ Society*. 4, 1 (2018), 2056305118763366.

[22] Fisher, R.J. 1993. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*. 20, 2 (1993), 303–315.

[23] Forrester Research 2010. *The ROI Of Personas*.

[24] Friess, E. 2012. Personas and Decision Making in the Design Process: An Ethnographic Case Study. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), 1209–1218.

[25] Gaiser, B. et al. 2006. Community Design-The Personas Approach. *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (2006), 520–525.

[26] Goodman-Deane, J. et al. 2018. Evaluating Inclusivity using Quantitative Personas. (Jun. 2018).

[27] Guo, A. and Ma, J. 2018. Archetype-Based Modeling of Persona for Comprehensive Personality Computing from Personal Big Data. *Sensors*. 18, 3 (Mar. 2018), 684. DOI:https://doi.org/10.3390/s18030684.

[28] Hajian, S. et al. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), 2125–2126.

[29] Hill, C.G. et al. 2017. Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways? *Proceedings of the 2017 CHI Conference* (Denver, Colorado, USA, 2017), 6658–6671.

[30] Hirskyj-Douglas, I. et al. 2017. Animal Personas: Representing Dog Stakeholders in Interaction Design. *Proceedings of the 31st British Computer Society Human Computer Interaction Conference* (Swindon, UK, 2017), 37:1–37:13.

[31] Hoffmann, A.L. and Jonas, A. 2016. Recasting justice for Internet and online industry research ethics. *Internet Research Ethics for the Social Age: New Cases and Challenges. M. Zimmer and K. Kinder-Kuranda (Eds.), np Bern, Switzerland: Peter Lang, Forthcoming*. (2016).

[32] Holden, R.J. et al. 2017. Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *International Journal of Medical Informatics*. 108, (Dec. 2017), 158–167. DOI:https://doi.org/10.1016/j.ijmedinf.2017.10.006.

[33] Holmgard, C. et al. 2014. Evolving personas for player decision modeling. *Computational Intelligence and Games (CIG), 2014 IEEE Conference on* (2014), 1–8.

[34] Huh, J. et al. 2016. Personas in online health communities. *Journal of Biomedical Informatics*. 63, (Oct. 2016), 212–225. DOI:https://doi.org/10.1016/j.jbi.2016.08.019.

[35] Ishii, R. et al. 2018. Monte-Carlo Tree Search Implementation of Fighting Game AIs Having Personas. *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (Maastricht, Aug. 2018), 1–8.

[36] Jansen, A. et al. 2017. Personas and Behavioral Theories: A Case Study Using Self-Determination Theory to Construct Overweight Personas. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA, 2017), 2127–2136.

[37] Jansen, B.J. et al. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*. 36, 2 (2000), 207–227.

[38] Jenkinson, A. 1994. Beyond segmentation. *Journal of targeting, measurement and analysis for marketing*. 3, 1 (1994), 60–72.

[39] Jung, S. et al. 2019. Personas Changing Over Time: Analyzing Variations of Data-Driven Personas During a Two-Year Period. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, UK, 2019), LBW2714:1–LBW2714:6.

[40] Kanno, T. et al. 2011. Integrating Human Modeling and Simulation with the Persona Method. *Universal Access in Human-Computer Interaction. Users Diversity* (2011), 51–60.

[41] Kim, H.M. and Wiggins, J. 2016. A Factor Analysis Approach to Persona Development using Survey Data. *Proceedings of the 2016 Library Assessment Conference* (2016), 11.

[42] Kwak, H. et al. 2018. What We Read, What We Search: Media Attention and Public Attention Among 193 Countries. *Proceedings of the Web Conference* (Lyon, France, 2018).

[43] Laporte, L. et al. 2012. Using Correspondence Analysis to Monitor the Persona Segmentation Process. *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (New York, NY, USA, 2012), 265–274.

[44] Leong, L.-Y. et al. 2017. Understanding impulse purchase in Facebook commerce: does Big Five matter? *Internet Research*. 27, 4 (2017), 786–818.

[45] Li, J. et al. 2016. A Persona-Based Neural Conversation Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany, Aug. 2016), 994–1003.

[46] Long, F. 2009. Real or imaginary: The effectiveness of using personas in product design. *Proceedings of the Irish Ergonomics Society Annual Conference* (2009).

[47] Luo, Y. et al. 2019. Co-Designing Food Trackers with Dietitians: Identifying Design Opportunities for Food Tracker Customization. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (Glasgow, Scotland Uk, 2019), 1–13.

[48] Mari, M. and Poggesi, S. 2013. Servicescape cues and customer behavior: a systematic literature review and research agenda. *The Service Industries Journal*. 33, 2 (2013), 171–199.

[49] Marsden, N. and Haag, M. 2016. Stereotypes and Politics: Reflections on Personas. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, USA, 2016), 4017–4031.

[50] Masiero, A.A. et al. 2011. Multidirectional Knowledge Extraction Process for Creating Behavioral Personas. *Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction* (Porto Alegre, Brazil, Brazil, 2011), 91–99.

[51] Matthews, T. et al. 2012. How Do Designers and User Experience Professionals Actually Perceive and Use Personas? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA, 2012), 1219–1228.

[52] McGinn, J.J. and Kotamraju, N. 2008. Data-driven persona development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, 2008), 1521–1524.

[53] Mesgari, M. et al. 2015. Affordance-based User Personas : A mixed-method Approach to Persona Development. *AMCIS 2015 Proceedings* (Jun. 2015).

[54] Miaskiewicz, T. et al. 2008. A latent semantic analysis methodology for the identification and creation of personas. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008), 1501–1510.

[55] Miaskiewicz, T. and Kozar, K.A. 2011. Personas and user-centered design: How can personas benefit product design processes? *Design Studies*. 32, 5 (2011), 417–430.

[56] Miaskiewicz, T. and Luxmoore, C. 2017. The Use of Data-Driven Personas to Facilitate Organizational Adoption–A Case Study. *The Design Journal*. 20, 3 (2017), 357–374.

[57] Mijač, T. et al. 2018. The potential and issues in data-driven development of web personas. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (May 2018), 1237–1242.

[58] Minichiello, A. et al. 2018. Bringing User Experience Design to Bear on STEM Education: A Narrative Literature Review. *Journal for STEM Education Research*. 1, 1–2 (2018), 7–33.

[59] Minichiello, A. et al. 2017. *Work In Progress: Methodological Considerations for Constructing Nontraditional Student Personas with Scenarios from Online Forum Usage Data in Calculus*. Technical Report #Paper ID #17980. American Society for Engineering Education.

[60] Mulder, S. and Yaar, Z. 2006. *The User is Always Right: A Practical Guide to Creating and Using Personas for the Web*. New Riders.

[61] Nielsen, L. 2019. *Personas - User Focused Design*. Springer.

[62] Pruitt, J. and Grudin, J. 2003. Personas: Practice and Theory. *Proceedings of the 2003 Conference on Designing for User Experiences* (San Francisco, California, USA, 2003), 1–15.

[63] Radjenović, D. et al. 2013. Software fault prediction metrics: A systematic literature review. *Information and software technology*. 55, 8 (2013), 1397–1418.

[64] Rahimi, M. and Cleland-Huang, J. 2014. Personas in the Middle: Automated Support for Creating Personas As Focal Points in Feature Gathering Forums. *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering* (New York, NY, USA, 2014), 479–484.

[65] Salminen, J. et al. 2018. Are personas done? Evaluating their usefulness in the age of digital analytics. *Persona Studies*. 4, 2 (Nov. 2018), 47–65. DOI:https://doi.org/10.21153/psj2018vol4no2art737.

[66] Salminen, J. et al. 2019. Automatic Persona Generation for Online Content Creators: Conceptual Rationale and a Research Agenda. *Personas - User Focused Design*. L. Nielsen, ed. Springer London. 135–160.

[67] Salminen, J. et al. 2019. Detecting Demographic Bias in Automatically Generated Personas. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), LBW0122:1–LBW0122:6.

[68] Salminen, J. et al. 2018. From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *First Monday*. 23, 6 (Jun. 2018). DOI:https://doi.org/10.5210/fm.v23i6.8415.

[69] Salminen, J. et al. 2018. Persona Perception Scale: Developing and Validating an Instrument for Human-Like Representations of Data. *CHI'18 Extended Abstracts: CHI Conference on Human Factors in Computing Systems Extended Abstracts Proceedings* (Montréal, Canada, 2018).

[70] Salminen, J. et al. 2019. Persona Transparency: Analyzing the Impact of Explanations on Perceptions of Data-Driven Personas. *International Journal of Human–Computer Interaction*. 0, 0 (Nov. 2019), 1–13. DOI:https://doi.org/10.1080/10447318.2019.1688946.

[71] Salminen, J. et al. 2019. The future of data-driven personas: A marriage of online analytics numbers and human attributes. *ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems* (Heraklion, Greece, Jan. 2019), 596–603.

[72] dos Santos, T.F. et al. 2014. Behavioral persona for human-robot interaction: a study based on pet robot. *International Conference on Human-Computer Interaction* (2014), 687–696.

[73] Siegel, D.A. 2010. The Mystique of Numbers: Belief in Quantitative Approaches to Segmentation and Persona Development. *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2010), 4721–4732.

[74] Stevenson, P.D. and Mattson, C.A. 2019. The Personification of Big Data. *Proceedings of the Design Society: International Conference on Engineering Design*. 1, 1 (Jul. 2019), 4019–4028. DOI:https://doi.org/10.1017/dsi.2019.409.

[75] Tanenbaum, M.L. et al. 2018. From Wary Wearers to d-Embracers: Personas of Readiness to Use Diabetes Devices. *Journal of Diabetes Science and Technology*. 12, 6 (Nov. 2018), 1101–1107. DOI:https://doi.org/10.1177/1932296818793756.

[76] Tempelman-Kluit, N. and Pearce, A. 2014. Invoking the User from Data to Design. *College & Research Libraries*. 75, 5 (Sep. 2014), 616–640. DOI:https://doi.org/10.5860/crl.75.5.616.

[77] Thoma, V. and Williams, B. 2009. Developing and Validating Personas in e-Commerce: A Heuristic Approach. *Human-Computer Interaction – INTERACT 2009* (2009), 524–527.

[78] Torgerson, C. 2003. *Systematic Reviews*. A&C Black.

[79] Tu, N. et al. 2010. Using cluster analysis in Persona development. *2010 8th International Conference on Supply Chain Management and Information* (Oct. 2010), 1–5.

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

[81] Tychsen, A. and Canossa, A. 2008. Defining Personas in Games Using Metrics. *Proceedings of the 2008 Conference on Future Play: Research, Play, Share* (Toronto, Ontario, Canada, 2008), 73–80.

[82] Van Laar, E. et al. 2017. The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in human behavior*. 72, (2017), 577–588.

[83] Vosbergen, S. et al. 2015. Using personas to tailor educational messages to the preferences of coronary heart disease patients. *Journal of Biomedical Informatics*. 53, (Feb. 2015), 100–112. DOI:https://doi.org/10.1016/j.jbi.2014.09.004.

[84] Wang, L. et al. 2018. Analysis of Regional Group Health Persona Based on Image Recognition. *2018 Sixth International Conference on Enterprise Systems (ES)* (Oct. 2018), 166–171.

[85] Watanabe, Y. et al. 2017. ID3P: Iterative Data-driven Development of Persona Based on Quantitative Evaluation and Revision. *Proceedings of the 10th International Workshop on Cooperative and Human Aspects of Software Engineering* (Piscataway, NJ, USA, 2017), 49–55.

[86] Williams, K.L. 2006. *Personas in the design process: a tool for understanding others*. Georgia Institute of Technology.

[87] Wöckl, B. et al. 2012. Basic Senior Personas: A Representative Design Tool Covering the Spectrum of European Older Adults. *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2012), 25–32.

[88] Zaugg, H. and Ziegenfuss, D.H. 2018. Comparison of personas between two academic libraries. *Performance Measurement and Metrics*. 19, 3 (Aug. 2018), 142–152. DOI:https://doi.org/10.1108/PMM-04-2018-0013.

[89] Zhang, X. et al. 2016. Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, 2016), 5350–5359.

[90] Zhu, H. et al. 2019. Creating Persona Skeletons from Imbalanced Datasets - A Case Study using U.S. Older Adults' Health Data. *Proceedings of the 2019 on Designing Interactive Systems Conference - DIS '19* (San Diego, CA, USA, 2019), 61–70.