IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 July 2026, accepted 6 September 2026, date of publication 10 September 2026, date of current version 16 September 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3732878

RESEARCH ARTICLE

Proposing a Quantitative Framework for Persona Set Diversity Measurement as a Pathway to Inclusive User Representation

DANIAL AMIN[1,2], JONI SALMINEN[1], AND BERNARD J. JANSEN[3,4]
[1]University of Vaasa, 65200 Vaasa, Finland
[2]Otherkind Tech, 00100 Helsinki, Finland
[3]Department of Information Management, Peking University, Beijing 100871, China
[4]School of Information and Communication, Nankai University, Tianjin 300350, China
Corresponding author: Danial Amin (danialam@uwasa.fi)

ABSTRACT Personas represent user segments to support design decisions. A persona set, the collection of personas built from a user population, should cover the range of segments in that population. However, quantifying persona-set diversity remains understudied. To address this shortcoming, we introduce the Persona Set Diversity Measurement (PSDM) task and systematically evaluate five diversity metrics for versatility, consistency, holism, and scalability. Our experiments across three diversity levels (low, medium, high) of simulated persona datasets indicate that Rao's Quadratic Entropy performed best among the five evaluated metrics under the tested conditions, a finding supported by additional tests with a real-world social media dataset on user attitudes. We make the source code available to quantify persona profile information from observed attributes and to measure diversity in any quantifiable persona set. This contribution provides researchers and practitioners with empirical tools to quantify and compare attribute variation across persona sets.

INDEX TERMS Data-driven, diversity, persona, quantitative, user representation.

I. INTRODUCTION

Persona is a fictional but realistic humanized representation of a user segment, originating from human-computer interaction (HCI) and widely used in design, product development, marketing, healthcare, and education [1], [2], [3], [4], among other areas. Personas represent real user segments, typically through *persona profiles* that comprise key attributes, such as demographics, behaviors, goals, and pain points, providing an empathetic decision-making instrument for designers and other stakeholders [1], [5]. Persona profiles, containing key attributes (see Fig. 1), are often generated as part of a collection called a *persona set*. In a persona set, each persona represents a segment of the overall user population, and it is important that different personas in the set represent distinct segments.

Persona set diversity is the range of variation in demographic and behavioral attributes among personas within a persona set that represents a specific user population [6]. A high-diversity persona set helps designers consider a wider range of user perspectives, circumstances, and needs, which supports inclusive design [7], [8]. Persona sets must therefore represent the diversity of the user base accurately to support inclusive design.

Despite this role, *research lacks a framework for quantifying diversity in persona sets using standardized metrics* [8], [9], while prior persona evaluation research has focused on metrics such as accuracy, validity, and credibility [10], [11]. Given that personas guide design decisions and diverse persona sets support inclusive design, standardized evaluation methods are needed to ensure that the persona set provides a diverse representation of the user base [9]. Thus, measuring persona set diversity can provide designers with empirical evidence of how much the observed attributes vary within and across persona sets [12], [13].

Therefore, there is a need for **P**ersona **S**et **D**iversity **M**easurement (**PSDM**), which quantifies the diversity of

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

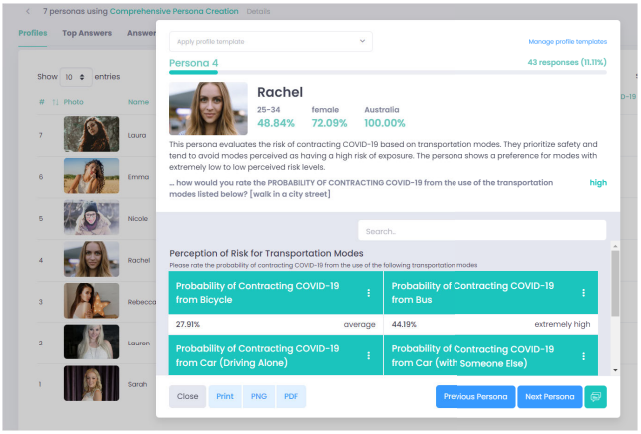


FIGURE 1. An example persona created from a survey dataset using Survey2Persona, a publicly available persona generation tool. The example illustrates that personas typically have a mixture of categorical and quantitative data. Multiple personas are created from the same dataset (seven in this example case), and diversity can exist within and between the personas at a set level.

a persona set by measuring the variation in the personas' attributes. Specifically, PSDM is defined as:

> **Exhibit 1. Persona Set Diversity Measurement (PSDM).** The quantitative measurement of the diversity of a persona set, where each persona comprises categorical, numerical, and ordinal attributes, yielding a value that is higher when attribute values diverge across personas and lower when they converge.

PSDM measures variation among the attributes observed in a persona set. It does not determine whether the source data include all relevant user groups or whether the personas adequately represent the target population. These questions require a separate assessment of the sampling frame, data collection, and persona-generation process. To the best of our knowledge, a standardized framework for measuring mixed-type attribute variation within persona sets is missing from the quantitative persona literature [14], [15], [16], [17], [18].

While persona sets can be defined in different ways, they can be modeled as information. Thus, *quantification of user representation* is necessary to evaluate the diversity of persona sets. When persona sets are modeled as information [15], this information can be analyzed to understand how well a persona set captures the range of attributes that appear in the user population [16]. There have been efforts to measure the diversity of attributes within a persona set. However, previous attempts to quantify persona set diversity tend to count the number of unique values across persona attributes [16], which becomes unreliable as cardinality grows. Moreover, persona profiles typically comprise multiple data types that reflect different user characteristics and behaviors [19]. These data types in a persona can be broadly categorized into *categorical* (e.g., gender, country, occupation, pain points), *numerical* (e.g., age, income, usage frequency), and *ordinal* (e.g., Likert scale responses to attitudes or preferences, etc.) variables.

Measuring persona set diversity, therefore, requires metrics that handle the quantitative nature of persona attributes. In the absence of persona-specific diversity metrics in HCI research, we adapt established metrics from ecology and economics, which have long traditions of measuring diversity in populations and species. Adapting these measures gives the assessment of persona sets against a tested methodological basis [20], which matters as design processes increasingly rely on representations of diverse user populations [21], [22]. To address the gap in quantitative measurement of diversity for persona sets, we pose three research questions (RQs):

- **RQ1:** *How to quantify persona set information for diversity measurement?*
- **RQ2:** *How to measure the attribute diversity of persona sets?*
- **RQ3:** *How to assess the suitability of metrics for persona set diversity measurement?*

Through systematic investigation of these questions, we (1) define the PSDM task, (2) develop a quantification framework for categorical, numerical, and ordinal attributes, (3) propose four requirements for evaluating metric suitability, and (4) compare one metric from each of five methodological families. We also apply the framework to personas generated from real survey data as a proof of concept. To support reproducibility, we share the implementation and persona datasets with controlled levels of attribute diversity as low, medium, and high. The resulting framework enables researchers and practitioners to quantify and compare attribute variation across persona sets generated from the same or comparable data.

II. RELATED WORK

The diversity of user representation is a prerequisite for inclusive design, that is, designing technologies that are equitable, inclusive, and relevant to a wide range of people [23]. Failing to consider the diversity and multifaceted nature of users can lead to technological exclusion [24], in which some segments of society cannot benefit from technological advancements. To combat this, researchers have developed a variety of quantitative and qualitative methods to assess diversity in user representation, including (1) demographic diversity [25] (e.g., age, gender, race, ethnicity, socioeconomic status, educational attainment, disability status, and geographic location), (2) social identity (e.g., membership in a non-demographic marginalized group), (3) (neuro)cognitive diversity (i.e., variety in how people think, perceive, and process information [26]), (4) behavioral diversity [27], [28] (i.e., differences in how users interact with a product or service), and (5) attitudinal and perceptual diversity [29] (how users' opinions, beliefs, preferences, and satisfaction levels vary).

Methods for measuring diversity range from statistical dispersion measures (e.g., standard deviation) to more elaborate

analyses, such as subgroup comparisons [30]. These are often considered in relation to the provision of findings about group-wise differences. However, some studies also focus on diversity itself, for example, using entropy-based metrics from information theory [31] to quantify uncertainty and variety among users. A higher entropy value indicates a more even distribution of users across groups, and thus greater diversity. In contrast, if most of the users belong to a single group, the entropy will be low.

Moreover, metrics can be assigned numerical values based on distances between groups (e.g., geographically closer user groups would contribute less to overall diversity than geographically farther groups). For example, ethnic fractionalization can be measured either as the simple proportion of different ethnic groups in a population or as the distance between groups.

Building on entropy- and distance-aware diversity measures, recent bias-aware screening systems show that modeling individual screening preferences can surface and mitigate cognitive biases in practice [32]. At the same time, studies comparing protected attributes with their proxy counterparts find that explanations can expose direct biases, while biases operating through correlated proxy attributes often persist [33]. These results point to the need for representational balance across multiple attributes and their interdependencies. For PSDM, these findings motivate considering multiple attributes and their interdependencies when evaluating persona-set diversity.

Persona diversity has been quantified in previous work by counting the number of unique attributes in a persona set, such as demographic groups or age categories [16]. This count can be converted to a ratio by comparing the unique groups in the persona set to the total unique groups in the baseline data, indicating how well the persona set covers existing or possible groups [11]. For conversational personas, researchers have measured the diversity of the personas' communicative expressions [11]. Similarly, researchers have measured the lexical diversity, i.e., the diversity of the language used in narrative personas [34]. These measures established that persona sets can be treated as information and evaluated quantitatively. They have generally focused on the range of attribute categories without considering how different those categories are from one another. That kind of diversity can be captured through distance-based measures, which we consider in this study.

To measure distance-aware diversity, we draw inspiration from ecology and economics, which measure diversity in populations composed of distinct members [35], [36]. The analogy applies to personas, as each persona in a persona set corresponds to a distinct member of the population. Thus, the measures used in ecology and economics can be used for PSDM. These measures can handle diverse members with mixed attributes, which also aligns well with the structure of a persona set. They were, however, designed for a different setting than PSDM requires.

Diversity has also been measured in other fields such as information retrieval and recommender systems, where the goal is to reduce repetition among the items recommended to a user [31]. PSDM differs in this respect because a persona set has no ranking and comprises distinct personas drawn from a single population. Similarly, in algorithmic fairness, diversity is measured through exposure across predefined groups, checking whether each group is represented fairly [12], [37]. While this is important for inclusive design, it does not apply to PSDM, as the groups in a persona set are not predefined. We therefore adapt population-level measures from ecology and economics that fit the mixed-attribute diversity required for a persona set.

Although these existing metrics form a nascent contribution to persona set measurement, they do not address pertinent issues in quantifying *persona sets* based on information in *persona profiles*, such as multiple data types in persona attributes, interdependencies between attributes, and coping with dimensionality when the number of attributes increases [15], [38]. Our research extends this work by introducing a methodological framework that addresses these issues through the adaptation of established diversity metrics.

The prevailing persona diversity measure counts the unique attribute values in a set and converts that count into a coverage ratio relative to the baseline data [11], [16]. This measure establishes that persona sets can be treated as information, but the count treats every distinct value as equally different from every other; for example, it registers that two persona sets contain different age brackets without indicating how far apart those brackets are. Two sets with the same number of distinct brackets score alike whether the ages span a decade or a lifetime. The count also becomes unstable as cardinality increases and does not combine cleanly when a persona includes categorical, numerical, and ordinal attributes at once [16]. A diversity measure for persona sets should register the magnitude of difference between attribute values, not only their number.

Distance-aware measures capture this magnitude, but measures developed in other fields were built for a different setting than the one a persona set presents. Ecology and economics measure diversity across populations of single-attribute members [35], [36], not across personas that carry correlated attributes from several data types. Information retrieval and recommender systems measure diversity to reduce repetition within a ranked list [31], whereas a persona set is unranked and its personas are meant to differ. Algorithmic fairness measures exposure across predefined protected groups [12], [37], whereas a persona set has no predefined groups to check exposure across. How distance-aware measures behave when adapted to the mixed-attribute, unranked, group-free structure of a persona set has not been systematically evaluated in persona research. We address this question in this study. Table 1 summarizes how PSDM relates to the main measurement approaches discussed above.

TABLE 1. PSDM in relation to related measurement approaches.

Approach	What it measures	Key difference from PSDM
Persona coverage counts [11], [16]	Presence or coverage of attribute values	PSDM measures variation among the observed attribute values, not only their presence.
Ranked-list diversity [31]	Diversity among ranked items	PSDM evaluates an unranked set of personas.
Predefined-group fairness [12], [37]	Representation or outcomes across specified groups	PSDM does not require predefined groups and measures variation among persona attributes.
Mixed-data dissimilarity	Pairwise differences between mixed-type records	PSDM produces a set-level diversity value rather than only pairwise distances.

III. METHODOLOGY

A. THE PSDM TASK

We measure persona set diversity by operationalizing the PSDM task as defined in Section I and shown in Exhibit 1. The input to the PSDM task is a set of personas, each with categorical, numerical, or ordinal attributes. The output of the PSDM task is a single diversity value that increases as attribute values vary across personas and decreases as they converge. The rest of this section specifies how we simulate persona sets, quantify their attributes, and select metrics to compute this value.

B. CONTROL AND BASELINE SIMULATIONS

To systematically approach the PSDM task, a controlled data simulation process (see Algorithm 1) was developed to create persona sets with predetermined levels of diversity. Each simulated persona consists of five key attributes, except when testing for the dimensionality requirement (see Section IV-D). These attributes (see Table 2) represent a mix of demographic and other characteristics common in persona profiles [39].

TABLE 2. Persona attributes used in our study.

Attribute	Type	Range/Categories
Age	Num.	18–80 years
Gender	Cat.	Male, Female, Non-binary
Occupation	Cat.	Student, Engineer, Teacher, Artist, Manager, Entrepreneur
Income	Num.	20,000–150,000
Tech Savviness	Ord.	5-point Likert scale

Algorithm 1 Simulated Persona Set Generation for PSDM Evaluation

Input: Diversity levels $L = \{\text{Low, Medium, High}\}$, Attributes A
Output: Persona sets $S = \{Set_1, \ldots, Set_{27}\}$
Set random seed for reproducibility
for all attribute $a \in A$ **do**
  for all diversity level $l \in L$ **do**
    Define sampling distribution $D_{a,l}$
  end for
end for
for all combinations c of diversity levels across A **do**
  Initialize empty set Set_c
  for $i = 1$ to 10 **do**
    Sample persona p using distributions in c
    Add p to Set_c
  end for
  Apply bootstrapping to adjust attribute variance in Set_c
  Assign unique ID to Set_c
end for
for all metrics that use reference distributions (e.g., KLD) **do**
  Define reference distributions per attribute type
end for
for all distance-based metrics (e.g., JSI, CS) **do**
  Generate reference persona set S_{ref}
  Perform stratified sampling to create $Set_{optimal}$
end for
return All persona sets S

Each persona set contains 10 simulated personas. The 27 combinations define the baseline diversity conditions, with one baseline persona set generated for each condition. Repeated persona sets were then generated from the same condition definitions for the evaluation experiments, as described in Sections IV–IV-D. Various persona sets for each evaluation were simulated for categorical diversity (gender and occupation), numerical diversity (age and income), and ordinal diversity (tech savviness rating). For each type of attribute, three levels of diversity were defined: low, medium, and high (controlling the attribute range and probabilities, resulting in $3^3 = 27$ experimental conditions. The three levels are considered to develop samples from populations with varying levels of diversity.

The target diversity levels were defined using the predefined attribute distributions and variance or probability parameters, independently of the five candidate diversity metrics. Bootstrapping was then used to adjust the simulated values to these predefined distributional targets, not to optimize any candidate metric. To ensure reproducibility, all random processes were controlled with a fixed seed value (and we make our code publicly available for reproduction

and further development of the persona set[1]). This approach allows for consistent results across multiple simulation runs while maintaining the stochastic nature of the data generation process [40]. Each simulated persona set was assigned a unique identifier ($Set_1, Set_2, \ldots, Set_{27}$) to facilitate clear reference and analysis throughout the study (see the online supplementary material for full datasets). This process allows for fine-tuning of diversity with realistic attribute distributions [41].

- **For categorical variables**, controlled random allocation was used to assign values based on predefined probability distributions corresponding to each diversity level.
- **For ordinal variables**, controlled random allocation was used to assign values based on predefined probability distributions corresponding to each diversity level.
- **For numerical variables**, the generation of random numbers from specified distributions (e.g., normal distribution for age, skewed normal distribution for income) was used with adjusted parameters to achieve desired diversity levels.
- **For ordinal variables** such as tech savviness in the low diversity condition, values 4 and 5 were intentionally assigned zero probability to create a strongly skewed distribution representing minimum diversity [42].

The use of simulated data offers several advantages in evaluating diversity metrics in the context of PSDM. First, it allows one to assess the metric sensitivity to variations in categorical, numerical, and ordinal diversity independently. Second, metric performance can be evaluated in both extreme and mixed-diversity scenarios. by introducing *controlled* levels of diversity. Third, by simulating the data and controlling at different levels (low/medium/high), the lack of variance in the underlying population is also eliminated, as the simulation provides better control of underlying variances.

Using simulated data, we can create scenarios with varying levels of diversity, spanning a wider range of applications. We argue that, in addition to the experiments we conduct on a real dataset (see Section IV-E), our simulated data provide a reasonable proxy for real user characteristics by incorporating established demographic distributions and attribute relationships from previous research [39]. For example, using skewed distributions for income variables mirrors economic patterns observed in population studies [43], while our age and technology proficiency distributions reflect documented correlations between these attributes in persona research [15], [44]. This approach aligns with persona validation, in which statistical analysis is often used to compare personas with actual user populations [45]. The simulation parameters were determined by the authors to span low, medium, and high diversity, and are not drawn from a single empirical population.

For metrics that compare diversity with respect to reference distributions (i.e., *Kullback-Leibler Divergence (KLD)*), the reference distributions were created as follows: categorical variables (`gender`, `occupation`) follow a uniform distribution; `age` follows a normal distribution of the normalized data; `annual income` follows a skewed normal distribution (income and wealth are generally represented by skewed distributions [43]) of the normalized data; and the ordinal variable `tech savviness` follows a uniform distribution. Similarly, for *distance-based metrics* (*Jaccard Similarity Index (JSI)* and *Cosine Similarity (CS)*), additional steps are performed. Using the reference distributions a large set of personas is created with personas ($10 \times n$ where n is the number of personas in a set of personas in the simulation (that is, $n = 10$)). From this larger persona set, stratified sampling is carried out to (1) achieve the optimal diversity level (the highest one), (2) reduce randomness, and (3) maximize similarity to the optimal personas. The baseline is defined as the optimal persona set ($Set_{optimal}$) for comparison, as it represents the maximum diversity within a persona set.

C. FRAMEWORK FOR PERSONA QUANTIFICATION

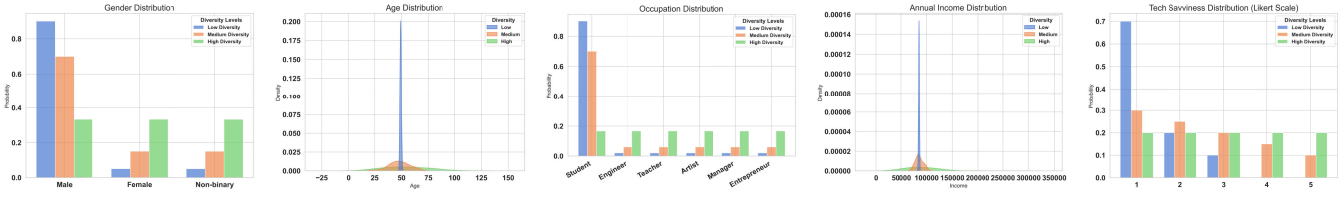
To effectively implement PSDM, it is essential to convert the qualitative attributes of personas into quantitative values. This framework outlines the data preprocessing steps to handle *categorical*, *numerical*, and *ordinal* attributes consistently to measure diversity metrics.

Categorical attributes (*gender* and *occupation*) are nominal variables without an inherent numerical value. To quantify these, we employ *one-hot encoding*, which transforms each category into a binary vector. In this method, each unique category is represented by a separate binary variable, where 1 indicates the presence of that category for a given persona and 0 indicates its absence.

Numerical attributes (*age* and *income*), often exist on different scales, which can disproportionately influence the results of diversity metrics if not standardized. To address this issue, we apply *min-max normalization* to scale the numerical values [46]. This normalization prevents larger-scale attributes from dominating diversity measurements. For highly skewed numerical attributes or extreme outliers, min-max normalization can compress most observed values. In such cases, an appropriate bounded or robust scaling approach should be applied before distance computation so that one numerical attribute does not dominate the resulting distance.

Ordinal attributes, such as a Likert-scale measure of *technological savviness*, contain an ordering but do not necessarily have equal distances between adjacent categories. Although the treatment of ordinal responses as equally spaced numerical values is debated [47], [48], it remains common in quantitative analysis [49]. We therefore assign category values from 1 to 5 and apply min-max normalization. This choice provides a consistent input for all five metrics but assumes that the distance between each pair of adjacent categories is equal. The resulting scores should be interpreted with this assumption in mind. Practitioners should use this treatment

[1] https://doi.org/10.17605/OSF.IO/3JADX



(a) Gender (b) Age (c) Occupation (d) Income (e) Tech Savviness

FIGURE 2. Attributes underlying distributions for different diversity levels. The figure shows the distributions of different attributes for Low (blue), Medium (orange), and High (green) diversity levels. Gender (2a) and Occupation (2c) are shown as grouped bar plots to illustrate the probabilities across categories. Age (2b) and Annual Income (2d) are visualized with KDE plots, illustrating their distributions. The Tech Savviness Likert scale (2e) is shown as a grouped bar plot to illustrate the probability distribution across Likert scale levels. (Reader is encouraged to zoom in.

only when the ordered scale has a numerical interpretation; otherwise, an ordinal-specific distance function is required.

After transforming the attributes, each persona is represented as a vector that combines binary variables for each category and normalized numerical values for numerical and ordinal attributes. This enables computing distances and applying information-theoretic methods (*RQE*, *JSI*, and *CS*). For distance-based diversity metrics such as *JSI* and *CS*, an *optimal persona set* is constructed to serve as a reference point. This set of optimal personas has the highest diversity–covering all possible categories and a full range of numerical values derived from the reference distributions. Comparing the analyzed persona set with the optimal persona set provides a measure of diversity relative to the optimal diversity.

D. SELECTION OF DIVERSITY METRICS

For the PSDM task, we consider a diversity metric that can quantify the variation of attributes. For example, a diverse persona set might include individuals aged 18 to 65, with multiple educational backgrounds and varying levels of technology adoption, whereas a non-diverse set might contain only university-educated users aged 25-30 with high technical proficiency. Compatible with this definition, statistical measures of diversity and similarity deployed in other domains, such as biology, ecology, economics, and population statistics [35], [36], [42], could work for measuring persona set diversity, as persona sets can be quantified in a manner similar to species and population abundance. To this end, we experiment with five metrics from distinct metric groups (i.e., *families*; Fig. 3) based on their methodological approaches; see the equations in Table 3.

First, **probabilistic diversity indices** measure *diversity based on the distribution of types within a set*. Some important metrics in this family include the *Shannon Entropy* [50], *Gini-Simpson Index (GSI)* [35], and *Margalef's Richness Index*. From this family, we select **GSI**, introduced by Simpson in 1949 [35]. *GSI* is widely used to quantify biodiversity in ecological studies [51], and it measures the probability that two individuals randomly selected from a data set belong to different types. Its applicability extends beyond ecology to any context in which there is variability of types. For example, considering a persona as both an individual and a set of attributes provides a straightforward interpretation of diversity based on the relative distribution of attribute values within the set.

Second, **the composite diversity measures** incorporate pairwise differences between types to assess diversity. Some metrics in this category include the *Bray-Curtis Dissimilarity*, *Gower Distance*, *Rao's Quadratic Entropy (RQE)*, and *Manhattan Distance*. We experiment with **RQE** [36] as a prominent metric in this family, because *RQE* accounts for the degree of dissimilarity within a set by considering both the relative abundances of persona attributes and their pairwise distances. This is particularly relevant for persona sets that differ in demographics, behaviors, or needs. Gower Distance is a pairwise dissimilarity measure, so we did not include it as a separate PSDM metric. It can be used to define the d_{ij} term in RQE when a mixed-data distance is required. We did not separately benchmark alternative distance functions within RQE because the present evaluation compares metric families at a higher level.

Third, **information-theoretic divergence measures** assess the difference between probability distributions. This family includes metrics such as the *Jensen-Shannon Divergence*, *KLD*, and the *Hellinger Distance*. We experiment with the **KLD** [52] as a key metric in this domain. KLD quantifies the divergence of the persona set from a reference distribution. In our analysis, *KLD* enables us to assess diversity relative to the reference distribution, which represents the highest level of diversity.

Fourth, **set-based similarity indices** measure overlap between sets of attributes. Some metrics in this category include the *Sørensen-Dice Coefficient*, *JSI*, and the *Overlap Coefficient*. From this family we take **JSI** [53], which assesses the similarity between two sets by dividing the size of their intersection by the size of their union. In the context of persona sets, this metric provides insight into the shared attributes between the observed persona set and the optimal persona sets. We compare the attributes of a persona set with those of the optimal persona set ($Set_{optimal}$), measuring the difference in diversity between the two sets.

Fifth, **vector-based similarity measures** evaluate similarity in a multidimensional attribute space. Some key metrics in this family include the *Pearson Correlation Coefficient*, *CS*, and the *Mahalanobis Distance*. We experiment with **CS**, a commonly used baseline metric for vector similarity [54]. *CS* measures the cosine of the angle between

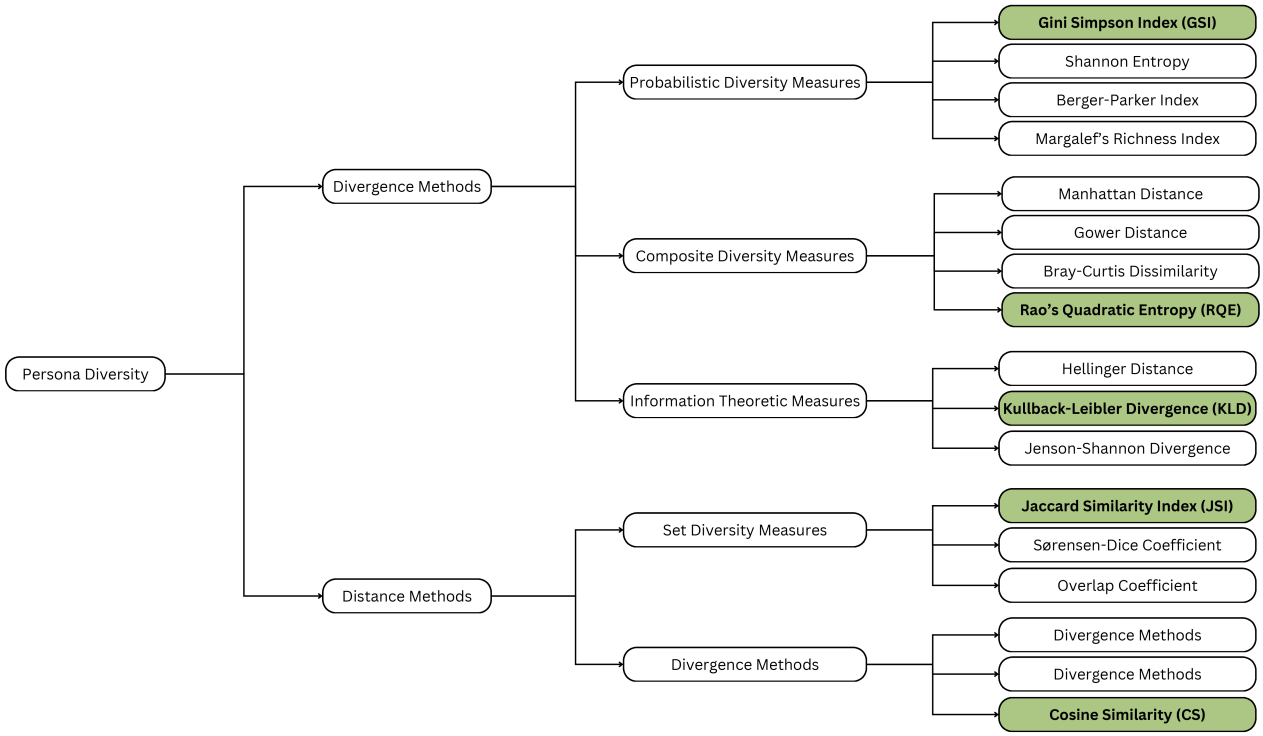


FIGURE 3. Hierarchical classification of persona diversity metrics. Highlighted nodes indicate selected metrics for our study.

two vectors, representing the similarity in the direction of the vectors regardless of their magnitude. Representing persona attributes as vectors is particularly useful for high-dimensional personas with many attributes.

Adopting these metrics allows us to test probabilistic, distance-based, information-theoretic, set-based, and vector-based measures. Using measures from different metric families allows their behavior to be evaluated under the same PSDM conditions. The adopted metrics can handle different types of persona-related data, including categorical attributes (e.g., demographics), continuous variables (e.g., usage frequency), and high-dimensional attribute spaces (e.g., behavioral patterns).

The selection of different metric families is based on comparative studies in diversity measurement. *GSI* is selected as it provides good stability even with uneven sampling [51] and is sensitive to abundance differences [42]. RQE integrates abundance patterns with dissimilarity relationships, which supports its applicability here [55]. For information-theoretic measures, *KLD* was selected based on its reported performance with continuous and mixed-type distributions [56]. Among similarity measures, JSI and CS handle mixed data types and high-dimensional spaces well [57]. So, there are plausible *theoretical* grounds for each metric to handle the PSDM task, although empirical comparisons remain to be completed.

Our design tests one representative from each family instead of every candidate metric. The experiment is built to detect between-family differences in behavior. Within each family, we selected representatives with documented properties in the comparative-evaluation literature [55], [57]. This design also clarifies why certain named alternatives were not tested. For example, Shannon Entropy is a monotonic transformation of GSI under our quantification and yields equivalent rankings of persona sets. Jensen-Shannon Divergence is the symmetric variant of KLD and depends on the same histogram binning of continuous attributes, so it does not address the limitation we observe with KLD. The Vendi Score depends on the choice of embedding, which is itself an open research question, and including it would conflate metric performance with embedding quality.

E. DESIDERATA FOR A PSDM METRIC

We propose that a "good" metric to measure the diversity of persona sets has certain requirements: (1) versatility, (2) consistency, (3) holism, and (4) scalability. These requirements directly follow from the definition of PSDM in 1, since the metric for this task should accomplish two objectives: (i) measure diversity and (ii) apply to persona sets. Thus, the requirements below are grounded in the literature on personas in HCI and on diversity measurement.

First, persona templates generally include attributes of different data types, because the qualities they capture about users are themselves of different kinds, including demographic categories, continuous quantities such as age, and ordered levels such as expertise [15], [39], [58]. Furthermore, different diversity metrics are used for different data types, and diversity metrics defined for one data type are typically not suitable for another unless the data are properly encoded [36], [59]. Thus, a PSDM metric should be

TABLE 3. Diversity metrics for persona measurement.

Metric	Equation	Variables
GSI	$\mathrm{GSI} = \frac{1}{m}\sum_{a=1}^{m}\left(1 - \sum_{v\in V_a} p_{av}^2\right)$	m: number of attributes V_a: values observed for attribute a p_{av}: proportion of personas with value v for attribute a
RQE	$RQE = \sum_{i=1}^{S}\sum_{j=1}^{S} p_i p_j d_{ij}$	S: distinct persona types $p_i,\ p_j$: relative abundances d_{ij}: distance between types i and j
KLD	$D_{KL}(P\,\|\,Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$	$P(i)$: distribution of evaluated set $Q(i)$: reference distribution
JSI	$JSI(A,B) = \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	A: attributes of personas B: attributes of optimal personas
CS	$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\lvert\mathbf{A}\rvert\,\lvert\mathbf{B}\rvert}$	$\mathbf{A}$: persona-attribute vector $\mathbf{B}$: optimal-persona vector

able to handle mixed types within a single set. Hence, **D01: Versatility** is required.

Second, persona sets are not generated once. As user segments are not static and user attributes change over time, persona sets are regenerated and compared throughout the design process [11], [39], so a good metric should return the same value whenever it is computed on the same data. Diversity-measurement research formalizes this as the stability of a metric under resampling: a metric whose value drifts across repeated samples from the same population cannot support comparison and thus is not a suitable metric for measuring diversity [42], [51]. We use these concepts to define our second requirement: consistency as the ability to produce stable results when repeatedly applied to persona sets with identical underlying diversity, tested via bootstrap resampling. Hence, **D02: Consistency** is required.

Third, most persona profiles include attributes that are often correlated [39]. For example, a persona might include attributes such as age, education, and mean annual income; education and income are mostly correlated. The same can be said about age and behavioral attributes such as technology acceptance. Mostly, these attributes are included in the persona precisely because they cohesively shape the understanding of persona users [15]. Several common diversity measures treat attributes as independent and can overstate diversity when that assumption does not hold [36], [55]. A PSDM metric must bridge these gaps by returning the same value regardless of whether the attributes are correlated. We measure this as stability across datasets with and without imposed correlations. Hence, **D03: Holism** is required.

Finally, scalability concerns the size of a persona set, which varies along two axes. The number of attributes per persona is a design decision, ranging from a handful in manual personas [39] to dozens extracted from users' online behavioral data, such as YouTube analytics [14], [44]. The number of personas in a set also varies widely, from the small sets a designer creates manually to the large populations now generated as simulated agents [45]. In diversity measurement, because calculations are compute-intensive, many measures are not very scalable, as they tend to return inflated or unstable values as the number of dimensions or items grows [60]. A PSDM metric should maintain consistent behavior across attributes of different scales. Hence, **D04: Scalability** is required.

These four properties were identifiable from the persona literature before any metric was tested, which protects against tailoring the requirements to a preferred metric. An empirical experimentation carried out (as detailed in Section IV) supports this.

Interpretability is also relevant to practical use, but we do not treat it as a tested requirement, as the present experiments do not involve practitioners. We consider it qualitatively through four properties: whether the metric has a fixed range, whether higher values consistently indicate greater diversity, whether a reference distribution is required, and whether the score can be explained in terms of differences among personas. We return to these properties in Section V.

We now apply these requirements to evaluate the candidate metrics. The experimental design for each requirement is summarized in Table 4.

TABLE 4. Summary of the experiments performed.

Desideratum	Experiment Design
D01: Versatility – Handles mixed data types	Diversity level controlled simulation for PSDM for each variable type
D02: Consistency – Is able to measure diversity consistently	Bootstrapped analysis of the parameters to observe stability and reliability
D03: Holism – Accounts for potential interdependencies between attributes	Diversity level controlled simulation for PSDM with one set with correlation and another without correlation
D04: Scalability – Accounts for high dimensionality	Diversity level controlled simulation for PSDM with higher-dimensional personas

IV. RESULTS

A. D01: VERSATILITY

To test the metrics for D01, a factorial design is employed with three factors: (1) *categorical diversity* with three levels (low, medium, high diversity in gender and occupation attributes), (2) *numerical diversity* with three levels (low, medium, high diversity in age and income attributes), and (3) *ordinal diversity* with three levels (low, medium, high diversity in tech savviness ratings), resulting in a $3^3 = 27$ experimental conditions. For D01, ten persona sets were created for each condition, and the metrics were calculated for each set. The corresponding replication procedure for D02–D04 is reported in each experiment below.

In the low- and medium-diversity conditions, where the data did not meet the assumptions for parametric tests, we employed the Kruskal-Wallis test. The results in Fig. 4 show significant effects for *GSI* (e.g., for low diversity, (H(2) = 25.82), (p < 0.001) with large effect sizes ($\epsilon^2 \approx 0.88$), particularly in distinguishing between data types at lower diversity levels. The large effect sizes observed in our diversity metric evaluations (e.g., $\epsilon^2 \approx 0.88$ for GSI) reflect the controlled simulation design, which deliberately introduces extreme parameter differences between attribute types. In the high diversity condition, where the data met parametric assumptions, the one-way ANOVA indicated significant differences (for *GSI* e.g., (F(2, 27) = 855.19), (p < 0.001), ($\eta^2 = 0.93$). The large eta-squared values (e.g., $\eta^2 = 0.93$ for GSI) reflect a controlled simulation with extreme parameter differences, rather than the effect sizes expected in applied persona development. However, a detailed examination reveals that the metrics' ability to distinguish diversity levels varies by data type. Specifically, *GSI* and *RQE* show reduced discrimination between medium and high diversity levels for numerical and ordinal data, *KLD* demonstrates limitations with ordinal data, and *CS* shows constraints with categorical and ordinal data types. Post-hoc analyses (see online supplementary material for details) show that while numerical data generally yield distinct diversity scores, the sensitivity of the metrics varied by data type and diversity level.

B. D02: CONSISTENCY

To evaluate how effectively diversity metrics capture persona diversity across different data types and levels, we conducted a bootstrap resampling experiment [61]. In line with the RQs, we designed a factorial experiment ($3^3 = 27$), varying three types of variables (categorical, numerical, and ordinal) at three levels of diversity (*low*, *medium*, *high*). For each combination, we generated 10 personas and calculated the diversity metrics. We repeated this process for $n = 1000$ bootstrap iterations to generate realistic variations in the personas while preserving the statistical properties and patterns of the baseline dataset.

The behavior of the diversity metrics at different diversity levels is shown in Fig. 5. *GSI* demonstrates the clearest separation between distributions at different diversity levels, followed by *RQE*, which shows moderate distribution overlap mainly between medium and high diversity levels. *JSI* exhibits more pronounced overlapping patterns, while *KLD* and *CS* show the most significant overlap in distributions, indicating decreased effectiveness at consistently distinguishing between different levels of diversity.

The statistical test results (see the full results in the online appendix) and the intraclass correlation coefficient (ICC) values support these observations. *GSI* ($t(28) = -5.02$, $p < .001$, $d = 1.83$) and *RQE* ($t(28) = -4.13$, $p < .001$, $d = 1.51$) show strong differences between diversity levels and have the highest ICC values (.931 and .911), which indicates reliable and consistent measurement (see Table 5). *KLD* is statistically significant ($U = 13.00$, $p < .01$, Cliff's $\delta = 0.67$) and has a moderate ICC (.817), though its direction runs counter to diversity, since it decreases as diversity grows. *JSI* reaches significance ($U = 8.00$, $p < .01$, Cliff's $\delta = 0.73$) but has a low ICC (.464), which indicates that its values are not stable across replicates even when it separates levels. *CS* shows the weakest differentiation ($U = 22.00$, $p = .08$, Cliff's $\delta = 0.47$) and the lowest ICC (.336).

TABLE 5. Intraclass correlation coefficients for each metric. ICC(2,1) with 95% confidence intervals and the F-test *p*-value, computed across bootstrap replicates. A higher ICC indicates a more consistent measurement of persona sets with the same underlying diversity.

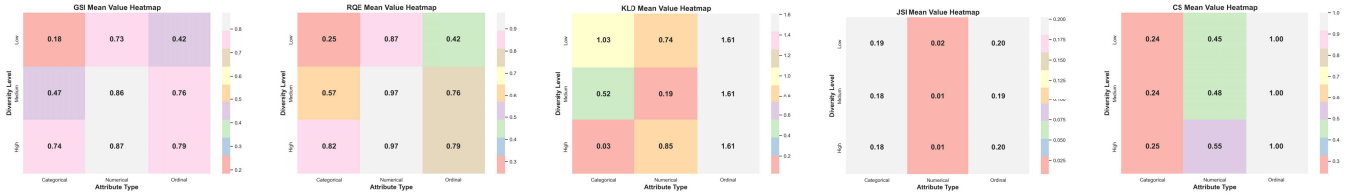
Metric	ICC(2,1)	95% CI	p
GSI	.931	[.780, 1.000]	< .001
RQE	.911	[.730, 1.000]	< .001
KLD	.817	[.550, .990]	< .001
JSI	.464	[.190, .970]	< .001
CS	.336	[.120, .950]	< .001

Consistency also requires that a metric behave in an interpretable way as the number of personas in a set changes. To test this, we used a 5 × 3 factorial design, varying the number of personas (10, 20, 30, 40, 50) and the diversity level (low, medium, high). We generated 15 persona sets for each condition and computed all five metrics for each set. We analyzed the effects of set size and diversity level on each metric using the Kruskal-Wallis test and Dunn's post hoc comparisons.

Table 6 reports the results. Diversity level had a significant effect on *GSI* ($H = 12.50$, $p = .002$), *RQE* ($H = 9.62$, $p = .008$), and *KLD* ($H = 9.98$, $p = .007$), with significant pairwise differences in Dunn's tests. The number of personas had no significant effect on any metric (all $p > .05$). *JSI* and *CS* showed no significant effect for either factor. The metrics, therefore, track the diversity of the set rather than its size within this range, which supports using 10 personas per set. *GSI* and *RQE* again separate diversity levels, while *JSI* and *CS* do not, consistent with the bootstrap consistency findings.

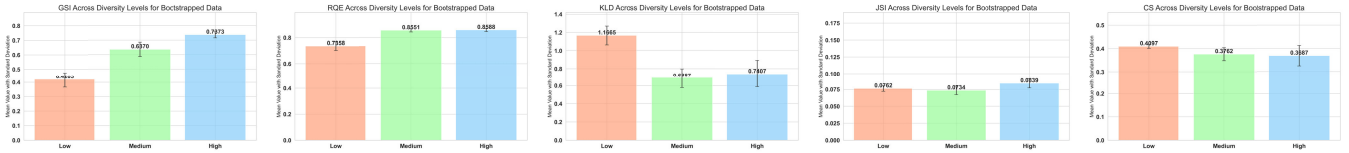
C. D03: HOLISM

To evaluate the impact of attribute interdependencies (i.e., correlations) on diversity metrics, we designed an experiment



(a) GSI (b) RQE (c) KLD (d) JSI (e) CS

FIGURE 4. Comparison of Diversity Metrics For Different Variable Types. ***GSI*** **[4a],** ***RQE*** **[4b], and** ***JSI*** **[4d] and** ***CS*** **[4e] show difference between different types of attributes.** ***KLD*** **[4c] show no difference between different attribute types. (The reader can zoom in or refer to the online supplementary material for better viewing.**



(a) GSI (b) RQE (c) KLD (d) JSI (e) CS

FIGURE 5. Diversity Metrics for Bootstrapped Persona Sets. ***GSI*** **[5a] and** ***RQE*** **[5b] increase monotonically from low to high diversity and separate the levels cleanly.** ***JSI*** **[5d] increases but with greater overlap between levels. In contrast,** ***KLD*** **[5c] counterintuitively decreases as diversity increases, while** ***CS*** **[5e] shows minimal differentiation between medium and high diversity levels.**

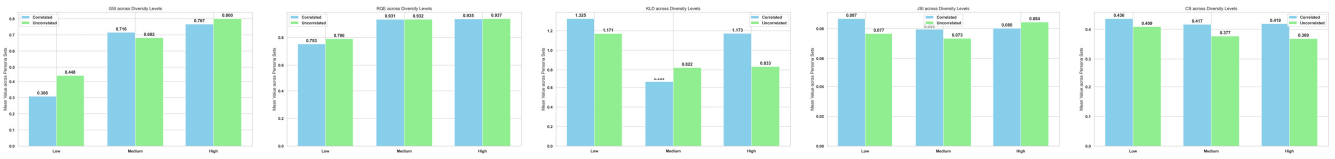
TABLE 6. Statistical results for the sample-size sensitivity analysis. Kruskal–Wallis *H* statistics and *p*-values are reported for diversity level and sample size. Post-hoc columns report pairwise Dunn's comparisons across diversity levels: low–medium (L–M), low–high (L–H), and medium–high (M–H). Non-significant comparisons are reported as $p > .05$.

Metric	Diversity effect	Sample-size effect	L–M	L–H	M–H
GSI	$H = 12.50, p = .002^{**}$	$H = 0.77, p = .943$	$p = .001^{**}$	$p > .05$	$p > .05$
RQE	$H = 9.62, p = .008^{**}$	$H = 3.90, p = .420$	$p = .011^{*}$	$p > .05$	$p = .049^{*}$
KLD	$H = 9.98, p = .007^{**}$	$H = 3.47, p = .483$	$p > .05$	$p > .05$	$p = .007^{**}$
JSI	$H = 0.46, p = .796$	$H = 8.11, p = .087$	$p > .05$	$p > .05$	$p > .05$
CS	$H = 4.22, p = .121$	$H = 5.00, p = .287$	$p > .05$	$p > .05$	$p > .05$

in which personas were generated both *with* and *without* interdependencies between attributes. For each case, we generated persona sets at three diversity levels: *low*, *medium*, and *high*, resulting in a total of $2 \times 3 = 6$ experimental conditions. For each diversity level and interdependency condition (*no correlation* and *moderate correlation* between variables), all five metrics were calculated for each persona set. To assess the relationship between *independent variables* (presence of interdependencies and diversity level) and *dependent variables* (diversity metrics), we conducted nonparametric (Kruskal-Wallis with Dunn's as a post-hoc test) or parametric tests (one-way ANOVA with Tukey HSD as a post-hoc test) for each diversity metric, based on data normality.

For this experiment, attribute dependence was induced through a formula-based scheme. Each persona's income was set as a linear function of age, scaled by an occupation-specific multiplier (for example, Engineer $\times 1.2$, Manager $\times 1.5$, Student $\times 0.5$). Occupation was made gender-dependent through shifted categorical distributions. This scheme imposes no explicit target correlations. The dependence it produces is measured after generation. Under medium diversity, the achieved correlations between age and income were $r \approx .58$ and income and occupation were $r \approx .36$. The uncorrelated sets generated for comparison showed near-zero correlations on the same pairs. The full achieved correlation values for each diversity level are provided in the supplementary material.

The results in Fig. 6 indicate that, for low diversity, all metrics except *RQE* show significant differences between correlated and uncorrelated groups. The effect sizes are notably large, particularly for *GSI* ($\epsilon^2 = 0.882$) and *CS* ($\epsilon^2 = 0.918$). In medium diversity, the results are mixed; while metrics such as *KLD* ($\eta^2 = 0.698$) and *CS* ($\eta^2 = 0.608$) continue to show significant differences, others like *RQE* ($p = 0.860$) indicate nonsignificant differences between the groups. This shift in non-significance for *RQE* suggests that medium diversity may reduce the observable differences between the correlated and uncorrelated groups. Finally, in high diversity, *GSI* ($\eta^2 = 0.937$) and *KLD* ($\eta^2 = 0.738$) maintain highly significant differences with large effect sizes,



(a) GSI (b) RQE (c) KLD (d) JSI (e) CS

FIGURE 6. Diversity Metrics for Correlated Attributes in Persona Sets. Only the *RQE* [6b] consistently handles both persona sets—with and without interdependence—across all diversity levels. Other metrics like *GSI* [6a], *KLD* [6c], *JSI* [6d], and *CS* [6e] do not show this consistent behavior. (The reader can zoom in or refer to the online supplementary material for better viewing.

while *RQE* remains non-significant ($p = 0.103$). While *GSI*, *KLD*, *JSI*, and *CS* exhibit significant sensitivity to attribute correlations across diversity levels (as demonstrated by varying measurement gaps between correlated and uncorrelated sets), *RQE* demonstrates consistency in its measurements between correlated and uncorrelated persona sets within each diversity level (see Fig. 6b).

D. D04: SCALABILITY

To assess the scalability of diversity metrics, we added attributes to generate higher-dimensional persona sets ($m = 39$ attributes). The number of attributes was configured with the context in mind—while high dimensionality in computer science often refers to hundreds or even thousands of dimensions, in the context of personas, there are often only a handful of attributes [39]. So, for *personas*, thirty-nine attributes would be unusually high and therefore represent a case of high dimensionality in this context.

In this high-dimensional setting, personas included additional categorical attributes (e.g., education level, marital status, employment sector, country), numerical attributes (e.g., years of experience, number of dependents), and ordinal attributes (e.g., satisfaction level, health status). This increased the data's dimensionality, especially after applying one-hot encoding per our quantification paradigm. For each diversity level, we calculated all metrics for each persona set. The results in Fig. 7 show that for higher-dimensional personas, the *GSI* exhibits an increase of 18% from low-dimensionality baseline personas at low diversity and 1% at higher diversity levels. *RQE* in high-dimensional personas increases by 15% at a low diversity level and 7%. However, the *KLD* shows a more significant decline in higher-dimensional personas, whereas in normal personas it decreases more moderately. Moreover, the shift in the medium is even lower than at the higher diversity level. Both *JSI* and *CS* show similar decreasing trends, but the drop is steeper in higher-dimensional personas.

We evaluated the scalability of metric computation using persona sets containing ($n \in 10, 20, 50, 100, 200, 500, 1000$) personas under the five-attribute baseline and 39-attribute high-dimensional configurations. The medium-diversity condition was kept constant to isolate the effects of set size and dimensionality, and each condition was evaluated over 30 runs using the same random seed and software environment. We measured metric computation time and peak additional memory after persona quantification, excluding data generation, and report the median runtime with the standard deviation across 30 runs. The benchmark was conducted on an Apple Silicon system running macOS 26.5, 15 CPU cores, 48 GB of RAM, and Python 3.9.6, with numerical library thread counts fixed at 1. GSI and KLD have approximately linear time complexity, (O(nm)), JSI and CS have ($O(nd)$) complexity in our implementation, and RQE has ($O(n^2d)$) time complexity and up to ($O(n^2)$) additional memory when the complete pairwise distance representation is retained. However, the full matrix does not have to be retained in memory, as blockwise or iterative distance computation can reduce the memory requirement, although the number of pairwise calculations remains quadratic. Thus, larger production-scale cohorts may therefore require batching, sampling, or approximate pairwise computation.

All five metrics remained computationally feasible within the tested range. With 1,000 personas and 39 attributes, the median runtime was 0.052ms for *JSI*, 0.053ms for *CS*, 2.70ms for *GSI*, 5.71ms for *KLD*, and 10.06ms for *RQE*. *RQE* had the highest runtime at the largest condition, consistent with its pairwise computation, increasing from 0.0022ms for 10 personas and five attributes to 10.06ms for 1,000 personas and 39 attributes. Nevertheless, all metrics completed in less than 11ms at the largest tested condition. *RQE* is therefore computationally practical within the evaluated range, although its quadratic growth may become consequential in population-scale persona simulations. The runtime results are reported in Table 7; the memory requirement depends on whether the complete pairwise distance representation is retained.

In sum, *RQE* performed best among the five evaluated metrics under the tested conditions (see Table 8).

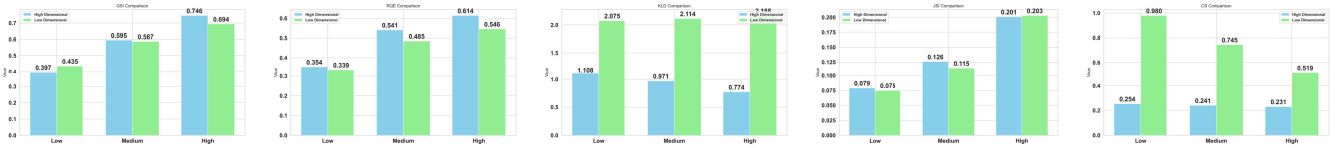
E. PRACTICAL APPLICATION TO A REAL-WORLD DATASET

To show that the framework applies beyond simulated data, we ran PSDM on personas generated from real survey data. This is a proof of concept instead of a validation study. A validation would require ground-truth diversity labels for real persona sets, which do not exist. The aim here is to demonstrate that the quantification framework and the metrics produce sensible values on actual data. We used the *Survey2Persona*[2] [62] software to create five personas

[2]An automated software to create personas from survey data https://s2p.qcri.org/

TABLE 7. Computational cost of the five PSDM metrics. Runtime cells report the median milliseconds per metric call ± standard deviation across 30 runs.

Metric	Time	$n = 10, m = 5$	$n = 100, m = 39$	$n = 1000, m = 39$
GSI	$O(nm)$	0.215 ± 0.025	1.64 ± 0.32	2.70 ± 0.44
RQE	$O(n^2 d)$	0.0022 ± 0.0002	0.078 ± 0.009	10.06 ± 1.24
KLD	$O(nm)$	0.427 ± 0.023	3.40 ± 0.63	5.71 ± 1.05
JSI	$O(nd)$	0.0059 ± 0.0011	0.0099 ± 0.0018	0.052 ± 0.008
CS	$O(nd)$	0.0067 ± 0.0008	0.011 ± 0.001	0.053 ± 0.005



(a) GSI (b) RQE (c) KLD (d) JSI (e) CS

FIGURE 7. Diversity Metrics for Higher Dimensionality in Persona Sets. ***GSI*** **[7a],** ***RQE*** **[7b], and** ***JSI*** **[7d] exhibit a monotonic increase when moving from low dimensionality to high dimensionality. In contrast,** ***KLD*** **[7c] and** ***CS*** **[7e] do not follow this trend. (The reader can zoom in or refer to the online supplementary material for better viewing.**

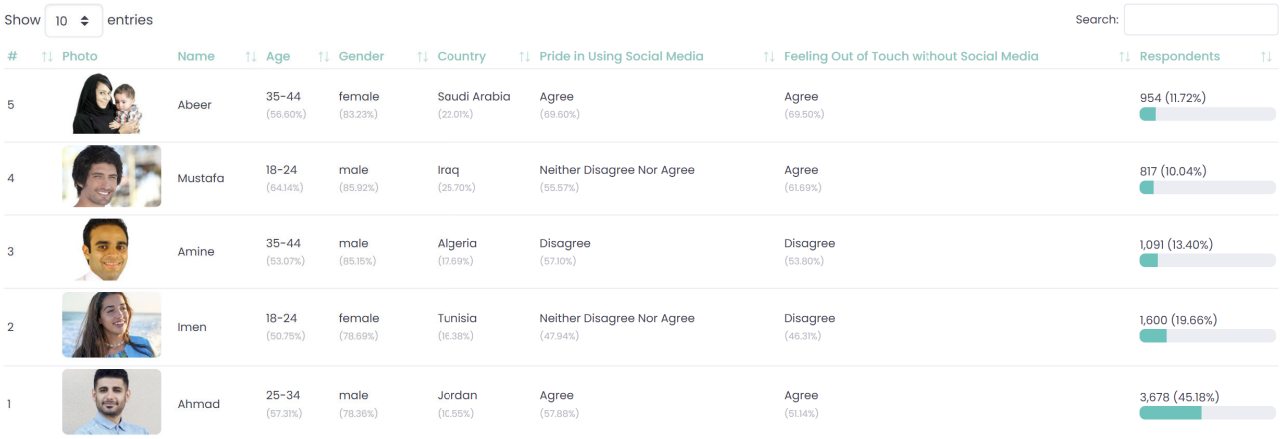


FIGURE 8. Sanity check with a real dataset. A persona set generated by Survey2Persona software from the MENA region dataset on social media preferences.

(see Fig. 8) from a publicly available social media preferences dataset collected from 8000 users from 16 countries in the Middle East and North Africa (MENA) region [63]. The personas represent realistic profiles based on users' self-reported attitudes and behaviors. We applied the same quantification framework and attribute transformations as in the simulation experiments (see Section III-C), then computed the five diversity metrics (*GSI*, *RQE*, *KLD*, *JSI*, and *CS*).

The attribute distributions indicate medium diversity within the persona set. Gender and education are concentrated, while country and occupation vary more widely. This structure appears in the persona responses, where shared attributes yield similar answers, while attributes such as country yield more distinct ones. The metrics partly reflect this structure. GSI and RQE both indicate moderate diversity (.547 and .544), close to what the attribute distributions suggest. KLD returns a low value (.143).

To gauge the stability of these point estimates on a single five-persona set, we computed percentile bootstrap 95% confidence intervals (GSI [.240, .680], RQE [.320, .630], KLD [.083, .255]). The intervals are wide, as expected for five personas. JSI and CS returned degenerate values under resampling (JSI $=$ 0, CS $=$ .500, both with zero-width intervals), because the five-persona set produced no attribute overlap variation for JSI and a constant value of CS. These two metrics are uninformative at this sample size, consistent with their relatively weak performance in the controlled experiments.

To examine whether RQE performs as expected on real data, we constructed three persona sets of increasing diversity from the same MENA survey data on privacy

TABLE 8. Summary of metric performance. A check indicates that the metric met the operational criterion for the corresponding desideratum: D01, applicability to categorical, numerical, and ordinal data; D02, stable repeated measurement with the expected diversity ordering; D03, stability under imposed attribute interdependencies; and D04, consistent behavior under increased dimensionality and computational feasibility within the tested range. A check for D01 does not imply equal discrimination across all data types. The asterisk indicates reduced separation between the medium- and high-diversity conditions. RQE performed best among the five evaluated metrics under the tested conditions.

Desideratum	GSI	RQE	KLD	JSI	CS
D01: Versatility	✔	✔	✔	✔	✔
D02: Consistency	✔	✔*	✘	✘	✘
D03: Holism	✘	✔	✘	✘	✘
D04: Scalability	✔	✔	✘	✔	✘

concerns [63]. The low-diversity set was restricted to respondents under 35 years of age from a single country (Egypt). The medium-diversity set retained the full set of unfiltered survey responses. For the high-diversity set, we generated a larger persona pool, which is known to increase diversity [64], and selected the five personas with the most varied attributes (see the online supplementary material). The resulting RQE values were 0.534 for low, 0.600 for medium, and 0.697 for high diversity (see Fig. 9). These values come from three deliberately constructed sets and are separate from the single-set estimate reported above. The values increase with expected diversity, supporting the monotonicity of RQE in the PSDM task on field data.

Overall, these results show that the PSDM framework works with real data-driven personas and that RQE behaves as expected.

V. DISCUSSION

A. THEORETICAL IMPLICATIONS

This study makes several contributions to the field of persona evaluation and, by extension, to inclusive design. First, we introduce and define the PSDM task. The definition of PSDM lays the foundation for a quantitative assessment of persona diversity, contributing to inclusive design with personas [9], [65] and addresses the problem of quantifying personas as information [15]. Although developed and evaluated for persona sets, the PSDM framework is equally applicable to other forms of user segmentation, such as user clusters or market segments, since these share the same attribute structure. This suggests that the results reported here may extend beyond personas to user representation methods more broadly.

Second, we develop a framework for the quantification of persona sets, which applies (1) one-hot encoding for categorical attributes, (2) min-max normalization for numerical attributes, and (3) consistent numerical assignments for ordinal data, as these approaches transform persona attributes into a quantitative format while preserving information. We show that pre-existing diversity metrics, including probabilistic and distance-based metrics, can be applied to the PSDM task once the persona set is quantified.

Third, we propose requirements for evaluating a persona set diversity metric, providing a starting point for assessing the suitability of various diversity metrics for the PSDM task. To facilitate experiments with more diversity metrics, we make baseline datasets available[3] with controlled diversity levels in categorical, numerical, and ordinal attributes. They can be used to evaluate and compare diversity metrics for PSDM.

Fourth, *RQE* performed best among the five evaluated metrics under the tested conditions. It handled the three attribute types, produced consistent values across bootstrap samples, retained the expected diversity ordering when attribute dependencies were introduced, and remained interpretable in the higher-dimensional condition. Its separation between medium and high diversity was weaker than its separation between low and medium diversity, so the result should not be interpreted as evidence that *RQE* is optimal for every persona dataset. *GSI* also performed consistently but was more affected by the imposed attribute dependencies. *JSI* and *CS* showed lower consistency, while KLD depended strongly on its reference distributions. Within the scope of the tested metrics and conditions, *RQE* performed most consistently across the four requirements. PSDM measures attribute variation and is separate from other persona evaluation properties such as *accuracy*, *credibility*, *validity*, *qualitative plausibility*, and *semantic coherence*. A high diversity score does not establish that a persona set performs well on these properties or that the source sample adequately represents the target population. These properties require separate evaluation.

B. PRACTICAL IMPLICATIONS

Practitioners can apply PSDM to quantify and compare attribute variation across candidate persona sets. Manually inspecting persona diversity is difficult, especially when considering multiple attributes at once. Thus, a metric such as RQE condenses that judgment into a single actionable signal. For example, practitioner can compute RQE for different persona sets and select the set with the higher value (see Fig. 10 for the algorithm).

In practice, the workflow combines the source-data audit in Section V-D with set-level and attribute-level diversity measurement. Check the source data first, then the set-level score and attribute-level values, and interpret the resulting score primarily relative to persona sets generated from the same data and attributes.

For metrics bounded in the [0,1] range, the score provides a common reporting scale, although the metrics do not have identical interpretations. GSI has a probability interpretation, RQE summarizes weighted pairwise dissimilarity, and JSI measures overlap with the reference set. We therefore use the bands in Table 9 as heuristic reporting aids rather than

[3]https://osf.io/3jadx/?view_only=1d083be1103e4898b44668c26e2be5d9

(a) (b)

FIGURE 9. **Real-world dataset application results. (a) RQE increases over the three constructed diversity levels, consistent with its expected behavior. (b) The five metrics on the actual persona set show GSI, RQE, and CS indicating moderate diversity, while JSI and KLD return extreme values.**

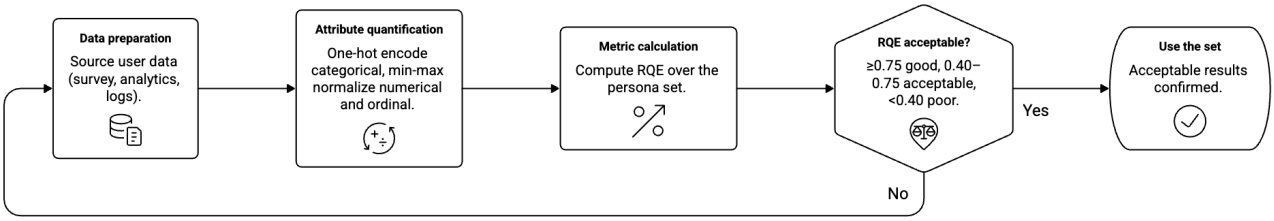


FIGURE 10. **Practitioner workflow for applying PSDM. A persona set is quantified, scored with RQE, interpreted using provisional bands or relative comparison, and then retained or regenerated.**

TABLE 9. **Heuristic interpretation bands for bounded diversity metrics (RQE, GSI, JSI). The cut points are provisional reporting aids rather than empirically derived thresholds. Comparison between persona sets built from the same data and attributes is the primary use.**

Metric value	Interpretation
≥ 0.75	Good diversity
0.40 to 0.75	Acceptable diversity
< 0.40	Poor diversity; inspect and augment

empirically calibrated thresholds. The cut points at 0.40 and 0.75 are interpretive conventions rather than values derived from outcome data. Thus, we interpret 0.75 or higher as good diversity, 0.40 to 0.75 as acceptable diversity, and below 0.40 as poor diversity. A diversity score also depends on which attributes are measured and how they are encoded. Comparison between persona sets built from the same user segments and evaluated with the same attributes should therefore remain the primary use.

Although the bands provide an approximate interpretation of an individual score, relative comparison between persona sets is more reliable. If two sets are drawn from the same source population and evaluated using the same attributes, the set with the higher *RQE* contains personas that are further apart under the selected distance function. Attribute-level *RQE* values can then identify which attributes contribute least to the set-level variation. A practitioner may use this information to inspect the segmentation or generate alternative candidate sets. The result does not, on its own, show that low-scoring attributes are underrepresented in the source population.

As an example, we constructed two persona sets from the MENA survey data used for experimentation in Section IV-E. The first set (Set_{lower}) was defined to only divide users under 35 in a single country and had an RQE of 0.534. The second set (Set_{higher}) was drawn from the full age range across countries and had a score of 0.697. As both persona sets are drawn from the same user population, the lower value of Set_{lower} indicates lower diversity. Similarly, the attribute-level values confirm this, as age and country were the least diverse attributes in the Set_{lower}. A practitioner interpreting the RQE value can recognize that the persona set lacks sufficient diversity and can take actions such as resampling the data to generate more diverse personas.

PSDM can also serve in LLM-assisted persona generation systems. As one application of the framework, we integrated RQE into the Persona Ecosystem Playground (PEP) to assess the attribute diversity of a persona set [66], [67]. PEP is a methodology and system for generating persona sets that represent stakeholder groups in a multi-agent simulation. In PEP, the user must decide which persona set best represents each stakeholder group before proceeding. The practitioner computes RQE scores for candidate sets generated from the same data under different clustering configurations and selects the set with the higher score, thereby securing broader attribute coverage before the personas enter the simulation (see Fig. 11). PSDM can therefore be integrated into a persona-generation pipeline as a comparative selection criterion or as a post hoc measure of attribute variation. In this study, the PEP integration demonstrates how the metric can be integrated into an existing workflow. We do not claim that the *RQE* selection step improves downstream design decisions, system performance, or user satisfaction because these outcomes were not evaluated.

In terms of limitations, a high RQE value reports variation among the personas in the set. It does not report whether the

source data covered the right people. A set built from a narrow sample can score high and still leave whole user groups out. The metric only sees the personas it is given, so it cannot flag the groups that were never created. This concerns the source data, not the metric, and we address it in Section V-D.

C. LIMITATIONS AND FUTURE WORK

Our study has some limitations. First, although we show that PSDM works on real-world data, real deployments face constraints that our experiments did not test. These include privacy concerns on personal data, as it can limit which attributes are collected and stored, and a narrower attribute set narrows what any diversity score can capture. PSDM therefore measures variation only among the attributes included in the persona set; it does not determine whether the source data adequately represent the population relevant to the design problem.

Another concern is computational cost. Several of the metrics, including RQE, compare every persona in a set against every other. The number of pairwise comparisons therefore grows with the square of the set size, and each comparison sums a distance over all attributes, so the cost scales as the square of the number of personas times the number of attributes. Thus, moving from ten personas to a thousand raises computation costs sharply. This cost is bounded in typical use, because persona sets are often kept small in practice. Most studies generate fewer than 10 personas, regardless of how heterogeneous the source data is [64], since larger persona sets increase cognitive load for designers, making it harder to relate to and empathize with the personas. For persona sets of this size, the pairwise cost of RQE is negligible. Our computational benchmark tested persona sets containing up to 1,000 personas and 39 attributes, but performance for substantially larger simulated populations remains untested. Such applications may require iterative distance calculation, batching, sampling, or approximate pairwise methods. The observed runtimes are also specific to the tested hardware, software environment, and metric implementations.

Second, we tested one representative metric from each of five families, leaving other families and measurement principles untested. New diversity metrics continue to appear, and some may rest on principles none of our five families capture. We list candidate metrics and their families in the supplementary material, along with the reason each was included or set aside, as a starting point for extending the comparison. The conclusion that RQE performed best therefore applies to the five selected metrics and the experimental conditions tested in this study. Other metrics within the same families, or metrics based on different measurement principles, may produce different results.

Third, our study focuses primarily on categorical, numerical, and ordinal data types commonly found in traditional persona attributes. However, in real applications, persona sets can include more complex data types, such as text descriptions, representative images, behavioral logs, or even visual representations [5], [68]. These complex data types would require additional processing steps to measure diversity. Developing appropriate diversity metrics for these additional data types could involve adapting techniques from natural language processing and computer vision to quantify semantic diversity in textual descriptions and visual diversity in persona images.

Fourth, our approach treats all attributes as equally important, yet some attributes matter more than others for a given design goal. Protected attributes such as gender, race, or disability status often carry weight that a plain count of variation does not capture. For example, a persona set that varies widely on income but not at all on disability status may score as diverse while missing the representation that matters most for an accessibility project. For such cases, weighting extends RQE without changing its structure, since RQE already sums pairwise distances over attributes. A weight w_k on attribute k enters through the distance term, so the pairwise distance becomes $d_{ij} = \sum_{k=1}^{m} w_k d_{ij}^{(k)}$ with $\sum_k w_k = 1$, where $d_{ij}^{(k)}$ is the per-attribute distance and equal weights recover the unweighted RQE used here.

However, the source of the weight values in this extended version of the RQE metric remains an open question. Designers could set them based on domain knowledge, which is direct but subjective, or derive the weights from the variance each attribute explains in downstream design outcomes, which grounds them in data but requires outcome measures that are often unavailable. Weights can also encode fairness constraints, thereby raising the contribution of attributes associated with historically underrepresented groups. Overall, weighting schemas can draw on studies in algorithmic fairness, in which weighting protected attributes and user preferences is a known way to steer outcomes toward fairer representation [12], [69]. Each source changes what a diversity score means, so the choice should be reported alongside the score. We leave the empirical study of weighting schemes for future work. The same applies to the preprocessing choices in our quantification framework. We applied one-hot encoding, min-max normalization, and equal-interval ordinal treatment throughout, and we did not test how alternative encoding or normalization techniques would affect the scores. The equal-interval treatment may distort distance calculations when the differences between adjacent ordinal categories cannot reasonably be assumed to be equal. Practitioners should therefore treat ordinal attributes as numerical only when the scale has a defensible equal-interval interpretation; otherwise, an ordinal-specific distance function is needed. A sensitivity analysis over these choices is a topic for future work.

Fifth, we primarily focus on diversity measurement in structured persona profiles. Personas can also be represented as text-based narratives or audiovisual personas. Applying PSDM to these formats would require additional representations of lexical, semantic, visual, or audiovisual diversity,

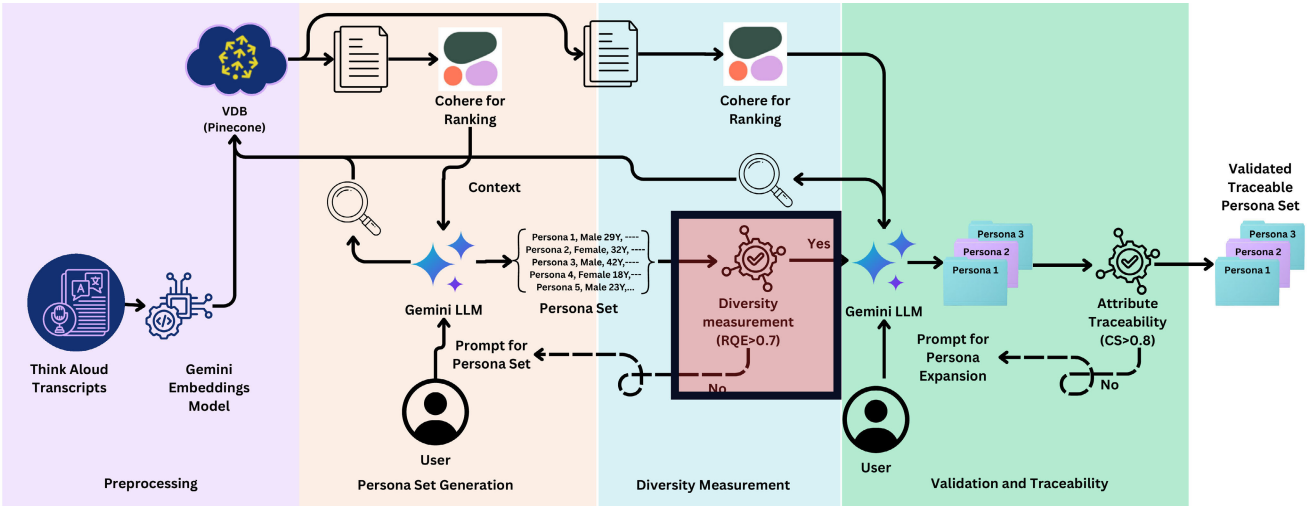


FIGURE 11. **RQE as a selection step in the PEP workflow [67]. Candidate persona sets are generated from the same data under different clustering configurations. RQE is computed for each candidate, and the set with the higher value is selected before the personas enter the multi-agent simulation. The RQE selection step is marked in red.**

so the present findings should not be assumed to transfer directly to these formats.

Sixth, the evidence here also rests on simulated sets and a single real-world dataset from a single domain. The controlled low-, medium-, and high-diversity conditions were deliberately separated to compare metric behavior, which likely contributed to the large effect sizes observed in the experiments. These effect sizes should therefore not be interpreted as estimates of differences expected in applied persona development. Furthermore, testing PSDM across several real persona datasets from different domains, which would show whether the relative performance of RQE observed here generalizes across datasets and domains, is left for future work.

Finally, the proposed interpretation bands are provisional reporting aids rather than thresholds derived from observed design outcomes. Similarly, the PEP integration demonstrates how PSDM can be incorporated into a persona-generation workflow, but it does not establish that selecting a higher-scoring persona set improves design decisions, system performance, or user satisfaction. These outcomes require separate empirical evaluation.

Future research should address these limitations by expanding the range of attributes in persona sets to include behavioral patterns, motivations, and goals, thereby enhancing the applicability of diversity assessments. More work could be done on technical improvements, where the computational complexity arising from point-point distance calculation in distance-based measures could be managed for more elaborate applications of data-driven personas (i.e., implementing the diversity measurement framework in real systems).

D. ETHICAL CONSIDERATIONS

A numeric diversity score measures the attribute variation present in a persona set, not whether the underlying data adequately represents all relevant user groups. If the data used to generate personas exclude marginalized groups due to sampling gaps, historical underrepresentation, or data-collection constraints, a high RQE score will not detect this absence. The metric reports what is in the data, not what is missing from it.

This boundary also makes RQE a useful diagnostic for further analysis of the underlying user data. A low value indicates that the data or the generation method has produced a less diverse set of personas. This is a prompt to revisit and qualitatively evaluate the source data or the clustering configuration and regenerate the persona set. RQE, or any diversity metric, therefore acts as a precursor to closer analysis rather than a verdict on its own. A high value carries less assurance, because the metric only measures the personas that were generated and says nothing about those that the data never produced. A set drawn from a skewed sample can vary internally and still score high. The score confirms that the personas differ from one another, but it cannot confirm that the sample behind them was complete. Whether the represented population is the right one depends on contextual factors the metric does not see, such as the sampling frame, the design goal, and the communities the product must serve. Before trusting a diversity score, a practitioner should therefore audit the source data against the following checks:

- **Sampling frame:** The frame is defined against the target population and reaches the groups the product must serve.
- **Known underrepresented groups:** Groups known to be underrepresented in the domain are present in the data instead of being absent from it.
- **Filtering effects:** No relevant group was removed during data cleaning or outlier filtering.
- **Attribute coverage:** The data spans the attributes that matter most for the design goal, including protected attributes.

A high score on data that fails these checks reports variation within a narrow sample and should be interpreted accordingly.

VI. CONCLUSION

Based on our findings, RQE performed best among the five evaluated metrics under the tested conditions, with the most consistent performance across our evaluation criteria. PSDM measures attribute variation within persona sets. However, it does not determine whether the source data adequately represent the target population. Future research should apply these metrics to real-world scenarios with complex attributes, investigate how diversity changes as user populations evolve over time, and develop weighted attribute approaches that reflect specific diversity priorities. As personas increasingly inform design decisions, measuring their diversity is one step toward representing users more inclusively in products and services developed using personas.

REFERENCES

[1] A. Cooper, *The Inmates Are Running The Asylum*, SAMS Publishing, 1999, p. 17.
[2] C. LeRouge, J. Ma, S. Sneha, and K. Tolle, "User profiles and personas in the design and development of consumer health technologies," *Int. J. Med. Informat.*, vol. 82, no. 11, pp. e251–e268, Nov. 2013.
[3] A. Kaur, A. Aird, H. Borman, A. Nicastro, A. Leontjeva, L. Pizzato, and D. Jermyn, "Synthetic voices: Evaluating the fidelity of LLM-generated personas in representing people's financial wellbeing," in *Proc. 33rd ACM Conf. User Model., Adaptation Personalization*, Jun. 2025, pp. 185–193.
[4] M. Almeshari, J. Dowell, and J. Nyhan, "Using personas to model museum visitors," in *Proc. 27th Conf. User Modeling, Adaptation Personalization*, Jun. 2019, pp. 401–405.
[5] L. Nielsen, *Personas In A More User-Focused World*. Cham, Switzerland: Springer, 2013, pp. 129–154.
[6] B. J. Jansen, S. Jung, and J. Salminen, "Creating manageable persona sets from large user populations," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–6.
[7] P. Turner and S. Turner, "Is stereotyping inevitable when designing with personas?" *Design Stud.*, vol. 32, no. 1, pp. 30–44, Jan. 2011.
[8] J. Goodman-Deane, S. Waller, D. Demin, A. G. D. Heredia, M. Bradley, and P. J. Clarkson, "Evaluating inclusivity using quantitative personas," in *Proc. Design Res. Soc. Conf.*, Jun. 2018, pp. 1828–1840.
[9] J. A.-L. Goodman-Deane, M. Bradley, S. Waller, and P. J. Clarkson, "Developing personas to help designers to understand digital exclusion," *Proc. Design Soc.*, vol. 1, pp. 1203–1212, Aug. 2021.
[10] C. N. Chapman and R. P. Milham, "The personas' new clothes: Methodological and practical arguments against a popular method," in *Proc. H*, 2006, pp. 634–636.
[11] J. Salminen, S.-G. Jung, and B. Jansen, "Developing persona analytics towards persona science," in *Proc. 27th Int. Conf. Intell. User Interfaces*, Mar. 2022, pp. 323–344.
[12] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, Nov. 2021, Art. no. 100336.
[13] N. Tomasev, K. R. McKee, J. Kay, and S. Mohamed, "Fairness for unobserved characteristics: Insights from technological impacts on queer communities," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2021, pp. 254–265.
[14] S.-G. Jung, J. Salminen, H. Kwak, J. An, and B. J. Jansen, "Automatic persona generation (APG): A rationale and demonstration," in *Proc. Conf. Human Inf. Interact. Retrieval*, 2018, pp. 321–324.
[15] C. N. Chapman, E. Love, R. P. Milham, P. ElRif, and J. L. Alford, "Quantitative evaluation of personas as information," in *Proc. Hum. Factors Ergonom. Soc. 52Nd Annu. Meeting*, 2008, pp. 1107–1111.
[16] J. Salminen, K. Chhirang, S. Jung, S. Thirumuruganathan, K. Guan, and B. J. Jansen, "Big data, small personas: How algorithms shape the demographic representation of data-driven user segments," in *Proc. Big Data*, vol. 10, 2022, pp. 313–336.
[17] J. J. McGinn and N. Kotamraju, "Data-driven persona development," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2008, pp. 1521–1524.
[18] H. Zhu, H. Wang, and J. M. Carroll, "Creating persona skeletons from imbalanced datasets—A case study using U.S. older Adults' health data," in *Proc. Designing Interact. Syst. Conf.*, CA. San Diego, CA, USA: ACM Press, Jun. 2019, pp. 61–70.
[19] F. Anvari, D. Richards, M. Hitchens, M. A. Babar, H. M. T. Tran, and P. Busch, "An empirical investigation of the influence of persona with personality traits on conceptual design," *J. Syst. Softw.*, vol. 134, pp. 324–339, Dec. 2017.
[20] S. D. Mitchell, "Perspectives, representation, and integration," in *Understanding Perspectivism: Scientific Challenges Methodological Prospects*. Evanston, IL, USA: Routledge, 2019, pp. 178–193.
[21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Excerpt from datasheets for datasets," in *Ethics of Data and Analytics: Concepts and Cases*. New York, NY, USA: Auerbach, 2022, pp. 148–156.
[22] A. Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, and O. Russakovsky, "REVISE: A tool for measuring and mitigating bias in visual datasets," *Int. J. Comput. Vis.*, vol. 130, no. 7, pp. 1790–1810, Jul. 2022.
[23] S. Yfantidou, P. Sermpezis, and A. Vakali, "12 years of self-tracking for promoting physical activity from a user diversity perspective: Taking stock & thinking ahead," in *Proc. 30th ACM Conf. User Model., Adaptation Personalization*, Jul. 2022, pp. 211–221.
[24] S. Gupta, Y.-C. Chen, and C. Tsai, "Utilizing large language models in tribal emergency management," in *Proc. 29th Int. Conf. Intell. User Interfaces*, Mar. 2024, pp. 1–6.
[25] E. Graells-Garrido, M. Lalmas, and R. Baeza-Yates, "Encouraging diversity- and representation-awareness in geographically centralized content," in *Proc. 21st Int. Conf. Intell. User Interfaces*, Mar. 2016, pp. 7–18.
[26] S. Khan, B. Knijnenburg, and E. Dixon, "Evaluating ADHD users' experience with recommender systems," in *Proc. 29th Int. Conf. Intell. User Interfaces*, Mar. 2024, pp. 84–88.
[27] D. Gotz and Z. Wen, "Behavior-driven visualization recommendation," in *Proc. 14th Int. Conf. Intell. User Interfaces*, Feb. 2009, pp. 315–324.
[28] A. Anderson, J. N. Guevara, F. Moussaoui, T. Li, M. Vorvoreanu, and M. Burnett, "Measuring user experience inclusivity in human-AI interaction via five user problem-solving styles," *ACM Trans. Interact. Intell. Syst.*, vol. 14, no. 3, pp. 1–90, Sep. 2024.
[29] S. Saisubramanian, S. C. Roberts, and S. Zilberstein, "Understanding user attitudes towards negative side effects of AI systems," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–6.
[30] V. Sivaraman, Z. Li, and A. Perer, "Divisi: Interactive search and visualization for scalable exploratory subgroup analysis," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2025, pp. 1–17.
[31] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. IT-37, no. 1, pp. 145–151, Jan. 1991.
[32] Q. Liu, H. Jiang, Z. Pan, Q. Han, Z. Peng, and Q. Li, "BiasEye: A bias-aware real-time interactive material screening system for impartial candidate assessment," in *Proc. 29th Int. Conf. Intell. User Interfaces*, 2024, pp. 325–343.
[33] S. Vaez Barenji, S. Parajuli, and M. D. Ekstrand, "User and recommender behavior over time: Contextualizing activity effectiveness diversity and fairness in book recommendation," in *Adjunct Proc. 33rd ACM Conf. User Model., Adaptation Personalization*, Jun. 2025, pp. 280–287.
[34] S. Sethi, J. Salminen, D. Amin, and B. J. Jansen, "'When AI writes personas': Analyzing lexical diversity in llm-generated persona descriptions," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, 2025, p. 8.
[35] E. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 4148, p. 688, 1949.
[36] C. R. Rao, "Diversity and dissimilarity coefficients: A unified approach," *Theor. Population Biol.*, vol. 21, no. 1, pp. 24–43, Feb. 1982.

[37] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–16.
[38] J. Brickey, S. Walczak, and T. Burgess, "Comparing semi-automated clustering methods for persona development," *IEEE Trans. Softw. Eng.*, vol. 38, no. 3, pp. 537–546, May 2012.
[39] L. Nielsen, K. S. Hansen, J. Stage, and J. Billestrup, "A template for design personas: Analysis of 47 persona descriptions from Danish industries and organizations," *Int. J. Sociotechnology Knowl. Develop.*, vol. 7, no. 1, pp. 45–61, Jan. 2015.
[40] J. E. Gentle, *Monte Carlo Methods*. Cham, Switzerland: Springer, 1998, pp. 131–150.
[41] B. Efron and R. J. Tibshirani, *An Introduction to The Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.
[42] M. O. Hill, "Diversity and evenness: A unifying notation and its consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, Mar. 1973.
[43] J. Benhabib, A. Bisin, and M. Luo, "Earnings inequality and other determinants of wealth inequality," *Amer. Econ. Rev.*, vol. 108, no. 9, pp. 2659–2678, 2018.
[44] S. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona generation from aggregated social media data," in *Proc. CHI Conf. Extended Abstr. Hum. Factors*, 2017, pp. 1748–1755.
[45] D. Park and J. Kang, "Constructing data-driven personas through an analysis of mobile application store data," *Appl. Sci.*, vol. 12, no. 6, p. 2869, Mar. 2022.
[46] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 563–582, Apr. 2001.
[47] G. M. Sullivan and A. R. Artino, "Analyzing and interpreting data from likert-type scales," *J. Graduate Med. Edu.*, vol. 5, no. 4, pp. 541–542, Dec. 2013.
[48] S. Jamieson, "Likert scales: How to (ab)use them," *Med. Educ.*, vol. 38, no. 12, pp. 1217–1218, Dec. 2004.
[49] G. Norm, "Likert scales, levels of measurement and the 'Laws' of statistics," *Adv. Health Sci. Educ.*, vol. 15, no. 5, pp. 625–632, 2010.
[50] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
[51] P. K. Sen, "Gini diversity index, Hamming distance, and curse of dimensionality," *Metron-Int. J. Statist.*, vol. 63, no. 3, pp. 329–349, 2005.
[52] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[53] P. Jaccard, "The distribution of the flora in the Alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912.
[54] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," in *Proc. Mexican Int. Conf. Artif. Intell.*, vol. 18, 2014, pp. 331–345.
[55] S. Pavoine, S. Ollier, and D. Pontier, "Measuring diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable?" *J. Theor. Biol.*, vol. 228, no. 4, pp. 637–646, 2005.
[56] F. Perez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1666–1669.
[57] S.-S. Choi, S.-H. Cha, and C. C. Tappert, "A survey of binary similarity and distance measures," *J. Systemics, Cybern. Informat.*, vol. 8, no. 1, pp. 43–48, 2010.
[58] J. Salminen, K. Guan, L. Nielsen, S. Jung, and B. J. Jansen, "A template for data-driven personas: Analyzing 31 quantitatively oriented persona profiles," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2020, pp. 125–144.
[59] A. Chao, C.-H. Chiu, and L. Jost, "Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers," *Annu. Rev. Ecology, Evol., Systematics*, vol. 45, no. 1, pp. 297–324, Nov. 2014.
[60] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.
[61] A. Gonzalez De Heredia, J. Goodman-Deane, S. Waller, P. J. Clarkson, D. Justel, I. Iriarte, and J. Hernández, "Personas for policy-making and healthcare design," in *Proc. Int. Design Conf.*, Feb. 2018, pp. 2645–2656.
[62] S.-G. Jung, J. Salminen, K. K. Aldous, and B. J. Jansen, "PersonaCraft: Leveraging language models for data-driven persona development," *Int. J. Hum.-Comput. Stud.*, vol. 197, Mar. 2025, Art. no. 103445.
[63] A. Farooq, J. Salminen, J. D. Martin, K. Aldous, S.-G. Jung, and B. J. Jansen, "Exploring social media privacy concerns: A comprehensive survey study across 16 middle eastern and north African countries," *IEEE Access*, vol. 12, pp. 147087–147105, 2024.
[64] J. Salminen, S.-G. Jung, L. Nielsen, and B. Jansen, "Creating more personas improves representation of demographically diverse populations: Implications towards interactive persona systems," in *Proc. Nordic Hum.-Comput. Interact. Conf.*, 2022, pp. 1–11.
[65] K. S. Fuglerud, T. Schulz, A. L. Janson, and A. Moen, "Co-creating persona scenarios with diverse users enriching inclusive design," in *Proc. I*, 2020, pp. 48–59.
[66] D. Amin, J. Salminen, B. J. Jansen, I. Kaate, and W. Akhtar, "Large language models (LLMs) in human–computer interaction: Using LLM-generated personas to model everything from minority views to entire ecosystems," in *Artificial Intelligence & Large Language Models: A Scientific Perspective*. Boca Raton, FL, USA: CRC Press, 2026.
[67] D. Amin, J. Salminen, and B. J. Jansen, "Introducing persona ecosystem playground (PEP): Generating behaviorally grounded personas from think-aloud sessions," in *Proc. 8th ACM Conf. Conversational User Interfaces*, Jul. 2026, pp. 1–6.
[68] J. Salminen, K. Guan, S.-G. Jung, and B. J. Jansen, "A survey of 15 years of data-driven persona development," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 7, pp. 601–636, 2020.
[69] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018.

DANIAL AMIN is currently a Researcher with the University of Vaasa, focusing on generative AI personas for social good and developing ethical and inclusive user representation that benefits marginalized communities in the Global South. He is also the Data and AI Lead with Otherkind Tech, a generative AI for social good impact laboratory. He has over eight years of experience consulting multinational organizations in artificial intelligence and data science for impactful projects.

JONI SALMINEN is currently an Associate Professor (tenure track) with the School of Marketing and Communication, University of Vaasa, Vaasa, Finland. His research focuses on data-driven personas, along with overlapping topics, such as human-centered AI, quantitative UX, interactive systems, user segmentation algorithms, and so on.

BERNARD J. JANSEN received the first master's degree in computer science from Texas A&M University and the second master's degree in international relations from Troy University (formerly Troy State University) and the Ph.D. degree in computer science from Texas A&M University. He served in U.S. Army as an infantry enlisted soldier and communication commissioned officer. He is currently a Visiting Professor with the Department of Information Management, Peking University, China.