Ground-truthing perspectives on highly subjective text: basic human values perceived in song lyrics

Anonymous ACL submission

Abstract

We present an interdisciplinary approach to creating a dataset on a highly subjective text annotation task. The task thus requires explicit insight into broader human annotator perspectives and perceptions, and conscious curation of what will be annotated. In this, with strong inspiration from best practices in the social sciences, we add to emerging and increasing calls for greater accountability with regard to data and its quality. For our task, we choose the annotation of perceived human values in song lyrics. Drawing from a representative US population sample, we present our strategy to select song lyrics to be annotated, estimate the amount of annotators needed, and assess data quality. Based on this, we obtain a dataset of 360 richly annotated song lyrics. We substantiate the benefit of having more annotators, and show how annotations show promising consistency with earlier insights on personal value proximity from a validated cross-cultural instrument study. Finally, we give a first illustration of how our data can be employed in connection to applied machine learning approaches.

1 Introduction

With growing interest in AI and the rising popularity of Large Language Models, AI advances appear to push for larger datasets to train models, which ideally need few human annotations. At the same time, language is a cultural phenomenon, in which human interpretation plays a key role in transmission and understanding.

In broader situations in which applied machine learning techniques may automate and scale up actions that formerly relied on human perception and judgement, the question of what makes for good data and 'ground truth' to depart from has been less articulated and appreciated than the promise of generalizability and scalability by the applied machine learning techniques (Birhane et al., 2022; Sambasivan et al., 2021). However, calls for data-centric AI have recently been emerging[1], and recognition that human annotator disagreement can be a meaningful signal, rather than noise suggesting unreliable annotation is increasing (Aroyo and Welty, 2015). Furthermore, awareness is rising on the need for more explicit data documentation, mostly from the perspective of higher accountability on responsible data handling and reporting, and out of concern for potential societally harmful consequences of irresponsible practice (Gebru et al., 2021; Mitchell et al., 2019; Geiger et al., 2020). Efforts to more strongly institutionalize responsible practice also are visible in the *ACL communities (Rogers et al., 2021), where the completion of a Responsible NLP Checklist presently has become a mandatory element of manuscript submissions.

Computational researchers historically may not have been trained to be aware of data quality considerations. As such, standardized checklists, forms and best practice 'rules of thumb' help lowering the threshold to report and discuss these. At the same time, a simple rule of thumb may not stimulate critical reflection on current common practice. For example, as for the question, "How many annotators would be needed for NLP corpus ground truth?", a well-cited book on natural language annotation for machine learning (Pustejovsky and Stubbs, 2013) suggests to "have your corpus annotated by at least two people (more is preferable, but not always practical)" before being ready to move on to gold standard data. This is a remarkably low number, without clear substantiation of whether this indeed would be sufficient.

Beyond the computational domain, other domains and disciplines have (typically for a much longer time than the computational domain) been building expertise on how to properly curate for data, and capture aspects of the data that may not trivially be measurable. For example, both

[1] see https://datacentricai.org/

in archives and museums, long-standing traditions of purposeful and well-documented curation exist (Jo and Gebru, 2020; Huang and Liem, 2022). Furthermore, in the quantitative social sciences, well-established best practices exist for situations in which *constructs*, i.e., phenomena that cannot directly physically be measured, are to be quantified in valid and reliable ways, based on human responses. For this, the discipline of psychometrics (Furr and Bacharach, 2014) is taught as an entry-level course to students in psychology, whereas long-standing expertise from survey science (Groves et al., 2009) can further assess in robust sampling, and designing for robust human responses. While this expertise has been referred to in several works targeting computationally oriented research communities (Welty et al., 2019; Jacobs and Wallach, 2021; Kern et al., 2023), to the best of our knowledge, institutionalized methodological uptake of the expertise remains rare.

The current paper resulted from work in an interdisciplinary team, including members with disciplinary backgrounds in the computational and social sciences. Departing from an interest in developing well-substantiated ground-truthing procedures for (highly) subjective annotation tasks, we describe the creation of a dataset with annotations of perceived human values in song lyrics. We approach this task in a way that social scientists would: we increase the odds of obtaining valid and reliable measurements by being purposeful about strata in data sampling, gaining annotations from a representative human population sample, explicitly investigating the impact of higher numbers of annotations per item, and relying on evidence-backed theory, in the sense that phenomena to be studied were shown in earlier literature to have scientific and general validity, and our results can be related to earlier established results.

2 Background

2.1 Perspectivist Ground Truthing

Automated systems often rely on manually annotated reference data for training and evaluation. Multiple labels from multiple annotators are gathered for reasons associated with the annotators themselves, e.g. a lack of trust in crowdsourcing or annotations from non-experts, or because there is an expectation that people will vary in their responses to the phenomenon of interest (Cabitza et al., 2023; Basile et al., 2020). These annotations are then aggregated to produce a single label that is used to train and / or evaluate systems, as it is often incumbent on automated systems to produce a singular response.

Thus, most problems are treated as 'classification' problems. Variance occurring in the reported annotations is removed, usually by taking the label chosen most often by the annotators. Although the sources that give rise to the variance observed in the data may legitimately vary (as is possible even in 'objective' problems where annotators are medical experts (Kompa et al., 2021)), variance is often treated as an error even when a case can be made that there are indeed multiple ways to interpret the phenomenon of interest (Aroyo and Welty, 2015), e.g., when different groups of annotators reliably label media differently (Prabhakaran et al., 2023; Homan et al., 2022), or when the task itself is ambiguous (Artstein and Poesio, 2008).

A growing movement in the field of ground-truthing has taken to viewing this variation in some instances as being a necessary part of the phenomenon of interest[2]. More specifically, it is argued that the annotation projects occur on a continuum: on one end are objective phenomena whose interpretation is not expected to vary based on the perspective of the annotator, and on the other are phenomena where it is indeed expected to vary based on the lived experience, feelings etc. of the annotator (Cabitza et al., 2023). In some instances, the expectation is that there will be multiple valid labels for an item, that will systematically vary based on the social group of the person who is labelling it e.g. (Prabhakaran et al., 2023).

Although determining the degree of subjectivity of a task is a challenge, and research is ongoing in terms of appropriate methods and metrics to extract, the Perspectivist approach advocates creating and reporting disaggregated data to allow for a continuous update as to knowledge on the dataset (Liem and Demetriou, 2023).

2.2 Human values

Basic human values can be used to describe people or groups: social science theory suggests that each person uses a hierarchical list of values as life-guiding principles (Rokeach, 1973). Most widely used in social and cultural psychology is the Schwartz theory, whose formal definition is that values "(1) are concepts or beliefs, (2) pertain

[2]Sometimes referred to as the Perspectivist manifesto.

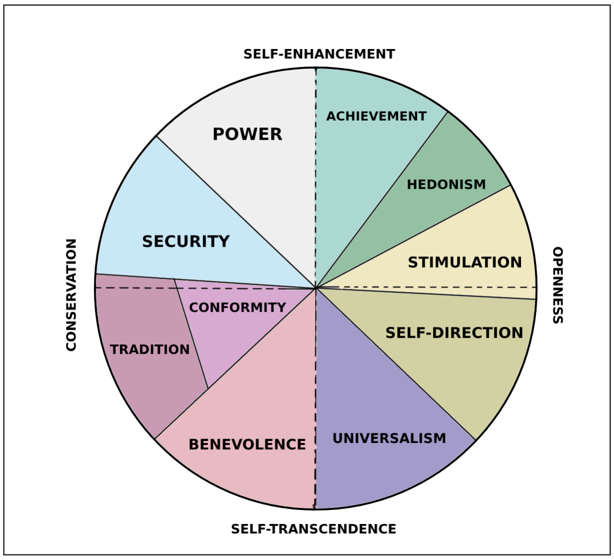

Figure 1: Visualization of the Schwartz 10-value inventory from (Schwartz, 1992) used in this paper, such that more abstract values of Conservation, vs. Openness to Change, and Self-transcendence vs. Self-enhancement form 4 higher-order abstract values. Illustration adapted from (Maio, 2010).

to desirable end states or behaviors, (3) transcend specific situations, (4) guide selection or evaluation of behavior and events, and (5) are ordered by relative importance" (Schwartz, 1992). Broadly speaking, values are abstract desirable goals that guide and motivate actions towards them across contexts (Sagiv and Schwartz, 2022).

The modern study of human values spans over 500 samples in nearly 100 countries over the past 30 years, and has shown a relatively stable structure (Sagiv and Schwartz, 2022), as illustrated in Figure 1. This has been observed across cultures in terms of the specific values present, and which values are prioritized together. Obtained scores across cultures also correlate with a broad range of impactful phenomena: e.g., cultures that value conservation and conforming to authority tend towards religiosity and away from openness and self-direction, altruistic behavior correlates with self-transcendent values like benevolence and universalism where competitiveness and unethical behavior correlate with self-enhancement goals like achievement and power, and right-wing political ideology correlates with tradition, conformity and security, where universalism values better predict left-wing ideology (Sagiv and Schwartz, 2022).

As such, the structure can be used to understand what individuals use to guide their actions, but also what entire populations prioritize when representative samples are aggregated. In addition, the relative stability of the structure allows for a convenient method to estimate the reliability and validity of measurements in novel contexts: these should, in principle, show similar structure.

2.3 Human Values in Text

Schwartz suggests that we communicate our values in order to gain cooperation and coordinate our efforts (Schwartz, 1992). As such, this communication will likely manifest in the form of words in speech and text (Boyd and Pennebaker, 2017). A vast amount of text and speech is produced and consumed: every minute in 2022 an estimated 1 million hours of content were streamed, and over 350,000 tweets were shared [3]. Thus, studying values perceived in text both is relevant from a social sciences perspective (in order to understand behaviors and priorities of diverse groups of people), and from a computational perspective (given how much text is available).

Although some work estimating the values of the authors of text has been conducted, work on how values in text are perceived is lacking: novel attempts have been made to measure the values of individuals who have written personal essays and social media posts e.g. (Maheshwari et al., 2017; Ponizovskiy et al., 2020), and in arguments abstracted from various forms of public facing text (Kiesel et al., 2022). However, we have not observed work on how to measure values perceived in text, nor work that treats the estimated values in text as a hierarchical list, in line with theory (Rokeach, 1973). Further, how language is perceived may vary substantially depending on what group is perceiving it: e.g. perceptions have been shown to vary widely by group in terms of what language is harmful (Solaiman et al., 2023; Prabhakaran et al., 2023), and how emotions are described even when there is a common structure (Jackson et al., 2019).

2.4 Music Lyrics

Music listening is an extremely popular activity. Over 616 million people subscribe to streaming services worldwide[4], and out of the music industry's reported 31.2 billion USD[5] revenue, more

[3] https://web-assets.domo.com/miyagi/images/product/product-feature-22-data-never-sleeps-10.png

[4] https://www.musicbusinessworldwide.com/files/2022/12/f23d5bc086957241e6177f054507e67b.png

[5] https://midiaresearch.com/blog/recorded-music-market-2022-reality-bites

than 17 billion comes from music streaming[6]. Out of over 1400 number-1 singles in the UK charts, only 30 were instrumental[7]. Lyrics were shown to be a salient component of music (Demetriou et al., 2018), and thus are likely to be a widely consumed form of text of importance to a broad audience. As reported in Appendix A.1, the responses of our annotation participants, who were drawn from a representative sample of the US population, quantitatively confirm the prevalence and importance of lyrics to them as music listeners, while lyrics at the same time indeed are a good data source for soliciting subjective judgements.

Although the relationship of values within the context of contemporary music remains relatively understudied, an interview study showed participant confirmation that personal values indeed play a role in people's music preferences (Manolios et al., 2019). In addition, (Gardikiotis and Baltzis, 2012) showed correlations between higher order value scores derived from the Schwartz Value Survey (Schwartz, 1992) and a dimension-reduced set of musical preference scores. Although their model explained only a small portion of the variance, (Preniqi et al., 2022) showed correlations between participants' self-reported moral values, and moral values estimated from the lyrical content of artists whose Facebook pages participants had liked.

3 Primary Lyric Data

Our aim is to collect a sample of lyric data where the lyrics are as accurate as possible, and our sample is as representative as possible. We sampled from the population of songs in the Million Playlist Dataset (MPD)[8] as it is large and recent compared to other similar datasets. The lyrics themselves were obtained through the API of Musixmatch[9], a lyrics and music language platform. Musixmatch lyrics are crowdsourced by users who add, correct, sync, and translate them. Musixmatch then engages in several steps to verify quality of content, including spam detection, formatting, spelling and translation checking, as well as manual verification by over 2000 community curators, and a local team of Musixmatch editors.

[6]https://cms.globalmusicreport.ifpi.org/uploads/Global_Music_Report_State_of_The_Industry_5650fff4fa.pdf

[7]https://en.wikipedia.org/wiki/List_of_instrumental_number_ones_on_the_UK_Singles_Chart

[8]https://research.atspotify.com/2020/09/the-million-playlist-dataset-remastered/

[9]https://www.musixmatch.com/

3.1 Fuzzy Stratified Song Sampling

An initial challenge is determining how to represent the population of songs when it is known to be very large[10]. An ideal scenario would be one in which we randomly sampled a known number of songs from a set of clearly defined strata (e.g., relevant subsets within the overall sample). However, for music, we do not know how many songs we would need to sample in order to reach saturation, what the relevant strata to randomly sample within should be, and how to measure relevant parameters from each stratum.

Some measurable strata that affect the use of language in the song lyrics are clear (e.g., the year of release, which may reflect different events or time-specific colloquial slang). Others are less clear: e.g., there is no single metric of popularity, although it can be estimated from various sources such as hit charts. Some may be subjective, such as genre, for which there may be some overlap of human labelling, but no clear taxonomy exists in the eyes of musicological domain experts (Liem et al., 2012). Based upon these considerations, we aim for a stratified random sampling procedure, based on strata that we acknowledge to be justifiable given our purpose, yet in some cases conceptually 'fuzzy': (1) release date; (2) popularity, as estimated via artist playlist frequency from the MPD (Chen et al., 2018); (3) genre, estimated from topic modeling on Million Song Dataset artist tags (Schindler et al., 2012); (4) topic, through a bag-of-words representation of the lyrics data.

To draw an initial subpopulation of songs, we first uniformly sub-sampled 60k out of 300k artists from the Million Playlist Dataset (MPD) (Chen et al., 2018). We then queried the Musixmatch API to determine if the lyrics for each of the songs of the 60k sample of artists was available.

3.2 Bias Correction

We expect that our dataset requires a bias correction. Given the skewness of data concentration with regard to several of our strata, songs that are recent and widely popular will most likely be drawn. To correct for this and get a more representative sample of an overall song catalogue, we oversample from less populated bins. For this, we use the maximum-a-posteriori (MAP) estimate of the categorical distribution of each stratum. The over-

[10]e.g., Spotify reports over 100 million songs in its cataloguehttps://newsroom.spotify.com/company-info/

sampling is controlled by concentration parameter α of the symmetric Dirichlet distribution. We heuristically set this parameter such that songs in under-populated bins still will make up up 5-10 % of our overall pool[11]. Through this method, we subsampled 2200 songs with lyrics.

3.3 Inclusion Criteria

As the annotation of highly subjective perceived values in lyrics has not been studied yet, it is unclear whether any valid and reliable annotations can be obtained from it. As such, together with the ambition to investigate many annotations from a representative population sample, it may be unwise to immediately annotate thousands of songs, but rather focus on rich insights on smaller well-curated data. For this, the following screening procedure was followed. Three members of the research team manually screened several hundreds of songs randomly sampled from our 2200 songs. They verified the match of songs to lyrics, the available metadata, and rejected songs that had words that were not English, contained very few words, were only onomatopoetic, or were only repetitions. As a consequence, we finally kept 380 songs.

4 Survey Measures

Our annotation procedure seeks to obtain human perceptions on perceived values in song lyrics. To obtain such perceptions, we design a survey, in similar fashion to how in psychometrics, a survey will be designed as a measurement instrument for psychological constructs.

With many lyrics being in the English language, we choose to obtain our annotations from representative samples of the US population in terms of self-reported sex, ethnicity and age. Such samples can be obtained through the Prolific[12] platform. We follow Prolific's guidelines on fair compensation to set our compensation rates. Survey design and data handling were pre-discussed with our institutional data management and research ethics advisors, we obtained formal data management plan and human research ethics approval, and participants give informed consent before proceeding with the survey.

4.1 Lyrics affinity and subjectivity perception

To gain further measurable evidence on the degree to which song lyrics are important yet subjective to a representative population sample, our survey starts with 16 general questions about song lyric preferences. Furthermore, after participants performed their annotations, we also ask them to rate how subjective they considered the task to be. As gaining a general understanding of these phenomena is not the main purpose of our current study though, but rather further substantiation supporting our current work, we will report on findings from these questions in Appendix A.1.

4.2 Short Schwartz Values Survey

Our primary annotations involve impressions of the values expressed in song lyrics. To this end, we adapted the Short Schwartz Values Survey (SSVS) (Lindeman and Verkasalo, 2005) to determine the wording of the questions, as it is the shortest instrument that has shown adequate reliability. The original wording of the questionnaire displays the name of the value being rated, followed by a number of words to describe it e.g. "POWER (social power, authority, wealth)"[13]. Original instructions can be found in Appendix A.2.

We made three adaptations to this questionnaire. First, we adjust the question text to ask not for ratings of life-guiding principles for the individual responding to the survey, but rather for the respondent's impressions of the 'speaker' of the lyrics. This 'speaker' is the someone or something whose perspective is reflected through the lyrics, and this may not be the author or artist expressing the lyrics. For example, the speaker in the song 'I gave you power' by the artist Nas is a gun, and the speaker in 'Rosetta Stoned' by the rock band Tool is a person hallucinating from psychedelics. In other words, the creator may use a persona in the writing of song lyrics for artistic purposes, which may not directly represent their values. As such, an annotator's impression of the creator may differ from their impression of the person represented by the lyrics. As we are interested in the values perceived in the text, we explicitly ask participants to respond with the perspective of the speaker in mind, and not the author. Further illustration of our explicit instructions is given in Appendix A.2.

Secondly, the original SSVS uses a 9 point Likert-Scale where 0="Opposed to My Principles", 1="Not Important", and 8="Of Supreme Importance". In order to gather continuous measure-

[11] Full code of our sampling procedure is at https://anonymous.4open.science/r/lyrics-value-estimators-CE33/1_stimulus_sampling/stratified_sampling.py

[12] https://prolific.co

[13] Actual wording of items was retrieved from https://blogs.helsinki.fi/everyday-thinking/files/2015/11/The-Short-Schwartzs-Value-Survey.docx.

ments, we aimed for a continuous scale where 0 essentially indicates that the value is either not discussed or otherwise not estimatable from the lyric text. Next to this, we balance the scale to have maximum opposition to a given value be at -100, where maximum importance will be at 100. As such, in contrast to the original SSVS, our scale is symmetric with a rating of 0 indicating neutrality.

Thirdly, (Cabitza et al., 2020) suggested that a rater's confidence is an indication of intra-rater reliability. Thus, we also asked participants "How confident are you in your ratings of these lyrics?", to which they responded on a scale of 0 (Not at all Confident) to 100 (Completely Confident).

4.3 Annotation Interface

The survey was implemented on an instance of the Qualtrics[14] platform. The annotations were collected using the response format shown as illustrated in Figure 10, following explicit instructions as discussed in Appendix A.2. More specifically, a set of lyrics are displayed, with a clickable interface below them. The interface contains brief descriptions of each of the 10 Schwartz values, followed by a vertical bar on which participants can indicate a continuous response, as described in Section 4.2. An option to select "Not Applicable" was also available for each value. We considered that "0" and "Not Applicable" responses both indicate that the importance of that given value to the speaker based on the lyrics could not be determined by the participant (i.e., they were either not discussed in the song lyrics, or were otherwise unclear). As we expect that not all songs will discuss all values, and most songs may discuss very few values, we initialize the rating bar at "0".

With this, we now have the setup to gather annotation data. In the remainder of this paper, we discuss how this was done to research three questions: (1) How many annotator ratings are needed for stable annotations to emerge? (2) Do our obtained value perception annotations relate to existing validated knowledge on stable structures among values? and (3) Can our refined annotations be used in computational NLP setups?

5 Number of Ratings

Our procedure to determine the number of ratings to gather was inspired by (DeBruine and Jones, 2018). Specifically, we first recruited a representative pilot sample (N=505), in which respondents used our interface to annotate perceived values for a fixed set of 20 songs. From these annotations, we computed canonical mean ratings per value, per song. For each of the values, we then estimated Cronbach's α for a range of subsample sizes (5 to 50 participants, in increments of 5), repeating this procedure 10 times per increment. Following this, we visually examined density plots of the distribution of Cronbach's α. In the social sciences, an $\alpha >= 0.7$ is commonly considered an acceptable level of reliability. Taking a conservative estimate, we choose to obtain 25 ratings per song lyric in our main study; for that amount of ratings, Cronbach's α in our pilot data would comfortably exceed 0.8.

From this, we perform our main study data collection. We recruit a new representative US population sample (N=600), where each participant goes through our survey questions, and receives 18 randomly selected song lyrics to annotate for perceived values. As a result, we obtained 22-30

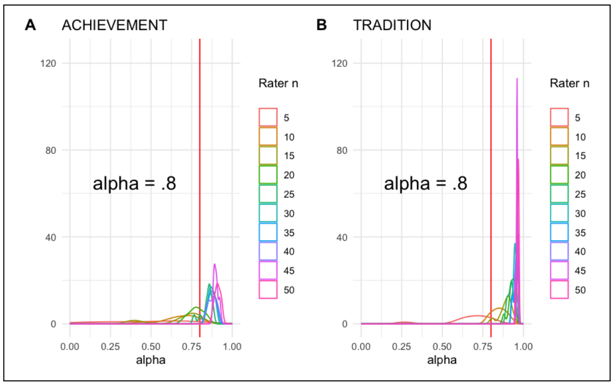

Figure 2: Distribution of Cronbach's α from a representative US Sample (N=505) rating 20 songs, for the values Achievement and Tradition. Vertical line represents the α threshold for comparison.

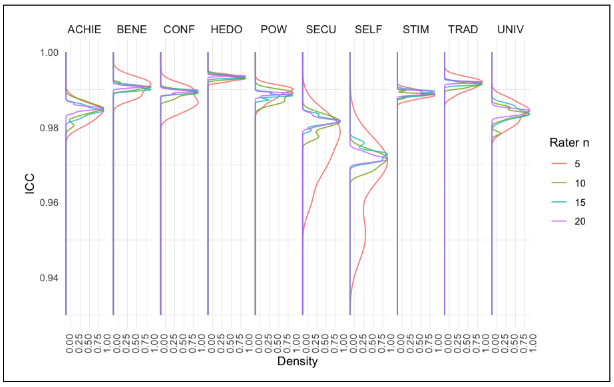

Figure 3: Rotated scaled density plots of ICC for subsamples from annotations on the 360 songs.

[14]https://www.qualtrics.com/

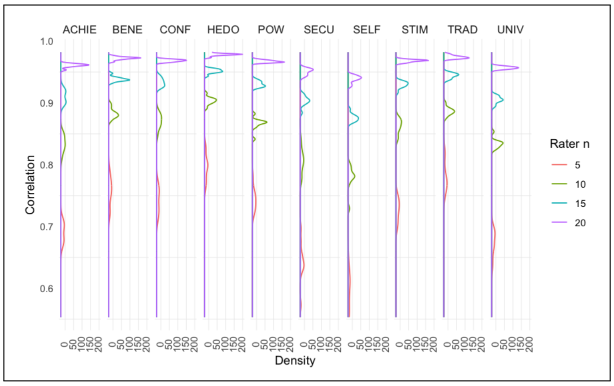

Figure 4: Rotated scaled density plots of Pearson correlations between canonical mean and subsample means, from a mean 27 ratings per each of the 360 songs.

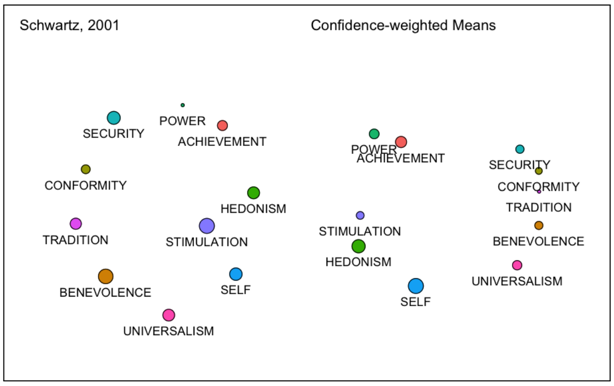

Figure 5: MDS plots derived from the correlation plot reported in (Schwartz et al., 2001), and our participant responses as confidence-weighted means[15].

	ACHIEVEMENT	BENEVOLENCE	CONFORMITY	HEDONISM	POWER	SECURITY	SELF	STIMULATION	TRADITION	UNIVERSALISM
word_count	-0.19	0.04	0.09	-0.17	-0.11	0.04	-0.1	-0.19	0.13	0.09
google_uni	-0.05	0.26	0.1	0.2	0.06	0.17	-0.05	0.09	0.29	0.12
google_idf	0	0.18	0.03	0.26	0.18	0.1	0.04	0.18	0.23	0.02
glove_uni	-0.07	0.28	0.24	0	-0.06	0.22	-0.07	-0.14	0.32	0.18
glove_idf	-0.02	0.17	0.17	0.18	0.01	0.1	0.01	0	0.25	0.08
far_uni	0.07	-0.18	-0.19	0.03	0.1	-0.25	0.01	0.05	-0.09	-0.02
far_idf	0.05	-0.12	-0.13	-0.05	0.1	-0.19	-0.04	-0.01	0	0.03
cv_uni	0.07	-0.18	-0.19	0.02	0.11	-0.25	0.02	0.06	-0.08	-0.01
cv_idf	0.05	-0.11	-0.12	-0.06	0.09	-0.19	-0.04	-0.01	0.01	0.04

correlation: 0.3, 0.2, 0.1, 0.0, -0.1, -0.2

Figure 6: Pearson correlations between NLP systems / word counts, and participant ratings of songs, by value.

annotations per song, with an average of 27.

From these, checking for the reliability of our annotations from this sample, we repeatedly subsample 5, 10, 15 and 20 ratings for each value within each song, and calculate and visualize intraclass correlations (ICC, Figure 3), as well as Person correlations between subsample means and canonical means (Figure 4). From this, we see that higher numbers of raters lead to higher ICCs and Pearson correlation, thus indicating more stable outcomes that approach the canonical mean that would be obtained from a large sample of annotators. Seeing Pearson correlation to the canonical mean already exceeds 0.9 for all values from 15 subsampled ratings, our target of 25 annotations per songs indeed can be considered too conservative, and in future work, 15 ratings on average will likely suffice.

For our further analysis, we will have to aggregate the subjective labels. Being unaware of a single ideal method to achieve this, for the purposes of this work, we report results using an aggregation method inspired by (Cabitza et al., 2020). Specifically, we estimate confidence-weights by dividing participant's self-reported confidence of a given rating by the highest possible response (100), and then compute aggregated means weighted by these.

6 Structural comparison

As a first attempt to assess the relative validity of our procedure, we depart from the earlier observations that cross-cultural stable structure was found on what values are likely to cluster together. We compare distances as derived from the upper triangle of a correlation matrix reported in (Schwartz et al., 2001) to those derived from proximity in ratings obtained in our study. For both, we generate a multi-dimensional scaling plot (MDS) for visual comparison, which has previously been used as method to assess confirmation of earlier theory (Ponizovskiy et al., 2020). From these plots (Figure 5, in as little as our 360 annotated lyrics, we surprisingly indeed see similar clusters and relative positioning relations emerging as those obtained from a formal cross-cultural study.

7 Comparisons with NLP models

Finally, as first step towards ways in which our data may be connected to computational NLP methods, we perform a preliminary comparison on how computational NLP-based value assessment of lyrics data compares to the way in which our annotators annotated perceived human values.

We depart from a dictionary of words associated with the 10 Schwartz values (Ponizovskiy et al., 2020). With this dictionary as reference, we computationally perform assessments to the degree to which each value would be reflected in the lyrics text according to traditional word counting (Poni-

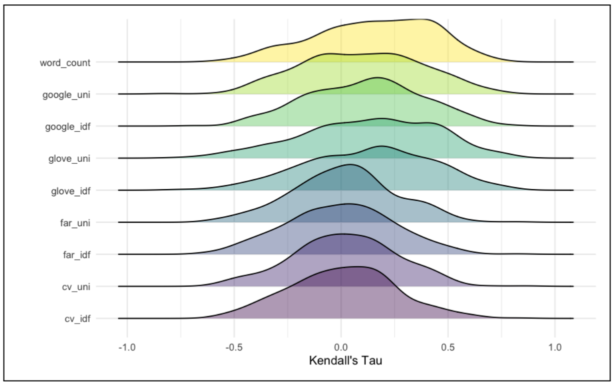

Figure 7: Rank correlations between NLP systems / word counts and confidence-weighted participant means transformed to rankings

zovskiy et al., 2020), as well as by assessing cosine similarity between dictionary words and lyrics texts using four classes of pre-trained word embeddings: `word2vec-google-new`, a generic English word embedding trained on Google News dataset (Mikolov et al., 2013); `glove-common-crawl`, another generic English word embedding trained on Common Crawl dataset (Pennington et al., 2014); `faruqui-mxm-[1∼10]`, trained on the collected initial lyrics candidate pool, employing the Glove model (Pennington et al., 2014) (using ten models populated from ten cross-validation folds, whose parameters are tuned based on English word similarity judgement data (Faruqui and Dyer, 2014).); and `cv-mxm-[1∼10]`, ten variants of lyrics based word-embeddings from cross-validation folds selected by Glove loss values on the validation set.

We weigh terms in the lyrics texts in two different ways: uniformly and weighted by Inverse Document Frequency (IDF). Then, we compare value assessments from these computational methods to the ones obtained from our annotators, in two ways. First, we consider Pearson correlations of computational value similarity assessments to our raw participant song value annotations (Figure 6). Second, we take the earlier-theorized the perspective that value assessments should be seen as ranked lists, and we consider rank correlations between the machine and human value assessments based on Kendall's τ (Figure 7).

In earlier work (Richard et al., 2003), Pearson correlations of 0.1-0.2 were considered as moderate evidence of the validity of a proposed dictionary in relation to a psychometrically valid instrument. As such, considering our data as good reference data, only the more generic Glove and Google news embeddings seem to reach those levels of correlation. From a rank correlation perspective, the word count methods and these two embedding models hint at slightly positive rank correlations. This may be promising in terms of the degree of specialization needed to assess values; at the same time, neither of the methods presented here have thoroughly been optimized, and as such, these results should not be seen as strong benchmarking evidence. Future work will be needed to more deeply connect computational NLP techniques with our data.

8 Limitations and future work

In this paper, we described our procedure for ground-truthing perspectives on highly subjective text. By paying attention to grounding in social sciences theory and purposeful sampling strategies, the discussion of how to get to 'good data' has been much more extensive than commonly is done in computational domains. With this, we hope to have illustrated how beyond (welcome) completion of checklists and data sheets, being purposeful about data can pro-actively shape annotation design.

As for limitations to our work, while we are committed to open science practices, we cannot share the primary lyric data due to copyright prohibitions. However, we do release metadata of the songs of interest, together with our participant annotations, and the code used for the analyses and plots in our paper [16]. We acknowledge our current sample of 360 lyrics is still small and may need expansion, an that, while we had a representative population sample, not every member of the sample rated every song. We thus did gather diverse opinions, but cannot claim they fully represent the target population. We also did not assess whether variations on the annotation instrument might result in substantial differences in the annotations we received (Kern et al., 2023), nor did we repeat our procedure (Inel et al., 2023). In addition, we can further connect our work to related research on examining how participants from different groups will annotate corpora (Homan et al., 2022; Prabhakaran et al., 2023). Finally, while we only provide a preliminary comparison to computational NLP methods, it will be worthwhile to use our data in the context of more sophisticated state-of-the-art NLP systems.

[16] https://anonymous.4open.science/r/values_in_lyrics-8F3F/

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: A new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2020. As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. *BMC Medical Informatics and Decision Making*, 20(1):1–21.

Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 527–528.

Lisa M DeBruine and Benedict C Jones. 2018. Determining the number of raters for reliable mean ratings.

Andrew Demetriou, Andreas Jansson, Aparna Kumar, and Rachel M Bittner. 2018. Vocals in music matter: the relevance of vocals in the minds of listeners. In *ISMIR*, pages 514–520.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 19–24. The Association for Computer Linguistics.

R. Michael. Furr and Verne R. Bacharach. 2014. *Psychometrics : an introduction*, second edition edition. SAGE Publications.

Antonis Gardikiotis and Alexandros Baltzis. 2012. 'rock music for myself and justice to the world!': Musical identity, values, and music preferences. *Psychology of Music*, 40(2):143–163.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey methodology*, volume 561. John Wiley & Sons.

Christopher Homan, Tharindu Cyril Weerasooriya, Lora Aroyo, and Chris Welty. 2022. Annotator response distributions as a sampling frame. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 56–65.

Han-Yin Huang and Cynthia C. S. Liem. 2022. Social inclusion in curated contexts: Insights from museum practices. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Oana Inel, Tim Draws, and Lora Aroyo. 2023. Collect, measure, repeat: Reliability factors for responsible ai data collection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 51–64.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability and Transparency (FAT '20) January 27-30 2020, Barcelona, Spain*. ACM, New York, NY, USA.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. *arXiv preprint arXiv:2311.14212*.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471. Association for Computational Linguistics.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4.

Cynthia C. S. Liem, Andreas Rauber, Thomas Lidy, Richard Lewis, Christopher Raphael, Joshua D. Reiss, Tim Crawford, and Alan Hanjalic. 2012. Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In *Dagstuhl Follow-Ups*, volume 3. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

Cynthia CS Liem and Andrew M Demetriou. 2023. Treat societally impactful scientific insights as open-source software artifacts. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 150–156. IEEE.

Marjaana Lindeman and Markku Verkasalo. 2005. Measuring values with the short schwartz's value survey. *Journal of personality assessment*, 85(2):170–178.

Tushar Maheshwari, Aishwarya N Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck, and Amitava Das. 2017. A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 731–741. Association for Computational Linguistics.

Gregory R Maio. 2010. Mental representations of social values. In *Advances in experimental social psychology*, volume 42, pages 1–43. Elsevier.

Sandy Manolios, Alan Hanjalic, and Cynthia CS Liem. 2019. The influence of personal values on music taste: towards value-based music recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 501–505.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality*, 34(5):885–902.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. A framework to assess (dis) agreement among diverse rater groups. *arXiv preprint arXiv:2311.05074*.

Vjosa Preniqi, Kyriaki Kalimeri, and Charalampos Saitis. 2022. " more than words": Linking music preferences and moral values through lyrics. *arXiv preprint arXiv:2209.01169*.

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning* , first edition edition. O'Reilly Media.

F Dan Richard, Charles F Bond Jr, and Juli J Stokes-Zoota. 2003. One hundred years of social psychology quantitatively described. *Review of general psychology*, 7(4):331–363.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Milton Rokeach. 1973. *The nature of human values.* Free press.

Lilach Sagiv and Shalom H Schwartz. 2022. Personal values across cultures. *Annual review of psychology*, 73:517–546.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Thomas Schäfer and Peter Sedlmeier. 2009. From the functions of music to music preference. *Psychology of Music*, 37(3):279–300.

Alexander Schindler, Rudolf Mayer, and Andreas Rauber. 2012. Facilitating comprehensive benchmarking experiments on the million song dataset. In *ISMIR*, pages 469–474. International Society for Music Information Retrieval.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, 32(5):519–542.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949*.

Chris Welty, Praveen K. Paritosh, and Lora Aroyo. 2019. Metrology for AI: from benchmarks to instruments. *arXiv preprint arXiv:1911.01875*.

A Appendix

A.1 Lyrics affinity and subjectivity perception

Our data collection protocols allowed us to gather self reports on the importance of lyrics. Our initial pool of questions was inspired by the Preference Intensity scale in (Schäfer and Sedlmeier, 2009), and consisted of Likert-type questions. We turn these into 16 question statements on the participants' relation to music lyrics by also adding our own suggested quetions. Participants respond to the questions using a 5-point Likert scale, which included the points “Strongly Disagree", “Somewhat Disagree", “Neither Agree nor Disagree", “Somewhat Agree", and “Strongly Agree". Percentages in the table below indicate the proportion of respondents that indicated either “Somewhat Agree" and “Strongly Agree".

We currently report on responses given to these questions from our two data collection rounds: our pilot study (N=505), whose primary aim was the estimation of the number of ratings needed per song lyric, and our actual annotation collection study (N=600) on which the main outcomes in our paper are reported.

In Figure 8, we visualize self-reported percentages of respondents' music libraries containing lyrics, for both our respondent samples. Here, we see that respondents' music overwhelmingly contains lyrics, with a median of 90%. Furthermore, in Table 1, for both samples of respondents, we indicate the percentages of users that indicated to somewhat or strongly agree with given statements.

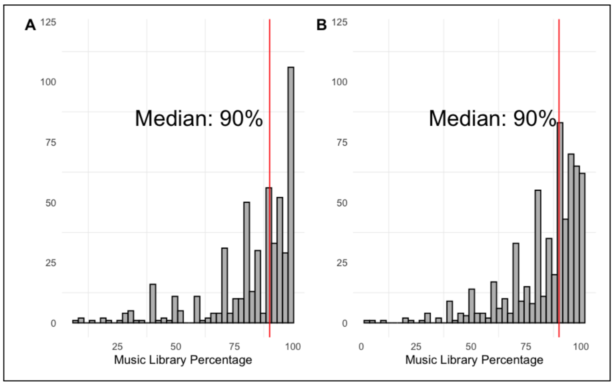

Figure 8: Distribution of self-reported percentage of music library containing lyrics from two representative US samples, N=505 and N=600 respectively.

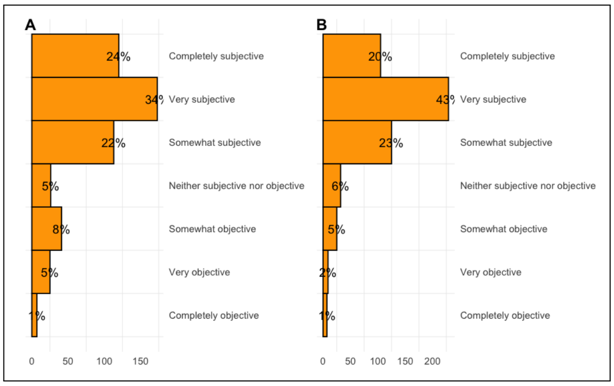

Figure 9: Distribution of self-reported subjectivity of lyric annotation task, N=505 and N=532 respectively.

From this, we again observe a strong preference for songs with lyrics (>70% on many of the statements).

Finally, at the end of our survey, we also ask participants to self-report a rating of the subjectivity of the lyrics annotation task we gave to them. Distributions are visualized in Figure 9. From these, we see confirmed the task indeed is perceived as highly subjective in the eyes of our sample population.

As gaining a general understanding of music lyrics affinity is not our current main goal, we chose not to iteratively validate and refine our questions as a formal psychometric instrument at this stage (for this, more explicit iterative analysis would be needed on the instrument being capable of distinguishing between different types of users by making use of the full scale). However, we did start analyzing to what extent the current questions may be used as an instrument, or at least as a way to further characterize sub-populations of human respondents. Here, given the large preference towards music that contains lyrics, asking for lyrics vs. non-

Question	Pilot	Main
`I prefer music that contains lyrics, as opposed to music that does not`	72%	72%
`I always pay attention to the lyrics of a song, if the song has them`	70%	72%
`If a song has lyrics that I don't like for any reason, I don't listen to it`	49%	43%
`If I am not sure about the lyrics of a song, I search them on the internet`	76%	77%
`I memorize the lyrics to the songs I listen to`	70%	75%

Table 1: Question wording, and proportion of respondents rounded to the nearest whole number, that indicated either 'somewhat agree' or 'strongly agree' in two surveys, N=505, and N=600 respectively.

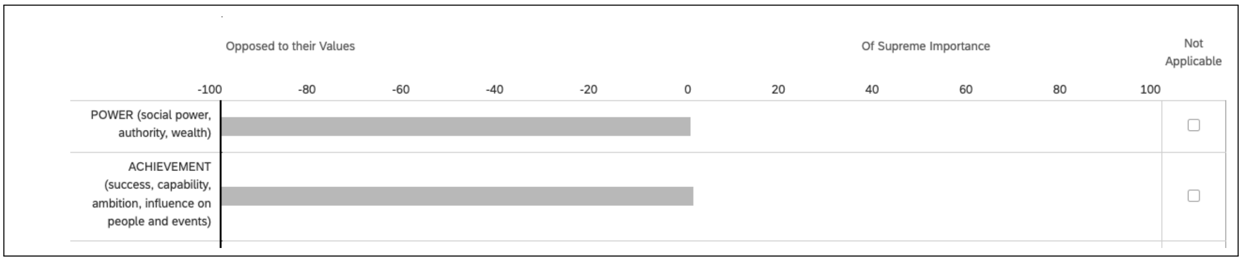

Figure 10: Visualization of the annotation interface on Qualtrics for two of ten annotated values

lyrics music preference will not allow for us to be able to distinguish between respondents. At the same time, responses to the degree to which a respondent pro-actively engages with lyrics (e.g. by actively searching for them, writing about them, or writing lyrics themselves) may yield interpretable factors on which respondents can be distinguished. However, we leave a deeper analysis of this for future work.

A.2 Adjusted Short Schwartz Value Survey

The original Schort Schwartz Value survey appears in (Lindeman and Verkasalo, 2005). The original question wording[17] was:

"Please, rate the importance of the following values as a life-guiding principle for you. Use the 8-point scale in which 0 indicates that the value is opposed to your principles, 1 indicates that the values is not important for you, 4 indicates that the values is important, and 8 indicates that the value is of supreme importance for you."

- POWER (social power, authority, wealth)
- ACHIEVEMENT (success, capability, ambition, influence on people and events)
- HEDONISM (gratification of desires, enjoyment in life, self-indulgence)
- STIMULATION (daring, a varied and challenging life, an exciting life)
- SELF-DIRECTION (creativity, freedom, curiosity, independence, choosing one's own goals)
- UNIVERSALISM (broad-mindedness, beauty of nature and arts, social justice, a world at peace, equality, wisdom, unity with nature, environmental protection)
- BENEVOLENCE (helpfulness, honesty, forgiveness, loyalty, responsibility)
- TRADITION (respect for tradition, humbleness, accepting one's portion in life, devotion, modesty)
- CONFORMITY (obedience, honoring parents and elders, self-discipline, politeness)
- SECURITY (national security, family security, social order, cleanliness, reciprocation of favors)

In our survey, participants were initially shown a set of instructions designed to explain how to use the instrument, and explain our working definitions of 'artist' as separate from the 'speaker' of the lyrics, see(Figure 11). We then presented our adjusted question wording:

"Between the quotation marks below are some song lyrics. Please take a moment to read them

[17] retrieved from https://blogs.helsinki.fi/everyday-thinking/files/2015/11/The-Short-Schwartzs-Value-Survey.docx.

and think about the SPEAKER the lyrics. Please remember that this SPEAKER might be a the AUTHOR themselves, or someone or something else:", after which lyrics were displayed, along with the annotation instrument.

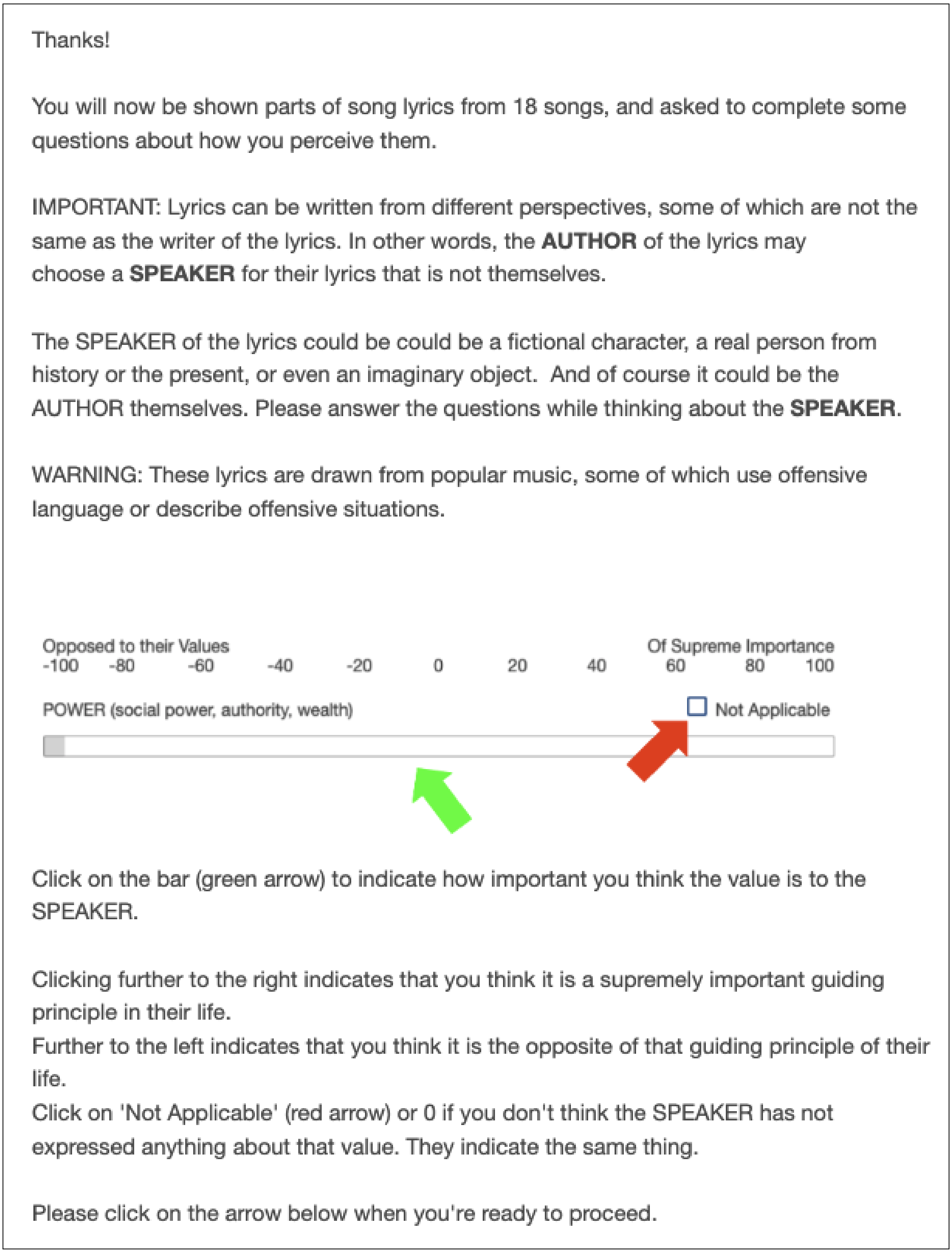

Figure 11: Visualization of the instructions page of the annotation interface on Qualtrics.