Representation in Corpus Studies of Music: Commentary on Shea's (2022) "A Demographic Sampling Model and Database for Addressing Racial, Ethnic, and Gender Bias in Popular-music Empirical Research"

TREVOR DE CLERCQ [1]
Middle Tennessee State University

ABSTRACT: This commentary responds to the article by Nicholas Shea in Vol. 17(1) of this journal, which offers a method to encode the demographic information of the artists in a corpus of popular music and then describes a resampling procedure to create a new corpus based on specific demographic targets. I argue that attempts to diversify the demographic distribution of popular music corpora should occur not at the sampling stage but instead at a prior stage when the researcher is determining the musical style or era to study. Once this musical style or era has been determined, empirical principles oblige the researcher to create a corpus that faithfully represents the statistical population.

Submitted 2023 December 3; accepted 2023 December 27.
Published 2024 June 7; https://doi.org/10.18061/emr.v18i2.9726

KEYWORDS: *popular music, race, ethnicity, gender, sampling, corpus studies*

AS is well known, the field of music theory has traditionally been focused almost exclusively on the music of White men. Ewell (2020), for instance, found that in the seven leading American music theory textbooks published between 2010 and 2018—which together represented 96% of the total market share at the time—only 1.67% of the musical examples were written by non-White composers. Subsequent research by Maust (2023) found that only 2.15% of these examples were composed by women. In response, music theorists have in recent years called for more demographic diversity regarding the musicians being discussed in an academic context—whether in textbooks, scholarly articles, or elsewhere—to better represent the population at large (Campbell et al., 2015; Ewell, 2023; Hisama, 2021; Palfy and Gilson, 2018).

To help foster demographic diversity within music theory discourse, Nicholas Shea (2022)—in his data report from Vol. 17(1) of this journal—offers an encoding scheme to compensate for what he identifies as racial, ethnic, and gender biases in corpus studies of popular music. He notes, for example, that existing corpora of popular music are, like most American music theory textbooks, comprised primarily of music written and performed by White men—citing in particular the McGill *Billboard* project (Burgoyne et al., 2011), the *Rolling Stone* magazine corpus (de Clercq & Temperley, 2011; Temperley & de Clercq, 2013), and his own sample from ultimate-guitar.com (Shea, 2020). There thus exists, Shea argues, "a continued need for researcher intervention to avoid amplifying real-world biases" (p. 49). After discussing the details of his encoding scheme and database, Shea describes in the final section of his report (entitled "Potential Applications") how his database could be used to reshape the demographic distribution of a corpus by using a quota-based sampling method. In the supplemental R script that Shea provides, a 200-song corpus is created by selecting 100 artists who belong to a minority racial or ethnic group, of which 50 are selected as also non-male.

Shea's database represents a great deal of time and effort, and I am sympathetic to the underlying goals that inspired him to propose this encoding scheme and sampling method. In earlier publications, I have made my own suggestions about how to address the lack of diversity within the field of music theory, particularly regarding the repertoire studied in the core undergraduate curriculum (e.g., de Clercq, 2019; de Clercq, 2020). I also acknowledge and accept the inherent challenges of assessing diversity using quantitative measures, and I believe the encoding scheme proposed by Shea is sufficient to give a general sense of the demographic distribution in a corpus when no operationalized method previously existed. Overall, I agree

with Shea that for any researcher conducting a corpus study, the issue of representation should be front and center.

But what should representation look like in the context of a corpus study? Shea is mostly silent on this issue, writing that his purpose is "not to determine a universal benchmark for artist diversity in corpus studies" (p. 54); each researcher, he says, will individually "need to determine which parameters best suit their representational needs" (p. 54). In this commentary, therefore, I hope to offer some advice in that regard, particularly given the standard assumptions of modern quantitative methods. The issues surrounding adequate representation are not new in statistical studies, of course. As I argue below, however, representation and diversity are two distinct and separate goals, which are often (if not usually) in direct opposition when conducting a corpus study of music. Moreover, it is representation, not diversity, that should be the main goal when creating a corpus, given a particular population of interest. This is not to say that the goal of diversity is not important or relevant to corpus work. On the contrary, the goal of diversity is the first and most important step in the creation of a corpus since it determines the population of interest.

REPRESENTATION AND DIVERSITY

Shea (2022) begins his paper by writing that "[a] core tenet of empirical research is that a robust sample offers a better approximation of tendencies within the broader target population" (p. 49). This is a fair statement and one with which I would certainly agree. In this context, the term "population" is used in the statistical sense, meaning some set of items or objects that belong to a class or group that a researcher would like to study. If we wanted to determine how many glass vials were cracked in a shipment of ten million glass vials, the statistical population would be the ten million glass vials. From this population of glass vials, we could select a random sample, since investigating each glass vial would be labor-intensive. Having collected our sample, we could then offer an estimate as to the number of cracked glass vials overall.

In his second sentence, Shea continues: "Despite this, datasets developed for music research do not always reflect real-world population diversity regarding an artist or composer's racial, ethnic, and/or gender identity" (p. 49). Now, Shea is using the word "population" in quite a different sense—referring to all the people living in a country, with "real-world population diversity" referring to the demographic makeup of those people. While the connection between Shea's opening two sentences may thus at first seem sound, since both discuss issues related to a population, notice that there has been a subtle but significant shift between two different meanings of the term "population," which sometimes overlap although more often do not. The population of glass vials described previously, for example, has nothing to do with the population of people living in a country.

What, then, is the statistical population for a corpus study of music? The answer may depend on the research question, but generally speaking, the statistical population for a corpus study of popular music is a set of songs, typically the set of songs belonging to some particular musical style or era. If we wanted to conduct a corpus study of drumbeats in disco music from 1975 to 1980, then the population (from a statistical perspective) is the set of all songs from 1975 to 1980 that we would consider to be disco music. Although it is easy enough to define this (or any other) population of songs in theory, it is more difficult to identify all the members of that population in practice. Presumably, songs such as "Stayin' Alive" by the Bee Gees (1977) and "I Will Survive" by Gloria Gaynor (1978) would count as disco music from this era, but other songs are more difficult to categorize. Is the song "Miss You" by the Rolling Stones (1978) truly disco? What about "Driver's Seat" by Sniff 'n' the Tears (1978)? Should we include songs from the old-school years of hip-hop, such as "The Breaks" by Kurtis Blow (1980) or "Rapper's Delight" by The Sugarhill Gang (1979)?

Ultimately, the decision of which songs to include in a corpus is something that requires serious consideration. After all, the value of any statistical summaries or inferences we derive from a corpus is predicated on how well that corpus represents the statistical population (as Shea reminds us in his opening sentence). If a corpus of blues music, for example, were to include only songs that the researcher who created the corpus knew how to play on guitar, this corpus would likely not represent blues music overall very well.

The choice of which songs to include in a corpus is thus a question of how to represent the statistical population most faithfully, and there have been different ways to approach this issue. Because it is difficult, if not impossible, to identify all the possible songs that belong to a particular style, music researchers have traditionally used some version of the *exemplar* sampling method (Bull, 2005), also known as *typical case* sampling (Patton, 2015). Following an exemplar method, the corpus is created by selecting songs that by some objective measure (i.e., not the researcher's subjective opinion) are considered to best represent the musical style. To create the *Rolling Stone* magazine corpus, for example, Temperley and I (2011, p. 51)

selected songs based on their critical acclaim, using a list of the "greatest rock & roll songs of all time" as compiled by 172 rock stars and leading authorities in 2004. The McGill/Billboard corpus, in contrast, was created by selecting songs based on their commercial success, as ranked on the *Billboard* "Hot 100" charts between 1958 and 1991 (Burgoyne et al., 2011). In my study of harmony in country music (2022), I used a hybrid approach, selecting songs through a combination of chart data as well as lists of award-winning songs. In each of these cases, the style is taken to be best represented by songs that are more well-known, commercially successful, or critically acclaimed. This method, admittedly, does not sample exhaustively or equally from the statistical population, which puts it in contrast to a classic simple random sample (described above in the discussion of glass vials). However, because the boundaries of any musical style are unclear, the exemplar sampling method has been the best available approach for music researchers to attempt to accurately represent the norms of a musical style or era. To be fair, music researchers may perhaps need to better wrestle with the idea of how to identify and randomly sample from all the possible songs that would represent a particular musical style or era, given the assumptions of standard statistical tests.

That said, the method that Shea offers in his article does not bring us closer to that ideal. Instead, I fear Shea's method brings us farther from it, since he suggests that a researcher should make decisions about the songs in a corpus based not on how well those songs represent a musical style or era but instead according to some diversity goal or quota (such as half of the songs being by artists that belong to a racial or ethnic minority). Unfortunately, those two objectives are likely in conflict. Consider, for example, the corpus of rap songs compiled by Condit-Shultz (2016), which includes rhythmic and lyric transcriptions of 124 rap songs sampled randomly from the *Billboard* charts between 1980 and 2015. As one might expect, the overwhelming majority of artists in this corpus are Black, with only a handful of White artists. Therefore, this corpus does not have much demographic diversity, nor should it, since including a significantly greater proportion of White artists would misrepresent rap music from this era. Similarly, my corpus study of country music (2022) includes harmonic transcriptions for 200 country songs spanning from 1933 to 2014, and the overwhelming majority of artists in this corpus are White, with only a handful of Black artists. Like Condit-Shultz's corpus of rap music, my corpus of country music does not have much demographic diversity, nor should it, since manipulating the demographic content of the corpus would misrepresent country music during this period.

We may wish or hope that rap music or country music (or some other musical style) would have more demographic diversity. But if we as researchers purport to be conducting empirical work, then our goal with any corpus should be to represent the musical style or era we wish to study in the most historically honest way possible. I do not disagree with Shea that popular music (or any style of music) is a byproduct of its cultural context and is therefore wrapped up with social norms and particular histories of racial and gender discrimination that were endemic at the time. But it's not a logical consequence to say that demographic skews within a particular musical style are necessarily the result of this discrimination. In fact, Shea's data appears to contradict this basic premise. The proportion of BIPOC artists in the *Rolling Stone* magazine and McGill *Billboard* corpora, for example, is substantially greater than 30% by any measure, as shown in Shea's Figure 2 and Table 2. This statistic far exceeds the proportion of BIPOC residents in any English-speaking country during the decades covered by these corpora, both of which center on the early 1970s. In the United States, for example, the proportion of the population that identified as non-White is significantly less than 30% around this time, as shown below in Table 1. More starkly, the proportion of United Kingdom residents who belonged to an ethnic minority in 1971 was less than three percent.[2] Shea's statements, then, that Black artists are underrepresented in rock music or popular music (p. 51 & 54) are thus not supported by his demographic analyses of these corpora.

Table 1. Available demographic data, as percentages by decade, for the United States, 1950–1990[3]

	1950	**1960**	**1970**	**1980**	**1990**
White	89.5	88.6	87.7	83.1	80.3
Non-Hispanic White	87.5	85.4	83.5	79.6	75.6
Hispanic (of any race)	2.1	3.2	4.4	6.4	9.0
Black	10.0	10.5	11.1	11.7	12.1
Native	0.2	0.3	0.4	0.6	0.8
Asian	0.2	0.5	0.8	1.5	2.9
Other Races	0.0	0.0	0.1	3.0	3.9

My point here is not to deny that discrimination was (or is) a part of history (or of music history). Rather, my point is that it is currently unknown to what extent, if any, discrimination or bias is responsible for any demographic skews we observe within a musical style or era, and we cannot assume that all demographic skews are necessarily the result of discriminatory behavior. It may instead be that different demographic groups have different preferences, musical or otherwise, with the observed demographic skews reflecting those preferences. If our goal is to study a particular style or era of music, in other words, part of what we are studying is the preferences, both musical and non-musical, of different races and genders.

LOOKING FORWARD

It may seem that my underlying argument is that there is nothing we as researchers can do—or rather, should do—to combat the documented sexism and racism of decades past (or today) when conducting a corpus study. From an anti-racist or anti-sexist perspective, this approach may seem problematic, since it does not compensate for any historical bias (known or unknown) that may have led to a particular demographic skew, perhaps resulting in the exclusion or marginalization of certain groups. Why, though, does a researcher need to study music that involves some particularly undesirable demographic skew?

If we hope to give a greater voice to a specific demographic group in academic research, then we should study the music created by and listened to by that group, making every attempt to best represent that style or era when creating a corpus. Shea, for example, notes that female Black musicians are "underrepresented in current corpora" (p. 54). I can imagine many ways to counteract this apparent trend. A researcher could, for example, create a corpus of music by "girl groups" from the 1960s, when songs by Motown artists dominated the charts. Or a researcher could create a corpus of contemporary R&B, which has a roughly equal split between genders and includes music primarily by Black artists.[4] Alternatively, a researcher could create a corpus that is explicitly designed to study Black female artists in the 1980s glam metal genre. Although exemplars of 1980s glam metal songs by Black female artists are more difficult to find than those by White men, this corpus clearly identifies its statistical population and then attempts (or should attempt) to represent that specific style in the most accurate way possible. If we identify the statistical population clearly (whatever it may be) and then try to represent that statistical population in a historically authentic way, then (and only then) will we have confidence that our statistical analyses have meaning and inferential value.

In contrast, it would be unclear what sort of inferences we would be able to draw from a corpus of popular music resampled according to the methodology proposed by Shea. What statistical population would that corpus represent? Instead of better representing a particular style or era, a corpus curated along those lines does not, I would argue, represent any statistical population at all, and as a result, any statistical analyses we might conduct with that corpus have no real inferential value. To be clear, I am not in principle opposed to quota-based sampling methods. The issue is rather that there is no reason to expect demographic distributions within a musical genre to match the general population (or any other target), since there is a strong correlation between artist demographics and style, with certain races and genders correlating positively with particular musical genres and negatively with others. Perhaps corpus researchers to date have focused an inordinate amount of attention on "rock" music (whether defined narrowly or broadly).[5] Fair enough; let us then study some other styles of popular music—particularly those styles that include significant representation by historically marginalized groups—and when we do, try to represent those styles as best we can.

While I thus appreciate the intentions that motivated Shea to propose his sampling methodology, I would recommend that researchers instead adopt more traditional methodologies in creating and curating a corpus of popular music. Following that approach, a researcher first identifies a research question (or set of questions) and then identifies the relevant statistical population—i.e., the set of songs from a particular musical style or era—that would best help answer that question (or set of questions). It is in this first stage that, in the words of Shea, "researcher intervention" (p. 49) should and can address issues of diversity, since choosing the statistical population determines the demographic distribution of artists. Once the statistical population has been identified, the main empirical objective should be to identify the songs that best represent that statistical population, not those songs that achieve a particular demographic quota.

In summary, I agree with Shea (2022) that we should make space in our empirical research for demographic groups that have been marginalized, and whose voices have been underrepresented in the academic study of music. It is not complicated to do that, though, and it does not require any bespoke sampling procedure. All we need to do is recognize that different musical styles associate strongly with

different demographic groups and then make a conscientious (if not concerted) effort to meaningfully study those styles. As I see it, that is the real way to foster real representation in empirical work on popular music.

ACKNOWLEDGEMENTS

This article was copyedited by Eve Merlini and layout edited by Jonathan Tang.

NOTES

[1] Correspondence can be addressed to: Trevor de Clercq, Department of Recording Industry, Middle Tennessee State University, 1301 East Main Street, Box 21, Murfreesboro, TN 37132, tdeclercq@mtsu.edu.

[2] See the table entitled "Estimates and census figures of the growth of the ethnic minority population in the United Kingdom" on the Wikipedia page entitled "Demography of the United Kingdom" (2023), based on Haug, Compton, & Courbage (2002).

[3] This table recasts the table entitled "Racial/Ethnic Demographics of the United States (1910–2020)" on the Wikipedia page entitled "Historical racial and ethnic demographics of the United States" (2023), based on Gibson & Jung (2005).

[4] On Allmusic.com, for example, the list of top songs in contemporary R&B includes 47 songs, half of which (23) are by female artists, the heavy majority of whom are Black. https://www.allmusic.com/style/contemporary-r-b-ma0000002969/songs (accessed September 16, 2023).

[5] To be clear, the *Rolling Stone* magazine corpus was intended to represent "rock" music in a broad way—in line with how the label was used by Moore (2001), Everett (1994), and Stephenson (2002)—not popular music overall (de Clercq & Temperley, 2011, p. 51). In more recent work, I deprecate the use of the term "rock" in a broad way, as explained in Footnote 1 of my 2021 *Music Theory Online* article (de Clercq, 2021), in an attempt to be more clear about the categories and boundaries of popular music genres.

REFERENCES

Bull, B. (2005). Exemplar sampling. *The American Statistician, 59*(2), 166-172. https://doi.org/10.1198/000313005X42886

Burgoyne, J. A., Wild, J., & Fujinaga, I. (2011). An expert ground truth set for audio chord recognition and music analysis. In A. Klapuri & C. Leider (Eds.), *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 633-38). Miami, Florida. https://ismir2011.ismir.net/papers/OS8-1.pdf

Campbell, P., Myers, D., & Sarath, E. (2014). Transforming music study from its foundations: A manifesto for progressive change in the undergraduate preparation of music majors. Report from the task force on the undergraduate music major, College Music Society. https://www.music.org/pdf/pubs/tfumm/TFUMM.pdf

Condit-Shultz, N. (2016). MCFlow: A digital corpus of rap transcriptions. *Empirical Musicology Review, 11*(2), 124–47. https://doi.org/10.18061/emr.v11i2.4961

de Clercq, T. (2019). A music theory curriculum for the 99%. *Engaging Students: Essays in Music Pedagogy,* 7. https://doi.org/10.18061/es.v7i0.7359

de Clercq, T. (2020). Popular music analysis too often neglects the analysis of popular music: Review of Ciro Scotto, Kenneth Smith, John Brackett, (Eds.), *The Routledge Companion to Popular Music Analysis: Expanding Approaches* (Routledge, 2019). *Popular Music, 39*(2), 339-344. https://doi.org/10.1017/S0261143020000173

de Clercq, T. (2021). The logic of six-based minor for harmonic analyses of popular music. *Music Theory Online 27*(4). https://doi.org/10.30535/mto.27.4.4

de Clercq, T. (2022). A corpus analysis of harmony in country music. In D. Shanahan, J. A. Burgoyne, and I. Quinn (Eds.), *The Oxford Handbook of Music and Corpus Studies*. https://doi.org/10.1093/oxfordhb/9780190945442.013.22

de Clercq, T. & Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music, 30*(1), 47-70. https://doi.org/10.1017/S026114301000067X

Demography of the United Kingdom. (2023, August 25). In *Wikipedia*. https://en.wikipedia.org/wiki/Demography_of_the_United_Kingdom

Everett, W. (2004). Making sense of rock's tonal systems. *Music Theory Online, 10*(4). https://doi.org/10.30535/mto.10.4.2

Ewell, P. (2020). Music theory and the white racial frame. *Music Theory Online, 26*(2). https://doi.org/10.30535/mto.26.2.4

Ewell, P. (2023). *On music theory and making music more welcoming for everyone*. Ann Arbor, MI: University of Michigan Press. https://doi.org/10.3998/mpub.12050329

Gibson, C. & Jung, K. (2005). Historical census statistics on population totals by race, 1790 to 1990, and by Hispanic origin, 1970 to 1990, for large cities and other urban places in the United States. Washington: U.S. Census Bureau.

Haug, W., Compton, P., & Courbage, Y., (Eds.) (2002). The demographic characteristics of immigrant populations. Strasbourg: Council of Europe.

Hisama, E. (2021). Getting to count. *Music Theory Spectrum, 43*(2), 349-63. https://doi.org/10.1093/mts/mtaa033

Historical racial and ethnic demographics of the United States. (2023, August 25). In *Wikipedia*. https://en.wikipedia.org/wiki/Historical_racial_and_ethnic_demographics_of_the_United_States.

Maust, P. (2023). Expanding the music theory canon: A collection of inclusive music theory examples. https://www.expandingthemusictheorycanon.com/about/.

Moore, A. (2001). *Rock: The primary text: Developing a musicology of rock*, 2nd ed. Aldershot: Ashgate.

Palfy, C., & Gilson, E. (2018). The hidden curriculum in the music theory classroom. *The Journal of Music Theory Pedagogy, 32*(1), 79-110. https://digitalcollections.lipscomb.edu/jmtp/vol32/iss1/5

Patton, M. Q. (2015). *Qualitative research & evaluation methods*, 4th ed. Los Angeles: Sage.

Shea, N. (2020). *Ecological models of musical structure in pop-rock, 1950–2019*. Doctoral dissertation, Ohio State University, OH. http://rave.ohiolink.edu/etdc/view?acc_num=osu158755665247824

Shea, N. (2023). A demographic sampling model and database for addressing racial, ethnic, and gender bias in popular-music empirical research. *Empirical Musicology Review, 17*(1), 49-58. https://doi.org/10.18061/emr.v17i1.8531

Stephenson, K. (2002). *What to listen for in rock: A stylistic analysis*. New Haven: Yale University Press. https://doi.org/10.12987/yale/9780300092394.001.0001

Temperley, D. & de Clercq, T. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research, 42*(3), 187-204. https://doi.org/10.1080/09298215.2013.788039