MTO A JOURNAL OF THE SOCIETY FOR MUSIC THEORY

What Does a Corpus of Music Represent? Commentary on "Diversity in Music Corpus Studies"

Trevor de Clercq

REFERENCE: https://mtosmt.org/issues/mto.24.30.1/mto.24.30.1.shea_et_al.html

KEYWORDS: corpus studies, race, ethnicity, gender, diversity, sampling, representation, statistics

ABSTRACT: In Volume 30 (1) of this journal, Shea et al. (2024) introduce a novel sampling method called the "Anti-Discriminatory Alignment System" (or ADAS), which is designed to adjust the demographic distribution of artists in a corpus of music. In this commentary, I consider what a corpus created using the ADAS represents. I begin with a general discussion of corpus work in the context of foundational statistical principles. I then posit two different types of corpora that could be created to model the real world: music as heard and music as produced. A corpus created with the ADAS, in contrast, appears to be neither of these two types; instead, it appears to be a corpus of music as imagined, which models only a hypothetical statistical population. As a result, statistical findings from a corpus created with the ADAS may have limited explanatory power. If a researcher hopes to diversify the demographic distribution in a corpus of music, I argue that these attempts should occur not at the sampling stage but rather at an earlier stage, when the researcher is determining the particular statistical population to study.

DOI: 10.30535/mto.31.4.9

Received July 2024

Volume 31, Number 4, December 2025

Introduction

[0.1] For at least the past few years, there has been a great deal of concern about what an academic degree in music should (or could) look like and what it means to study music theory in that context. This issue has been front of mind because, as many authors have noted (e.g., Campbell, Myers, and Sarath 2014; Ewell 2020), music departments and music theory courses have traditionally focused heavily (if not exclusively) on European classical music—and thus on the music of dead White men—which seems problematic given the demographic diversity of students in present-day America that a music department presumably hopes to serve. Moreover, the analytical tools that music theorists have traditionally used were developed and honed within this environment (i.e., the music of dead White men), and therefore these tools may need to be

overhauled, if not scrapped entirely, for music theorists to do justice to a wider and more diverse repertoire (see, e.g., Ewell 2023; Palfy and Gilson 2018).

[0.2] Corpus study is one tool that has recently been identified as being in need of such recalibration. According to the authors of "Diversity in Music Corpus Studies" (Shea et al. 2024), researchers using corpus methods "run the risk of perpetuating harm to marginalized populations through ongoing studies grounded in inequitable corpora" ([0.2]). Over the course of their article, the seven authors argue that discriminatory social forces will manifest themselves in the demographic distribution of artists within a corpus of music. And so, if a music theorist conducting a corpus study hopes to be more socially responsible, that researcher should counteract those forces in some way, such as manually reshaping and rebalancing the demographic distribution of artists. Because heavy metal music is dominated by White male artists, for example, the authors suggest reducing the number of songs by White men and concomitantly increasing the number of songs by artists who are non-White or non-male. This, in a nutshell, describes the authors' *Anti-Discriminatory Alignment System* (or ADAS).

[0.3] To someone without training in empirical methods, the recommendations made by Shea and his coauthors may seem like a timely proposal, channeling the zeitgeist of contemporary music theory to bring about a more equitable academic environment. Indeed, I agree that corpus studies of music should make more space for historically marginalized voices that have been underrepresented in the academic study of music, and I have argued in previous work for ways to address the lack of diversity in the field (e.g., de Clercq 2019, 2020). But on closer examination, the specific methodology that Shea et al. put forth—the ADAS itself—can be seen as introducing its own set of problems. In particular, a corpus created using the ADAS may distort or misrepresent a musical style, such that it is not clear whether research findings from a corpus created using the ADAS have much if any explanatory power beyond the limited scope of the corpus itself. In other words, the authors' recalibration of the corpus study may strip it of any value as a research tool.

[0.4] In what follows, I consider some of the problems with existing corpus work, beginning with a back-to-basics discussion of statistical principles. I agree with Shea and his coauthors that corpus work should be clearer in its representation goals, but I see those goals as being constrained by the requirements of statistical theory. In essence, I argue that the creation of a corpus should have a single, overriding goal, which is to represent a particular *statistical population* (defined below) in the most accurate way possible. This statistical population might be a genre of music as heard by some particular group of listeners, for example, or a genre of music as produced by some particular group of artists. The statistical population that a corpus created with the ADAS represents, however, is much less clear, and it may represent only a figment of the researcher's imagination.

1. Music as Heard

[1.1] The Anti-Discriminatory Alignment System is, by the authors' own description, a novel sampling method for corpus research (Shea et al. 2024, [0.3]). Because Shea and his coauthors offer no significant discussion of existing sampling methods in their article, a brief review here will help put the ADAS in context. To begin, one fundamental and powerful tenet of statistics is that a researcher does not have to investigate every single member of a population to study that population. Here, I use the term "population" in the statistical sense, meaning the set of objects or items that are of interest for some research question (Miller and Miller 2014, 231); to avoid confusion, I will refer to this as the *statistical population*. A *sample* is a subset of this larger statistical population, and each member of this sample is called an *observation*. If, for example, the owner of a widget factory wanted to determine the proportion of broken widgets produced yesterday by the factory, the owner could—instead of inspecting every widget produced yesterday—simply take a sample (of, say, 150 widgets) and from that sample estimate the proportion of broken widgets overall. In this case, the statistical population is all the widgets produced yesterday. Note that the type of object in the statistical population and the sample are the same—here, widgets—and that the statistical population often has nothing to do with the term "population" in the everyday sense (e.g., as the group of people living in a country).

[1.2] Ideally, each observation is chosen in a manner that makes the sample sufficiently representative of the larger statistical population; otherwise, the sample may have *selection bias* (Lohr 2022, 6), which is a distinct concept from other types of bias (e.g., racial or gender bias). If, for instance, the factory manager chooses to sample widgets from only the day shift (in order, say, to avoid coming in at night), this sampling process may result in a problematic selection bias, particularly if production quality differs from day to night. One standard way to avoid selection bias is for observations to be drawn randomly from the entire statistical population, which helps avoid the (conscious or unconscious) influence of a human researcher on selecting observations. Truly random sampling can be difficult to achieve in practice, however, and so other sampling methods are often used.

[1.3] In corpus studies of popular music, for instance, researchers have often relied on lists of critically acclaimed songs or *Billboard* chart data to compile the corpus. What, then, is the statistical population that this sampling method is intended to represent? I would argue that the intended statistical population in this case is the musical genre as experienced by some specific group of human listeners (e.g., the US market), which I will refer to as a corpus of *music as heard*. In more concrete terms, the statistical population for music as heard is the complete set of encounters between a group of listeners and a song during a particular period of time, i.e., a set of listening experiences. Ideally, we would sample from this statistical population by having a list of songs in a genre, with each song weighted by the total listenership for that song given a particular market and timespan. For more recent years, this could be done quite accurately using data from Luminate, which is the company that currently provides the *Billboard* charts with sales, streaming, and radio metrics.[(1)] A researcher could, for instance, select the top 10,000 songs in a genre and have the probability of selecting a song proportional to its year-to-date album-equivalent sales in the US.

[1.4] Unfortunately, comprehensive Luminate data currently extends back only to 2014, so a truly random sample of listening experiences would be possible only for recent music history.[(2)] For earlier eras, researchers tacitly assume a correlation between *Billboard* chart rank and listening frequency. Consider, for example, that if we were to randomly sample the output of various radio stations, listener streaming data, and song sales in the US for a year, the most likely song to be sampled would be the number one song on the US *Billboard* chart that year, the second most likely song would be the number two song, and so on, since this is exactly what the *Billboard* charts measure. That said, *Billboard* charts for specific genres are typically not available for the entire history of a genre. *Billboard*, for example, only first published a chart dedicated exclusively to rock music on March 21, 1981, long after the beginning of rock music by any definition (Whitburn 2008). In these cases, a researcher tacitly assumes a correlation between rank on a list of critically acclaimed songs (such as the *Rolling Stone* magazine list) and listening frequency. Thus, although using chart data or a list of critically acclaimed songs does not directly involve random sampling, the assumption is that it offers the best approximation of the probability of getting a particular song in the corpus based on a random sample of listening experiences given the limitations of data availability. Nonetheless, the extent to which these assumptions result in a sample that is similar to a truly random sample remains underexplored, and I can imagine a tidy paper that empirically investigates this issue.

[1.5] Assessing how well a list of top songs models music as heard is somewhat tangential here, however, because music as heard is not the statistical population that Shea et al. are attempting to represent with the ADAS. Indeed, they directly acknowledge that their sampling method "cannot claim to capture the *most likely* listening experience of a casual mainstream listener. . . . [since] this experience would be best represented by a corpus of the most-popular, most-purchased, and most-respected music in some genre" (2024, [8.11]). In other words, Shea and his coauthors are not claiming that a corpus created with a list of top songs involves selection bias if the statistical population is music as heard. Instead, Shea and his coauthors are simply not interested in studying this statistical population (at least for their current purposes), since they believe these listening experiences are shaped by discriminatory social forces.

[1.6] What, then, is the statistical population that Shea et al. are attempting to represent with the ADAS? That is not entirely clear, as far as I can tell. They write only that a corpus created with the

ADAS "is an array of music that *could* have been heard by an audience being exposed to a particular genre" (2024, [8.11]). The word "could" here (italics in the original) does not tell us much; it says only that the sample includes observations from the statistical population. But a sample should be more than that; in particular, a sample should be an accurate representation of the distribution of a statistical population, whatever that might be.(3) Only with a representative sample can we have some confidence that the statistical results we draw from the sample (i.e., the corpus) have value beyond the limited scope of the sample itself.

[1.7] In short, having a clear idea of the statistical population that a corpus is meant to represent is critical to understanding the value of that corpus. If we had a sample of voters, for instance, it would be essential to know whether this sample was drawn from (and thus meant to represent) all eligible voters, all registered voters, all likely voters, or some other statistical population, since any statistics from the sample could only be properly interpreted within the context of understanding its source. To understand what value the ADAS has as a sampling method, therefore, we need to think more precisely about what it is meant to represent.

2. Music as Produced

[2.1] As discussed above, a corpus of music as heard would weight the probability of selecting a song by its relative popularity (i.e., scaled total listens) within a particular market and time period. Alternatively, we could create a corpus of songs in which each song has the same (unweighted) probability of being selected. In this corpus, a chart-topping hit by Taylor Swift, for example, would have the same chance of being sampled as a deep album cut from a much lesser known artist. The statistical population in this case would be all the songs written and recorded in some musical style during some window of time, which I will refer to as a corpus of *music as produced*.(4) Ideally, we would have a list of all the songs recorded in some musical style during some time frame, and then we would sample randomly from that list, with no song being more likely to be selected than another. For more recent years, this would be fairly easy to do using data from Luminate, which can output a list of up to 10,000 songs for any core genre. For earlier periods of music history, we would need a different approach. One option, for instance, would be to compile a list of all the songs from the entire catalog of a record label, which includes both the hits and the flops of an era, and then randomly sample from that.

[2.2] The difference in findings between a corpus of music as heard and a corpus of music as produced could reveal interesting aspects of popular music, particularly with regard to what features correlate with song success. If, for example, intro sections are shorter in a corpus of music as heard than intro sections in a corpus of music as produced (assuming a similar style and era), we might hypothesize that intro length is a factor that distinguishes more successful songs from less successful songs.(5) Note that both of these statistical populations—music as heard and music as produced—are aspects of the real world that we are attempting to model with our corpus, and each gives us a different perspective on reality.

[2.3] Although Shea and his coauthors do not intend for their corpus to represent music as heard (as quoted above in [1.5]), perhaps their aim with the ADAS is to represent music as produced. For example, they write that the ADAS "provides a better understanding of the internal musical characteristics of the genre by including the musical practices of a wider variety of humans who contributed to that tradition" (2024, [8.9]). Similar statements about the goal of increasing the "variety" within a corpus are found elsewhere in their article ([0.2]; [8.8]). One way to interpret these statements is that they refer to the degree to which the corpus represents the broader collection of artists in the genre being studied. Taken this way, the authors' statements could be seen as a reasonable critique of corpus studies that rely on lists of critically acclaimed songs or top hits, since songs by the most famous artists in a genre might be different in some important way from songs by less famous artists. Of course, to study a genre of music as produced, we should have some idea of the distribution of the broader statistical population of songs and thus artists who have contributed to a genre. Shea and his coauthors never explore this with any data, and so it will be informative here to do a small investigation in order to examine how the distribution of artists in a genre overall compares to the distribution of artists in a list of top songs.

[2.4] To this end, **Example 1** shows different measurements of the percentage of female artists in two genres of popular music, rock and pop, based on total sales in 2023.(6) The rightmost column, for example, shows the percentage of female artists in the top 25 best-selling artists of that genre; the column to the left ("Top 50") shows the percentage of female artists in the top 50 best-selling artists of that genre (which includes the top 25 artists); and so on.(7) In contrast, the column labeled "Random 200" shows a random sample of 200 observations taken from the top 10,000 best-selling artists of that genre, excluding the top 200 best-selling artists.(8) In other words, the "Random 200" column represents an assessment of all the artists in the genre, including the long tail of much less successful artists.(9)

[2.5] Notice, for example, that a random sample of the top 10,000 artists in rock music estimates that 17.5% of artists in the genre overall are female. In contrast, only 6.5% of the top 100 artists in rock are female, which represent a significant difference from the estimate of 17.5% for rock overall. In other words, there is evidence that rock music skews more male the more we focus on the more commercially successful artists. If we wanted to better represent the wider variety of people who have composed in the rock tradition, therefore, we would want to sample from a larger population of rock songs that would include a greater proportion of women than found in a list of top rock songs.

[2.6] Now, however, consider the corresponding data for the pop genre. Based on a random sample from the top 10,000 pop artists, it is estimated that 34.3% of artists in pop overall are female. As we look more narrowly at the most successful pop artists, however, we find an increasingly larger proportion of women, with 64.0% of the top 25 pop artists as female. If our goal is to better represent the variety of people who have composed in the pop tradition, therefore, we would want to sample from a larger population of pop songs that would include a greater proportion of men than found in a list of top pop songs.

[2.7] Overall, the data in Example 1 shows that, indeed, a corpus of music as heard would likely differ from a corpus of music as produced in terms of the proportion of artists who belong to some particular identity. Nonetheless, the data also show that the difference in demographic distribution between music as heard and music as produced is not consistently skewed in the same direction. Women, for example, are not always less well represented in a list of top songs as compared to their proportion in a genre overall. To be clear, achieving (or approximating) a particular demographic distribution is *not* the actual research goal when assembling a corpus of music as produced. Instead, the goal is to model a different statistical population, which is the broader variety of *songs* written within a musical style, as compared to just the hit songs. The broader variety of artist identities (whatever that might be) is simply a byproduct of the true research goal.

[2.8] In contrast, Shea and his coauthors do not appear to be attempting to create a corpus that represents the broader variety of songs in a musical style. This becomes clear from the mechanics of their sampling system: The ADAS starts with a "parent" corpus—which is compiled using a traditional method, such as selecting from a list of critically acclaimed songs or songs ranked by commercial success—and then the distribution of artists in that corpus is adjusted according to gender and race/ethnicity to better match some target distribution (such as the US population) to create a final "child" corpus by resampling from this same list of songs.(10) In other words, all of the songs in the final child corpus are still top songs, since the observations were pulled from a list of top songs. This method, therefore, will never do a good job of modeling music as produced (no matter what adjustments the researcher makes), since that would require the researcher instead to sample from all the songs produced in a genre, including those that were successful, those that were unsuccessful, and everything in between. Shea et al. are not, therefore, creating a corpus of music as heard or a corpus of music as produced. What, then, are they creating?

3. Music as Imagined

[3.1] Shea and his coauthors are very clear about their goals in creating a corpus using the ADAS. Specifically, they write that:

> Our current study aimed to make our corpora more representative of the overall population of the United States during the chronological window represented by the corpus . . . We therefore set target proportions for gender and for race/ethnicity by calculating the geometric mean—an averaging function designed for proportions—between the proportion in the parent corpus and the relevant US Census data for that time period. (2024, [5.5])

In other words, a list of top country songs (the parent corpus) might not include very many Black artists, for example, but since people identifying as Black comprise a significant proportion of the US population, the child corpus (i.e., post-ADAS) would include more Black artists so that it would more closely mirror the US population. Similarly, a list of top hip-hop songs might not include very many female artists, but since women comprise about half of the US population, the post-ADAS corpus would include more female artists.

[3.2] Before considering what this sampling method represents, let us first assess whether the authors in fact achieved their stated goal of making their corpora more representative of the overall population of the United States. In that regard, **Example 2** shows the percentage of nonmale and BIHAP (i.e., non-White) artists in each of their four corpora—metal, hip-hop, country, and pop—both before the implementation of the ADAS (the parent corpora) and after (the child corpora).[(11)] For reference, Example 2 also shows the population of the United States during the 1990s, which is the decade represented by the authors' corpora.[(12)] Notice that the parent corpus of country music includes only 1.5% BIHAP artists, whereas the final child corpus includes 14.0% BIHAP artists. The implementation of the ADAS, therefore, has indeed made the demographic distribution of artists in the country corpus more representative of the US population.

[3.3] Now, however, look at the before-and-after results for the hip-hop corpus. The parent corpus includes 98.2% BIHAP artists, whereas the final child corpus includes 99.0% BIHAP artists. In this case, the ADAS has not made the hip-hop corpus more representative of the US population, as the authors claimed was their goal. To do so, of course, would require the addition of a significant percentage of White artists. Similarly, the proportion of women in the child pop corpus (62.0%) is less representative of the US population than the original parent corpus (60.4%). As we look at other genres in Example 2, it becomes clear that the ADAS does not consistently make a corpus more representative of the overall population of the United States, as the authors claimed was their intention. That said, it would not be too difficult to posit a sampling method that does more closely align the demographic distribution of artists in a corpus with the demographic distribution of the US population, if that were a researcher's intention, even though the ADAS does not actually accomplish that.

[3.4] As Example 2 reveals, Shea et al.'s real goal appears to be to increase the proportion of female and BIHAP artists in a corpus, no matter what the original distribution, whenever it falls below a population target. (If the proportion is above the population target, they essentially leave it as is.) This is certainly an understandable goal, given the various types of discrimination that women and people of color have experienced in the United States, both inside and outside of the music industry, as Shea and his coauthors document throughout their article. But good intentions do not necessarily make for good research methodologies, which becomes clear now that we have a better sense of what a corpus created with the ADAS represents.

[3.5] A corpus created with the ADAS is not music as heard (as Shea et al. concede, noted above), nor is it a corpus of music as produced (as discussed above) since it does not explore a musical genre beyond the hit songs. Rather, the ADAS creates what I will refer to as a corpus of *music as imagined*. The statistical population for music as imagined is not something that exists in the real world but rather something that exists only in the researcher's imagination, a sort of fictional alternate reality that the researcher envisions. As evidence of this, consider the third step of the ADAS, which is to establish benchmarks for the child corpus. These user-defined benchmarks, according to the authors, "should reflect the priorities and goals of a particular corpus study," and it is against these benchmarks that "the demographic distribution of the parent corpus can then be measured" (Shea et al. 2024, [5.5]). In other words, a researcher is free to choose whatever target value for representation they desire and then adjust the corpus as they see fit to match that target

value. There is no requirement to assess the reality of the demographic distribution in the statistical population (whether taken as a weighted or unweighted collection of songs), because the goal is not to approximate reality but rather to imagine an alternative world. This is not to say that the songs or artists in a corpus of music as imagined are not real; without question, the corpus includes real songs by real artists. But the statistical population that the corpus represents is only hypothetical.

[3.6] To be clear, there is nothing inherently wrong with imagining a hypothetical statistical population. We might wonder, for example, what the outcome of some particular past election would have been had, say, Black voters turned out in greater numbers. But when we create this hypothetical statistical population, its value is only as compared to the real world. My concern is that Shea and his coauthors seem to have convinced themselves—and appear to be trying to convince their readers—that a corpus created with the ADAS actually better approximates the real world than a corpus created using other methods. They write, for example, that a child corpus "may better represent the musical diversity within some style or genre" (2024, [0.3]). It is clear, however, that the ADAS does not better approximate music as heard (by the authors' own admission), and it is also clear that the ADAS does not better approximate music as produced, since it still relies on a list of top songs, without any attempt to objectively assess the distribution of songs in the genre more broadly.

[3.7] I should also clarify that I am not arguing that a researcher cannot, for instance, study women in hip-hop or, say, Black artists in country music. It is rather that if a researcher wants to study one of these subgroups, they should do so more directly, with a statistical population that is more clearly defined. A researcher could, for example, create a corpus of Black country artists (either as heard or as produced). This is a corpus of a specific statistical population that exists in the real world. Alternatively, a researcher could use listenership data from women in Nebraska and examine what country music they listen to. This would be a corpus of country music as heard by women in Nebraska, which certainly also exists in the real world. (Whether such a corpus would include more or fewer female artists than a corpus of country music as generally heard by US listeners is unknown.) Or a researcher could make a corpus of R&B music, which I would expect would include a high proportion of Black and female artists. In contrast, a corpus of country music that includes more Black or female artists using the mechanics of the ADAS is not, as far as I can tell, a corpus that is intended to model any statistical population in the real world.

[3.8] Ultimately—and this is a critical point—a corpus of music as imagined has very limited explanatory power. If what it represents is simply a more diverse conception of a genre, then any statistics we derive from that corpus only are estimates for the parameters of an imagined statistical population. And there is something fundamentally contradictory about that effort: a researcher is using powerful and fairly complicated statistical tools to measure some aspect of music, reported with the objective veneer of mathematical precision, yet based on their subjective interpretation of what the demographic distribution of artists in a genre of music could or should be, which anyone is free to define in whatever way they see fit.

Conclusion

[4.1] To be clear, the statistical population that a corpus represents may be something other than music as heard, music as produced, or music as imagined. We might, for instance, be interested in studying all the songs used in a music appreciation course, which could be a corpus intended to represent *music as taught*. Just as there may be interesting differences between music as heard and music as produced, there may be interesting differences between music as taught and other statistical populations of music. Christopher **White**, for example, has shown that augmented-sixth chords comprise a higher proportion of chords in music theory textbooks than in common-practice era music overall, thereby potentially revealing a disconnect between the more limited repertoire that music theorists have traditionally taught and the wider practice of European classical music (2021). Moreover, we do not necessarily need to think about what a corpus represents. Someone, for instance, could put together a corpus of interesting musical examples on some topic—say, piano sonatas that use augmented-sixth chords or chamber works by women composers—to act as

a repository for other people to find useful examples for use in the classroom. But we should not understand a corpus compiled in this way to represent the genre at large (i.e., piano sonatas or chamber works in general).

[4.2] I also want to clarify that I am not arguing that a corpus researcher should never set targets for subgroups in their sampling outcomes. In a *stratified sampling method*, for example, "the population is partitioned into regions or strata, and a sample is selected by some design within each strata" (Thompson 2012, 141). Returning to the example of the widget factory above (in [1.1]): If we know that 30% of the total daily output comes from the night shift and the remaining 70% comes from the day shift, we may want to take 30% of our random samples from the night shift and 70% from the day shift (rather than sampling randomly from the total daily output) to ensure that our final sample is proportionally representative of the different subgroups (i.e., night-shift versus day-shift widgets) in the population overall. My sense is that Shea and his coauthors have too much slippage around the term "population," often conflating its everyday meaning with its more technical definition as used by statisticians (see de Clercq 2024). To stratify a sample by demographic category, we would need to have some idea of the true distribution of those categories within the statistical population. Yet Shea et al. do not appear to make any attempt to assess the distribution of different demographic categories within a genre (whether as heard, as produced, or otherwise), and it seems unreasonable to assume that the distribution of demographic categories within a genre will match the distribution of demographic categories in the US population.

[4.3] In fairness to Shea and his coauthors, corpus studies of music have not always clearly defined the particular statistical population that the corpus is meant to represent. A researcher, for example, might simply say that they are creating a corpus of rock music or a corpus of country music and leave it at that. As a result, it may not appear to matter which songs are included in the corpus as long as those songs belong to that genre. For example, David Temperley and I—in our corpus study of rock music—wrote that, "Whether or not the songs on the [*Rolling Stone* magazine] list are in fact the greatest rock songs (whatever that might mean) is not important for present purposes; all that matters is that they are rock songs" (de Clercq and Temperley 2011, 51). This statement is incorrect. What matters is not just that the corpus includes rock songs (although that's one important factor); as discussed above (in [1.6]), it also matters that the collection of songs is representative of the distribution of the statistical population. It's not obvious what statistical population Temperley and I had in mind for our corpus (and recent discussions between us of the issue have revealed that we may have had different ideas). If our goal were to model music as produced, then using a list of greatest rock songs would not be a good strategy (except under the dubious assumption that the greatest rock songs are representative of all rock songs, with respect to the features being studied). In contrast, if our goal were to model music as heard, then using a list of greatest rock songs would be understandable, although then it *would* actually be important whether the songs on the list are in fact the greatest rock songs (under the assumption that the greatest rock songs are the most listened to).

[4.4] Corpus work of the past (including my own), in other words, has admittedly been conducted with some problematic looseness in terms of the sampling method and the statistical population under study. But rather than attempting to tighten up corpus methodology, Shea and his coauthors rend it asunder, unraveling any apparent restrictions or efforts to represent a particular statistical population in the real world. Rather than better representing music as heard or music as produced (or any other view of music in reality), I fear a corpus curated with the ADAS does not represent anything real at all, and as a result, any statistics we might derive from this corpus do not allow us to meaningfully infer anything at all.

[4.5] Shea and his coauthors conclude their essay by saying that they hope it "will challenge the field of music theory to be more intentional and accountable when selecting musical datasets for research purposes" (2024, [8.12]). For me, this has definitely been the case, and I hope their essay spurs other corpus researchers to think more deeply about what a corpus of music is intended to represent and how to best do so when compiling it. I also agree with Shea et al. that corpus work should address the extensive discrimination that marginalized groups have faced in our society. Indeed, we should make more serious efforts to study the music of Black artists, the music of

women, and the music of other marginalized groups in our corpus research. But we don't need the ADAS to do that. We need only to clearly identify what aspect of the real world we would like to study and then make our best attempt to accurately represent that in our work.

Trevor de Clercq
Middle Tennessee State University
Department of Recording Industry
1301 East Main Street, MTSU Box 21
Murfreesboro, TN 37132
tdeclercq@mtsu.edu

Works Cited

Biamonte, Nicole, Lindsey Reymore, Ben Duinker, Leigh VanHandel, Nicholas Shea, Matthew Zeller, Christopher William White, Jeremy Tatar, Jade Roth, and Kelsey Lussier. 2023. "The Timbre in Popular Song (TiPS) Corpus: An Interactive Report." Zenodo. https://doi.org/10.5281/zenodo.11176578.

Burgess, Richard James. 2013. *The Art of Music Production: The Theory and Practice*. 4th ed. Oxford University Press.

Campbell, Patricia Shehan, David Myers, and Ed Sarath. 2014. "Transforming Music Study from its Foundations: A Manifesto for Progressive Change in the Undergraduate Preparation of Music Majors. Report of the Task Force on the Undergraduate Music Major." The College Music Society. https://www.music.org/pdf/pubs/tfumm/TFUMM.pdf.

de Clercq, Trevor. 2019. "A Music Theory Curriculum for the 99%." *Engaging Students: Essays in Music Pedagogy* 7. https://doi.org/10.18061/es.v7i0.7359.

————————. 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): 339–44. https://doi.org/10.1017/S0261143020000173.

————————. 2024. "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." *Empirical Musicology Review* 18 (2): 175–80. https://doi.org/10.18061/emr.v18i2.9726.

de Clercq, Trevor, and David Temperley. 2011. "A Corpus Analysis of Rock Harmony." *Popular Music* 30 (1): 47–70. https://doi.org/10.1017/S026114301000067X.

Devore, Jay. 2012. *Probability and Statistics for Engineering and the Sciences*. 8th ed. Cengage.

Ewell, Philip. 2020. "Music Theory and the White Racial Frame." *Music Theory Online* 26 (2). https://doi.org/10.30535/mto.26.2.4.

————————. 2023. *On Music Theory, and Making Music More Welcoming for Everyone*. University of Michigan Press. https://doi.org/10.3998/mpub.12050329.

Lohr, Sharon. 2022. *Sampling: Design and Analysis*. 3rd ed. CRC Press. https://doi.org/10.1201/9780429298899.

Luminate. 2024. "Music Connect." Accessed May 8, 2024. https://musicconnect.mrc-data.com/.

Miller, Irwin, and Marylees Miller. 2014. *John E. Freund's Mathematical Statistics with Applications*. 8th ed. Pearson.

Palfy, Cora, and Eric Gilson. 2018. “The Hidden Curriculum in the Music Theory Classroom.” *Journal of Music Theory Pedagogy* 32 (1): 79–110. https://doi.org/10.71156/2994-7073.1194.

Seufitelli, Danilo, Gabriel Oliveira, Mariana Silva, Clarisse Scofield, and Mirella Moro. 2023. “Hit Song Science: A Comprehensive Survey and Research Directions.” *Journal of New Music Research* 52 (1): 41–72. https://doi.org/10.1080/09298215.2023.2282999.

Shea, Nicholas. 2022. “A Demographic Sampling Model and Database for Addressing Racial, Ethnic, and Gender Bias in Popular-Music Empirical Research.” *Empirical Musicology Review* 17 (2): 49–58. https://doi.org/10.18061/emr.v17i1.8531.

Shea, Nicholas, Lindsey Reymore, Christopher Wm. White, Leigh VanHandel, Ben Duinker, Matthew Zeller, and Nicole Biamonte. 2024. “Diversity in Music Corpus Studies.” *Music Theory Online* 30 (1). https://doi.org/10.30535/mto.30.1.8.

Thompson, Steven K. 2012. *Sampling*. 3rd ed. Wiley.

Whitburn, Joel. 2008. *Joel Whitburn Presents Rock Tracks 1981–2008*. Record Research.

White, Christopher William. 2021. “Some Aspects of Pedagogical Corpora.” *Empirical Musicology Review* 16 (1): 154–65. https://doi.org/10.18061/emr.v16i1.7785.

Wikipedia. 2009. “Non-Hispanic whites.” Last modified May 4, 2024. https://en.wikipedia.org/wiki/Non-Hispanic_whites.

Footnotes

1. See https://luminatedata.com/about/ for more information about Luminate (previously SoundScan and Nielsen Music). To analyze Luminate data requires an active “Luminate Data” account, which allows access to the subscription-based service called “Music Connect.” Service costs are not publicly posted and depend on the size of the organization and number of users. For more information, see https://support.mrc-data.com/portal/en/kb/articles/how-do-i-access-music-connect.
Return to text

2. See https://support.luminatedata.com/portal/en/kb/articles/country-list for more information about data availability within Luminate by year and market.
Return to text

3. More technically, the sample should consist of *independent* and *identically distributed* (or “i.i.d.”) random variables (Devore 2012), which means that the probability of selecting a particular member of the statistical population is the same for each observation. To be clear, different members of the statistical population may have different probabilities of being selected, such as more popular songs being more likely to be sampled than less popular songs in a corpus of music as heard; the point is that the probability distribution from which the next observation will be chosen is not affected by and is the same as that for the other observations.
Return to text

4. The word “produced” can have various meanings in the context of music, especially popular music (see Burgess 2013). The person who produced a song, for example, could be the person who funded the project, the person who booked the studio time or session musicians, the person who arranged or orchestrated the song, or the person who created the backing track for the vocalist. With regard to *music as produced*, I use the term in the economic sense, where a song has gone through the means of production to become a product available to the general public for consumption.
Return to text

5. Research that explores the differences between top songs and less popular songs is often referred to as “Hit Song Science.” For a recent survey of this work, see Seufitelli et al. 2023.

Return to text

6. The data in Example 1 was obtained from Luminate (2024). For the sake of encoding, male artists or all-male groups (as determined from photographic evidence) were assigned a value of zero (0), female artists or all-female groups were assigned a value of one (1), and any group that included at least one woman and at least one man was assigned a value of 0.5 to reflect that the ensemble included both men and women.
Return to text

7. Genre classifications in the Luminate data (2024) are initially determined by the artist (or label), but "Billboard makes all final decisions on genre flagging" (https://support.mrc-data.com/portal/en/kb/articles/how-can-i-assign-the-correct-genre-to-a-song-16-8-2023). An artist can release different songs in different genres, but each song can have only one core genre classification (https://support.mrc-data.com/portal/en/kb/articles/what-do-the-different-acronyms-mean). Taylor Swift's ranking as a country artist, for example, would thus be based only on the songs she has released that are classified as country songs.
Return to text

8. The 99% confidence interval for the random sample uses the normal approximation for an estimate based on a sample proportion.
Return to text

9. The choice to use the top-selling 10,000 artists, which is the current search limit on the Luminate interface for a ranked list, obviously does not include the entire pool of artists within a genre, but it arguably gives a very good snapshot of less successful artists. For example, the best-selling pop artist in 2023 was Taylor Swift, who had over 13 million album-equivalent sales in the pop category, as compared to the 10,000th best-selling artist in the pop category (unnamed here for privacy), who had only 413 album-equivalent sales in 2023.
Return to text

10. If the parent corpus does not include a sufficient number of artists from some gender or race/ethnicity to assemble the child corpus, Shea and his coauthors will turn to other lists of top songs or the next-highest ranking songs on the list from which the parent corpus was derived, as described in section [6] of their paper (2024). For an earlier discussion of this sampling methodology, see Shea 2022.
Return to text

11. The demographic distribution of artists in the authors' corpora is extracted from their original article (2024) as well as from the more explicit discussion found in Biamonte et al. (2023). Note that BIHAP is the term Shea et al. use to refer to Black, Indigenous, Hispanic, Asian, and other People of Color (2024, [2.2]).
Return to text

12. The value of 27.7% BIHAP was calculated using a quadratic regression that modeled the proportion of the US population who identified as BIHAP based on the data from the Wikipedia (2009) page https://en.wikipedia.org/wiki/Non-Hispanic_whites (accessed May 14, 2024), specifically the table entitled "Historical population by state or territory," which collates information from various US census reports. The value of 51.1% for the proportion of women is an average of the proportion from 1990 (51.3%) and 2000 (50.9%) based on US census data (https://www2.census.gov/library/publications/decennial/2000/briefs/c2kbr01-09.pdf).
Return to text