Improving the Measurement of Shared Cultural Schemas with

Correlational Class Analysis: Theory and Method*

Andrei Boutyline

University of California, Berkeley

July 25, 2016

Conditionally accepted for publication at Sociological Science.

* This research was supported by a fellowship from the National Science Foundation Graduate Research Fellowship Program. I thank Ronald Breiger, Neil Fligstein, John Flournoy, Amir Goldberg, Monica Lee, Valden Kamph, Fabiana Silva, Matthew Stimpson, Stephen Vaisey, Robb Willer, and participants of the Berkeley Mathematical, Analytical and Experimental Sociology workshop for feedback on the paper. I am also grateful to Amir Goldberg for generously discussing RCA and making its software implementation available online. Direct all correspondence to Andrei Boutyline at Department of Sociology, 410 Barrows Hall, University of California, Berkeley, CA 94720. Email: boutyline@berkeley.edu
ABSTRACT

The measurement of shared cultural schemas is a central methodological challenge for the sociology of culture. Relational Class Analysis (RCA) is a recently developed technique for identifying such schemas in survey data. However, existing work lacks a clear definition of such schemas, leaving RCA’s accuracy largely unknown. Here, I build on the theoretical intuitions behind RCA to arrive at this definition. I demonstrate that shared schemas should result in linear dependencies between survey rows—the relationship usually measured with Pearson’s correlation. I thus modify RCA into a “Correlational Class Analysis” (CCA). When I compare the two methods using a broad set of simulations, results show that CCA is reliably more accurate at detecting shared schemas than RCA, even in scenarios that substantially violate the assumptions behind CCA. I find no evidence of theoretical settings where RCA would be more accurate. I then revisit a prior RCA analysis of the 1993 GSS musical tastes module. While RCA had partitioned these data into three schematic classes, CCA partitions them into four. I compare these results with a multiple groups analysis in SEM, finding that CCA’s partition yields greatly improved model fit. I conclude with a parsimonious framework to guide future work.
Improving the Measurement of Shared Cultural Schemas with Correlational Class Analysis: Theory and Method

The task of revealing intelligible structures of meaning beneath complex collections of cultural data is among the most central methodological challenges posed by the sociology of culture (Mohr 1998; Mohr and Rawlings 2012). From the perspective of culture and cognition, this task is a search for shared “cultural schemas”—abstract cognitive structures which specify relationships between cultural elements. In a high-profile recent work, Goldberg (2011) proposes an innovative methodology for identifying groups of survey respondents who share such cultural schemas, which he terms Relational Class Analysis (RCA). RCA has rightfully garnered a substantial amount of attention across diverse domains of study including cultural tastes (Goldberg 2011), public opinion (Baldassarri and Goldberg 2014; Wu 2014), organizational behavior (Miranda, Summers, and Kim 2012), and economic sociology (DiMaggio and Goldberg 2010). However, existing work has not yet provided a clear definition of shared cultural schemas, the central concept under investigation. This is a crucial limitation as, without such a definition, RCA’s accuracy at locating such schemas cannot be convincingly measured.

Goldberg (2011) introduces RCA using a survey of musical tastes as his case study. Taking a cue from relational theories of meaning (e.g., Saussure 1916 [2013]; Lévi-Strauss 1963), RCA searches the data for schemas that define not the musical tastes themselves, but the relationships between these tastes—that is, which genre tastes are perceived as similar and which as opposed. This kind of schema can be found in the implicit agreement between an individual who likes musical genres A and B but dislikes genre C, and another who dislikes A and B but likes C: though the two hold no tastes in common, they nonetheless agree that A “goes with” B, whereas C is “opposed to” A and B. On the other hand, an individual who likes all three of the
genres A, B and C has two tastes in common with the first individual, but does not agree that C is the opposite of A and B. Thus, under this conception, the first two individuals arrange their tastes according to the same cultural schema, while the third one does not. The goal of RCA is to partition the survey population into classes of respondents that share such cultural schemas—a novel methodological task which is, in itself, a bold conceptual innovation.

At the core of RCA’s approach is a novel similarity measure termed “relationality.” Goldberg contends that relationality can quantify the extent to which two respondents organize their attitudes according to a shared cultural schema. However, he does not provide a clear definition of such shared schemas. With this key link between theory and measurement missing, relationality’s ability to successfully capture such schemas cannot be convincingly shown. In this paper, I examine the theoretical intuitions implicit in current work to arrive at this missing definition. I demonstrate that, in order to detect shared cultural schemas like those in Goldberg’s (2011:1404-1405) motivating example, relationality must measure the degree of linear dependency between two individuals’ responses. This lends itself to a simple, intuitively plausible formal model of schematic similarity as linear dependence between response vectors, and suggests that Pearson’s correlation may already provide a solution for the task that relationality sets out to solve. When I reexamine Goldberg’s motivating example with correlation, I find that it indeed yields more accurate results than relationality. Its results match Goldberg’s own description of the data, while relationality’s do not.

I then use simulations to verify that this difference in accuracy generalizes more broadly. First, I simulate 10,000 test cases where shared schemas result in linear dependencies between

---

1 In Martin’s (2002) terminology, such schemas thus underlie the “tightness” rather than the “consensus” of a system of attitudes. If one imagines attitudes as an abstract space where each dimension represents a like/dislike of a given musical genre, such schemas would specify an axis or plane along which culturally valid tastes can be arranged rather than a specific point in space at which tastes should be located.
responses, as they do in my model and in Goldberg’s example. I analyze each simulated dataset with both relationality and correlation, and compare their results to the true schematic structure of the simulated data, which is known by design. The results confirm that this switch from “Relational” to “Correlational” Class Analysis (CCA; see Appendix A for details) reliably increases the accuracy of the technique, so much so that, when the two disagree, the odds that CCA’s results are more accurate approach 17:1. I also document that RCA introduces a strong distributional assumption which has not previously been noted in published work. When it is violated, the odds in favor of CCA’s results further increase to 23:1.

But what if the socio-cognitive processes giving rise to shared cultural schemas are actually non-linear in character, contrary the theoretical intuitions formalized in this paper? Though the motivating examples behind relationality do not entail such a possibility, Goldberg nonetheless designed RCA to function even in the presence of complex nonlinearities in the data (Goldberg 2011:1433). Given that the theory of shared cultural schemas is still in its infancy, this may be a prudent goal. I thus conduct further simulations to determine whether RCA is preferable over CCA when shared cultural schemas depart from the theorized linearity. In the first three sets of simulations, which I term “sub-linear,” I investigate the possibility that some of the basic theorized relationships between schemas—scaling, shift or inversion—do not take place. In the second three sets, which I call “super-linear,” I examine the possibility of schematic transformations that are more complex than those envisioned by the theory: high-degree polynomial transformations, multi-way interactions, and independence between parts of the schema. The simulations show that CCA remains reliably more accurate than RCA under every theoretical scenario examined. While substantial departures from linearity cause CCA’s average
accuracy to decrease, RCA’s average accuracy always decreases to an even lower level. I find no evidence of settings where RCA would be more accurate than CCA.

Finally, to explore the substantive consequences of switching from relationality to correlation, I revisit Goldberg’s (2011) RCA analysis of the 1993 General Social Survey (GSS) musical taste data. The RCA analyses had detected three schematic classes: “Omnivore – Univore,” “Contemporary – Traditional,” and “Highbrow – Lowbrow.” The CCA results confirm the first two of these schemas, but contain two other schemas in place of RCA’s third. I term these additional schemas “Anything (but) Country” and “Anything (but) Heavy Metal.” These schemas more closely resemble the patterns of exclusionary omnivorousness documented by Bryson (1996) than the hierarchical “Highbrow – Lowbrow” logic described by Goldberg. Thus, the CCA analyses confirm broad outlines of RCA findings, but also contain potentially important differences. I further compare these results via a widely-known multiple groups analysis technique from structural equation modeling (SEM), which can be used to provide a goodness-of-fit measure for inductively located heterogeneity (Author et al., under review). The multiple groups analysis indicates that CCA’s classes have a far better fit to the GSS data than RCA’s. I conclude by discussing how future work can further advance the methodological and theoretical project stemming from Goldberg’s (2011) deeply innovative contribution.

**SHARED SCHEMAS**

RCA’s conception of shared cultural schemas builds on relational theories of meaning in the tradition of Saussure and Lévi-Strauss (e.g., Cerulo 1993, Emirbayer 1997). Such theories hold that the meaning of symbols in a cultural system rests not in the signs themselves, but in the relationships among them. So, for example, the concept “hot” acquires its meaning from its opposition to “cold” rather than from any innate properties of the word itself (Saussure 1916
The concepts that make up a single cultural domain then exist in semantic networks, where each element gains its significance from its relationships to the other elements. For example, each of the concepts “mother,” “father,” “son,” and “daughter” is defined in part through its distinction from the other three concepts, and cannot be fully understood without understanding them as well (Lévi-Strauss 1963). Such a configurations of concepts within a single cultural domain can be thought of a shared cultural schema—an abstract cognitive structure that individuals acquire through experience or acculturation.

As I noted above, Goldberg (2011) does not formally define what it means for a set of survey respondents to share such a cultural schema. He instead illustrates this relationship by way of a motivating example and accompanying diagram\(^2\), which I recreate as Figure 1.

---

**Figure 1.** Musical tastes of four respondents, with evaluations ranging from 1 (strongly dislike) to 5 (strongly like) for each genre. Respondents A, B and C follow the same taste schema, while D does not. This figure recreates the contents of Goldberg’s Figure 1A (2011:1405).

---

\(^2\) Goldberg describes cultural schemas as “complex structures of mental representation […] that are built up incrementally through interaction with the environment [that] embody our taken-for-granted assumptions about the world,” and further states that “schemas are not clear sets of behavioral rules but rather implicit recognition procedures that emerge from intricate associational links among salient aspects of our cognitively represented
Describing this figure, Goldberg states that A and B have “identical” patterns of musical tastes, C’s pattern is “almost a mirror image” of A’s and B’s, and D’s is “different but not antithetical.” He thus concludes that “respondents A, B, and C exhibit the same logic of musical taste construction ... as they all exhibit the same structure of relevance and opposition” (Goldberg 2011:1405), whereas respondent D does not.

To arrive at a formal definition of schematic similarity, I expand Goldberg’s discussion of this introductory example by writing out the implied algebraic operations. Respondent A likes pop, blues and rock, strongly likes classical and opera, and is indifferent towards bluegrass and country: $A = [4,4,4,5,5,3,3]$. Respondent B, on the other hand, dislikes pop, blues and rock, is indifferent towards classical and opera, and strongly dislikes bluegrass and country: $B = [2,2,2,3,3,1,1]$. Except for an overall downward shift in the appraisal of all the genres, this pattern of tastes is identical to that of the first respondent: $B = A - 2$.

In contrast to A and B, respondent C is indifferent towards pop, blues and rock, strongly dislikes classical and opera, and strongly likes bluegrass and country: $C = [3,3,3,1,1,5,5]$. These tastes again follow the same relative pattern as A and B, except all tastes are vertically shifted, inverted and amplified: $C = 2 \times (-1) \times A + 11$, or, equivalently, $C = 2 \times (-1) \times B + 7$.

Finally, respondent D strongly dislikes pop and rock, strongly likes blues, likes classical, opera and bluegrass, and dislikes country: $D = [1,5,1,4,4,4,2]$. Unlike A, B and C, this respondent construes an opposition between bluegrass and country, but not between bluegrass and opera. No series of inversions, multiplications or shifts of this pattern can transform it into the one experiences.” (2011:1401). These descriptions provide intuitions about the functional role schemas play in culture and cognition. However, they are not the clear definition needed to assess whether relationality can accurately measure such schemas.
exhibited by A, B and C. We thus conclude that, while respondents A, B and C follow the same schema, respondent D does not.

From this example, we can surmise that two respondents follow exactly the same schema if (i) their attitudes are identical, (ii) their attitudes are exact inverses of each other’s, (iii) the attitudes of either respondent are uniformly more extreme than those of the other, (iv) the attitudes of either respondent are uniformly more positive than of the other, or (v) any combination of (ii), (iii) and (iv). These conditions specify the mathematical operations of identity ($Y = X$), inversion ($Y = -X$), scaling ($Y = kX$) and vertical shift ($Y = X + b$) (see Figure 2). They can thus be captured by a single algebraic statement: two respondents X and Y follow exactly the same schema if and only if there exists a linear transformation that can produce one vector of responses from the other one, or, more formally, if there exist such constants $b$ and $k \neq 0$ that $Y = kX + b$. It is therefore intuitively clear that any measure of schematic similarity between two respondents should obtain its maximum value when such $k$ and $b$ exist, and should otherwise capture the degree to which one pattern can be approximated by linear transformations of the other.

Because of the simplicity of these operations, it is easy to imagine the social processes which could bring about such schematic transformations. For example, a highly opinionated person may follow the same pattern of genre tastes as a less opinionated one, but turn all “likes” into “strong likes” and all “dislikes” into “strong dislikes,” thus yielding situation (iii); a music lover may begin with a pattern of musical tastes common to her social group, but shift all of them upward, yielding situation (iv); a rebellious teenager could begin with his or her parents’ musical tastes and invert all of them, yielding situation (ii), etc.
Since many aspects of physical reality are linear (or approximately linear) in character, similar algebraic transformations can also be found in other cognitive domains. For example, in visual object perception, individuals are capable of recognizing the same stimulus placed in different positions and orientations (e.g., an apple that is rotated onto its side and moved further away, making it appear smaller). This process has been theorized to involve the perceiver effectively carrying out matrix-algebraic transformations of the perceived object so as to match it with representations of objects stored in memory (Ullman 1989; Hummel 2013). The theorized transformations of taste schemas captured in this model thus appear both socially and cognitively plausible.

Figure 2. Original schema and three basic linear schematic transformations.
Relationality

Following his motivating example, Goldberg (2011) offers relationality $R_{ij}$ as a measure of this schematic similarity. It is computed by first taking the row vector containing the attitudes belonging to a respondent, and calculating the differences between each pair of that respondent’s attitudes by subtracting them from one another. Each survey row $i$ is thus transformed into a square matrix $X_i$ of pairwise arithmetic differences between variables in that row. Then, to calculate the relationality between a pair of respondents $i$ and $j$, the absolute values of their respective difference matrices $X_i$ and $X_j$ are element-wise subtracted from each other. Each element of the resulting matrix $D_{ij}$ is assigned a sign based on whether the corresponding entries of $X_i$ and $X_j$ were in the same or in opposite directions. Finally, the elements of matrix $D_{ij}$ are summed together to yield the relationality $R_{ij}$, which is rescaled to range from 1 to $-1$.

The distinction between positive and negative relationalities is not useful for RCA, as either extreme indicates that respondents $i$ and $j$ follow the same schema. Thus, following a bias-reduction step that I examine later, RCA uses only the absolute values of relationality $|R_{ij}|$, which range from 1 (same schema) to 0 (unrelated schemas). RCA interprets the absolute relationalities as an adjacency matrix for a weighted network, with respondents as nodes and their pairwise absolute relationalities as ties. Finally, it uses a modularity maximization algorithm (Newman 2006) to partition this network into groups of respondents who have relatively high absolute relationalities $|R_{ij}|$.

---

3 Formally, Goldberg (2011) defines the relationality between two respondents $i$ and $j$ to equal $R_{ij} = \frac{2}{K(K-1)} \sum_{k=1}^{K} \sum_{l=k+1}^{K} (\delta_{ij}^{kl} \star \delta_{ij}^{kl})$, where $\delta_{ij}^{kl} = 1 - |\Delta X_i^k - \Delta X_i^l| + |\Delta X_j^k - \Delta X_j^l|$ and where $\Delta X_i^k = X_i^k - X_i^l$ is the difference between the values of the variables $k$ and $l$ for respondent $i$, and $\delta_{ij}^{kl} = 1$ if $\Delta X_i^k$ and $\Delta X_j^k$ have the same sign or are both zero, and $\delta_{ij}^{kl} = -1$ otherwise.
Goldberg’s (2011) argument that relationality measures schematic similarity is discursive rather than formal. Drawing on the structuralist idea that symbols acquire their meaning from their distinctions with other symbols, he argues that “[c]omparing how two individuals organize meaning therefore requires examining the associations between their attitudes. This calls for a method that looks at the extent of dissimilarity between the pairwise differences between their individual opinions” (1403). He therefore constructs relationality around an across-respondent comparison of within-respondent arithmetic differences in genre ratings, apparently reasoning that, since the shared schemas specify the relative value individuals assign to musical genres, they would be captured by such subtractions. But that reasoning is based on a conflation of distinction (semantic difference) and subtraction (arithmetic difference)—two concepts that are related in some contexts, but substantially different in others. To illustrate the problem with this conflation, consider an analogy to melody perception, which is another cultural domain where individual elements are defined relative to one another. Instead of the question “are two individuals who report different musical tastes nonetheless following the same schema in different ways?” the question here becomes “are two individuals who play different notes nonetheless performing the same melody in different keys?”

As is commonly known, most people perceive only relative rather than absolute pitch: that is, rather than discerning the absolute sound frequency of a musical note, they perceive its frequency relative to other notes played. Thus, if one were to change the key a melody is performed in—i.e., to multiply all the tone frequencies in the melody by the same constant—most listeners would perceive the result to be identifiably the same melody (Radocy and Boyle 2012). Algebraically, if an audible frequency \( k \) is taken to be the note \( A \) in an equal-tempered
music scale, then the following two notes in the scale ($B\flat$ and $B$) would have frequencies $k \times \sqrt[12]{2}$, and $k \times \sqrt[12]{2} \times \sqrt[12]{2} = k \times \sqrt[6]{2}$, respectively. For any melody, changing the value of $k > 0$ would alter the scale’s key, but keep the melody recognizably the same (since, e.g., for any $k \neq 0, B/A = \sqrt[6]{2}$). Therefore, within the constraints of human hearing, two tone sequences $(x_1, y_1, z_1)$ and $(x_2, y_2, z_2)$ would be recognizable as the same melody as long as $y_1/x_1 = y_2/x_2$ and $z_1/y_1 = z_2/y_2$. An algorithm that determines whether two individuals are performing the same melody in different keys would thus need to effectively carry out an across-performer comparison of within-performer ratios (geometric differences) between the tone frequencies. If it were to instead compare the melodies by subtracting the tones from one another, as RCA does with musical tastes, it would generally get the comparison wrong.

The goal of this analogy is not to argue that subtraction-based relationality is necessarily the wrong measure for schematic similarity. Rather, it is to point out that distinctions between elements in a relative system may not be subtractive in character. We thus cannot conclude that relationality $|R_{ij}|$ is a valid measure of schematic similarity simply because it compares within-respondent arithmetic differences in tastes. Its ability to correctly detect shared schemas must instead be tested directly, by applying it to sets of response patterns that either do or do not follow the same schema, and examining whether it can correctly determine which are which. Goldberg’s introductory example discussed above can be used as such a test case.

---

5 Goldberg (2011) reports some simulation tests of RCA’s accuracy in his online Appendix B. However, his simulations consist entirely of individual taste vectors $Y = X$ and their exact inverses $Y = -X$, plus noise. He reports an RCA accuracy of 100% in test cases without noise, and 97.8% with noise. But these simulations are too easy: they cover only one out of the three kinds of schematic transformation, missing $Y = kX$ and $Y = X + b$. They also grant RCA’s strong assumption that inversion probability equals 50% (see “Distributional Assumptions” below). To test performance under the full range of schematic similarity patterns in the motivating example, simulations would need to cover a substantially broader variety of scenarios. The simulations I report in the following sections cover this full range. On these more realistic tests, neither RCA nor CCA can reach such a high accuracy.
Figure 3. Absolute values of pairwise correlations (solid black) and relationalities (diagonal hatch) between the four patterns from Figure 1. The goal is to correctly determine that A, B and C belong to one schematic class, but D does not. The “desired” arrows indicate the values of each comparison that would best lend themselves to this correct answer. Remaining bars depict $|R_{ij}|$ after bias adjustment if each pattern occurs once (dotted), and if patterns A and B occur three additional times (dashed).
Recall that, in the introductory example, respondents A, B and C follow exactly the same schematic pattern, while D follows a different one (Goldberg 2011:1405). The values $|R_{ij}|$ obtains in the introductory example should thus clearly identify that A, B and C share a schema, but D does not. However, relationality’s difficulties at this task are evident in Figure 1B of the original paper (Goldberg 2011:1405). I depict the relevant parts of this diagram\(^6\) in Figure 3, where diagonally shaded bars represent $|R_{ij}|$. Relationality achieves its maximum value for the respondent pair A and B (AB = 1.00), thus clearly indicating that the two follow the same schema. Conversely, the relationality between the pair A and D is approximately 0.2, which is appropriately low as A and D follow different schemas. Since C follows the same schema as A and B, the absolute relationalities AC and BC should optimally be equal to the same value as AB (1.00). Unfortunately, this is not the case: both AC and BC have absolute relationalities of approximately 0.3, which is far closer to the relationality of the unrelated pair AD (0.2). Thus, relationality appears to grossly understate the schematic similarity between respondent C and respondents A and B.

To determine whether this inaccuracy affects the solution yielded by RCA, I created a dataset consisting entirely of rows A, B, C and D, each repeated 200 times for a total of 800 rows. I then analyzed it with the RCA software provided by Goldberg (Goldberg and Zhai 2013). To produce the correct solution, RCA would have to partition this population in two classes, the first containing all the copies of rows A, B and C (600 rows total), and the second all copies of D (200 rows). However, RCA instead produced an erroneous solution consisting of three distinct

---

\(^6\) Respondents are assigned to classes based on the absolute values of the relationalities $|R_{ij}|$, so I plot the values $|R_{ij}|$ instead of $R_{ij}$ to enable easier visual comparison of their magnitudes. I also omit the Euclidean distances, which are not relevant to the present discussion. These changes make the limited dynamic range of relationality values more visually apparent. Since Goldberg does not present the numerical relationality scores for this example, I reproduce these values by measuring the bar lengths in his figure.
classes, with the copies of C incorrectly assigned to their own class, separate from copies of A and B\textsuperscript{7}. Since Goldberg uses this example to introduce relationality as a tool for detecting shared schemas, RCA’s inability to do so here is troubling.\textsuperscript{8}

**Correlation**

This shortcoming means that there is merit in trying out a different measure of schematic similarity. Recall that two respondents X and Y exactly follow the same schema if there exist such constants \( k \neq 0 \) and \( b \) such that \( Y = kX + b \). Thus, provided that X and Y have a finite non-zero variance, it can be easily shown that the absolute Pearson’s correlation \(|r|\) between X and Y equals 1 if and only if they follow exactly the same schema. As the two responses become more and more linearly independent of one another—that is, as the best possible linear transformation of X leaves an ever larger percentage of Y’s variance unexplained—the value of \(|r|\) decreases monotonically towards 0. Finally, \(|r|\) will be equal to 0 if and only if \( k = 0 \) gives the best linear approximation of Y, or, in other words, if the best linear approximation of Y ignores the contents of X altogether. This is why Pearson’s correlation is often interpreted as the “measure of the degree of linear relationship between two variables” (Stockburger 2007; see Rodgers and Nicewander 1988 for a detailed treatment). Thus, Pearson’s correlation appears to be a perfect candidate for this task.

\textsuperscript{7} When I ran RCA with default parameters, it partitioned the population into 800 separate classes, thus assigning even identical rows to different classes. This obviously faulty solution appears to be due to the pseudo-significance testing RCA uses to filter weak relationalities, which is based on strong assumptions about how relationalities are distributed in the data. Disabling it produced the substantially more realistic solution I report above. (As I discuss below, this filter appears to generally decrease the average accuracy of RCA.)

\textsuperscript{8} Because of its complexity, I have thus far omitted discussion of the bias adjustment procedure that RCA performs on the relationality scores before taking their absolute value. I will return to this introductory example later in the paper to describe this procedure in due detail. In particular, I will show that the bias reduction procedure relies on strong distributional assumptions that may be violated in much empirical data, and can cause substantial problems when these assumptions do not hold. See “Distributional Assumptions” section below. Please also note that the RCA software did not offer an option to disable the bias adjustment step. All the RCA analyses I report in this paper thus include this bias adjustment.
The solid bars in Figure 3 demonstrate the results obtained by applying Pearson’s correlation to the same problem. The absolute correlations AB, AC and BC all equal 1, whereas AD, BD and CD equal 0.25. The absolute correlations between responses that follow the same schema are thus at their theoretical maximum, while the ones between members of different schematic classes are closer to their minimum. Thus, correlation appears to produce a far clearer depiction of the schematic relationships between these respondents than relationality. To examine whether this improvement results in a correct partition into classes, I adapted Goldberg’s technique to use row correlations in place of relationalities. I term the resulting algorithm Correlational Class Analysis (CCA; see Appendix A for details). And indeed, when I applied CCA to the same 800-row dataset, it correctly assigned all the rows into the two schematic classes present in the data. Thus, while RCA failed to correctly recover the schematic classes in Goldberg’s example, CCA produced a perfect answer.

SIMULATION
The above analysis suggests that CCA is a more accurate tool than RCA for detecting groups of respondents whose tastes follow the same cultural schema. However, one may rightly object that a single example does not provide a sufficient basis for drawing such a broad conclusion. To rule out the possibility that CCA’s apparently superior performance is due to features specific to Goldberg’s introductory example, I turn to simulation to carry out a more thorough analysis of the accuracy of both methods. In each of the 10,000 simulation runs reported in this section, I create a dataset where simulated respondents arrange their tastes according to one of a set of randomly generated shared cultural schemas by a process consistent with the theoretical model described above (in later sections, I examine simulations that violate the assumptions of this model). Since the schematic class membership of each respondent is known by design, these
simulations enable me to test how accurately CCA and RCA assign respondents to schematic classes under the conditions described by the theory.

The formal definition of shared schemas can serve as the basis for such simulations. Since two response vectors exactly follow the same schema if and only if they are linear transformations of one another, the schema specifying relationships between \( N \) tastes can itself be specified with a vector \( \rho = [\rho_1, \rho_2, \ldots, \rho_N] \).\(^9\) Such a schema-specifying vector (hereafter “schema”) can be randomly generated by drawing a vector of integers from an appropriate probability distribution. In turn, each response vector \( X \) that exactly follows this schema can be generated by randomly drawing a pair of linear transformation constants \( k \) and \( b \), which can invert, rescale or shift the pattern in ways consistent with the theory discussed earlier. Since real survey respondents do not perfectly reproduce cultural schemas, a simulated response must also include substantial stochastic deviations from the schema. These can be introduced via an independent error vector \( \epsilon \) of the same length as the original pattern, so that \( X = k\rho + b + \epsilon \). This is the basic formula behind the first two sets of simulations I examine.

To ensure that the simulations cover a wide range of potential cases, the simulation procedure contains three randomization steps. The first step of each run randomizes the broad characteristics of the simulated dataset, such as the ranges and variances that will be used to generate the values of \( \rho, k, b \) and \( \epsilon \), as well as the number of distinct taste schemas behind the responses. The second step generates these schemas using the variance parameters produced in the first step. Finally, the third step generates a random number of respondents following each of these schemas by applying random linear transformation and adding random noise, both

\(^9\) For example, the schema shared by A, B and C in Figure 1 can be specified as \( \rho = (0,0,0,1,1,-1,-1) \), so that \( A = \rho + 4, B = \rho + 2, \) and \( C = -2\rho + 3 \). The schemas are thus also themselves defined up to a linear transformation.
generated using the ranges and variances set in the first step. I repeat the entire procedure 5000 times\textsuperscript{10}, creating simulated datasets that widely differ in the ranges of simulated variables, variance of individual responses, signal to noise ratio, and many other parameters. A more detailed description of the simulation procedure can be found in Appendix B.

**Figure 4.** Three simulated schemas (solid black) for one simulation run, with a small sample of responses derived from each schema depicted behind each one in the same plot (dashed). In this run, the pattern variance was 0.51, and the maximum noise variance for individual responses was 1.05. This moderately high ratio of noise to schema variance creates a relatively difficult classification task.

\textsuperscript{10} Because of an apparent bug, the RCA software repeatedly crashed for 69 (1.3\%) of these simulated datasets, producing no results. I excluded these cases from the analysis.
Figure 4 illustrates a single simulation run. The randomly determined parameters from the first step of the run set the schema variance to 0.51, number of schemas to 3, and the maximum error variance, shift and scaling to 1.02, 1 and 2, respectively. Thus, to make the schemas $A, B$ and $C$, the simulation drew vectors from the Normal distribution with $\mu = 0$ and $\sigma^2 = 0.51$, and rounded to them the nearest integer. The resulting schemas are depicted with solid black lines, one per plot. For each schema, the simulation created a set of followers by randomly picking a value of shift $b$ from $\{-1,1\}$, scaling and inversion factor $k$ from $\{-2,-1,1,2\}$, and a noise vector $\epsilon$ drawn from the normal distribution with variance of no more than 1.02. A small sample of such respondents is depicted in dashed lines behind the appropriate schema.

**Measuring Accuracy**

Each of the 5000 simulated datasets generated by this procedure consists of randomly generated respondents who arrange their tastes according to one of a set of randomly generated taste schemas in a fashion consistent with the formal model I developed here. Thus, as in Goldberg’s (2011) introductory example, the true schematic class membership for each simulated respondent is known by design. The simulation’s goal is to assess the accuracy with which the group assignments made by the two algorithms correspond to this known membership. If two respondents were generated from the same schema, they should be assigned to the same group; if they were created from different schemas, they should belong to different groups.

For each run, I measured this classification accuracy of each algorithm with Normalized Mutual Information (NMI):
where $C$ is the vector of true class memberships for every respondent, $\Omega$ contains the corresponding class assignments made by the algorithm, $I$ is mutual information, and $H$ is Shannon entropy. NMI is an established criterion for measuring the accuracy of network partitioning algorithms (e.g., Danon et al. 2005; Lancichinetti, Fortunato, and Kertész 2009). It ranges from 1 when the estimate $\Omega$ perfectly recreates the true membership structure $C$, to 0 when $\Omega$ is independent with respect to these true schematic classes (Manning, Raghavan, and Schütze 2008).

### Table 1: Comparison of RCA and CCA accuracy in 10,000 simulation runs

<table>
<thead>
<tr>
<th>Measure</th>
<th>Simulation 1 (5000 runs)</th>
<th>Simulation 2 (5000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relationality (RCA)</td>
<td>Correlation (CCA)</td>
</tr>
<tr>
<td>Overall accuracy (median NMI)</td>
<td>0.74</td>
<td>0.87</td>
</tr>
<tr>
<td>Accuracy, interquartile range (25% to 75%)</td>
<td>(0.54, 0.88)</td>
<td>(0.69, 0.97)</td>
</tr>
<tr>
<td>Runs with near-perfect accuracy (NMI &gt; 0.95)</td>
<td>13.2%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Runs with near-complete inaccuracy (NMI &lt; 0.05)</td>
<td>2.8%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Runs with higher accuracy than other method</td>
<td>5.2%</td>
<td>88.1%</td>
</tr>
<tr>
<td>Odds of higher CCA accuracy in a given run</td>
<td>16.9 : 1</td>
<td></td>
</tr>
</tbody>
</table>

*Notes: Results from simulations with randomly varying schema variances, noise amounts, and ranges of linear transformations. In Simulation 2, inversion odds also randomly varied.*
Simulation Results

The results of these 5000 tests are presented in Table 1 under the heading “Simulation 1”. The median accuracy of CCA (0.87) is higher than that of RCA\(^{11}\) (0.74), a difference that is highly significant statistically (Wilcoxon \(W = 8585271, p < 0.0001\)). The interquartile range (IQR) of CCA’s accuracy extends from 0.69 to 0.97, while RCA’s extends 0.54 to 0.88. Thus, while CCA’s 75\(^{th}\) percentile is just shy of a perfect accuracy, RCA’s 75\(^{th}\) percentile barely surpasses CCA’s median. The substantive significance of these differences is clearer when the CCA accuracies (\(Y\)) are plotted against the RCA accuracies (\(X\)) in Figure 5. While the accuracies are strongly associated (\(R^2 = 0.79\)), CCA is more accurate than RCA in the vast majority of cases (88.1%). In contrast, RCA is more accurate than CCA in only 5.2% of the cases. Thus, when RCA and CCA disagree, which they do in 93.3% of the cases, the odds that CCA’s result is more accurate than RCA’s equal 17:1.

To determine if the results point to any classes of data where RCA would nonetheless be preferable to CCA, I disaggregated them by schema variance and noise variance, which are the parameters most responsible for the difficulty of the classification task. Lower schema variances or higher noise variances result in more challenging signal to noise ratios, which should make the performance of both algorithms poorer. The loess curves demonstrating this effect are presented in Figure 6.

\(^{11}\) RCA software contains a user-configurable filtering step based on pseudo-significance testing, where weak relationalities are dropped prior to partitioning to reduce potential noise. I examined how filtering affected RCA’s performance using 250 simulation runs. Filtering increased accuracy in 56% of the cases but and decreased it in 43%. However, the average decrease (-0.33) was three times greater than the average increase (0.11). Overall, disabling the filter substantially raised RCA’s median accuracy, from 0.56 when enabled to 0.69 when disabled. Additionally, in 10% of the cases with filtering, RCA encountered an error and yielded no solution at all (as compared to 1% without filtering). Thus, to increase RCA’s accuracy and avoid potential bias from substantial missing results, I disabled the filter for all the simulations reported in this paper.
On both plots, the accuracies of the two algorithms overwhelmingly rise and fall together, thus suggesting that the algorithms find the same cases challenging and the same cases easy. In spite of this similarity, CCA’s accuracy remains reliably above RCA’s throughout the full ranges of both variances, with a median gap of roughly 0.10 in favor of CCA. This accuracy gap remains remarkably stable throughout most of the two ranges. For example, in simulations within the top 5% of noise variance, the median accuracies equal 0.61 for CCA and 0.51 for RCA. Within the bottom 5% of noise variance, they equal 1.00 and 0.91, respectively. CCA thus retains the same advantage over RCA under both the least and the most challenging noise conditions examined.

**Figure 5.** CCA accuracy compared to RCA accuracy for 5000 simulation runs. Each point is a single run. Runs where CCA was more accurate are above the $Y = X$ diagonal (in gray), while those where RCA was more accurate are below. Note the absence of points in the bottom-right corner.
Figure 6. Loess curves comparing RCA and CCA accuracy, based on 5000 simulation runs.
The only instance where the two accuracy curves substantially deviate from each other occurs when the schema variance is low. In such situations, the pairwise correlations between responses tend towards zero, while their relationalities approach one. Therefore, as Goldberg points out, the relationality between low-variance respondents is systematically higher than the correlation. When Goldberg briefly considers using correlation in his online appendix, he dismisses it on the basis of this difference. However, Goldberg incorrectly interprets this difference to mean “relationality does a better job at examining relationships between respondents whose responses have relatively low variance” (Goldberg 2011:online Appendix A). The fact that relationality produces higher values than correlation does not imply that it produces more accurate values. And indeed, as can be seen on the left side of Figure 6A, the opposite appears to be true. When schema variance is in the lowest 5% of its simulated range, CCA’s median accuracy decreases to 0.31. However, RCA’s accuracy experiences a disproportionally large drop, decreasing to 0.07. This low accuracy indicates that RCA results for low-variance schemas contain almost no information about the true membership structure of the data. It further suggests that relationality may exhibit an upward bias for low-variance observations that is substantially more damaging to its performance than correlation’s downward bias.\(^{12}\)

**Distributional Assumptions**

Relationality also introduces a strong distributional assumption which may further degrade its accuracy when violated. While a correlation of zero always indicates an absence of a linear relationship, the equivalent “null value” of relationality differs from dataset to dataset and is generally skewed above zero (Goldberg 2011:Appendix A). RCA attempts to compensate for

---

\(^{12}\) Those rare respondents with a variance of absolute zero require special treatment. See Appendix E for detailed discussion.
this bias by re-centering the matrix of relationalities by its mean. This bias adjustment procedure

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


This bias adjustment procedure crucially rests on the assumption that the true mean relationality between all the rows in the data is zero, or, equivalently, that the relationality values are distributed symmetrically around their null value. This would normally be the case only if the proportion of respondents following a schema without inverting it equals the proportion following its inverse—a quantity I call inversion probability. For example, when this probability is 50%, the number of highbrow respondents following the schema “like classical, like opera, dislike rock, dislike country” would equal the number of lowbrow respondents following its inverse, “dislike classical, dislike opera, like rock, like country”. However, since in reality the number of highbrow respondents can differ greatly from the number of lowbrow respondents, there is no reason to expect that the inversion probability generally equals 50%. This suggests that RCA’s symmetry assumption may be frequently violated by empirical data.

To illustrate the magnitude of potential error that occurs when this assumption is violated, I return to the introductory example. In Figures 1 and 3, the mean relationality between the four respondents equals 

\[
R_{obs} = \frac{AB + AC + AD + BC + BD + CD}{6} \approx 0.15.
\]

Subtracting this value from the other relationalities yields the adjusted relationalities \(|R - R_{obs}|\), depicted in dotted lines in Figure 3. In this adjusted result, the related pairs AC and BC are assigned higher relationalities than before, whereas the unrelated pairs AD, BD and CD are decreased. Aside from the relationality of the related pair AB, which is also decreased by the bias adjustment, the bias-adjusted scores \(|R - R_{obs}|\) appear a clear improvement over the unadjusted values of \(|R|\).

However, the bias adjustment procedure can easily have the opposite effect. Consider an alternate dataset which is just like the one in the introductory example, but with the addition of
three more copies each of patterns A and B. In this alternate dataset, the mean relationality would then be $\bar{R}_{\text{alt}} \approx 0.6$. The relationality matrix that is adjusted for this new mean, $|R - \bar{R}_{\text{alt}}|$, is depicted in dashed lines in Figure 3. This adjustment correctly brings the relationalities between AC and BC closer to 1. However, it also causes a large decrease in the relationalities between the related pair A and B, which are identical except for a vertical shift. It further substantially increases the estimated relationalities between unrelated pairs AD, BD and CD. These completely unrelated patterns are now erroneously assigned a higher relationality than the closely related pair AB. As this example illustrates, RCA’s bias adjustment procedure crucially depends on the relative number of times each pattern appears in the observed dataset—a quantity that is fundamentally arbitrary.

All the simulations presented above have granted RCA’s assumption of symmetric distribution of relationalities by keeping the inversion probability fixed at 50%. To examine the effects of relaxing this assumption, I created a second simulation where the inversion probability is instead drawn from a uniform distribution over its full range, and varies between each of the 5000 simulation runs. The results of this second simulation are reported on the right side of Table 1. As expected, both the median (0.86) and the interquartile range (0.69 to 0.97) of CCA accuracies remain unchanged from the first simulation. On the other hand, the median accuracy of RCA drops to 0.67, significantly lower than its prior median of 0.74 (Wilcoxon $W = 13802245$, $p < 0.0001$), and below CCA’s 25th percentile of accuracy. CCA is now more than 3 times as likely as RCA to produce a nearly perfect answer (NMI > 0.95), while RCA is more than 20 times as likely to produce an almost completely incorrect one (NMI < 0.05). Thus, RCA’s accuracy appears to suffer a significant further drop when its assumption of symmetrical

---

14 See Appendix B (“Random Inversion Probability”) for details of procedure.
distribution is violated, as may generally happen in empirical applications. CCA remains unaffected by this change.\footnote{For a point of comparison, note that Pearson’s correlation is only guaranteed to be unbiased and asymptotically efficient if the vectors are bivariate normal in distribution. To examine whether this is the case with my primary simulations, I applied the Royston $H$ test for multivariate normality (Royston 1983) to 425 datasets randomly generated by the full linear simulation procedure from Appendix B. In each one of the 425 datasets, the hypothesis of multivariate normality was rejected at $p < 0.0001$. I then partitioned each simulated dataset by schematic class membership, so that respondents in each resulting dataset would come from exactly one schematic class. In each one of these 1728 single-class datasets, the hypothesis of multivariate normality was again rejected at $p < 0.0001$. The data produced by my linear simulation procedure thus violate correlation’s bivariate normality assumption. Note that CCA was nonetheless able to analyze many of the simulated linear datasets with near-perfect accuracy.}

The results of both simulations thus reinforce my earlier suppositions. Though the accuracies of RCA and CCA were highly correlated, they nonetheless differed in almost 95% of the 10,000 combined simulation runs. In simulation 1, which obeyed RCA’s symmetry assumption, the odds that CCA’s result was more accurate equaled 17:1. In simulation 2, where this assumption was relaxed, these odds further rose to 23:1. CCA was more accurate over the full range of schema and noise variances examined. The only major deviation from this otherwise stable gap in accuracies occurred in simulations with low schema variance. In those simulations, RCA performance suffered a significant additional drop relative to CCA’s. Overall, the two simulations provided no evidence of any use cases where RCA would be preferable to CCA.

**DEVIATIONS FROM THE MODEL**

The 10,000 simulations I described above demonstrate that CCA yields substantially more accurate results than RCA when the shared cultural schemas result in linear dependencies between rows in the survey dataset. Such linear dependencies are consistent with the motivating examples of schematic similarity in Goldberg’s (2011) work, as well as the formal reasoning behind CCA I laid out here. As I argue above, they are also socially and cognitively plausible.
Nonetheless, schematic cultural cognition is still little understood, and whether this model of schematic similarity as linearity is correct is not known. For this reason, the performance of both algorithms when the model is violated is also of interest. Another reason for this interest stems from the close conceptual relationship between correlation and linearity, which could potentially lead correlation-based CCA to perform disproportionally well in simulations that obey the linearity assumption. Since Goldberg may not have been aware that the examples he presented in his work were linear in character, he did not specifically design RCA to detect linear transformations. Simulations that violate the model would thus yield a more conservative test of CCA’s accuracy, and could potentially point to use cases where RCA has an advantage over CCA. For these reasons, I now turn to simulations where respondents’ patterns are produced through processes that differ from the linear transformations theorized here.

There are two fundamentally distinct ways that schematic transformations could deviate from the theorized model. First, schematic transformations could in reality be based on only a subset of the possible linear operations (shift, scaling and inversion). Though such cases would technically remain linear, the two algorithms may perform differently on these “edge cases” of linearity than they do with all three operations present\(^\text{16}\). And second, schematic transformations could be based on operations which are non-linear in character, such as in the presence of multi-way interactions between variables. I will use the shorthands “sub-linear” and “super-linear” to refer to these two broad kinds of deviations from the model. In this section, I examine six distinct sets of simulations to investigate both of these possibilities.

\(^{16}\) Among other reasons, such differences in accuracy may arise if two measures have different amounts of sensitivity to inversion, scaling and shift. *Ceteris paribus*, the measure that is more likely to identify that \(X\) and its noisy inverted copy \(Y = -1 \times X + \epsilon\) follow the same schema would probably yield a more accurate solution in the linear simulations examined above. However, the same property would then make it more likely to yield false positives in the “no inversion” scenario, where two patterns that appear to be inverses of one another would actually be produced by two different schemas.
Sub-linear Transformations

Recall that musical taste schemas are theorized to consist of relative evaluations of different musical genres. Because these evaluations are relative rather than absolute, individuals can use one schema to produce a variety of contrasting taste patterns. Specifically, they can invert the schema, turning likes into dislikes; they can scale it, making all the appraisals uniformly more or less extreme; and they can vertically shift it, making all the appraisals more positive or negative. All three of these transformations intuitively appear highly plausible. However, in the absence of empirical work attesting to these theorized socio-cognitive processes, it is difficult to say for certain that every one of these schematic transformations is pervasive enough to affect the overall distribution of taste patterns on a society-wide scale. Given the noisiness of much survey data, a method that searches for rare or nonexistent schematic transformations could generate substantial amounts of false positives, thus yielding inaccurate results. Since the simulations reported above generally featured all three of the transformations, their results could conceal such poor performance. In this section, I examine three sets of simulations where each of these transformations is absent to investigate this possibility.

I begin by briefly sketching some potential reasons to question the prevalence of the basic linear operations of scaling, inversion or shift. My goal here is not to argue that any of these operations are likely to be absent, but rather to point out that such an absence is plausible enough to warrant further attention. To do so, I return to Figure 2, which illustrates each of the three basic linear transformations in isolation.

The patterns in this figure are all derived from one shared schema (labeled “original”). Respondent P follows this schema without any transformations, and thus likes both rock and rap

17 I describe these operations in terms of individual action largely for the sake of readability. The actual socio-cognitive processes of schematic transformation could well be supra-individual.
(+2) and mildly dislikes country, folks and oldies (−1). Respondent Q uses the same schema as P, but inverts it to create the opposite pattern of tastes (Q = −1 ∗ P). The cognitive demands of such an inversion are minimal, as it simply requires taking a pattern of tastes and replacing the “likes” with “dislikes”. However, whether two social actors with very different taste patterns like “highbrow” and “lowlbrow” would actually employ exactly the same cultural schema is uncertain. Since cultural taste is intertwined with social position, individuals with dramatic cultural differences are also likely to live in different kinds of places, consume media from different sources, and have network ties to distinct kinds of alters (Bourdieu 1987; Fararo and Skvoretz 1987; McPherson 2004; McPherson, Smith-Lovin, and Cook 2001). This social distance suggests that, rather than inverting the same taste schema, two groups with contrasting tastes may employ separate cultural schemas that maintain only an indirect, imprecise, symbolic or accidental opposition to one another.

It is also possible to raise questions regarding the empirical prevalence of vertical shifts like the one employed by respondent R (= P − 3). While the original schema consisted of a pattern of contrasting positive and negative appraisals, R’s pattern contains only genres she dislikes a little (rock and rap = −1) and those she dislikes a great deal (country, folk and oldies = −4). Such a transformation could occur if respondent R had internalized the same set of distinctions as P but simply did not enjoy listening to music, thus giving lower appraisal to all the genres. However, internalizing cultural distinctions takes effort and time (e.g., Bourdieu 1984). It is thus reasonable to wonder if respondents who dislike music would generally put in the effort to internalize the same complex set of genre contrasts as music lovers, who presumably have a far greater intrinsic motivation to expose themselves to music.
The case of respondents with multiplicatively scaled response patterns like \( S (= P * 2) \) could raise similar questions. Unlike \( P \), who perceives relatively mild distinctions between musical styles, \( S \)'s tastes feature dramatic contrasts between very strong positive assessments of some genres and very negative assessments of others. This again invites the same concern regarding generally unequally levels of motivation.

These questions about the empirical prevalence of the three kinds of linear transformations will need to be settled by future empirical work. In the meantime, they merit investigating how CCA and RCA would perform in the scenario that each of these transformations does not take place. Among the second set of 5000 simulations I reported earlier in the paper, 1185 featured no vertical shift \((b \in \{0\})\) and 1700 featured no scaling \((k \in \{1\})\). To examine the performance of CCA and RCA in the absence of these transformations, I reanalyzed these cases in isolation. I also performed 1000 further simulations to examine the performance of the two algorithms in the absence of schematic inversion \((\zeta = 0)\).

I report the accuracy of CCA and RCA in the three sub-linear scenarios in Table 2. Among the “no shift”, “no scaling” and “no inversion” simulations, CCA had accuracies of 0.84, 0.82 and 0.87, respectively. These are slightly lower on the average than its prior accuracy of 0.87 (Table 1, simulation 2), though the difference is minor. With RCA, the “no shift,” “no scaling” and “no inversion” simulations yielded accuracies of 0.68, 0.66 and 0.41, respectively. While the first two of these scores are close to RCA’s prior accuracy of 0.67, its score in the “no inversion” scenario is almost 40% lower. RCA was thus capable of retaining its prior accuracy in the “no shift” and “no scaling” scenarios, but not the “no inversion” scenario. This large drop in accuracy may again originate with RCA’s strong assumption of symmetric distribution discussed above.
Comparing the CCA and RCA accuracies on each individual run showed that CCA was more accurate in 88%, 88.5% and 99% of the scenarios that omitted shift, scaling and inversion, respectively. Thus, correlation-based CCA remains the preferred choice of method if there is doubt regarding the empirical prevalence of some of the theorized linear transformations, and especially so if these doubts concern schematic inversion. In Appendix D, I also show how Pearson’s formula can be altered to produce modified coefficients that specifically reflect each of these three sub-linear scenarios. The three modified coefficients treat inversion, shift, or scaling as schematic difference rather than schematic similarity. Though CCA performed well in each of these settings, it may be possible to use such “specially tuned” correlation coefficients to increase its accuracy even further while retaining the method’s speed and simplicity.

Table 2: Comparison of RCA and CCA accuracy in six simulations with departures from theorized model

<table>
<thead>
<tr>
<th>Simulation</th>
<th>RCA accuracy (median &amp; IQR)</th>
<th>CCA accuracy (median &amp; IQR)</th>
<th>RCA more accurate (% of runs)</th>
<th>CCA more accurate (% of runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sub-linear:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No shift</td>
<td>0.68 (0.46, 0.86)</td>
<td>0.84 (0.65, 0.96)</td>
<td>6.6%</td>
<td>88.0%</td>
</tr>
<tr>
<td>No scaling</td>
<td>0.66 (0.44, 0.84)</td>
<td>0.82 (0.61, 0.95)</td>
<td>6.4%</td>
<td>88.5%</td>
</tr>
<tr>
<td>No inversion</td>
<td>0.41 (0.21, 0.57)</td>
<td>0.87 (0.69, 0.97)</td>
<td>1.0%</td>
<td>98.7%</td>
</tr>
<tr>
<td><strong>Super-linear:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent subschemas</td>
<td>0.36 (0.15, 0.58)</td>
<td>0.54 (0.28, 0.75)</td>
<td>14.1%</td>
<td>85.0%</td>
</tr>
<tr>
<td>Multi-way interactions</td>
<td>0.23 (0.12, 0.42)</td>
<td>0.52 (0.26, 0.79)</td>
<td>4.1%</td>
<td>95.8%</td>
</tr>
<tr>
<td>Random polynomial functional form</td>
<td>0.27 (0.16, 0.38)</td>
<td>0.42 (0.31, 0.57)</td>
<td>5.7%</td>
<td>94.2%</td>
</tr>
</tbody>
</table>
Super-linear Transformations

The theoretical uncertainty surrounding the sociocognitive dynamics of cultural schemas mean that it is also conceivable that shared schemas undergo transformations that are more complex than shift, scaling and inversion. Goldberg intended relationality to function even in the presence of such non-linear relationships between variables (e.g., Goldberg 2011:1433), arguing that the complex way that relationality is computed makes it “more sensitive to interdependencies” between variables than correlation (Goldberg 2011:Appendix A). In this section, I use further simulations to investigate whether relationality’s complexity in fact makes it more accurate than correlation in the presence of such “super-linear” transformations.

The three further sets of simulations I examine in this section correspond to three broadly distinct kinds of non-linearity: independent subschemas, polynomial (rather than linear) functional form, and multi-way interactions between tastes. All three of these scenarios contradict the theory of schematic similarity developed here. All three also violate different assumptions made by Pearson’s correlation. They may thus reveal potential use cases where relationality-based RCA would be preferable to correlation-based CCA.

Independent Subschemas. All the reported simulations have thus far assumed that each cultural schema specifies the relationships between all the attitude objects. However, it is easily conceivable that different sets of tastes may be organized according to distinct cultural schemas. For example, one schema may specify the relationship between genres to which most Americans are frequently exposed on radio and television, while a different schema specifies the relationships between genres encountered only in more specialized settings. In this case, schematic transformations could affect different parts of the taste vector separately from each
other, since, e.g., a respondent could invert the taste schema for popular tastes without inverting the taste schema for specialized tastes.

When different transformations are applied to different parts of a schema, they combine into overall schematic transformations that cannot be captured in terms of simultaneous deviations from each vector’s single common mean. Since the correlation coefficient is computed from these deviations, the schematic classes defined by such transformations should be substantially more challenging for correlation to detect. In contrast, relationality compares the variations of variables around each other rather than around a common mean (Goldberg 2011: Appendix A), which may give RCA an advantage under this kind of non-linearity. To examine whether this is the case, I simulated 3000 datasets in which individual responses were derived from the schemas by applying different random transformations to separate sections of the taste schema (see Appendix C for details of simulation procedure).

I report the results of these “independent subschemas” analyses in Table 2. In these simulations, the median accuracy of CCA was 0.54, which is 38% below its prior accuracy of 0.87. The median accuracy of RCA, however, was only 0.36, or 46% below its prior accuracy of 0.67. Thus, for both CCA and RCA, these simulations proved substantially more challenging than the linear ones. Overall, RCA produced more accurate results than CCA in 14% of these simulation runs, while CCA produced more accurate results in 85% of the runs.

In Figure 7, I disaggregate these results by the count of independent subschemas in each simulation run. This count tracks the number of separate transformations that are applied to different tastes in the schema, and thus the extent to which a given run deviates from the assumption of variance around a single common mean. In the runs where this count equals 1, all the tastes were assigned to a single subschema. Such runs are thus analogous to the linear
simulations I analyzed above (albeit with different signal-to-noise ratios). CCA’s median accuracy in these single-subschema runs was 0.80, and RCA’s was 0.65. As the number of subschemas grew, the accuracy of both algorithms degraded at nearly the same gradual rate, falling to 0.38 for CCA and 0.25 for RCA when this quantity reached the simulation maximum of 4—a drop of 0.42 and 0.40, respectively. CCA remained consistently more accurate than RCA throughout the full range. Thus, differences in how the two coefficients are computed did not appear to make relationality any more robust to the absence of a common mean than Pearson’s correlation.

Functional form. I now turn to deviations from the model that can come from more complex functional forms. It is conceivable that schematic transformations are based on operations other than inversion, scaling and shift that I discussed in the first part of the paper. For example, it is possible to imagine a “polarizing” schematic transformation that further increases the extreme tastes in a schema (“like a lot”) but leaves the moderate tastes (“like a little”) relatively

![Graph showing accuracy for 2000 simulations with independent subschemas.](image)

**Figure 7.** Accuracy for 2000 simulations with independent subschemas.
unaffected (e.g., $X = \rho^3, -10 \leq \rho \leq 10$). It is also possible to imagine a transformation that retains the extremity of the tastes but ignores the valence, thus treating “like a lot” and “dislike a lot” as equivalent (e.g., $X = \rho^2, -10 \leq \rho \leq 10$). While intuitively less plausible as a general model of shared cultural schemas than the simple linear schematic transformations I described above, these scenarios are of interest because they test the methods’ robustness to unexpected violations of the model. They may also reveal potential use cases for RCA. I therefore use 2000 further simulations to investigate the performance of CCA and RCA in the presence of non-linear functional forms.

In these simulations, I replace linear transformations with randomly generated polynomials of the form:

$$X = g(\rho) = \sum_{i=0}^{10} a_i b_i \rho^i = a_0 b_0 + a_1 b_1 \rho + a_2 b_2 \rho^2 + \cdots + a_{10} b_{10} \rho^{10},$$

where $a_i \in \{-1,0,1\}$ and $b_i > 0$. The $a_i$ coefficients determine whether a polynomial term is present as well as whether it is inverted, and the $b_i$ coefficients determine its relative weight (see Appendix C for further details). High-degree polynomials such as these can express a wide range of distinct non-linear transformations while remaining easy to generate at random.

The results of these simulations are reported in Table 2 with the label “Random functional form”. In these simulations, CCA obtained a median accuracy of 0.43, which is roughly half of its prior median accuracy of 0.87. RCA obtained a median accuracy of 0.27, or roughly 40% of its prior accuracy of 0.67. CCA yielded more accurate results in 93.9% of the runs, while RCA yielded more accurate results in the remaining 6.1%. Thus, while polynomial transformations proved challenging for correlation, they appeared no less challenging for relationality.
In Figure 8, I disaggregate these results by the degree (maximum exponent) of the random polynomial transformation in the simulation run. This count determines the maximum number of minima and maxima of the function, and thus the complexity of the transformation. Some simulations were randomly assigned first-degree polynomial transformations, which are analogous to the linear transformations used before (albeit again with different signal-to-noise ratios). The figure shows that CCA as well as RCA had their highest accuracies with these first-degree transformations (0.74 and 0.64, respectively). The accuracy of both algorithms then dropped as the degree of polynomial increased, though the rate of this change slowed by the time the polynomial degree reached roughly 4. Throughout the entire range, CCA remained reliably more accurate than RCA, with the gap in favor of its accuracy growing slightly as the degree of polynomial increased. For $10^{th}$ degree polynomials, the CCA and RCA accuracies measured 0.44 and 0.27, respectively—a drop of 0.30 for CCA (41%) and 0.37 for RCA (58%). There is thus no

**Figure 8.** Accuracy for 3000 simulations with polynomial transformations.
evidence that RCA’s computational complexity improves its ability to detect more complex
functional forms of schematic similarity.

Multi-way interactions. I now turn to potential departures from the model caused by interactions
between variables in the taste schema. In the examples presented by Goldberg (2011) and the
simulations reported above, each elements of a respondent’s vector \( R \) was always produced from
transformations of exactly one element of the schema \( \rho \). For example, consider a respondent \( R \)
who begins with a taste schema \( S = [S_{\text{rock}} = 2, S_{\text{rap}} = x, S_{\text{classical}} = y, S_{\text{opera}} = z] \), then
inverts it, scales it by 2 and finally shifts it by +1. This respondent is, in effect, applying a
univariate function \( R_k = f(S_k) = (-1) \ast 2 \ast S_k + 1 \) to each element of \( S \). No matter what
values \( x, y \) and \( z \) assume, \( R \)'s attitude to rock would equal \( f(2) = -3 \). The only relationships
between the tastes are those described by the schema vector.

In contrast, in the presence of interactions between attitudes, a given respondent’s attitude
regarding a single genre may be a function of two or more tastes in the schema, e.g., \( R_{\text{rock}} =
\)
\( g(S_{\text{rock}}, S_{\text{rap}}, S_{\text{classical}}) = (S_{\text{rock}}^2 \ast S_{\text{rap}} \ast S_{\text{classical}})^{\frac{1}{3}} = \sqrt[3]{4xy}. \) The presence of such multi-way
interactions would entail intricate and perhaps unintuitive relationships between cultural schemas
and the responses they generate\(^{18}\). In contrast to the other schematic transformations I described
here, it is difficult to imagine any plausible socio-cognitive mechanism that would consistently
yield such intricate multi-way interactions. However, in the interest of examining both
algorithms under a maximally broad range of unexpected departures from linearity, I use further

\[^{18}\text{A population of respondents who use transformations such as these would be introducing cultural logics that are not part of the schema. Consider, for example, a schema like } S = [S_{\text{rock}} = 1, S_{\text{rap}} = 1, S_{\text{classical}} = 4, S_{\text{opera}} = 4] \text{ on a scale from } 1 = \text{like} \text{ to } 4 = \text{dislike, and a respondent } Z \text{ using the transformations } Z_{\text{rock}} = (S_{\text{rock}}^2 \ast S_{\text{opera}} \ast S_{\text{classical}})^{\frac{1}{25}} = 2 \text{ and } Z_{\text{classical}} = (S_{\text{classical}}^2 \ast S_{\text{rock}} \ast S_{\text{rap}})^{\frac{1}{25}} = 2. \text{ Though the lowbrow/highbrow schema } S \text{ placed rock in opposition to classical, and } Z \text{ produced her tastes without any error, she nonetheless ended up with an omnivorous taste pattern.}\]
simulations to study the performance of RCA and CCA in the presence of such multi-way interactions.

In this simulation procedure, each simulated respondent partitions a shared schema into groups of tastes I term “interaction blocks”, and then generates each genre taste by calculating its randomly weighted geometric mean with other tastes in its interaction block\(^19\). This means that each respondent uses a random transformation of the form similar to \(g\) above, e.g., \(R_k = g(k, m, n, p) = \sqrt[k]{k^{3mnp}}\). Since each respondent is assigned idiosyncratic interaction blocks, each response vector consists of multiplicative interactions between different sets of tastes. To create a range of complexities, the maximum number of tastes per interaction block differs at random from run to run (see Appendix C for detailed procedure).

I report the results of 2000 runs of this simulation in Table 2 (“multi-way interactions”). The median accuracy was 0.52 for CCA and 0.23 for RCA, indicating that correlation again yielded more accurate results than relationality. To examine how this difference in accuracy differed by complexity of interactions, I disaggregated the simulation results by maximum number of tastes per interaction block in the given run (see Figure 9). When this number equals 1, the transformations do not feature any interactions between terms, and respondents simply reproduce the original taste schema with the addition of noise but without any further changes. As expected, this is the easiest scenario for both CCA and RCA, which achieve accuracies of 0.72 and 0.42, respectively. These simulations become increasingly more challenging for both CCA and RCA as the number of tastes per interaction increases. By the time the interaction size

\(^{19}\) The general functional form of the transformation that produces a respondent’s taste \(R_k\) in interaction block \([R_k, R_m, R_n]\) from schema \(\rho\) is the weighted geometric mean \(R_k = g(\rho_k, \rho_m, \rho_n) = \sqrt[k]{k^{3mnp}}\). I attach weight \(a\) to the focal element \(\rho_k\) because without weights the respondent’s tastes for all genres in the interaction block would be identical: \(\sqrt[k]{k^{3mnp}} = \sqrt[k]{k^{3mnp}} = \sqrt[k]{k^{3mnp}}\). I override this unintuitive behavior in roughly 75% of the runs by randomly drawing \(a\sim U[1,2,3,4]\). When \(a > 1\), \(\rho_k\) has a stronger effect on \(R_k\) than do \(\rho_m\) or \(\rho_n\). See Appendix C for detailed simulation procedure.
reaches its maximum of 5, both CCA and RCA see their initial accuracies drop by roughly half, to 0.39 and 0.20, respectively. Throughout the 2000 runs of this simulation, CCA yielded a more accurate result than RCA on 96% of the runs, thus indicating that it is the preferable technique even in the presence of complex interactions between tastes.

The simulations I examined in this section compared three broad ways in which relationships between taste vectors can violate the assumptions of linearity. I examined situations where some parts of the taste schema were transformed independently from others; where linear transformation functions were replaced by polynomials; and where transformations of the taste schema involved multi-way interactions between separate tastes. While Goldberg (2011) contended that relationality’s computational complexity makes is better equipped than Pearson’s correlation to handle complex relationships between variables, these simulations pointed to the opposite conclusion. Correlational Class Analysis yielded more accurate results than Relational

![Graph](image)

**Figure 9.** Accuracy for 2000 simulations with multi-way interactions.
Class Analysis in 85% of simulations with independent subschemas, 94% of simulations with random polynomial functional forms, and 96% of simulations with multi-way interactions (see Table 2).

When I disaggregated each scenario by difficulty, I found that RCA’s accuracy trailed but closely paralleled CCA’s across all three non-linear scenarios (Figures 7-9), just as it had before for the linear ones (Figure 6). Even in Figure 8, where the relationship between polynomial degree and classification accuracy was non-monotonic, RCA’s accuracy repeatedly rose and fell in close synchrony with CCA’s. This remarkably persistent pattern suggests that relationality may simply be detecting the same relationships as correlation, albeit at a lower accuracy. It furthermore yields no evidence of cases—linear or nonlinear—where RCA would be preferable to CCA.

**EMPIRICAL EXAMPLE: MUSICAL TASTES**

To compare the results produced by the two methods in an empirical setting, I applied CCA to the 1993 GSS music tastes module previously analyzed with RCA (Goldberg 2011). This dataset contains 1532 respondents’ evaluations of 17 musical genres. Each respondent rated each genre using a five-point Likert scale that ranges from “like very much” to “dislike very much”. For comparability, I followed exactly the same coding procedures as Goldberg (2011). Below, I contrast the substantive contents of the results. I then compare them using a multiple groups analysis technique from structural equation modeling.

The RCA analyses partitioned the survey population into three schematic classes, which Goldberg labeled “Omnivore – Univore”, “Highbrow – Lowbrow”, and “Contemporary – Traditional.” For respondents in the “Omnivore – Univore” class, most genre tastes were positively correlated among each other. Goldberg interpreted this as evidence of a culturally
omnivorous taste schema, in which no genres are perceived as opposites, but rather a high appraisal of most genres is opposed to a low appraisal of most genres. In the “Highbrow – Lowbrow” class, tastes for “elitist” genres such as opera and classical music were positively correlated among each other, but negatively correlated with most tastes for popular genres. Finally, Goldberg characterized the third schematic class as “Contemporary – Traditional.” Here, a cluster of positively correlated tastes for well-established musical genres including gospel, bluegrass, and country is negatively correlated to tastes for arguably more contemporary genres, including heavy metal, pop and rap, as well as oldies and jazz.

The persistence of the Highbrow – Lowbrow taste schema was perhaps Goldberg’s most surprising finding, as much contemporary work has argued that omnivorousness has replaced highbrow tastes as a marker of high status in the contemporary United States (Peterson and Kern 1996; Peterson 1997, 2005; see Goldberg 2011 for a more detailed discussion). The CCA analyses of these data, however, do not replicate this finding. While RCA identified three classes in these data, CCA identified four, which are presented in Figure 10. The first two of these closely resemble those located by RCA. The first class features practically no negative correlations between the genres, suggesting that respondents in this class perceive little opposition between different musical styles. In this population, positive appraisal of any one genre generally “goes with” positive appraisal of any other genre, suggesting an undiscriminating logic of taste that ranges between near-uniformly positive appraisals of all genres on one extreme, and a near-uniformly negative appraisal of all genres on the other. This is the same omnivorous logic as behind the Omnivore – Univore class identified by RCA.
Figure 10. Networks illustrating the four schematic classes identified by CCA. Dashed lines indicate negative correlations. Weak correlations ($|r| < 0.05$) not plotted.
The second class located by CCA appears to be defined by an opposition between rock, rap and metal on one extreme, and gospel, country, folk and bluegrass on the other. This suggests a bifurcation of respondents into those who prefer newer musical genres and those who prefer more established ones, which closely resembles the logic of the Contemporary – Traditional class identified by RCA. The two methods, however, deviate in their classification of blues and jazz, which RCA had categorized as “contemporary” rather than “traditional”. In contrast, CCA analyses instead suggest that blues belongs to the “traditional” side of the divide, while jazz, along with latin, straddles the two sides without clearly belonging to either. Since both blues and jazz were already well-established genres by the first decades of the 20th century (Henderson and Stacey 2014), the Contemporary – Traditional class identified by CCA fits better with the intuitive chronological understanding of those terms, and may thus have greater face validity.

The biggest apparent difference between the CCA and RCA analyses concerns the remaining group of respondents. RCA had placed the remaining respondents into a single class, which appeared to follow a traditional hierarchical logic with “highbrow” genres such as opera and classical music on the one extreme, and popular “lowbrow” genres on the other. In contrast, CCA separated the remaining population into two further schematic classes which appear to follow a different set of logics (see panels C and D of Figure 10). In both CCA-identified classes, the majority of genres are tied together in a dense cluster of positive correlations, thus suggesting that both are variants of an omnivorous taste schema. These results resemble previous findings that documented the existence of multiple distinct logics of omnivorousness (e.g., Tampubolon 2008).
However, the omnivorousness of respondents in these last two classes features clear exceptions. In the class depicted on the left, a higher appraisal of most genres is generally accompanied by a lower appraisal of heavy metal (and frequently also rap music.) The class on the right exhibits a nearly identical structure, except country and gospel music occupy the same position of exclusion as metal and rap did in the class on the left. These patterns closely echo the analyses of Bryson (1996), who famously showed that omnivores may retain a symbolic boundary against genres most closely associated with low education: heavy metal, rap, country and gospel music. Thus, drawing on the title of Bryson’s work, I term these latter two classes “Anything (but) Heavy Metal” and “Anything (but) Country”.

Table 3 cross-tabulates the group assignments made by the two algorithms. A plurality (though not a majority) of respondents that RCA grouped into its first two classes remain grouped together in the CCA results. The third class identified by RCA, however, does not have

Table 3: Cross-tabulation of Estimated Schematic Class Memberships in 1993 GSS Music Tastes Data

<table>
<thead>
<tr>
<th>RCA class</th>
<th>CCA class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Omnivore – Univore</td>
</tr>
<tr>
<td>Omnivore – Univore (673)</td>
<td>281</td>
</tr>
<tr>
<td>Contemporary – Traditional (394)</td>
<td>60</td>
</tr>
<tr>
<td>Highbrow – Lowbrow (461)</td>
<td>142</td>
</tr>
<tr>
<td>Total (N=1528)</td>
<td>483</td>
</tr>
</tbody>
</table>

Notes: Cross-tabulation of class memberships estimated by CCA (columns) and RCA (rows). RCA class sizes are indicated in parentheses. Four out of 1532 respondents had no response variance and were omitted from this analysis.

The parentheses around “but” in “Anything (but) Country” are meant to signify that the class could equally be named “Anything But Country” or “Anything Country”—i.e., that the logic opposes “any music that isn’t country” to “any music that is country.” The same logic applies to “Anything (but) Heavy Metal.”
an analogous relationship to any one of CCA’s classes. Such a divergence between CCA and RCA class assignments is consistent with the simulation analyses above, where the results of the two methods were correlated but rarely the same. While the simulations had demonstrated a reliable difference in accuracy between the two methods, these empirical analyses confirm that the two methods can point to substantively different conclusions.

**Multiple Groups Analysis**

Both RCA and CCA are theoretically understood as methods for partitioning the population into groups of respondents who follow different cultural schemas. However, from a more practical perspective, their task is to detect unobserved heterogeneity in how attitudes or tastes correlate for different groups. This heterogeneity is exemplified by the hypothetical situation where a pair of attitudes are correlated positively in one subgroup and negatively in another, leading them to appear uncorrelated in the overall sample (e.g., Baldassarri and Goldberg 2014:56). RCA and CCA reveal these latent logics of taste organization by inductively partitioning the population into groups where the same pairs of tastes correlate differently with one another. The sets of logics these methods locate are thus subsample correlation matrixes which were mixed together to yield the overall sample.

Their success at locating this kind of unobserved heterogeneity can be measured by how well these resulting subsample correlation matrixes describe the observed distribution of tastes in the population. A well-known multiple groups testing technique from structural equation modeling can provide such a measure (Bollen 1989; review in MacCallum and Austin 2000). Multiple groups analysis is widely used in psychology and public health to ascertain whether survey scales “have the same meaning across groups” (Gregorich 2006:S78), such as across age
cochains or education levels. Here, we are interested instead in ascertaining that the subgroups located by one of the methods show evidence of substantial differences in meaning.

To test the extent of this difference, I use SEM to evaluate whether using separate correlation matrixes for each subgroup meaningfully improves model fit over a single correlation matrix for the whole population, taking parsimony into account. The differences in AIC and BIC between the single-matrix and multiple-matrix models can be used to quantify the “statistical significance” of the heterogeneity detected by RCA or CCA. Roughly speaking, these indicators answer the question “are the groups located by these methods distinct enough in their logics of taste organization to make partitioning the population worth it?” Since neither AIC nor BIC require the models to be nested, the same approach can also be used to compare the fit of RCA and CCA partitions to one another.

I use SEM to fit three models to the musical tastes data: a homogeneity model where a single correlation matrix describes the whole population; a model with three correlation matrixes corresponding to RCA’s partition of the population; and a model with four correlation matrixes corresponding to CCA’s partition. The log likelihoods are -35376.85 (df=153) for the single-matrix model; -34054.2 (df=459) for the RCA partition model; and -33381.32 (df=612) for the CCA partition model. Comparing their model fit via AIC and BIC, I find that both the RCA (∆ AIC = -2033.3; ∆ BIC = -401.8) and CCA (∆ AIC = -3073.07; ∆ BIC = -625.81) partitions are strongly preferred to the single-matrix model. This is consistent with the existence of significant heterogeneity in the schemas that different subgroups use to organize their musical tastes. Both indicators also show that the CCA-based partition fits the data far better than the RCA-based partition (∆ AIC = -1039.76 and ∆ BIC = -224.01). The multiple groups analysis thus suggests that CCA offers a better description of the heterogeneity present in this data than RCA.
DISCUSSION

In this paper, I introduced Correlational Class Analysis (CCA), a technique that builds on Goldberg’s (2011) Relational Class Analysis (RCA) methodology. Both methods aim to partition a survey population into groups of respondents who arrange their tastes according to shared cultural schemas. RCA uses an eponymous relationality measure to quantify the extent that two respondents appear to use such a “shared schema”—a central concept that has lacked a clear definition. By formalizing and making explicit the intuitions about such schemas implicit in Goldberg’s work, I was able to substantially advance the clarity of this approach. Furthermore, my formalization showed that shared schemas should manifest themselves in linear dependencies between pairs of response vectors—the same measurement task for which Pearson’s correlation has long been an established solution. When I applied Pearson’s correlation to the same example that Goldberg (2011) used to introduce relationality, I found that correlation yielded substantially more accurate results. I thus proposed CCA as a correlation-based alternative to RCA.

I then used simulations to generalize and broaden the comparison between the two methods. In the first 10,000 simulations, I generated random taste schemas and then used random noisy linear transformations to simulate respondents following these schemas. I measured the ability of both methods to correctly determine which sets of respondents were produced from the same schema, and which from different ones. I found that the accuracy of correlation-based CCA remained reliably higher than that of relationality-based RCA across the full range of simulation parameters. I also demonstrated that RCA relies on a strong distributional assumption. When the two methods produced different results, the odds that CCA was more accurate than RCA were 17:1 in runs that obeyed RCA’s assumption, and rose to 23:1 in the runs that violated it.
Because of the theoretical uncertainty surrounding schematic cognition, and because Goldberg meant RCA to capture non-linear dependencies between tastes, I then investigated the performance of both methods when the model of schematic similarity as linearity was violated. I used simulations to examine six broad ways that schematic transformations could deviate from the theorized model. In the first three sets of simulations, I examined how the algorithms would perform if some basic linear operation (scaling, inversion or shift) did not take place. In the second three sets, I examined performance in the presence of independent subschemas, schematic transformations with polynomial functional forms, and multi-way interactions between genre preferences. Even though these simulations violated the assumptions of my model, CCA remained consistently more accurate than RCA throughout the six sets of simulations, indicating that relationality was not better than correlation at detecting any of these alternate patterns of schematic similarity. Overall, I found no evidence of any use cases where RCA would be preferable over CCA. On the contrary, the accuracy gap in favor of CCA remained stable throughout the broad range of simulations I examined. Such a gap is consistent with the one we would expect to observe if relationality captured largely the same relationships as correlation, but did so at a persistently lower accuracy.

I concluded with a re-analysis of the 1993 GSS musical tastes module previously analyzed by Goldberg (2011). While RCA had partitioned this population into three classes, CCA partitioned it into four. Two of the four classes resembled those found by RCA, while the other two did not. This confirmed that the two methods can yield substantively different conclusions in empirical settings. I then examined their partitions of the GSS population with a multiple groups analysis in SEM, which indicated that CCA’s results yielded far model better fit to the GSS data than did RCA’s.
**Limitations and Future Directions**

To highlight further avenues for methodological improvement, it is useful to locate RCA and CCA as two instances of a more theoretical and methodological framework of *schematic class analysis*, which further generalizes the approach introduced by Goldberg (2011). This framework begins with two theoretical propositions:

1) Latent *shared cultural schemas* specify relative evaluations of tastes vis-à-vis each other in a cultural domain.

2) Individuals’ taste patterns are manifestations of such schemas, created from them via one of a finite set of *schematic transformations*. The individuals who created their taste patterns from the same schema make up a *schematic class*.

Then, methodologically,

3) The extent to which two respondents $i$ and $j$ appear to follow the same latent schema can be measured via some *schematic similarity measure* $S_{ij}$, which yields a similarity matrix $S$ when applied to all pairs of respondents.

4) Finally, it is possible to partition the matrix $S$ into zones of greater and lesser similarity via some *partitioning method* $\Phi(S)$, thus yielding an estimate of schematic classes present in the data. These can then be used to estimate the contents of the shared cultural schemas.

In the present work, I showed that the intuitions behind the theory of shared cultural schemas imply that the schemas (1) can be captured as vectors of real numbers, and that the transformations (2) are linear in character. I therefore proposed to use Pearson’s correlation as schematic similarity measure $S_{ij}$. I further demonstrated that correlation-based CCA can yield informative results even when true schematic transformations depart from the linear form.
assumed in the model. Nonetheless, if and when future empirical investigations into the socio-
cognitive dynamics of cultural schemas cause us to update our theoretical model, a new measure
$S_{ij}$ specifically constructed to take this knowledge into account will likely be able to perform
better.

To maximize analytical clarity, correlation can be used as a starting point for many such
alterations. In the 128 years since Galton proposed an index of “co-relation” and 121 years since
Pearson introduced the $r$ coefficient that is still in use today, a substantial number of other
measures of correlation, such as Spearman’s $\rho$ or point-biserial correlation, have been developed
to capture this relationship under different special cases of data (Rodgers and Nicewander 1988).
Many of these can be used as “drop-in replacements” for Pearson’s correlation, thus making
CCA easily adaptable to situations for which specialized alternatives to correlation already exist.

Moreover, the simple algebraic form of Pearson’s $r$ makes it possible to derive “custom”
correlation coefficients for many other potential situations. For example, above, I examined three
“sub-linear” scenarios where the basic linear schematic transformations of inversion, scaling or
shift were theorized not to occur. In Appendix D, I demonstrate how Pearson’s $r$ can be modified
to yield three new indexes $r_-, r_\times$ and $r_+$, each of which makes the coefficient account for these
theoretical changes.

Datasets. Other avenues for methodological development lie in adapting schematic class analysis
to new types of data, such as the datasets from review websites like Yelp!, Epinions.com and
Amazon.com, which individuals use to publicly share their ratings of businesses, services or
products. Such data, which are private but regularly made available to researchers, greatly
exceed traditional survey data in scale and richness of detail, and thus provide important new
opportunities for students of culture. Since Pearson’s correlation is easy to compute efficiently and fast modularity maximization algorithms are available, scaling CCA to a large $N$ is straightforward. However, to make use of the full richness of these datasets, future methodological work may need to develop alternate schematic similarity measures that are better able to deal with such sparse data.

**Schematic classes.** While this paper has formally defined the concept of shared cultural schemas, the concept of schematic classes remains a second black box to be opened. How far can an individual’s responses deviate from a schema for him to still count as a member of a schematic class, and how similar do two schemas have to be to actually count as one and the same schema? The answers to these questions fall to the partitioning method $\Phi(S)$. Presently, CCA, like RCA, uses modularity maximization as its partitioning method. As the simulations in this paper demonstrate, it does its task with acceptable accuracy. But, while modularity maximization is among the most widely used network analytic techniques across many disciplines (e.g., Neal 2014; Shwed and Bearman 2010; Porter, Onnela, and Mucha 2009), its use may implicitly introduce undesired theoretical assumptions about the two questions I lay out above. A substantial literature notes that modularity maximization suffers from a “resolution limit” that biases it against detecting both very small and very large modules in many empirical settings, joining even very different modules together when they are too small in proportion to the whole network, and conversely breaking apart modules that are too large (Fortunato and Barthélemy 2007; Lancichinetti and Fortunato 2011; for a thorough overview, see Good, de Montjoye, and Clauset 2010).

---

21 A straightforward implementation of $|r(X, Y)|$ would iterate twice over the $N$ dataset rows: first to calculate the means of $X$ and $Y$, and second their covariance and variances. This yields a computational complexity of $O(N)$. 
By relying on modularity maximization, RCA and CCA may thus unintentionally introduce an assumption that each member of the survey population belongs to one of a moderate number of schematic classes, and may produce misleading results when this is not the case—e.g., in situations where many respondents follow their own idiosyncratic response patterns, or where the whole population belongs to only one schematic class. Future methodological work should develop a theoretically-informed partitioning method $\Phi(S)$ that avoids these biases\textsuperscript{22}. Before this is done, RCA and CCA results should be examined to verify that they indeed contain a significant amount of heterogeneity, much as the results of other methods are tested for statistical significance or goodness of fit. In the GSS case study above, I show how this can be accomplished via a multiple groups analysis in SEM.

CONCLUSION

In this paper, I demonstrated that Pearson’s correlation can be used to accurately measure schematic similarity between survey responses. Across the broad range of possible theoretical scenarios I examined, correlation-based CCA proved reliably more accurate at partitioning survey populations by shared cultural schema than relationality-based RCA. The switch from relationality to correlation brings a number of further benefits. Relationality is substantially more computationally costly to calculate than correlation, and also requires bias correction and extensive bootstrapping for significance testing. Correlation obviates the need for these further steps. This leaves an algorithm that is clear, fast and easy to implement (see Appendix A). It also clarifies and standardizes the method, thus placing it in conversation with existing methodologies in other disciplines—e.g., the correlation network approaches in bioinformatics which employ

\textsuperscript{22} For an example of a partitioning technique with a domain-specific fix to problems stemming from modularity’s resolution limit, see Sohn and colleagues (2011).
similar analytical steps in a different empirical domain. Future improvements of CCA can draw insights from these existing literatures, thus helping further build on Goldberg’s methodological innovations.
APPENDIX A. CCA Algorithm

Correlational class analysis can be implemented in minutes in any programming environment which supports network partitioning by modularity maximization. It consists of only four steps:

1. Create a matrix $G$ of absolute row correlations between survey respondents.
2. Set statistically insignificant correlations to 0 to reduce noise (e.g., using $t$-tests\(^\text{23}\)).
3. Import $G$ into a network analysis package, treating it as an adjacency matrix.
4. Use the existing network partitioning routines to produce the class assignments.

In the $R$ statistical environment with the $igraph 0.7$ library, this can be implemented as:

```r
CCA <- function (dataset, min.significant.row.cor = 0.60) {
  C <- abs(cor(t(dataset)))                   # 1st step
  C[C < min.significant.row.cor] <- 0         # 2nd step
  G <- graph.adjacency(C, mode="undirected",
                       weighted = TRUE, diag = FALSE)    # 3rd step
  leading.eigenvector.community(G)$membership # 4th step
}
```

A more full-featured implementation of the method is available on CRAN, and can be installed in $R$ with `install.packages("corclass")`.

See Appendix E for discussion of how to treat respondents with zero variance.

APPENDIX B. Simulating the Theorized Model

This appendix contains the simulation procedures for the first two sets of 5000 simulations I report in the paper.

Procedure 1—Initial linear simulation procedure (fixed inversion probability)

---

\(^{23}\) My exploratory results suggest that more stringent cutoffs may produce more accurate results as long as they are not so extreme as to turn some nodes into isolates. I used $\alpha = 0.05$ as the cutoff for the simulations reported above, and $\alpha = 0.01$ for the GSS analyses. A $\text{min.significant.row.cor}$ of 0.60 approximates a $t$-test at $\alpha = 0.01$ for rows of 17 variables.
Step 1 (parameters): I first randomly set the maximum ranges of various broad simulation parameters by drawing them from the uniform distribution: schema variance $v \sim U[0.3,3]$, noise variance $\varepsilon \sim U[0,3]$, maximum shift $\delta \sim U\{0, \ldots, 3\}$ maximum scaling $\gamma \sim U\{1, \ldots, 3\}$, and number of schematic classes $c \sim U\{2, \ldots, 6\}$.

Step 2 (schemas): Then, for each of the $c$ classes, I randomly generate a schema vector $\rho = [\rho_1, \ldots, \rho_{10}]$ by drawing from the Normal distribution, $\rho_i \sim N(\mu = 0, \sigma^2 = v)$, and rounding to the nearest integer. Any duplicate vectors are discarded, and new vectors generated in their place, until I have $c$ unique vectors. Then, I randomly set the counts $n_1, \ldots, n_c$ of respondents in each schematic class, $n_i \sim U\{100,101, \ldots, 500\}$.

Step 2b (range limits): The range of the 10 taste variables is then limited to $z_i = \pm[\max(|\rho_i|) \times \gamma + \delta] \cap \mathbb{Z}$ (this limit is enforced at the end of Step 3).

Step 3 (responses): Finally, for each respondent $f \in [1, n = n_p]$ following schema $\rho$, I generate the 10-element response vector $X_f = (k_f \times \rho) + \delta_f + \varepsilon_f$ by first drawing the values of the vertical shift $\delta_f \sim U\{-\delta, \ldots, \delta\}$ and the scaling and inversion factor $k_f \sim U\{[-\gamma, \ldots, -1] \cup [1, \ldots, \gamma]\}$. I then generate each respondent’s noise vector $\varepsilon_f$ by first determining $f$’s individual noise variance $E_f \sim U(0, \varepsilon)$, and then drawing each $i^{th}$ element $\varepsilon_{fi} \sim N(\mu = 0, \sigma^2 = E_f)$, $i \in [1, \ldots, 10]$, rounded to the nearest integer. If any $X_{fi} \notin z_i$, where $X_{fi}$ is the $i^{th}$ value of $X_f$, set it to the nearest value in $z_i$ to enforce the range of the variable.

Procedure 2—Full linear simulation procedure (Random inversion probability)

For the second set of 5000 simulations, the changes to the simulation procedure described above are as follows:
Step 1. I now also draw a random inversion probability: \( \zeta \sim U[0,0.5] \).

Step 3: I now draw a random inversion factor \( z_f \in \{1,-1\} \), with \( P(z_f = -1) = \zeta \). Since factor \( k_f \) now controls the scaling but not the inversion, I now restrict it to positive values:

\[
\begin{align*}
  k_f & \sim U[1,y].
  \end{align*}
\]
Each respondent \( f \) following schema \( \rho \) is generated by:

\[
X_f = (z_f \cdot k_f \cdot \rho) + \delta_f + \epsilon_f.
\]

**APPENDIX C. Simulation Procedures with Deviations from the Model**

This appendix contains the simulation procedures for the three “super-linear” simulations.

**Procedure 3: Partial independence**

I begin with Procedure 2 from Appendix B, add step 1b after step 1, and replace step 3:

Step 1b (subschemas): I set \( i = 0 \) and \( j = 1 \). Then, while \( j \leq 4 \) and \( i < 10 \), I draw a count of genres \( m_j \sim U[1,10 - i] \); assign genres \( \{i, i + m_j\} \) to subschema \( j \); increment \( i \) by \( m_j \); and increment \( j \) by 1. If at the end of this procedure \( j > 4 \) and \( i < 10 \), I assign genres \( \{i, 10\} \) to subschema \( j - 1 \), and decrement \( j \) by 1 so that \( j \) equals the number of different subschemas.

Step 3: I will use the notation \( \rho_{ij} \) to refer to the elements of schema \( i \) that were assigned to subschema \( j \). For each respondent \( f \in [1,n_\rho] \) following schema \( \rho \), I first draw an individual noise variance \( \varepsilon_f \sim U(0,\varepsilon) \). Then, for each subschema \( j \in [1,j] \), I simulate a partial response vector containing \( m_j \) cells, \( X_{fj} = (z_{fj} \cdot k_{fj} \cdot \rho_{pj}) + \delta_{fj} + \epsilon_{fj} \), by drawing the values of the subschema vertical shift \( \delta_{fj} \sim U[-\delta,...,\delta] \) the subschema scaling factor \( k_{fj} \sim U[1,y] \), and the subschema inversion factor \( z_{fj} \in \{1,-1\} \) with \( P(z_{fj} = -1) = \zeta \). I then draw each element of the subschema noise vector \( \epsilon_{fji} \sim N(\mu = 0,\sigma^2 = \epsilon_f) \), \( i \in [1,...,m_j] \), rounded to the nearest integer. At the end of this step, I produce the respondent’s vector \( X_f \) by concatenating together these \( j \) partial response vectors in the order they were generated.
**Procedure 4: Polynomial functional form**

I begin with Procedure 1 from Appendix B, omit step 2b, alter step 1, and replace step 3:

**Step 1:** I begin as before, though I no longer draw \( \delta \) or \( \gamma \). Instead, to determine the general form of the polynomial for this simulation, I first draw the count of polynomial terms \( d \sim U\{1, ..., 10\} \).

If \( d \geq 2 \), I then draw each exponent \( D_i, i \in \{1, ..., d\} \), from the set \( \{0, ..., 10\} \), by sampling without replacement. If \( d = 1 \), I instead draw the sole exponent \( D_1 \sim U\{1, ..., 10\} \). I then sort \( D \) in ascending order. Then, for each term in \( D \), I generate a maximum scaling factor \( \gamma_i \sim U[0,1] \).

**Step 3:** For each respondent \( f \in [1,n_f] \) following schema \( \rho \), I first I draw an exponent inversion factor \( z_{fi} \sim U\{1, -1\} \), and an exponent scaling factor \( k_{fi} \sim U[1, \gamma_i] \) for each exponent \( D_i \). I then generate the vector

\[
X'_f = \mathbb{I}(D_1 = 0) * z_{f0} * k_{f0} + \sum_{i:d_i \neq 0} [z_{fi} * (k_{fi} * \rho)^{D_i}],
\]

where \( \mathbb{I} \) is the indicator function. Since this procedure is prone to generate extremely wide ranges of variables that are not realistic for survey settings, before adding noise, I set all the values of \( X'_f \) that are below the 10\(^{th}\) percentile or above the 90\(^{th}\) percentile to equal the value at that percentile. Then I take the modified vector \( X'_f \), linearly translate it to the [1,10] range, and round each element to the nearest integer. I then generate each element of their noise vector \( \epsilon_{fi} \sim N(\mu = 0, \sigma^2 = \epsilon) \) for \( i \in [1, ..., 10] \), rounding it to the nearest integer, and finally produce \( f \)'s response vector \( X_f = X'_f + \epsilon_f \).

**Procedure 5: Multi-way interactions**

I again begin with Procedure 1. I then remove step 2b, alter steps 1 and 2, and replace 3:

**Step 1:** I begin as before, though I omit drawing \( \delta \) and \( \gamma \). I instead draw the maximum number of terms per interaction \( d \sim U\{1, ..., 4\} \), and the focal term weight \( D_1 \sim U\{1, ..., 4\} \).
Step 2: At the end of step 2, I now linearly translate all the schemas to the range \([1, \max(|\rho_i|) - \min(|\rho_i|) + 1]\).

Step 3: For each respondent \(f \in [1, n_p]\) following schema \(\rho\), I first draw an individual noise variance \(E_f \sim U(0, \epsilon)\). To determine which of \(f\)'s tastes interact with each other (\(f\)'s “interaction groups”), I first set \(k = 0\) and \(l = 1\). Then, while \(k < 10\), I draw a count of genres \(m \sim U[1, \min(10 - k, d)]\): assign genres \((k, k + m)\) to interaction group \(l\); increment \(k\) by \(m\); and increment \(l\) by 1. To generate each \(i\)th taste \(X_{fl}\) belonging to respondent \(f\), I calculate a weighted geometric mean of tastes in the same interaction group. To do this, I begin with the \(i\)th element of \(\rho\), exponentiate it to power \(D\), multiply it with every \(j\)th element of \(\rho\) belonging to the same interaction group \(m\) as \(i\), and finds the \(k\)th root of the product, where \(k\) equals \(D - 1 + |m|\), and \(|m|\) is the number of elements in \(m\). I then generate \(X_{fl}\) by taking the result of this procedure and adding random noise \(\epsilon_{fl} \sim N(\mu = 0, \sigma^2 = E_f)\), rounded to the nearest integer.

APPENDIX D: Theory-Driven Changes to Pearson’s Correlation Coefficient

Consider the absolute value of Pearson’s correlation coefficient \(|r(X, Y)| = \left| \frac{Cov(X, Y)}{\sqrt{Var[X] \cdot Var[Y]}} \right| = \left| \frac{E[(X - \bar{X})(Y - \bar{Y})]}{\sqrt{E[(X - \bar{X})^2] \cdot \sqrt{E[(Y - \bar{Y})^2]}}} \right|

Different components of this formula make the coefficient invariant to inversion, scaling and shift in the vector. Each of the three “sub-linear” scenarios I described earlier can be specifically accommodated by altering the relevant component (and rescaling the result to the \([0,1]\) range if
needed). While I leave a fuller methodological treatment of this topic for future work, I derive some basic formulas below as an example of this approach.

No inversion. Most obviously, $|r(X, Y)|$ is invariant to inversion because of the absolute value operator. If inversion is to be interpreted as maximum schematic difference rather than schematic similarity, i.e., $r_-(X, Y) = 1 \rightarrow r_-(X, -Y) = 0$, the absolute value operator can simply be removed, with the resulting formula shifted and rescaled to $[0,1]:$

$$r_-(X, Y) = 0.5 \times \left( \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]} \sqrt{\text{Var}[Y]}} + 1 \right)$$

No scaling. To create a version of the coefficient that is sensitive to scaling, it is useful to note that correlation between a variable $X$ and its multiple $Y = kX$ equals 1 because both the numerator and denominator of $r(X, Y)$ scale along with $k$:

$$r(X, kX) = \frac{\text{Cov}(X, kX)}{\sqrt{\text{Var}[X]} \sqrt{\text{Var}[kX]}} = \frac{k \cdot \text{Var}[X]}{\sqrt{\text{Var}[X]} \sqrt{k^2 \text{Var}[X]}} = \frac{k}{k} = 1$$

It is possible to transform this mechanism into one that penalizes differences in scaling in proportion to the multiplier $k$:

$$r_\times(X, Y) = \left| \frac{\text{Cov}(X, Y)}{\max(\text{Var}[X], \text{Var}[Y])} \right|$$

If the variances of $X$ and $Y$ are equal, $r_\times(X, Y) = |r(X, Y)|$. However, if $\text{Var}[X] > \text{Var}[Y]$,

$$r_\times(X, Y) = \left| \frac{\text{Cov}(X, Y)}{\text{Var}[X]} \right| = |r(X, Y)| \times \frac{\sqrt{\text{Var}[Y]}}{\sqrt{\text{Var}[X]}} = |r(X, Y)| \times \frac{\sigma_Y}{\sigma_X}.$$ Thus, as desired, $r_\times(X, kX) = \frac{|1|}{|k|}$ if $|k| > 1$, and $|k|$ otherwise. More broadly, for any $X \neq Y$, $r_\times(X, Y)$ will penalize their correlation in proportion to the ratio of their standard deviations.
No shift. Finally, it is possible to modify the correlation coefficient to penalize vertical shifts by replacing the variances in the denominator with the variables’ second moments around the grand mean $\bar{M} = 0.5(\bar{X} + \bar{Y})$, yielding

$$r_+(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{E}[(X - \bar{M})^2] \ast \sqrt{\text{E}[(Y - \bar{M})^2]}}}$$

Since $\text{Var}[X] = \text{E}[(X - \bar{X})^2] < \text{E}[(X - \bar{M})^2], \forall \bar{M} \neq \bar{X} \in R$, this quantity will equal $|r(X, Y)|$ only if the two variables have the same mean, and will otherwise penalize it towards 0.

**APPENDIX E: Zero-variance responses**

Since correlation is normalized by the product of variances, it is undefined when the variance of either respondent is at absolute zero. The minimal implementation in Appendix A requires that such respondents be dropped from the analysis. Since zero-variance respondents are relatively rare in empirical survey data (e.g., out of the 1532 respondents in the GSS analyses, there was a total of only 4 with zero variance), dropping them is likely the most pragmatic solution.

The theory of cultural schemas does not appear to suggest any clear way of dealing with zero-variance respondents. Since such respondents express the same attitude towards all musical styles, their tastes literally contain no distinctions between any pair of genres. This lack of cultural judgment means that they are, in a sense, sitting out the game of distinction altogether. Thus, depending on theoretical context, an “undefined” schematic class membership may be justified. On the other hand, this lack of distinctions could be interpreted as a kind of “null schema” that specifies no contrasts between any genres. Following this logic, correlations between two zero-variance respondents could be set to 1, and their correlations with others to 0. To keep CCA’s accuracy directly comparable to RCA’s, the algorithm I use to analyze the simulated datasets took this latter approach.
In other theoretical settings, other solutions to the classification problem may be preferable. However, as long as multiplicative scaling \( Y = kX \) is one of the theorized schematic transformations, many intuitively appealing solutions may prove theoretically problematic. For example, if zero-variance responses were treated as extreme cases of a low-variance schema, it may be tempting to assign them to the schematic class containing respondents with the lowest average estimated variance. But variance, unlike correlation, is not invariant to multiplication: if \( Y = kX \), \( \text{Var}[Y] = k^2 \cdot \text{Var}[X] \). Thus, with multiplicative scaling, any non-null schema could produce both low-variance respondents when the multiplier is small, and higher-variance respondents when it is large. This means that any normal schematic class could contain a mixture of respondents with a range of different variances, making “amount of variance” into a potentially problematic class assignment criteria.
REFERENCES


