Mode effects in cultural domain analysis: comparing pile sort data collected via internet versus face-to-face interviews

Clarence C. Gravlee, Chad R. Maxwell, Aryeh Jacobsohn & H. Russell Bernard

To cite this article: Clarence C. Gravlee, Chad R. Maxwell, Aryeh Jacobsohn & H. Russell Bernard (2017): Mode effects in cultural domain analysis: comparing pile sort data collected via internet versus face-to-face interviews, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2017.1341187

To link to this article: http://dx.doi.org/10.1080/13645579.2017.1341187

Published online: 25 Jun 2017.

Article views: 33
Mode effects in cultural domain analysis: comparing pile sort data collected via internet versus face-to-face interviews

Clarence C. Gravleea, Chad R. Maxwella, Aryeh Jacobsohna and H. Russell Bernarda,b

aDepartment of Anthropology, University of Florida, Gainesville, FL, USA; bInstitute for Social Science Research, Arizona State University, Tempe, AZ, USA

ABSTRACT
This article tests whether collecting pile-sort data online produces results similar to those obtained with face-to-face methods. We collected pile sorts from 227 university students in the cultural domain of emotions. To test for mode and design effects, we randomly assigned participants to face-to-face or internet modes and to either a 15- or 30-item sort. We found no evidence of mode effects in semantic structure, but the level of agreement among respondents varied by mode, depending on the number of items in the task. Agreement was higher among online respondents in the 15-item task but among face-to-face respondents in the 30-item one. Because the central purpose of pile sorting is to elicit semantic structure, we conclude that online data collection is a viable option for most researchers, but the mode effect in level of agreement implies that other design elements may affect comparability of face-to-face and web-based methods.

Social scientists now collect many types of data online – a trend that is only likely to grow. The development of web-based alternatives to traditional face-to-face methods raises the question of mode effects: that is, whether collecting similar types of data in different ways produces comparable results. The literature on mode effects is extensive for survey research (Jäckle, Roberts, & Lynn, 2010) and growing for other types of data collection, including psychological experiments (Antoun, Zhang, Conrad, & Schober, 2016), qualitative interviewing (Hanna, 2012; Vogl, 2013), and free listing (Gravlee et al., 2013). In this article, we extend the mode effects literature to the collection of pile sort data, a staple of cultural domain analysis.

Cultural domain analysis is an approach derived from cognitive anthropology to describe the contents, structure, and distribution of knowledge in organized spheres of experience, or cultural domains (D’Andrade, 1995). Culture, in this approach, is understood as a series of dynamic, interlinked cognitive models that provide a schematic outline of how the world works in various domains (Ross, 2004). Cultural models in the domain of ‘illness,’ for example, encode knowledge about the kinds of sickness one may acquire; how those illnesses are related to one another; and which signs, symptoms, and treatments are associated with each illness. Cultural models of illness, in turn, are linked to models in closely related cultural domains, such as ‘health care providers.’

Cognitive anthropologists have developed a suite of structured ethnographic methods to study cultural domains (Weller & Romney, 1988). These methods are tailored to one of three questions we
can ask about any cultural domain: (1) what are the contents and boundaries of the domain? (2) how are the relationships among items in the domain organized in terms of semantic structure? and (3) how is knowledge about the domain shared and distributed within a social group? Pile sorting is the most widely used technique to answer the second question and can also inform the third.

In its most common form, the free or unconstrained pile sort, items from a domain are written on index cards or slips of paper and respondents are asked to sort them into groups that belong together (Weller & Romney, 1988). The resulting piles are converted into a proximity matrix – one for each respondent – in which the cells indicate whether any two items were placed in the same pile. These individual proximity matrices are then aggregated across respondents and analyzed using exploratory visualization methods such as multidimensional scaling (MDS) and hierarchical cluster analysis to detect underlying dimensions of semantic structure (Burton & Romney, 1975). The purpose of this method is to elicit perceptions of the similarities among items in a cultural domain. It is not to be confused with the Q-sort technique (Stephenson, 1936), a scaling method in which respondents are asked to rank-order items along some dimension.

Pile sorting has a long history across the social sciences (Rosenberg, Nelson, & Vivekananthan, 1968) and is widely used today. Pile sorts have been used, for example, to examine conceptual models of jealousy (Sobraske, Boster, & Gaulin, 2013), motivations for having sex among low-income African American women (Deardorff et al., 2013), hunters’ prey choices in Nicaragua (Koster, Hodgen, Venegas, & Copeland, 2010), gender norms in Ethiopia and Kenya (Bourey, Stephenson, Bartel, & Rubardt, 2012), barriers to cancer screening in American Indian communities (Yeh et al., 2013), conceptions of HIV/AIDS in Bolivia and Chile (Torres López, Reynaldos Quinteros, Lozano González, & Munguía Cortés, 2010), food choices among Mexican working mothers (Rodríguez-Oliveros, Bisogni, & Frongillo, 2014), reasons for farmers’ suicides in India (Dongre & Deshmukh, 2012), ethnic classification in Puerto Rico (Gravlee, 2005), and sacred places and gods in Japan (Roberts, Morita, & Brown, 1986).

Online tools for collecting pile sorts have emerged in response to demand from information architecture and user-design research, where the technique is known as card sorting (Righi et al., 2013). Several commercial providers cater to this market (e.g. OptimalSort, CardZort Zone, Socratic CardSort). The potential benefits of collecting pile sorts online, however, apply across the social sciences for the same reasons that web-based surveys have become commonplace: reduced cost per respondent, potential for automation, reduced data entry, ability to incorporate multimedia, and increased access to geographically dispersed respondents (Couper, 2005).

Despite the growth of online pile sorting, only two previous studies have compared web- vs. paper-based collection of pile sorts. Bussolon, Russi, and Missier (2006) find high concordance between web- and paper-based methods, whereas Greve (2014) reported significant differences between modes. Our study adds new evidence about mode effects that may help explain this discrepancy and provides guidance on further development of web-based tools for collecting pile sorts and related kinds of data online.

**Methods**

**Study design**

We tested for mode effects in a nonprobability sample of university students in the United States. Students were assigned to one of two data collection modes (face-to-face interview or web-based survey) and to either a short (15-item) or long (30-item) pile sort task. The purpose of including short and long versions was to test whether the presence and magnitude of mode effects depends on the complexity of the task. We hypothesized that potential limitations of online data collection (e.g. screen size) may become more salient as the number of items increases.

We chose the cultural domain of emotions because it has been fertile ground for theoretical and methodological developments in cognitive anthropology and related fields (D'Andrade, 1981; Lutz,
INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY

To facilitate comparison with previous work, we selected 15 emotion terms that Rusch (1996) elicited from speakers of English and Japanese. Subsequent studies used this same set of emotions to examine cross-cultural variation in semantic structure among English-, Japanese-, and Chinese-speaking students (Moore et al., 1999; Romney, 2000; Romney et al., 1997). The full list of emotion terms, including 15 we added for the longer task, is given in Table 1.

For the short pile sort task, respondents were randomly assigned to complete either a ranking (with paired comparisons) or a rating task, in addition to the pile sort. All respondents assigned to the longer pile sort task subsequently completed the ratings exercise; because paired comparisons have a relatively high respondent burden, we did not combine them with the longer pile sort task. We plan to compare the ranking and rating data in a later paper. Here we note only that the pile sort data were collected prior to the rankings and ratings, such that the pile sort data are not biased by which of the subsequent tasks respondents completed. We present data separately for the subsamples of 15-item pile sorts to illustrate variation attributable to sampling error rather than mode.

Sampling

For each study condition, we selected a quota sample of university students with roughly equal numbers of men and women and of people who identified themselves as Black, White, or Hispanic. Based on theory (Weller, 2007) and prior empirical applications (Romney et al., 1997), we aimed to recruit 8–10 respondents for each of the six race-gender categories (e.g. Hispanic women), for a total of 48–60 respondents per mode and length of task. In the end, we obtained a total of 227 pile sorts (see Table 2). Although subsample sizes for the 15-item pile sorts fell short of our initial quotas, they are still consistent with standard recommendations in cultural domain analysis (Weller & Romney, 1988).

Interviewers recruited respondents from a variety of public places on campus where we expected to maximize heterogeneity among students (e.g. business school vs. liberal arts and sciences, student union, select student organizations, classrooms). All respondents were recruited face-to-face and randomly assigned to a study condition (mode, length of pile sort, and either paired comparisons

---

**Table 1. List of emotion terms used in 15-item and 30-item pile sort tasks.**

<table>
<thead>
<tr>
<th>First 15</th>
<th>Second 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>joy</td>
</tr>
<tr>
<td>anguish</td>
<td>jealousy</td>
</tr>
<tr>
<td>anxious</td>
<td>frustration</td>
</tr>
<tr>
<td>bored</td>
<td>confidence</td>
</tr>
<tr>
<td>disgust</td>
<td>grief</td>
</tr>
<tr>
<td>envy</td>
<td>pride</td>
</tr>
<tr>
<td>excitement</td>
<td>disappointment</td>
</tr>
<tr>
<td>fear</td>
<td>surprise</td>
</tr>
<tr>
<td>happy</td>
<td>apathy</td>
</tr>
<tr>
<td>hate</td>
<td>compassion</td>
</tr>
<tr>
<td>lonely</td>
<td>guilt</td>
</tr>
<tr>
<td>love</td>
<td>lust</td>
</tr>
<tr>
<td>sad</td>
<td>mad</td>
</tr>
<tr>
<td>shame</td>
<td>doubt</td>
</tr>
<tr>
<td>tired</td>
<td>depression</td>
</tr>
</tbody>
</table>

**Table 2. Subsample sizes for six study conditions testing mode and design effects in collection of pile sort data.**

<table>
<thead>
<tr>
<th></th>
<th>Web</th>
<th>Face-to-face</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-item pile sort + ratings</td>
<td>34</td>
<td>27</td>
<td>61</td>
</tr>
<tr>
<td>15-item pile sort + paired comparisons</td>
<td>33</td>
<td>24</td>
<td>57</td>
</tr>
<tr>
<td>30-item pile sort + ratings</td>
<td>56</td>
<td>53</td>
<td>109</td>
</tr>
</tbody>
</table>

1982; Moore, Romney, Hsia, & Rusch, 1999; Romney, 2000; Romney, Moore, & Rusch, 1997; Schrauf & Sanchez, 2008). To facilitate comparison with previous work, we selected 15 emotion terms that Rusch (1996) elicited from speakers of English and Japanese. Subsequent studies used this same set of emotions to examine cross-cultural variation in semantic structure among English-, Japanese-, and Chinese-speaking students (Moore et al., 1999; Romney, 2000; Romney et al., 1997). The full list of emotion terms, including 15 we added for the longer task, is given in Table 1.
or ratings). It was challenging to get people to respond to the web-based survey. We had hoped for a viral effect, since people could take the survey on their own time and recruit others, but our reliance on face-to-face recruiting counteracted this effect. Instead, it proved to be more efficient simply to ask people to complete the task on the spot.

**Procedure**

For face-to-face interviews, we followed the conventional practice of printing the name of each emotion on 3×5” index cards. We prepared decks of both 15 and 30 items. Interviewers shuffled the appropriate deck and asked respondents to place the cards into piles they thought went together. We did not limit the number of piles respondents could create, except that we ruled out placing all items in one pile or every item in its own pile. After respondents finished sorting the cards, we asked them to label each pile and tell us how the items in it went together.

We built the online pile sort task using the open-source Educara SURVEY software (freely available at https://sourceforge.net/projects/educarasurvey/). Educara SURVEY supports web-based data collection using structured ethnographic methods commonly used in cultural domain analysis, including free lists, pile sorts, triad tests with lambda designs, frame substitution, and paired comparisons with lambda designs (Weller & Romney, 1988). Data collected with Educara SURVEY are exported as Microsoft Excel® or XML files. English and Spanish localizations of the software are available, and the built-in templates for the user interface can be customized.

For online pile sorts, Educara SURVEY presents respondents with virtual representations of cards, each with the name of an item from the domain (in our case, emotions) on it. Respondents move the cards using the mouse into piles of ‘things that go together.’ After completing the pile sort, respondents see a list of the piles they made and are given an opportunity to make changes. Once satisfied, they are then asked to name each pile. Screenshots of the online pile sort we created in Educara SURVEY are available as supplementary figures.

**Analysis**

We imported pile sort data from online and face-to-face interviews into ANTHROPAC software (Borgatti, 1996). For each study combination of mode, length of list, and ratings or paired comparison, ANTHROPAC created individual and aggregate proximity matrices of the similarities among emotions. We visualized the six resulting aggregate proximity matrices using non-metric multidimensional scaling (MDS; Romney, Shepard, & Nerlove, 1972) to compare semantic structure across study conditions. For a formal test of mode effects, we used the quadratic assignment procedure (QAP) to measure pairwise correlations among the proximity matrices resulting from each study condition (Borgatti, 2002). In both the exploratory visualizations and the formal test, we included data provided by Moore et al. (1999) to enable comparisons with previous work.

Next, we tested whether paper- and web-based pile sorts produced similar evidence for the level of agreement and distribution of cultural knowledge about emotions. Following Handwerker (2002), we first appended the individual proximity matrices for the 118 15-item pile sorts and, separately, the 109 30-item pile sorts. This step resulted in a 118-row by 105-column (15 × 14 ÷ 2) matrix for the 15-item task and a 109-row by 435-column (30 × 29 ÷ 2) matrix for the 30-item task. We then factor analyzed the rows of each matrix to test whether respondents drew on a single, shared cultural model of emotions and to estimate respondents’ knowledge of the shared model. Based on cultural consensus theory (Romney, Batchelder, & Weller, 1987), we infer that there is a single, shared model if there is a one-factor solution to the similarities among respondents. If so, the first factor is taken to represent the shared component of cultural knowledge among respondents; respondents’ loadings on the first factor, then, provide estimates of how well each individual knows the shared cultural model. The second factor reflects intracultural variation in knowledge about the domain, often referred to as residual agreement (Boster & Johnson, 1989; Dressler, Balieiro, & dos Santos, 2015).
Unconstrained pile sort data violate the strict assumptions of formal consensus analysis because respondents are not required to make the same number of piles (Weller, 2007, p. 353). However, factor analysis of respondent similarities is still useful for our purposes because it allows us to estimate whether mode and number of items are associated with patterns of agreement among respondents (Handwerker, 2002). We formally test for mode effects in the estimated level of individual cultural knowledge using a two-way factorial ANOVA with mode and the number of items as predictors of knowledge scores.

Results

Figure 1 shows MDS plots for the six study conditions. Because the axes are arbitrary, we have reversed the direction in some cases to facilitate comparison across study conditions. Panels (a) and (b) compare results for face-to-face vs. web-based pile sorts with 15 items and paired comparisons; (c) and (d) do the same for 15 items and ratings. Panels (e) and (f) compare results from the face-to-face vs. web-based versions, respectively, of the 30-item task. A comparable plot for data from Moore et al. (1999) is available in the Supplementary Material.

Overall, the structure of the plots is similar. In all four of the 15-item plots (Figure 1(a)–(d)), three items (happy, love, and excitement) appear in a single cluster apart from the rest, and the distribution of the remaining 12 items is also similar: Sad, lonely, bored, and tired form a cluster in all four plots, as do hate, envy, anger, disgust, anguish, shame, and fear. All six plots – including both 15- and 30-item tasks – suggest similar dimensions of meaning. The horizontal dimension appears to range from unpleasant emotions (e.g. sad, shame, grief, depression) to pleasant ones (e.g. joy, happy, love, excitement). The vertical dimension arrays emotions directed toward others (e.g. anger, envy, love) vs. emotions that one can experience alone (e.g. bored, tired, apathy). Other substantive interpretations of the dimensions are possible. The important point, for our purposes, is that all six plots suggest a similar interpretation. Data from Moore et al. (1999) exhibit a similar pattern (see Supplementary Material).

Table 3 summarizes formal tests of similarity across study conditions using QAP to compare judged similarities among the six pairs of 15-item pile sort tasks and the single pair of 30-item tasks. Correlations across study conditions are high, ranging from .902 to .963. Table 3 provides no evidence of a mode effect. Indeed, the two highest correlations we observe (.963 and .940) are between rather than within modes, suggesting that sampling error is greater than mode effects.

The bottom row of Table 3 shows that the aggregate, judged similarity of emotion terms in all four 15-item pile sort tests also correlates strongly with data from Moore and colleagues – even though they collected data more than 10 years earlier using a different data collection method (triad tests vs. pile sorts) in a different part of the country (Southern California vs. Florida) with native speakers of three different languages (English, Japanese, and Chinese vs. English-only). Given these design differences, it is not surprising that correlations between studies are slightly lower than are correlations among conditions within our study. Still, the strong correlations between studies attests to the stable semantic structure of the domain.

Figure 2 examines the pattern of agreement among respondents, by mode, for 15-item and 30-item pile sort tasks. The overall pattern reflects a moderate level consensus – both among respondents and between modes. For both modes, there appears to be a one-factor solution, indicated by the roughly 4:1 ratio of the eigenvalues for factors 1 and 2 and the greater than 80% of variance explained by the first factor in both analyses. Drawing on cultural consensus theory (Romney et al., 1987), we interpret the first factor as evidence of a shared cultural model of the emotions domain. Despite this evidence for consensus, the average factor loadings for the first factor are only moderately high (.55 ± .13 for the 15-item task and .46 ± .13 for the 30-item task). When we analyze the four subsamples separately, the variance explained by the second factor ranges from about 13% to about 23%. This range of 13–23% appears to be a sampling effect, not a mode effect, because we see about the same difference in variance explained by the second factor for the 51 face-to-face pile sorts matched with either paired comparisons
Figure 1. Multidimensional scaling (MDS) of pile sort data for the cultural domain of emotions in six study conditions.

Table 3. Pairwise correlations among aggregate proximity matrices for six study conditions plus Moore et al. (1999), quadratic assignment procedure (QAP).

<table>
<thead>
<tr>
<th></th>
<th>F2F 15 + PC</th>
<th>F2F 15 + R</th>
<th>Web 15 + PC</th>
<th>Web 15 + R</th>
<th>Web 30 + R</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2F 15 + PC</td>
<td>.922</td>
<td>.902</td>
<td>.940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2F 15 + R</td>
<td></td>
<td>.918</td>
<td>.963</td>
<td>.913</td>
<td></td>
</tr>
<tr>
<td>Web 15 + PC</td>
<td></td>
<td></td>
<td>.848</td>
<td>.904</td>
<td>.913</td>
</tr>
<tr>
<td>Web 15 + R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2F 30 + R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.913</td>
</tr>
<tr>
<td>Moore et al. (1999)</td>
<td>.904</td>
<td>.891</td>
<td>.848</td>
<td>.904</td>
<td></td>
</tr>
</tbody>
</table>
or ratings. The relatively large proportion of variance explained by the second factor may be because unconstrained pile sort data do not strictly conform to assumptions of cultural consensus analysis.

**Figure 2.** Cultural consensus and intracultural variation, as assessed by respondent loadings on factor analysis of similarities in pile sort responses, by number of items in pile sort task.

**Table 4.** Two-way analysis of variance of cultural knowledge by mode of data collection (online vs. face-to-face) and number of items (15 vs. 30).

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>.0004</td>
<td>1</td>
<td>.0004</td>
<td>.03</td>
<td>.8683</td>
</tr>
<tr>
<td>Number of items</td>
<td>.3591</td>
<td>1</td>
<td>.3591</td>
<td>23.34</td>
<td>.0000</td>
</tr>
<tr>
<td>Mode × Number of items</td>
<td>.2004</td>
<td>1</td>
<td>.2004</td>
<td>13.03</td>
<td>.0004</td>
</tr>
<tr>
<td>Error</td>
<td>3.4311</td>
<td>223</td>
<td>.0154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4.0423</td>
<td>226</td>
<td>.0179</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $N = 227$, $R^2 = .151$.

**Figure 3.** Interaction effect between mode of data collection and number of items in pile sort task, 95% confidence intervals ($N = 227$).
Panels (a) and (b) of Figure 2 suggest slightly different patterns in the distribution of cultural knowledge, by modes, for the 15-item and 30-item tasks. For both plots, the most important signal is the overlap among respondents. Yet for the 15-item task, higher factor loadings tend to be from web-based respondents, whereas for the 30-item task higher factor loadings tend to come from face-to-face respondents (see Supplementary Material for a further illustration of this trend). Table 4 tests whether this apparent difference is real. The ANOVA confirms a small but statistically significant interaction between mode and number of items in relation to the level of agreement among respondents, as illustrated in Figure 3. Holding the number of items constant, face-to-face and web-based pile sorts provide statistically distinguishable estimates of cultural knowledge, but the direction of the difference depends on the number of items. The change in direction is due to the fact that the level of consensus (mean cultural knowledge) in web-based pile sorts is substantially lower in the 30-item pile sort, as compared to the 15-item one, $F(1, 223) = 38.71, p < .001$. In face-to-face pile sorts, by contrast, the number of items is not associated with differences in the level of agreement among respondents, $F(1, 223) = .69, p = .407$.

**Discussion**

The goal of this study was to determine whether web-based collection of pile sort data produced results that are comparable to traditional, face-to-face methods. The answer is yes and no. We found no evidence that collecting data online changes our understanding of the semantic structure in a consensual cultural domain like *emotions*. Because the central purpose of collecting pile sorts is to elicit semantic structure, our findings suggest that online data collection is a viable option for most researchers. Modes did vary, however, in estimates of cultural consensus, as measured by the level of agreement among respondents. The small but significant mode effect in agreement among respondents warrants caution and implies that other design elements, including the number of items in a pile-sort task, may affect comparability of face-to-face and web-based methods.

Web-based data collection has increasing appeal across the social sciences (Couper, 2005). Writing about pile sorting in particular, Ford (2013) noted three potential advantages of online methods. First, collecting pile sorts via the internet eliminates data entry, which potentially reduces errors and saves time, effort, and money. Second, web-based pile sorts may be less threatening to respondents and therefore improve data quality because respondents are able to participate in the study on their own computer, in a comfortable place, and at a time of their choosing. Third, collecting pile sorts online increases access to a broader range of potential participants because the task is remote and asynchronous. It becomes unnecessary to travel to respondents or to coordinate schedules between respondents and fieldwork staff. Ethnographers could leave the study in the hands of local assistants or collaborators who can continue to collect data after the ethnographer leaves the field.

Yet online data collection also has disadvantages. First, as Ford (2013) observed, collecting pile sorts via the web may entail the loss of qualitative data that can be obtained more easily in face-to-face studies by asking respondents to think aloud or to explain their choices about how to group items. Such data facilitate interpretation of output from visualization methods like MDS or cluster analysis.

Second, screen size may impose greater limits, as compared to paper methods, on the number of items that could be included in a pile sort task. This limitation is important, as one advantage of paper-based pile sorting, as compared to other methods for eliciting similarity data (e.g. triad tests), is that respondents can generally sort dozens of items. Researchers commonly include 30–60 items in pile sort studies, and Weller and Romney (1988) note that traditional pile sorting is possible with more than 100 items. Indeed, Killworth and Bernard (1974) included 150 items in pile sorts to elicit cognitive social networks from prison inmates and staff.

Last, web-based data collection raises questions about the potential effects of the user interface and aesthetic design of online pile sorts. We know that seemingly subtle differences in the design of web-based surveys can alter results (Christian & Dillman, 2004; Diaz de Rada & Domínguez, 2014; Maeda, 2013). Online pile sorts may be even more susceptible to such design effects because respondents
are asked to interact with and manipulate digital representations of the cards. How these cards are
designed (e.g. font size, length of text included on each card, ease of grouping cards) may have subtle
or substantial effects on respondents’ choices.

Given the possible pitfalls, it is striking that we obtained similar evidence from web- and paper-
based pile sorts regarding the semantic structure of emotions. Both visual inspection of MDS plots
and a formal test of aggregate similarities across study conditions suggests that, for this domain, the
results are effectively the same no matter how we constructed the pile sort task. What remains for future
research is to determine whether characteristics of the cultural domain contribute to the presence
or absence or mode effects. We deliberately chose a domain with a well-known, stable, and possibly
universal semantic structure (Romney et al., 1997). This choice was appropriate for an initial test of
mode effects because it enabled direct comparisons with previous research. We cannot rule out the
possibility of mode effects, however, in cultural domains with less stable or consensual structures.

Even in the consensual domain of emotions, we find evidence that online and face-to-face pile
sorts produce different estimates of cultural consensus, as measured by the level of agreement among
respondents. The mode effect is small but statistically significant and depends on the number of items
in the pile sort task. For pile sorts with only 15 items, mean cultural knowledge is highest among web-
based respondents, but with 30 items it is highest among face-to-face respondents. This interaction
can be attributed to a significant difference within web-based pile sorts for 15- vs. 30-item tasks; for
paper-based pile sorts, the number of items made no difference.

We can think of four reasons why the interaction between mode and number of items might occur.
First, it could be random noise. For example, in our analysis of agreement among respondents, we
observed substantial variability in residual agreement, or the amount of variance explained by the
second factor, across study conditions. It is possible that our dependent variable – individual cultural
knowledge about the domain of emotions, as estimated by loadings on the first factor in cultural con-
sensus analysis – is affected by this sampling error, such that the ANOVA results could be an artifact.
We think this possibility is unlikely to be the whole story, but it can be ruled out only with replication.

Second, online respondents who faced the 30-item pile sort may have taken the task less seriously
than did the face-to-face respondents, who had to deal with interviewers. This interpretation would be
consistent with literature from survey research showing that respondents in self-administered modes
tend to engage in more satisficing behavior (Heerwegh & Loosveldt, 2009).

Third, the 30-item pile sort may have magnified the effect of limited screen space. Even with 15
items, it becomes difficult on smaller screens to separate the piles into groups, and variation in mon-
itor size and resolution may alter the number and types of piles that people make. This possibility is
also consistent with evidence from online survey research, where variability in screen size, resolution,
and rendering of web questionnaires pose enduring challenges to measurement validity (Dillman &
Bowker, 2001; Tourangeau, Couper, & Conrad, 2013). It would be worthwhile in future research directly
to test hardware and software effects on pile sort results – especially as more and more users access
the web through mobile tablets and handheld devices, rather than on desktop or laptop computers
(de Bruijne & Wijnant, 2013).

Fourth, relatively low consensus in 30-item web-based pile sorts may reflect the lumper-splitter
problem. The lumper-splitter problem refers to the tendency of some people to make few piles and
others to make many (Weller & Romney, 1988, p. 22). With larger sets of items, the number of possible
groupings grows, such that variation in the number of piles people make increases and agreement
among respondents decreases. Indeed, variation in the number of piles respondents are permitted
to make is why unconstrained pile sort data do not conform strictly to the assumptions of cultural
consensus theory. The relatively large proportion of variance accounted for by the second factor in
our analyses (13–23%) may reflect lumper-splitter differences, rather than true variation in cultural
knowledge about emotions. For the purpose of testing for mode effects, however, measurement error
due to the lumper-splitter problem is signal rather than noise. It may be that limited screen size interacts
with measurement error associated with the increasing number of items to amplify lumper-splitter
differences on the web, resulting in the lowest level of agreement among online respondents for the
30-item task. Future work could test this possibility by adding even more items or by using methods such as the successive pile sort (Boster, 1994), which eliminates lumper-splitter differences and permits formal cultural consensus analysis.

In the meantime, our results provide new context for interpreting the seemingly contradictory results of two prior studies that compared online and face-to-face methods for collecting pile sorts. Bussolon et al. (2006) reported high concordance between online and paper card sorts, whereas Greve (2014) found significant differences between modes. Yet these two studies differed in one crucial detail: Bussolon et al. (2006) used only 15 items, whereas Greve (2014) used 51 items. Thus, the previous study with relatively few items – corresponding to the condition in our study with the highest level of agreement among respondents – showed that paper and web methods produced comparable results. The study with many items – even more than the condition that produced the lowest agreement among respondents in our study – suggested a mode effect. The contrasting results from these two studies is consistent with the interaction we observed between mode and number of items. Together, the evidence suggests that web-based pile sorts are more likely to diverge from face-to-face methods as the number of items respondents are asked to sort grows.

The more general point is that we cannot speak simply of mode effects related to the internet vs. conventional methods. Other aspects of research design (question construction, choice of cultural domain, sample composition, etc.) are likely to alter the effect of mode. This general point is also evident in results we previously reported comparing elicitation of free lists via face-to-face interviews, self-administered paper questionnaires, and self-administered web-based surveys (Gravlee et al., 2013). In particular, we found that evidence for mode effects depended on (a) whether the task included supplementary probes and (b) how easily contents of the cultural domain could be listed. The implications are that researchers considering online data collection for cultural domain analysis should carefully consider how other elements of research design may minimize or exacerbate potential limitations of self-administered questionnaires via the web. Moreover, future research on mode effects for cultural domain analysis should systematically test how variation in other aspects of research design affects the comparability of results across modes.

Our conclusions are limited to the context of our study. It remains important to test for mode effects (1) in populations other than university students who grew up with the web, (2) controlling for technical factors such as screen size, (3) in other cultural domains that may be less consensual, and (4) with other methods for collecting similarity data. We will address this last limitation in future analyses of mode effects for triad tests. We hope that others will seize the opportunity to test for mode effects in the collection of pile sort data in other populations and cultural domains.

Acknowledgments
We thank Jessica Pisano, Gina Eubanks, and Veronica McClain for assistance in data collection.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This material is based upon work supported by the National Science Foundation under [grant number BCS-0244104]. Development of Educara software was supported by grants to the University of Florida from the Ford Motor Company and the National Science Foundation.

Notes on contributors
Clarence C. Gravlee is Associate Professor in the Department of Anthropology at the University of Florida, with affiliate appointments in the Center for Latin American Studies, African American Studies, and the Department of Behavioral
Science and Community Health. His research interests include ethnicity and racism, health inequalities, and research methodology.

Chad R. Maxwell works in design anthropology as well as consumer research for marketing advertising. His professional experience has primarily been with digital marketing agencies where he works as an integrated measurement and people-centered research lead.

Aryeh Jacobsohn is Senior Product Manager at Pivotal Labs. He enables engineers to launch successful products by practicing lean thinking and user-centered design on balanced teams.

H. Russell Bernard is Professor Emeritus of Anthropology at the University of Florida and Director of the Institute for Social Science Research at Arizona State University. His areas of research include technology and social change, language death, and social network analysis.

References


