Analyzing larger sample sets with rapid methods: Incomplete-block designs with free-sorting and free-linking tasks

Marlon Ac-Pangan
Marino Tejedor-Romero
Kyra Swatko
David Orden
Jacob Lahne*

[1]Food Science & Technology Department, Virginia Tech, 1230 Washington St SW, Blacksburg, VA 24060, USA
[2]Departamento de Física y Matemáticas, Universidad de Alcalá, Campus Universitario, Ctra. Madrid-Barcelona, Km. 33.600, 28805 Alcalá de Henares, Spain

*Corresponding Author: jlahne@vt.edu, +1-540-231-7428

Abstract

As rapid, holistic methods for similarity and description—such as sorting and projective mapping—have grown in popularity, a limiting factor is the number of samples that can be presented to subjects: more than 25 food samples decreases the quality and stability of results. While incomplete-block designs could address this, their use has not been developed for these holistic methods. In this paper we present an empirical investigation into the use of incomplete-block designs with free sorting and the newer free linking. We compare these two methods because while their results are comparable, the cognitive tasks are different, and thus their suitability for incomplete-block designs may differ. We evaluated the effects of incomplete-block designs in two studies. In Study 1, 20 subjects evaluated 6/10 chocolate bars by free linking in an incomplete-block design, with each subject completing 2 blocks; results were compared to a complete-block evaluation of the 10 bars by free sorting and free linking. In Study 2, a total of 90 subjects evaluated 62 terms from a chocolate flavor-wheel in 3 conditions (between subjects): free sorting with complete blocks (N = 30, all 62 terms) and free sorting (N = 30) or free linking (N = 30) with 3 incomplete blocks of 16/62 terms. We introduce a novel method to evaluate stability for the incomplete-block designs that we call "pairwise simulation." From Study 1, we find that pairwise simulation provides adequate stability estimates and that, with sufficient pairwise cooccurrences, free linking with incomplete blocks produces results that are comparable to free sorting or linking with complete blocks. From Study 2, we demonstrate that free linking with incomplete blocks can produce high quality results from a large sample set, maintaining the increased discrimination capacity that marks free linking in general, and that with incomplete blocks, free linking is likely to be more stable than free sorting. This research demonstrates that incomplete-block designs can be used with free linking, and also provides a new, effective method through pairwise simulation for evaluating stability with incomplete-block designs, which cannot be resampled using standard bootstrapping approaches.

1. Introduction

Over the last several decades there has been a strong tendency in sensory science towards the use of rapid, "holistic" methods for understanding similarities and even sensory properties in a set of samples (e.g., Delarue & Lawlor, 2023; Valentin et al., 2012; Varela & Ares, 2014). These methods have many advantages: typically they require only a single data-collection session per subject, they produce data that is often quite rich, and their analysis is reasonably straightforward. Among the many rapid methods that have seen their popularity grow, free sorting and projective mapping (Valentin et al., 2018) are among the most popular because of their simplicity: they estimate an overall (dis)similarity structure from a comparatively large sample set in a single session, and can even generate rough descriptive profiles and identify groups among the samples using various clustering and partitioning analyses.

1.1. Holistic designs and incomplete blocks

A key challenge in the use of rapid, holistic methods, like free sorting or projective mapping for sensory evaluation of food is that these methods typically present all samples to each subject simultaneously. In cognitive psychology, where many of these methods originated (Coxon, 1999; Faye et al., 2004), this is not a major problem: there are reports of the successful application of free sorting to sample sets of more than 100 words or images (Gaillard et al., 2011). However, for food samples, which must be tasted or smelled, panelist performance and consistency quickly degrade when more than about 20 samples are presented (Chollet et al., 2011; Kessinger et al., 2020; Lahne et al., 2022). While no studies have formally explored this, it is reasonable to expect that the difference stems from the qualitative differences in sensory modalities between visual/lexical stimuli (in cognitive psychology) and aroma/taste/flavor stimuli (in sensory science): in the first case subjects have an easy ability to remind themselves of the characteristics of the samples they have already evaluated by visually returning to them,

whereas in the second case the chemical senses do not allow ready simultaneous comparison of multiple samples or non-fatiguing sample re-evaluation (Lawless & Heymann, 2010). Put very simply, panelists cannot make good judgments group-wise of similarity when they are overwhelmed with food samples.

Rapid, holistic methods are often suggested as alternatives to methods of Descriptive Analysis (DA; e.g., Faye et al., 2004; Moussaoui & Varela, 2010; Nestrud & Lawless, 2010; Valentin et al., 2012, 2018). In DA, studies will sometimes include more samples than can be evaluated by a single subject in one session, but this problem is dealt with by the application of balanced incomplete-block designs to the presentation order of samples to subjects (Heymann et al., 2014; Lawless & Heymann, 2010). For DA, this presents no problems, because subjects are comparing the samples' attributes to the reference standards on which they have been trained, rather than to other samples. This is not the case for the rapid, holistic methods we consider here, in which subjects, without prior training, only make comparisons between the samples that have been presented. To our knowledge, there are no publications systematically exploring the application of incomplete-block designs to sample presentation for rapid, holistic methods, although some recent publications point towards the possibility of such an approach (Courcoux et al., 2023). Moreover, an early study on the application of incomplete designs to an earlier method in cognitive psychology (on so-called "triad tests" for similarity) yielded promising results, indicating that incomplete-block designs for sample presentation may be compatible with studies of holistic similarity (Burton, 2003; Burton & Nerlove, 1976).

Two key problems need to be considered in the application of incomplete-block designs to rapid, holistic methods: the problem of "prototypical" samples and the problem of data structure. By "prototype" we mean samples that would be expected to define a new category (here we are using "prototype" to refer to a more "representative" object in the sense of Mervis & Rosch, 1981): a simple example is given in Figure 1a, where in a visual sort of 10 samples, 9 black and 1 red, the red sample clearly defines a new category by itself. A more realistic example might be in a study of chocolate bars, in which all but one bar are dark chocolate: the single milk-chocolate bar would be prototypical of a different category. In an incomplete-block design this would present a potential problem, because we would expect subjects presented with blocks without the prototypical sample to make qualitatively different judgments of underlying groups (categories) formed by the presented samples than those who receive blocks containing the prototype (See Figure 1a for a schematic example of this paradigm).

In principle, a fully-balanced incomplete-block design (such as those described in Gacula Jr. et al., 2009), which ensures that all pairs of samples are presented equally often across the entire design, would avoid this problem for the overall data structure, but in practice these designs are rarely feasible for two reasons. First, they require a great increase in the number of subjects required for the study (Chollet et al., 2011; Gacula Jr. et al., 2009), negating one of the key advantages of these rapid holistic methods (Courcoux et al., 2023; Valentin et al., 2012, 2018). Simply put, for most large-sample studies for which incomplete blocks might be of interest to the sensory analyst, the required number of subjects for a fully balanced incomplete-block design will often not be practical for reasons of time or budget. Second, and perhaps even more importantly, the basic combinatorics underlying such designs guarantee that a fully balanced design does not exist for all (or even many) possible study designs (Kuehl, 2000), and so a partially balanced design must be used instead (Gacula Jr. et al., 2009; Kuehl, 2000), which will be subject to the problem of prototypes described above.

The second problem is one of data structure. Typically, holistic designs generate data that represent the pairwise relationships between each presented sample. In the case of free sorting, for every pair of samples there is a binary indicator of whether they are in the same group; for projective mapping, for every pair of samples there is a Euclidean distance, where distance represents dissimilarity. When a subject is presented with an incomplete block, their data will then include missing values for when one or both of a pair of samples is not presented

in that block, and it is not clear what should be done with this missing data. Should it be treated as dissimilarity (a "0" in free sorting, or an infinite distance in projective mapping)? Almost surely not. Should these missing data somehow be imputed? Because the generative process for sorting data remains poorly explored (Hamilton & Lahne, 2020), this is not an easy task. Should they simply be dropped? Unless the incomplete blocks are perfectly balanced, which is unlikely for realistic sample sizes (Chollet et al., 2014), this seems likely to bias the analysis in unpredictable ways.

1.2. Free linking and incomplete blocks

Recently, we proposed a new rapid, holistic method which we dubbed the "free-linking task" (Lahne et al., 2022). This task resembles the free-sorting task, except that instead of making groups of similar samples, the subject is asked to consider similarity *between each pair of samples* (see schematic example in Figure 1c), by *linking* similar pairs in a physical or virtual diagram representing the samples. The data generated by free linking is an undirected similarity graph (Arney & Horton, 2013; Gross et al., 2013); in the previous study, we described how this can be seen as a generalization of the dissimilarity matrix in free sorting, and can be transformed into a dissimilarity matrix by a calculation on the graph distance (Chartrand & Zhang, 2014; Lahne et al., 2022). In free sorting, (dis)similarity is a fully transitive and binary property: if A is similar to B, and B is similar to C, then A must be similar to C *in the same degree*. In free linking, by contrast, similarity is at least ordinal, as linking A to B and B to C means there is an *indirect* link between A and C, not that A is *directly* similar to C. We demonstrated that, despite these differences, free linking produces results that are overall comparable to free sorting, but which represent a more realistic model of cognitive judgments around similarity (Lahne et al., 2022). Thus, free linking can also be seen as a single-pass alternative to multiple- or hierarchical-free sorting, which generate equivalent cophenetic distances among samples (Courcoux et al., 2012; Dehlholm, 2015; Lahne et al., 2022).

Here, we propose that these two key differences between free linking and sorting—pairwise similarity and the consequent lack of transitivity in similarity judgments—should make free linking a better candidate for application with incomplete-block designs. First off, the cognitive task of pairwise comparison should ameliorate the problem of prototypes: the presence or absence of a prototypical sample should not distort other pairwise similarity judgments, or at least should distort these judgments to a much lesser degree than in free sorting (see Figure 1c).

Second, the missing-value problem is less serious because the data produced by free linking is a less restricted form of *graph* data than that produced by free sorting. While both types of study produce individual and consensus results that can be treated as graph data, the results on the individual level are produced with different restrictions (Lahne, 2020; Lahne et al., 2022; Orden et al., 2019). In free linking, an individuals' results can be represented as a graph with edges representing a similarity judgment; the form of the graph is unrestricted by the method, and so, simply put, the union of two graphs from free linking that contain some (but not all) of the same nodes (samples) is simply another (weighted) graph, which can then be analyzed in the same way (Gross et al., 2013). On the other hand, free sorting produces *partition* data, which when represented as a graph is either a fully connected graph or a set of disjoint cliques. In other words, an individual's free-sorting result is constrained to a particular form: an adjacency matrix made up of the direct sum of all-1 or all-0 matrices (Abdi et al., 2007; Lipschutz & Lipson, 2017). The sum of two or more distinct partitions does not retain this particular structure (a direct sum of all-1 or all-0 matrices), and so is no longer a partition of disjoint cliques (see Figure 1b-c).

Therefore, we believe that free linking may be a better candidate for application with incomplete-block designs than free sorting. Specifically, in the context of incomplete-block designs, we would expect that groups of samples identified by free linking would be more stable

than those from free sorting—because the step of aggregating the individual results produces less distortion when the constraints imposed by free sorting are relaxed—and that the groups would discriminate better both quantitatively (finding more groups) and qualitatively (having groups that are better explained by the samples' intrinsic qualities).

1.3. Objective and aims

To explore the potential usefulness of incomplete blocks for rapid, holistic methods, and the potential suitability of the novel free-linking task for this particular problem, we present two studies that apply free sorting and free linking to incomplete blocks. In Study 1, we investigate the comparative *stability* of the free-linking task with incomplete blocks against the free-linking and free-sorting tasks with complete blocks, using a set of K = 10 chocolate samples evaluated by mouth by N = 20 subjects. In Study 2, we compare the actual sample resolution (discrimination) and stability of the free-linking and free-sorting tasks with the same incomplete blocks against the free-sorting task with complete blocks in a much larger sample set of K = 62 *descriptive terms for chocolate* from the Cocoa Wheel of Excellence (Cocoa of Excellence Technical Committee, 2021; Seguine & Sukha, 2015) with N = 90 subjects. Across these two experiments we will compare incomplete free-linking to incomplete free-sorting as well as to complete free-linking and complete free-sorting, using graph statistics (Gross et al., 2013; Lahne et al., 2022), Jaccard stability (Yu et al., 2019), and additive-tree partitioning (Koenig et al., 2021).

2. Materials and Methods

Below we report two studies that, while substantially similar in that they compared free-linking with incomplete blocks to other holistic designs, varied substantially in other respects. One key common aspect of both studies was that, in each study, subjects in the incomplete-block conditions completed multiple blocks in order to reach the required number of blocks for each partially balanced incomplete-block design (Chollet et al., 2011; Kuehl, 2000). While this may have familiarized subjects with samples, previous research has shown that replication of sorting tasks with the same subjects and samples in fact improves results (Lahne et al., 2016), and therefore it is very likely that the same result hold for free linking.

Beyond this shared methodology, we report each study separately to help the reader follow what was done for each.

2.1. Study 1: Chocolate

The goal of the first, smaller study was to determine how a rapid, holistic method with an incomplete-block design would compare to a standard, rapid holistic method. In order to do this, we repeated the free-linking task with the same chocolate bars reported in Lahne et al. (2022) and new subjects, but applied an incomplete-block design to the sample presentation, as described below. This provided data on a small sample set which could feasibly be analyzed by both complete and incomplete-block designs (Chollet et al., 2014).

We compare the data from the free linking with incomplete blocks to the results reported in Lahne et al. (2022).

2.1.1. Study 1: Subjects

In the original study, 63 subjects (49 female, 14 male, average age 34 years: Lahne et al., 2022) completed a free-linking and a free-sorting task, and their data is reused here. An additional 20 subjects were recruited from the Virginia Tech population in Blacksburg, VA (5 male, 13 female, 2 nonreporting, average age 22.5 years). Subjects were not given monetary compensation.

Both the new study and the original study were approved by the Virginia Tech Human Research Protection Program Institutional Review Board (IRB #s 19-1030, 21-858).

2.1.2. Study 1: Samples

Samples were the same 10 chocolate bars as used in the original study (Lahne et al., 2022), and are reported in Table 1 here.

2.1.3. Study 1: Experimental design

A partially balanced incomplete-block design was generated using the `crossdes::find.BIB()` function for `R` (R Core Team, 2023; Sailer, 2022). Specifically, we generated a design with the 10 chocolate samples as treatments, 40 blocks, and 6 treatments per block. In this design each sample was evaluated 24 times, and each pair of samples cooccurred between 12 and 14 times. Each subject completed 2 blocks in a single session in order to complete the required partially balanced incomplete-block design.

Each subject completed a free-linking task with each block of 6 chocolate samples. Samples in a block were presented simultaneously, in randomized order on the sampling tray, without identifying characteristics, in sensory booths under standard conditions for sensory evaluations (Lawless & Heymann, 2010). Subjects were provided with water and unsalted crackers *ad libitum*, and asked to refresh their palates in-between blocks. Subjects recorded their responses by drawing lines connecting similar samples in pencil on paper templates provided to them, and the results were manually transcribed into edgelists (Kolaczyk & Csárdi, 2014) by the researchers.

The experimental design for the data from the complete free-linking and free-sorting tasks with the same chocolate-bar samples are described in Lahne et al. (2022).

2.2. Study 2: Terms from the Cocoa Wheel of Excellence

The goal of the second study was to determine the applicability and stability of rapid, holistic methods (free sorting and linking) in a realistically large sample set. In order to accomplish this, we used words as samples (following Koenig et al., 2020, 2021) so as to allow for comparison of the results of sorting and linking with incomplete and complete sample blocks.

In this study, we used 62 terms from a “flavor wheel” for chocolate quality evaluation used in industry (Cocoa of Excellence Technical Committee, 2021; Eskes et al., 2012; Seguine & Sukha, 2015). This gave us a large sample set of terms that had a “true” structure against which we could compare our results. We also were confident that the terms from the wheel would not exceed subjects’ ability to evaluate using free sorting (Koenig et al., 2021).

2.2.1. Study 2: Subjects

Subjects were recruited online through mailing lists maintained by the Virginia Tech Food Science & Technology Department, as well as through snowball sampling. Subjects were not given monetary or other compensation. Subjects were screened for frequency of chocolate consumption, and had to report consuming chocolate at least once per month in order to participate.

In total, 255 subjects passed this initial screening and were directed into the study (166 female, 40 male, 4 non-binary, and 3 preferred not to answer). However, due to unexpected reboots affecting the SensoGraph hosting server, we will analyze complete data for only 30 subjects for each of the three possible treatments detailed below in section 2.2.3 (total N = 90, N = 30 per condition), and cannot directly link the demographic data to the sensory data. We believe that given the nature of this study this link is not important, and because the data loss was completely at random the reported demographic data is representative.

2.2.2. Study 2: Samples

Samples were the terms listed on the Cocoa Wheel of Excellence (Seguine & Sukha, 2015), the standard sensory-lexicon wheel for cocoa-quality judging. We used the terms from

the wheel directly, but split segments of the wheel with multiple terms into individual terms (e.g., "Earthy / Mushroom / Moss / Woodsy" was treated as 4 terms: "Earthy", "Mushroom", etc.) We also clarified terms that did not make sense in isolation from the wheel structure (e.g., changed "Yellow" as a subcategory of the "Fruity" category to "Yellow Fruit"). This yielded a total of 62 terms, listed in Table 2. Samples were presented through the various software terminals as plain terms, with no added definition or context.

2.2.3. Study 2: Experimental design

Through an online recruitment survey in Qualtrics (Provo, UT, USA), subjects were randomly assigned to one of 3 possible treatments: a free-sorting task with a complete-block design, a free-sorting task with an incomplete-block design, or a free-linking task with an incomplete-block design. We did not conduct a free-linking task with a complete-block design, because previous research suggests results would be very similar to the free sorting with complete blocks at the cost of much higher subject fatigue (Lahne et al., 2022).

In the free sorting with a complete-block design, subjects completed a typical free-sorting task (Chollet et al., 2014) with all 62 terms from the Cocoa of Excellence Wheel in a single block. Subjects completed the task online through Compusense Cloud (Guelph, ON, Canada). Subjects were constrained to make at least 2 groups of samples and no more than 61 groups, but otherwise could use their own criteria.

For the two treatments with incomplete-block designs, the same sample-presentation design was used. A partially balanced incomplete-block design was generated using the `crossdes::find.BIB()` function for `R`. Specifically, we generated a design for the 62 samples with 93 blocks of 16 samples each. Each subject completed 3 blocks of 16 samples. Each block was a single sorting or linking task (dependent on the between-subjects condition), in which subjects received all 16 samples from the specific block simultaneously, and evaluated their similarities according to the appropriate methodology.

Because of the data loss described above (section 2.2.1), in incomplete linking and incomplete sorting each sample was presented a minimum of 21 and a maximum of 24 times, and each pair of samples co-occurred between 3 and 8 times. A plot of the pairwise co-occurrences is presented in Figure 2. For complete sorting, of course, each sample was presented 30 times and each pair co-occurred 30 times.

For the free-sorting with incomplete blocks, subjects completed a total of 3 free-sorting tasks (1 for each block of 16 samples) using Compusense Cloud. Subjects were constrained to make at least 2 groups of samples and no more than 15 groups, but could otherwise use their own criteria.

For the free-linking task with incomplete blocks, subjects completed a total of 3 free-linking tasks (1 for each block of 16 samples) using SensoGraph (Orden et al., 2019; Orden & Tejedor-Romero, n.d.). Subjects were constrained to draw at least 1 link between 2 samples, but could otherwise use their own criteria.

2.3. Data analysis

While Studies 1 and 2 had slightly different goals, the overall analytical strategy for both was quite similar. First, all results were converted into graphs representing judgments of similarity. Then, additive trees (Abdi, 1990) representing the consensus similarities were partitioned (Koenig et al., 2021) to determine groupings. Summary statistics for similarity graphs and additive trees were examined as overall indicators of consensus quality. Finally, resampling and/or simulation approaches (Yu et al., 2019) were used to determine the stability of the groupings found through additive-tree partitioning. All data analysis was conducted in `R` 4.3.0 (R Core Team, 2023). Details for each step of the process follow.

As described above in sections 1.1 and 1.2, free-sorting and free-linking tasks generate data that can be treated as undirected graphs (Gross et al., 2013; Kolaczyk & Csárdi, 2014). A

graph, in this context, is a data structure in which the samples are represented as "nodes" (vertices in the graph) that are connected by "links" (edges in the graph); the links give information about whether a subject considers any given pair of samples to be similar (Lahne, 2020; Lahne et al., 2022; Orden et al., 2019). If a pair of samples is shown as two nodes that are not directly connected by a link, they were never judged as similar by any subjects in the study. Individual subjects' graphs can be summed to give a consensus, *weighted* graph with edge weights representing the number of times two samples are judged to be similar. Key graph statistics for understanding quality in sorting and linking tasks are *degree*—representing the number of samples each sample is considered similar to—and the number of disjoint components (disconnected subgraphs) in each subjects' response, which gives insight into the overall transitivity of a subjects' similarity judgments (see section 1.2) (Gross et al., 2013); we do not use a calculation of local or general transitivity directly because those measures are, by definition, either 0 or 1 for free-sorting data (Lahne et al., 2022). These analyses were conducted using the `igraph` (Csardi & Nepusz, 2006)and `tidygraph` packages (Pedersen, 2023).

In our previous paper, we proposed that consensus graph (dis)similarities could be calculated using a function derived from graph distance on individual subjects' similarity matrices followed by a summation across all subjects to get a consensus "graph dissimilarity" (Lahne et al., 2022). In brief, graph distance is defined as the number of links between any 2 nodes in a graph, and is in the range $[1, \infty)$, with the two extremes being achieved when two nodes are directly connected or when there is no path between nodes, respectively. In Lahne et al. (2022) we showed how to convert this distance into a dissimilarity in the range of $[0,1]$ appropriate for analysis by typical methods for rapid, holistic methods. In this paper we use these dissimilarities as the basis for additive-tree representations (Abdi, 1990) of the consensus using the `ape::nj()` function (Paradis & Schliep, 2019). Additive trees are an effective representation of free-sorting data that provide a graph representation of distance emphasizing groupings of samples, which is the key focus of the current study (Abdi, 1990; Chollet et al., 2011; Koenig et al., 2020, 2021). From these additive trees we generate consensus partitions using the additive-tree recursive partitioning algorithm using the `AddDistTreeSplit::recursive_partioning()` function (Koenig et al., 2021), with the default *LengthRatio* criterion of 0.6 (Koenig et al., 2021).

In additive trees based on judgments of (dis)similarity, total edge length between two nodes represents the observed similarity between those two samples (Abdi, 1990): longer total edge lengths indicate more dissimilar samples. Trees can be classified as "caterpillar" or "star" shapes: caterpillar-shaped trees have less well-defined branches, often consisting of many groups formed "step-wise", whereas star-shaped trees represent a more binary, "group of groups" formation (Mir et al., 2013). This shape tendency can be quantified through the Total Cophenetic Index (TCI), which for a tree with K nodes (samples), ranges from 0 to $\binom{K}{3}$, where more caterpillar-shaped graphs have higher TCIs. For free sorting and linking, a star-shaped tree represents more well-defined (discriminative) groupings.

In their original paper on additive-tree recursive partitioning, Koenig et al (2021) proposed using measures of *cohesion* and *isolation* to evaluate the stability of recursive partitioning, but while conceptually these measures are sensible, we were unable to replicate the calculations given by the original authors. Here, we use Jaccard stability instead (Yu et al., 2019). Briefly, Jaccard stability measures partitioning stability from *a sample/item-wise perspective* based on the original partition of the data. In the original partition, each sample is in a partition (group) with some other samples. If the partition is stable, after either resampling (as in bootstrapping) or another simulation method (such as our new approach in section 2.3.1), each sample should ideally be in a group *with and only with* the same samples as in the original partition. Jaccard stability gives a numerical representation of this quality bounded within $[0,1]$:

stability of 0 means that a sample is never with the same samples in different simulations, while 1 means that it is always with only the same samples. Thus, Jaccard stability combines the same information as Koenig et al.'s (2021) *cohesion* and *isolation*, but should be less vulnerable to bias from too many true negatives than another combined statistic they suggest originally (i.e., the Rand Index: Qannari et al., 2014). However, before it is possible to calculate Jaccard (or any other) stability measures, we must propose a method to effectively resample or (more accurately) simulate the data from incomplete blocks.

2.3.1. Simulation strategy

The problem of resampling data from partially balanced incomplete-block designs is not trivial, and to our knowledge there is no one, accepted method for doing so (Berry et al., 2021). In a typical bootstrap on data with a block structure (e.g., the one proposed by Koenig et al. (2021) or Yu et al. (2019)), a resampled dataset is generated by randomly drawing a sample of *complete* blocks from the original dataset. In the case of sensory evaluation typically the unit which is considered as a "block" is the subject, so that sensory scientists resample on the subjects: a resampled dataset will include 0, 1, or many replicates of a subjects' original responses (Abdi et al., 2009). However, when the data are generated from an incomplete-block design this is no longer feasible for two, related reasons. First, resampling on subjects obviously does not guarantee the sample-wise or pairwise balance from the original incomplete-block design, and second, as a consequence, the missing values (described in section 1.1), which stem from a missing pairwise co-occurrence and therefore are compensated for in the original design, will propagate through some non-zero number of the resampled designs. Thus, it was necessary to design a novel approach to simulating data that would generate plausible datasets from the original data while still respecting the original incomplete-block design.

As described in section 2.3, the consensus (dis)similarity data from free sorting or linking is representable as a graph, with samples as nodes and the observed similarity between samples as links. More specifically, the links can be given *weights* equal to the proportion of observed similarity judgments to total presented co-occurrences, making them *weighted edges* (Kolaczyk & Csárdi, 2014). Such a graph, representing the observed results of the incomplete linking from Study 1, is shown in Figure 3a.

This representation allows for simulating from these incomplete blocks in a fashion that will avoid the paired problems described above. Specifically, in order to simulate from the observed data, represented as the weighted graph in Figure 3a, we follow a two-step procedure: selected nodes (samples) and simulating edges between the selected nodes (similarity judgments). Nodes *could* be selected completely at random, but the problem of respecting the generated incomplete-block design can be solved by selecting *according to the original incomplete-block design*: for each simulated dataset of the same size as the original we will select nodes *according to each block in the incomplete-block design*. This will guarantee that we have the same sample occurrence and pairwise cooccurrence in each simulation as in the original design.

Once nodes are selected for a particular block (Figure 3b), for each edge in the graph a random draw from the uniform distribution (over $[0,1]$, for convenience call it p) is compared to the observed weight of that edge (the proportion of times the two samples were judged to be similar, for convenience call it p_0). If $p \leq p_0$, that edge will be retained in the simulated similarity graph (see Figure 3c). Thus, each possible similarity (edge) between two co-occuring samples (nodes) will be observed in a simulated block with the same probability as the observed results. A full simulation of the same size as the original data is obtained by running through the entire incomplete-block design; the entire process can then be repeated some large number of times (typically $i = 1000$) in order to generate a set of simulated datasets while respecting the incomplete-block design.

An example of this process applied to Study 1, with the nodes labeled with the blinding codes assigned to the samples used for readability, is given in Figure 3.

2.4. Data and code availability

All data analysis was conducted in R. Data and code to replicate the results shown in this manuscript are available at https://github.com/jlahne/incomplete-linking.

3. Results

3.1. Study 1 results

In Study 1, we extended the experiment with chocolate samples described in Lahne et al. (2022) with 20 new subjects using incomplete-block presentations. The overall results from the new data, as well as the original results, are presented in Figure 4 as additive trees. The free-linking task with incomplete blocks (Figure 4c) gave results that were quite similar to the results from the original free-linking and free-sorting with complete blocks (respectively, Figures 4b and 4a). This can be seen by comparing cluster membership and overall structure in Figure 4, and is confirmed by using Generalized Procrustes Analyses (GPA) to check the alignment of the 3 study configurations from Multidimensional Scaling (*RV* coefficients between the 3 configurations range between 0.92 and 0.98; full results not shown, but available in the manuscript code, see section 2.4). A notable exception is the Endangered Species Milk Chocolate sample, which appears to be appropriate given that this sample was an outlier in the sample set: it is a milk chocolate with an unusually high cocoa-content (48%) that was judged highly dissimilar to all other samples in the original study (Lahne et al., 2022, Figure 5). The Jaccard stability, discussed below and in Figure 6, helps explain and enrich this observation. We might consider this sample to be an example of a category "prototype" as discussed in section 1.1.

The graph statistics for the (dis)similarity graphs generated by subjects in Study 1 are given in Table 3 on a per-block basis. As described in section 2.3, the *degree* of a graph indicates the tendency, on a per-node basis, for samples to be sorted or linked together with more other samples, and the number of components (disconnected subgraphs) gives an idea of how many distinct groups subjects tend to make. From Table 3 it is apparent that free linking with incomplete blocks maintains the advantages of free linking in general, as reported in Lahne et al. (2022): degree remains lower for incomplete linking than either sorting method, whereas the number of distinct components remains the same or higher.

We evaluated the stability of the results from both the new and the original results using the Jaccard stability metric (Yu et al., 2019). In addition, we compared two methods for generating the data for stability calculations: "classic" bootstrapping on a per-subject basis (Abdi et al., 2009), which can only be applied to the data from complete-block designs, and our novel "pairwise" simulation approach (described in section 2.3.1), which can be applied to both complete and incomplete-block designs. As Yu et al. (2019) note, it is also possible to assess stability on a per-group level by taking the arithmetic mean within groups of a particular partition; individual stabilities are shown in Figure 5, but given the low number of samples it is possible for the reader to ascertain group stabilities visually.

First, it is worthwhile to compare the "bootstrap" and "pairwise" stabilities for the complete-block designs in Figure 5. For free linking with complete blocks, the pairwise simulation produces per-sample and per-groups stabilities that are extremely similar to classical resampling estimates with bootstrapping; for complete sorting, the estimated per-sample and per-group stability tends to be somewhat higher with the pairwise simulation except for a few samples. The observed pattern—better correspondence for stability estimates between the two strategies for free linking than for free sorting—follows predictably from a limitation of the pairwise simulation (see section 2.3.1): because the chance of each edge being selected for a

simulation is independent of all other edges, the actual, expected correlations between edge co-occurrence is not captured. This makes simulations generated by this approach much less representative for free sorting, because in free sorting all edges are part of *cliques* (fully connected subgraphs, Gross et al., 2013), implying a strong correlation between edges in observed data. However, as we see here, pairwise simulation performs quite well for free linking, precisely because transitivity is not a requirement (arguably, an artifact) of the data-collection methodology (see section 1.2). Therefore, we can conclude that our stability estimates for stability for our incomplete linking results, which can *only* be simulated pairwise, may have an upward bias in terms of absolute values (as observed in the free-sorting results), but adequately allow for within-study comparisons of stability among samples and groups.

With this in mind it is worth discussing the observable decrease in stability estimates for both complete methods for some samples—the dark chocolate samples with low cocoa-content—when simulations are generated through the pairwise method. These samples are dark chocolate but appear to share sensory properties with the "dark milk" Endangered Species Milk Chocolate, not the other dark chocolates. We can then compare *only* the pairwise simulation stabilities for the free linking with incomplete blocks to the free linking and free sorting with complete blocks: for the *other* dark-chocolate samples (grouped together in all three studies, with dark blue points in Figure 5) the stabilities are very close, with the incomplete-block design performing on the whole either better or only slightly worse than the complete-block designs; for the milk-chocolate samples, the *incomplete*-block design performs notably better. By examining the groups in Figure 5, it is apparent that this performance improvement comes from the hypothesized prototypical outlier (Endangered Species Milk Chocolate) leaving the milk-chocolate cluster and grouping with the cluster of dark chocolates that had similar cocoa contents.

3.2. Study 2 results

In Study 2, we examined a large sample set—62 terms from the Cocoa Wheel of Excellence (see Table 2). There were three main goals of this study: to evaluate the stability of results with incomplete-block designs in a real-world sample, to compare the stability of results from free-linking and free-sorting tasks with incomplete blocks, and to evaluate the discriminative *quality* of the results from these tasks by comparing between methods and to the original configuration.

Jaccard stability was calculated for each treatment, using pairwise simulation (see section 2.3.1) for the free sorting and linking with incomplete blocks and both pairwise simulation and classic bootstrapping for free sorting with complete blocks. The results for stability on an individual sample/term basis and on a group basis are given in Figure 6. From these results, it is clear that for almost all samples incomplete free-linking is more stable than incomplete free-sorting, and is reasonably stable even compared to complete free sorting (Figure 6). This stability is encouraging because the data loss from the incomplete linking and incomplete sorting (see section 2.2.3) is expected to reduce overall stability: each pair of samples is seen together much less frequently than planned (with a minimum of 3 pairwise cooccurrences for a small number of pairs, see figure 2).

Graph statistics for the various methods applied to the 62 samples in this study show the same pattern observed in Study 1: incomplete linking produces results that are overall of lower degree and with more components per individual subjects' graph (Table 3). It is interesting to note that the median number of components produced by *incomplete* sorting is in fact even lower than for complete sorting, and the median observed sample degree is in fact higher (Table 3). This indicates that incomplete sorting does not improve on complete sorting in terms of producing more nuanced or fine-grained results, while incomplete linking appears to retain the advantage of complete linking observed in Lahne et al. (Lahne et al., 2022). Overall, this also

supports the interpretation that free-linking, even with incomplete blocks, produces results that may better discriminate among samples and groups of samples.

The consensus solution for each of the three treatments is visualized as an additive tree with groups identified by the recursive partitioning algorithm in Figure 7 (Koenig et al., 2021). Groups are colored so that "similar" groups (determined again by an application of the Jaccard coefficient) are assigned similar colors across the studies. Based on the Total Cophenetic Index the complete sorting is the most "star-shaped" (Figure 7a), and thus discriminating among groups of samples, whereas the incomplete sorting is the most "caterpillar-shaped" (Figure 7c), and thus less discriminating, with the incomplete linking in-between (Figure 7b). However, the range of observed TCIs is quite small and all are a small fraction of the maximum TCI: $\binom{62}{3} = 37{,}820$.

Finally, the quality of the groups produced by the different methods under recursive partitioning can be examined in terms of subjective quality (how sensible they appear) and matching to the original taxonomy provided by the Cocoa of Excellence Wheel (Cocoa of Excellence Technical Committee, 2021). Recursive partitioning of the incomplete linking results identifies more groups (23) than does complete or incomplete sorting (17 and 19, respectively). In Figure 7 and Table 2 these groups can be more closely examined, and it is apparent that qualitatively, incomplete linking produces generally sensible, useful partitions. For example, incomplete linking results propose to partition the "fresh fruit" group in a manner that appears sensible (into stone-fruit and yellow-fruit/tropical groups), whereas both sorting methods do not partition the group. Alignment of these groups with the original wheel group is achieved through another application of Jaccard similarity: for each group of terms in the original wheel, Jaccard similarity is calculated against all proposed groups stemming from each experimental method, and the group(s) with the highest similarity score are presented in Table 2. Two encouraging results are found: first, incomplete linking results produces fewer non-matching groups (represented with NA in Table 2). As noted above, incomplete linking also tends to propose sensible partitions of the groups from the original wheel. While both of these are qualitative observations, it indicates that incomplete free-linking produces results of a quality that are comparable to complete free-sorting, even in the presence of serious data-loss.

4. Discussion

These studies' results provide insight into the application of free-sorting and free-linking approaches with partially balanced incomplete-block designs. In order to keep the discussion focused, we focus on the stability of free sorting and free linking with, respectively, complete- and incomplete-block designs, on the quality of the groups proposed by each methodology, and on the implications for method selection for sensory scientists interested in using incomplete-block designs with holistic methods.

4.1. Stability

Results from both Study 1 and Study 2 provide insight into the stability of free linking with incomplete blocks using the Jaccard index method suggested by Yu et al. (2019). In a previous study comparing free sorting, we had observed that free linking was stable at relatively low subject counts, but that in general its stability was always slightly lower than that of free sorting (Lahne et al., 2022, Figure 6) at the same subject count. Therefore, it is interesting to observe in the results from Study 1 in the current paper (Figure 5) that the Jaccard stability assessed by pairwise simulation for free linking with incomplete blocks was at least as high as that of free linking with complete blocks simulated in the same way. It must, however be noted that the method used to assess stability of results is different between this study and Lahne et al. (2022). In the previous work, the *RVb* coefficient was used to assess overall configuration stability at different subject counts (Blancher et al., 2012); in the current study the Jaccard index

was used instead, as the question of interest was not just overall configural stability, but grouping/partitioning stability (as defined by Koenig et al., 2021), and because the *RVb* cannot be applied directly to these partially balanced designs because of the way in which it is generated. However, if we compare stabilities estimated for the free linking and sorting results with complete blocks using both bootstrapping and pairwise simulation, the replication of the pattern observed in Lahne et al. (2022) bolsters our confidence that the stability results observed with *only* pairwise simulation for incomplete linking is reasonable. We speculate that this difference in stabilities can be explained partially by the method of generating the estimates, as discussed in section 3.2, and partially by the decreased presentation of a "prototypical" sample, the Endangered Species Milk Chocolate, as discussed in section 3.1. In regards to the former, it is likely that the pairwise simulation approach, because it does not constrain simulations to have correlated edges in the same way as bootstrapping, produces simulated results that are on the whole "more similar" to each other, and therefore more similar. In regards to the latter, it is possible that an unanticipated side effect of the incomplete-block design is the generation of a consensus that "averages out" the sample similarities in the presence and absence of such unusual, prototypical samples.

Study 1 does provide evidence that our novel method of graph-based, pairwise simulation is effective and produces results that, for free-linking, are comparable to more classic, block-wise bootstrapping (Abdi et al., 2009, 2012). In Figure 5, free linking with complete blocks show parallel patterns of sample- and group-wise Jaccard stability when resampled using bootstrapping or our novel pairwise simulation. In addition, for both free sorting and free linking with complete blocks, the direction of difference between stability results is not overall consistent, so it is not clear that there is a consistent bias. This gives us confidence that, for free linking with incomplete blocks, which cannot be resampled using a classical bootstrap, the estimate of stability using the novel pairwise simulation is a valid estimate, especially for intra-study results.

Examining the stability results from Study 2, however, provides a reason for some caution: the only comparison between the two simulation approaches possible in this set of results is in the free sorting with complete blocks, and in this case there is a dramatic difference in stability estimates between the classical bootstrapping and the pairwise simulation. Given the now well-established appropriateness of bootstrapping for complete-block designs (Abdi et al., 2009; Berry et al., 2021; Efron & Tibshirani, 1994), we must conclude that the pairwise simulation approach is producing estimates of stability that are very negatively biased. We believe that this can be explained by revisiting the generative procedure described in section 2.3.1: for free linking, the weighted graph representation is the natural data structure (see Figures 1 and 2), but for free sorting the weighted graph doesn't capture the disjoint and transitive structure of the individual subjects' results. Therefore, when simulated graphs are generated from the weighted-graph representation of the sorting results, the absence of correlations between edge presences described in sections 2.3.1 and 3.1 is likely to create more unrealistic results: in the current simulation, out of $i = 1000$ complete sorting results simulated, 52% of the individual subjects' simulated results have only a single component in the similarity graphs. A result with a single component not only indicates that a subject did not find any samples different from each other in free sorting; it is explicitly forbidden in the instructions given to subjects in free sorting (Valentin et al., 2018)! This degenerate case is, counterintuitively, a result of the transitivity assumption from sorting: as sorting results when represented as a graph consist of fully connected cliques, the pairwise simulation will always result in an erroneously high number of links that indicate similarity (see Figure 3), in contrast to the more realistic simulations for incomplete linking results, where far fewer links are observed (see Table 3).

Therefore, for Study 2, we should probably not consider pairwise simulation to be an acceptable estimate of stability for free sorting (either complete or incomplete); unfortunately,

therefore, while we are confident in this pairwise-simulation approach for incomplete *linking* (cf. Figure 6) and the bootstrap-resampling approach for complete *sorting*, we do not feel confident that the stability estimate available for *incomplete sorting* based on pairwise simulation is reflective of the method's true stability. However, we can investigate the quality of the proposed partitions derived from free sorting and free linking with incomplete blocks.

4.2. Quality

For applications of free sorting, a key outcome is the grouping structure derived from the consensus (dis)similarity judgments of the subjects: which samples are most similar to each other, and what groups of samples can be hypothesized from this pairwise similarity (Valentin et al., 2012, 2018)? We have argued here and elsewhere (Lahne et al., 2022) that more effective holistic methods identify more groups, rather than fewer, when all else remains the same. In the current studies, we see that free linking with incomplete blocks identified at least as many groups as free sorting with complete groups (Study 1, Figure 4), or indeed identified more (Study 2, Figure 6 and Table 2). This empirical measure of quality corresponds with the observed graph statistics reported in Table 2: in both studies, free linking with incomplete blocks produced similarity graphs that were less tightly connected and less transitive, which allows group-identifying (e.g., partitioning or clustering) algorithms like recursive partitioning to find better refined structures (Koenig et al., 2021).

Of course, if groups are meaninglessly differentiated by a methodology, then number of groups is a poor metric for quality. In this case we can examine the actual group memberships qualitatively for each study, as recursive partitioning is a form of unsupervised learning (Hastie et al., 2009). Proposed groups across the three methods in Study 1 are substantially similar, with the exception of the placement of the Milk Endangered Species chocolate, as discussed in section 3.1. In Study 2 the large number of samples provides the opportunity for many different group arrangements. We would argue that *none* of the methods produced an obviously unacceptable partition of the sample space (i.e., no proposed groups for any method were on-their-face ridiculous: Table 2). In this situation, then, the increased nuance and detail provided by the incomplete linking method, which identified 23 groups, is more useful: an end user of this lexicon would probably benefit from the separation of the fruity terms into two sub-groups, for example, or the effective partitioning of the general, single "vegetal" group into three subgroups.

4.3. Free linking vs free sorting

Complete free-sorting performs quite well in both Study 1 and Study 2, and as previously demonstrated complete free-linking performs well in Study 1 (Lahne et al., 2022); however, the real question posed in the current work is whether free sorting and free linking can be used when the number of samples for evaluation necessitates the use of a partially balanced incomplete-block design. When we consider both the stability and the quality of the results, it is apparent that free sorting's ability to produce stable partitions that explain the underlying similarity structure in a sample set may be worse than free linking when samples are presented in incomplete blocks. The results that we observe are consistent with our main hypothesis that the underlying cognitive task of free linking—pairwise comparison—and the data structure that best represents it—a similarity graph—are not challenged by incomplete-block designs in the same way as the cognitive task and data structure of free sorting.

4.4. Limitations and future research

A key limitation of the current research is the use of sample sets that could be easily evaluated in complete blocks when tasted (in Study 1) or without tasting (in Study 2). This constraint was necessary for the study design, but it does limit our ability to generalize these results. Specifically, will the observed performance benefits of free linking over free sorting in incomplete blocks persist either when subjects have to taste a larger number of samples, or

when many more subjects must be included to fill the blocks of a larger incomplete-block design with fewer samples per block? Only by conducting these studies can we answer these questions.

We propose, therefore, that future research should focus on applying incomplete-block designs with free linking to investigate a realistically large set of samples to be evaluated by taste or smell (past results indicate quality of free sorting results decays around 20 samples: Chollet et al., 2011; Kessinger et al., 2020). In order to assess replicability, one or more blind replicates can be included in the sample set without increasing the number of blocks required unduly. Such studies will validate the stability and the empirical applicability of this approach.

A second key limitation to our conclusions' generality was the unplanned data loss due to unexpected server reboots in Study 2. This forced us to compare very asymmetric situations between the free-linking and free-sorting tasks in terms of pairwise cooccurrence (a difference of an order of magnitude, in some cases), when the original design had controlled for this difference more effectively. However, in this case we feel that the reduced data for the incomplete-block designs represent a "worst case" scenario; because we have the results from Study 1 and the methodological "gold standard" of complete free-sorting in Study 2, we know that the results with the data loss in the incomplete-block studies have *reduced* quality; it is very improbable that the results will appear to be erroneously *more* stable or have *higher* quality. Therefore, we are more confident in our results, as the unbalanced design should only bias our estimates of stability in the incomplete-block designs downward. Given the relative success of the incomplete-linking in this unfavorable design, we hypothesize that in a better-balanced experiment, free linking with incomplete blocks will perform extremely well (as was indeed the case in Study 1).

A third limitation, but also an opportunity for future research and improvement, is the specific mechanism we have proposed for pairwise simulation of similarity results. We have discussed extensively throughout that, while this approach resolves the problem of implicit missing data in partially balanced incomplete-block designs, it does not fully capture the generating process of the original data, and so it will create results that only resemble the original data in the long run and aggregate (the consensus results), not in the individual subject results. This is particularly apparent in the dramatic differences between pairwise simulation and bootstrapping when applied to free-sorting results, as compared to the more modest differences for free-linking results. We propose that further methodological development could go into this pairwise-simulation proposal; this approach is to our knowledge novel, and so its properties should be further evaluated to examine the precise nature of how it may bias (different) stability estimates and to support extension to other data types. One key area of interest is developing an approach that is more "fair" to free sorting—that is, a method that does not disrupt transitivity, while still allowing for resampling or at least simulation of incomplete-block designs.

Finally, a key area for future research is an investigation into the effects of the number of samples and the number of subjects on the stability of similarity judgments derived from free linking (or indeed free sorting) with incomplete-block designs. From the current results, we can speculate that with a small number of samples and a relatively large number of subjects, as in Study 1, the results of incomplete-block designs will remain relatively stable; and, from Study 2, with the effects of the unplanned data loss, we can also speculate that with a large number of samples and a relatively small number of subjects, stability is clearly negatively affected.

5. Conclusions

Rapid, holistic methods for ascertaining the similarities, differences, and groups among a set of samples are increasingly widely used and popular in sensory evaluation. These methods—whether free sorting, free linking, projective mapping, or other variations—are all

limited by the number of samples a single subject can assess in one sitting. The current work presents the results of two studies that evaluate the application of the free-sorting and free-linking tasks to incomplete-block designs for sample presentation. The use of incomplete blocks allows sensory scientists to avoid the current limitation on the size of sample sets, and would increase the usability and impact of these methods. We hypothesized that it would be possible to use incomplete-block designs with these methods, but that free linking, because of its cognitive framework and data structure, would prove to be a better fit for the use of these designs.

Indeed, we found that, in two studies, free linking with an incomplete-block design was able to provide results that were consistent with results from a complete-block design with free sorting or free linking. Results from this incomplete free-linking also maintained the advantages of the free-linking task, producing results that were more nuanced than those from free sorting. We also found that free sorting with incomplete-block designs, as hypothesized, was less stable and less discriminating in identifying groups of samples than free linking. Therefore, we are able to conclude that incomplete-block designs are feasible with these holistic methods, and that of the two methods compared, free linking was much better suited for use with incomplete-block designs. These results should enable sensory scientists to employ rapid, holistic methods with larger sample sets and more fatiguing samples.

Acknowledgments

David Orden is partially supported by Project Project PID2019-104129GB-I00/ MCIN/ AEI/ 10.13039/501100011033 of the Spanish Ministry of Science and Innovation. This research did not receive any additional external funding. We would like to thank the volunteers who participated in this study, as well the undergraduate research volunteers at Virginia Tech who helped to coordinate data collection.

We would like to thank the editor and two anonymous reviewers who engaged deeply with the material in this manuscript and whose suggestions have improved the analysis and visualizations herein immensely.

References

Abdi, H. (1990). Additive-Tree Representations. *Lecture Notes in Biomathematics*, *84*, 43–59.

Abdi, H., Dunlop, J. P., & Williams, L. J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). *NeuroImage*, *45*(1), 89–95. https://doi.org/10.1016/j.neuroimage.2008.11.008

Abdi, H., Valentin, D., Chollet, S., & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, *18*(4), 627–640. https://doi.org/10.1016/j.foodqual.2006.09.003

Abdi, H., Williams, L. J., Valentin, D., & Bennani-Dosse, M. (2012). STATIS and DISTATIS: Optimum multitable principal component analysis and three way metric multidimensional scaling: STATIS and DISTATIS. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 124–167. https://doi.org/10.1002/wics.198

Arney, D., & Horton, S. (2013). Network Science for Graph Theorists. In J. Gross, J. Yellen, & P. Zhang (Eds.), *Handbook of graph theory*. CRC Press.

Berry, K. J., Kvamme, K. L., Johnston, J. E., & Mielke Jr., P. W. (2021). *Permutation Statistical Methods with R*. Springer.

Blancher, G., Clavier, B., Egoroff, C., Duineveld, K., & Parcon, J. (2012). A method to investigate the stability of a sorting map. *Food Quality and Preference*, *23*(1), 36–43. https://doi.org/10.1016/j.foodqual.2011.06.010

Burton, M. L. (2003). Too Many Questions? The Uses of Incomplete Cyclic Designs for Paired Comparisons. *Field Methods*, *15*(2), 115–130. https://doi.org/10.1177/1525822X03015002001

Burton, M. L., & Nerlove, S. B. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research*, *5*(3), 247–267. https://doi.org/10.1016/0049-089X(76)90002-8

Chartrand, G., & Zhang, P. (2014). Distance in Graphs. In J. Gross, J. Yellen, & P. Zhang (Eds.), *Handbook of graph theory*. CRC Press.

Chollet, S., Lelièvre, M., Abdi, H., & Valentin, D. (2011). Sort and beer: Everything you wanted to know about the sorting task but did not dare to ask. *Food Quality and Preference*, *22*(6), 507–520. https://doi.org/10.1016/j.foodqual.2011.02.004

Chollet, S., Valentin, D., & Abdi, H. (2014). Free Sorting Task. In P. Varela & G. Ares (Eds.), *Novel Techniques in Sensory Characterization and Consumer Profiling* (pp. 207–228). CRC Press.

Cocoa of Excellence Technical Committee. (2021). *Cocoa of Excellence Flavour Wheel* [Whitepaper]. Bioversity International.

Courcoux, P., Faye, P., & Qannari, E. M. (2023). Free sorting as a sensory profiling technique for product development. In J. Delarue & J. B. Lawlor (Eds.), *Rapid Sensory Profiling Techniques* (2nd ed.). Woodhead.

Courcoux, P., Qannari, E. M., Taylor, Y., Buck, D., & Greenhoff, K. (2012). Taxonomic free sorting. *Food Quality and Preference*, *23*(1), 30–35. https://doi.org/10.1016/j.foodqual.2011.04.001

Coxon, A. P. M. (1999). *Sorting data: Collection and analysis* (Vol. 127). Sage.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695.

Dehlholm, C. (2015). Free multiple sorting as a sensory profiling technique. In *Rapid Sensory Profiling Techniques* (pp. 187–196). Elsevier. https://doi.org/10.1533/9781782422587.2.187

Delarue, J., & Lawlor, J. B. (Eds.). (2023). *Rapid Sensory Profiling Techniques* (Second Edition). Woodhead.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
Eskes, A., Assemat, S., Jeantet, F., Seguine, E., Sukha, D. A., Weise, S., Thiriet, J., Rond, J., Laliberté, B., & Barel, M. (2012). *The cocoa of excellence and international cocoa awards initiatives: Rewarding diversity and excellence in producing high-quality cocoa origins*.
Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., & Nicod, H. (2004). Perceptive free sorting and verbalization tasks with naive subjects: An alternative to descriptive mappings. *Food Quality and Preference*, *15*(7–8), 781–791. https://doi.org/10.1016/j.foodqual.2004.04.009
Gacula Jr., M., Singh, J., Bi, J., & Altan, S. (2009). *Statistical Methods in Food and Consumer Research*. Academic Press.
Gaillard, A., Urdapilleta, I., Houix, O., & Manetta, C. (2011). Effects of task and category membership on representation stability. *Psicológica*, *32*, 31–48.
Gross, J., Yellen, J., & Zhang, P. (Eds.). (2013). *Handbook of graph theory*. CRC Press.
Hamilton, L. M., & Lahne, J. (2020). Assessment of Instructions on Panelist Cognitive Framework and Free Sorting Task Results: A Case Study of Cold Brew Coffee. *Food Quality and Preference*, 103889. https://doi.org/10.1016/j.foodqual.2020.103889
Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.
Heymann, H., King, E. S., & Hopfer, H. (2014). Classical Descriptive Analysis. In P. Varela & G. Ares (Eds.), *Novel Techniques in Sensory Characterization and Consumer Profiling* (pp. 9–40). CRC Press.
Kessinger, J., Earnhart, G., Hamilton, L., Phetxumphou, K., Neill, C., Stewart, A. C., & Lahne, J. (2020). Exploring Perceptions and Categorization of Virginia Hard Ciders through the Application of Sorting Tasks. *Journal of the American Society of Brewing Chemists*, 1–14.
Koenig, L., Cariou, V., Symoneaux, R., Coulon-Leroy, C., & Vigneau, E. (2021). Additive trees for the categorization of a large number of objects, with bootstrapping strategy for stability assessment. Application to the free sorting of wine odor terms. *Food Quality and Preference*, *89*, 104137. https://doi.org/10.1016/j.foodqual.2020.104137
Koenig, L., Coulon-Leroy, C., Symoneaux, R., Cariou, V., & Vigneau, E. (2020). Influence of expertise on semantic categorization of wine odors. *Food Quality and Preference*, *83*, 103923. https://doi.org/10.1016/j.foodqual.2020.103923
Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R*. Springer.
Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis* (2nd ed.). Duxbury/Thomson Learning; WorldCat.org. http://catdir.loc.gov/catdir/enhancements/fy1215/99023308-t.html
Lahne, J. (2020). Sorting Backbone Analysis: A network-based method of extracting key actionable information from free-sorting task results. *Food Quality and Preference*, 103870. https://doi.org/10.1016/j.foodqual.2020.103870
Lahne, J., Collins, T. S., & Heymann, H. (2016). Replication Improves Sorting-Task Results Analyzed by DISTATIS in a Consumer Study of American Bourbon and Rye Whiskeys: Replicated sorting of American whiskeys …. *Journal of Food Science*, *81*(5), S1263–S1271. https://doi.org/10.1111/1750-3841.13301
Lahne, J., Phetxumphou, K., Tejedor-Romero, M., & Orden, D. (2022). The free-linking task: A graph-inspired method for generating non-disjoint similarity data with food products. *Food Quality and Preference*, *95*, 104355. https://doi.org/10.1016/j.foodqual.2021.104355
Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practices* (Second). Springer.

Lipschutz, S., & Lipson, M. (2017). *Schaum's Outline of Linear Algebra, Sixth Edition*. McGraw-Hill Education. https://books.google.com/books?id=Lix_vgAACAAJ

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89–115. aph.

Mir, A., Rosselló, F., & Rotger, L. (2013). A new balance index for phylogenetic trees. *Mathematical Biosciences*, *241*(1), 125–136. https://doi.org/10.1016/j.mbs.2012.10.005

Moussaoui, K. A., & Varela, P. (2010). Exploring consumer product profiling techniques and their linkage to a quantitative descriptive analysis. *Food Quality and Preference*, *21*(8), 1088–1099. https://doi.org/10.1016/j.foodqual.2010.09.005

Nestrud, M. A., & Lawless, H. T. (2010). PERCEPTUAL MAPPING OF APPLES AND CHEESES USING PROJECTIVE MAPPING AND SORTING: PROJECTIVE MAPPING. *Journal of Sensory Studies*, *25*(3), 390–405. https://doi.org/10.1111/j.1745-459X.2009.00266.x

Orden, D., Fernández-Fernández, E., Rodríguez-Nogales, J. M., & Vila-Crespo, J. (2019). Testing SensoGraph, a geometric approach for fast sensory evaluation. *Food Quality and Preference*, *72*, 1–9. https://doi.org/10.1016/j.foodqual.2018.09.005

Orden, D., & Tejedor-Romero, M. (n.d.). *SensoGraph* (Spanish General Register of Intellectual Property Patent 16/2020/5028). https://sensograph.it/

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pedersen, T. L. (2023). *tidygraph: A tidy API for graph manipulation* [Manual]. https://CRAN.R-project.org/package=tidygraph

Qannari, E. M., Courcoux, P., & Faye, P. (2014). Significance test of the adjusted Rand index. Application to the free sorting task. *Food Quality and Preference*, *32*, 93–97. https://doi.org/10.1016/j.foodqual.2013.05.005

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sailer, M. O. (2022). *crossdes: Construction of Crossover Designs*. https://CRAN.R-project.org/package=crossdes

Seguine, E., & Sukha, D. (2015). *Flavour wheel with main categories and subcategories for both liquor and chocolates* [Whitepaper]. Cocoa Research Centre.

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science & Technology*, *47*(8), 1563–1578. https://doi.org/10.1111/j.1365-2621.2012.03022.x

Valentin, D., Chollet, S., Nestrud, M., & Abdi, H. (2018). Projective mapping and sorting tasks. In J. Hort, S. Kemp, & T. Hollowood (Eds.), *Descriptive Analysis in Sensory Evaluation*. Wiley-Blackwell.

Varela, P., & Ares, G. (Eds.). (2014). *Novel Techniques in Sensory Characterization and Consumer Profiling*. CRC Press. https://doi.org/10.1201/b16853

Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M., & Blair, R. H. (2019). Bootstrapping estimates of stability for clusters, observations and model selection. *Computational Statistics*, *34*(1), 349–372. https://doi.org/10.1007/s00180-018-0830-y

Tables

Table 1. Sample information for Study 1 samples (from Lahne et al., 2022)

Manufacturer	Chocolate type	Cocoa content
Cadbury	*Dark*	*35%[?]*
Hershey's	*Dark*	*45%[?]*
Green & Black's	*Dark*	*70%*
Endangered Species	*Dark*	*72%*
Green & Black's	*Dark*	*85%*
Pascha	*Dark*	*85%*
Cadbury	*Milk*	*26%[?]*
Hershey's	*Milk*	*30%[?]*
Green & Black's	*Milk*	*34%*
Endangered Species	*Milk*	*48%*

[?]information gathered indirectly from manufacturer's website rather than packaging.

Table 2. Sample information for Study 2: Cocoa of Excellence groups and terms. Also shown are results from Study 2 with samples matched against best-fitting groups from sorting and linking studies (as determined by the Jaccard index, Yu et al., 2019).

Cocoa of Excellence Descriptors (Cocoa of Excellence Technical Committee, 2021)		**Lexicon groups determined from sorting and linking studies**		
Group Name	**Group Terms**	**Complete Sorting Groups**	**Incomplete Linking Groups**	**Incomplete Sorting Groups**
acidity	*total acidity, fruity, acetic, lactic, mineral, rancid butter*	*acetic, lactic, total acidity*	*acetic, over-fermented*	*acetic, lactic, total acidity*
animal	*meaty, dirty animal / farmyard, leather*	*leather, mineral, tobacco*	*leather, mineral* *lactic, meaty, mushroom, umami*	*leather, moss* *meaty, mushroom*
astringency	*mouth-drying*	*bitterness, mouth-drying*	*dusty, mouth-drying*	*bitterness, dusty, mouth-drying*
bitter	*bitterness*	*NA*	*bitterness, total acidity*	*NA*
browned fruit	*browned fruit, dried, brown, over ripe*	*browned fruit, over ripe* *brown, dried*	*browned fruit, over ripe, rotten fruit*	*browned fruit, panela*
caramel	*caramelized sugar, caramel, brown sugar, panela*	*brown sugar, caramel, caramelized sugar, sweetness* *panela, resin*	*brown, brown sugar, caramel, caramelized sugar*	*brown sugar, caramel, caramelized sugar, sweetness*
cocoa	*cocoa*	*cocoa, spice, spices*	*cocoa, panela*	*cocoa, dried*
earthy	*earthy, mushroom, moss, woodsy*	*dark wood, earthy, light wood, moss, mushroom, woodsy, woody*	*NA*	*dark wood, earthy, mineral, smoky, tobacco, woodsy, woody*
floral	*floral, orange blossom, flowers*	*floral, flowers*	*floral, flowers, orange blossom*	*floral, flowers, orange blossom*
fresh fruit	*fresh fruit, berry, citrus, cherry / plum, peach, apricot, banana, tropical*	*apricot, banana, berry, cherry / plum, citrus, fresh fruit, fruity, orange blossom, peach, tropical*	*apricot, berry, cherry / plum* *banana, fresh fruit, peach*	*apricot, banana, berry, cherry / plum, citrus, fresh fruit, fruity, peach, tropical*
nutty	*nutty, nut skin, nut flesh*	*nut flesh, nut skin, nutty*	*nut flesh, nutty* *nut skin, woodsy*	*light wood, nut flesh, nut skin*
off-flavors	*off-flavors, dirty, dusty, musty, mouldy, over-fermented, rotten fruit, smoky*	*dirty, dirty animal / farmyard, dusty, manure, mouldy, musty* *off-flavors, over-fermented, putrid, rancid butter, rotten fruit*	*mouldy, off-flavors, putrid, rancid butter* *dirty, dirty animal / farmyard, manure, musty*	*off-flavors, over-fermented, putrid, rancid butter* *dirty, dirty animal / farmyard, mouldy, musty*
putrid	*putrid, manure*	*NA*	*NA*	*manure, over ripe, rotten fruit*
roast degree	*roast degree*	*roast degree, smoky*	*roast degree, smoky, tobacco*	*brown, nutty, roast degree*
spice	*spice, savory, umami, tobacco, spices*	*meaty, savory, umami*	*resin, savory, spice* *citrus, spices*	*savory, umami* *spice, spices*
sweetness	*sweetness*	*NA*	*fruity, sweetness, tropical*	*NA*

vegetal	*grassy, green vegetal, herbal*	*grassy, green vegetal, herbal*	*dried, herbal* *grassy, moss* *earthy, green vegetal*	*grassy, resin* *green vegetal, herbal*
woody	*woody, resin, dark wood, light wood*	*NA*	*dark wood, light wood, woody*	*NA*

910

911

Table 3. Summary graph statistics for individual subjects' (dis)similarity graphs from Study 1 and Study 2.

Methodology	# Components*	Degree*
Study 1		
Complete linking	*4 (1, 5)*	*0.11 (0, 0.44)*
Complete sorting	*4 (2, 7)*	*0.22 (0, 0.56)*
Incomplete linking	*3 (1, 3)*	*0.2 (0, 0.60)*
Study 2		
Complete sorting	*7 (3, 21)*	*0.16 (0.02, 0.51)*
Incomplete linking	*8 (1, 14)*	*0.07 (0, 0.27)*
Incomplete sorting	*5 (2, 12)*	*0.2 (0, 0.53)*

*All graph statistics are reported as Median (95% quantile).

912
913

Figures

Figure 1. Schematic representation of sample presentation and design for rapid, holistic methods, with (A) two possible incomplete blocks presented, showing a "prototypical" sample in red, and with non-presented samples in any block shown as "greyed out". In (B) two possible results of free sorting are shown, and the combined results are shown to no longer be a partition of the full sample; in contrast, in (C) two possible results of free linking are shown, and the combined results remain a (weighted) similarity graph.

Figure 2. Pairwise co-occurrence counts for samples for the incomplete-block designs in Study 2 after data loss.

Figure 3. A schematic representation of the novel, graph-based simulation strategy described in more detail in section 2.3.1. Briefly, in (A) all observed results of Study 1 are represented as a *weighted graph*: nodes represent samples (labeled with blinding codes), each edge represents the co-occurrence of two samples in the entire study, and the thickness (weight) of the edge represents the proportion of times in which a pair of samples, when presented together in a block, were judged to be similar. In (B), the same row of the incomplete-block design is simulated; for each simulation, all co-occurrences are assigned a probability of similarity based on the observed proportion in the actual study. Finally, in (C) all nodes not in the block and all edges which were not selected as "similar" in the simulation are dropped, giving two simulated, unweighted similarity graphs for the same row of the incomplete-block design. The entire process can be repeated a large number of times (in this case $i = 1000$) to produce a simulated dataset suitable for stability and other calculations.

Figure 4. Additive tree representations of similarities among 10 chocolate samples, as determined by (A) complete sorting, (B) complete linking, and (C) incomplete linking. Colors represent groups determined by recursive partitioning of additive trees (Koenig et al., 2021).

Figure 5. Jaccard stabilities for samples from Study 1, by complete linking, complete sorting, and incomplete linking (indicated by point shapes). Solid or dashed lines indicate "classical" bootstrapping or the novel, pairwise simulation described in Section 2.3.1, respectively. Points are colored according to the same group memberships shown in Figure 4. Group stability would be calculated by averaging the individual members, but is easily visible with only 10 samples and so is not drawn separately.

Figure 6. Jaccard stabilities for samples (top) and groups (bottom) from Study 2, by complete sorting, incomplete linking, and incomplete sorting. Dark and light green lines and points represent complete and incomplete sorting (respectively), and tan line and points represent incomplete linking. Solid lines represent "classical" bootstrapping, and dashed lines represent the novel, pairwise simulation described in Section 2.3.1. The faint, horizontal lines represent the average stability for the entire study in colors and line types as described above.

Figure 7. Additive tree representations of similarities among 62 terms from the Cocoa Wheel of Excellence (Cocoa of Excellence Technical Committee, 2021), as determined by (A) complete sorting, (B) incomplete linking, and (C) incomplete sorting. Colors represent groups determined by recursive partitioning of additive trees (Koenig et al., 2021).

962 **Figure 1**
963

964 **Figure 2**
965

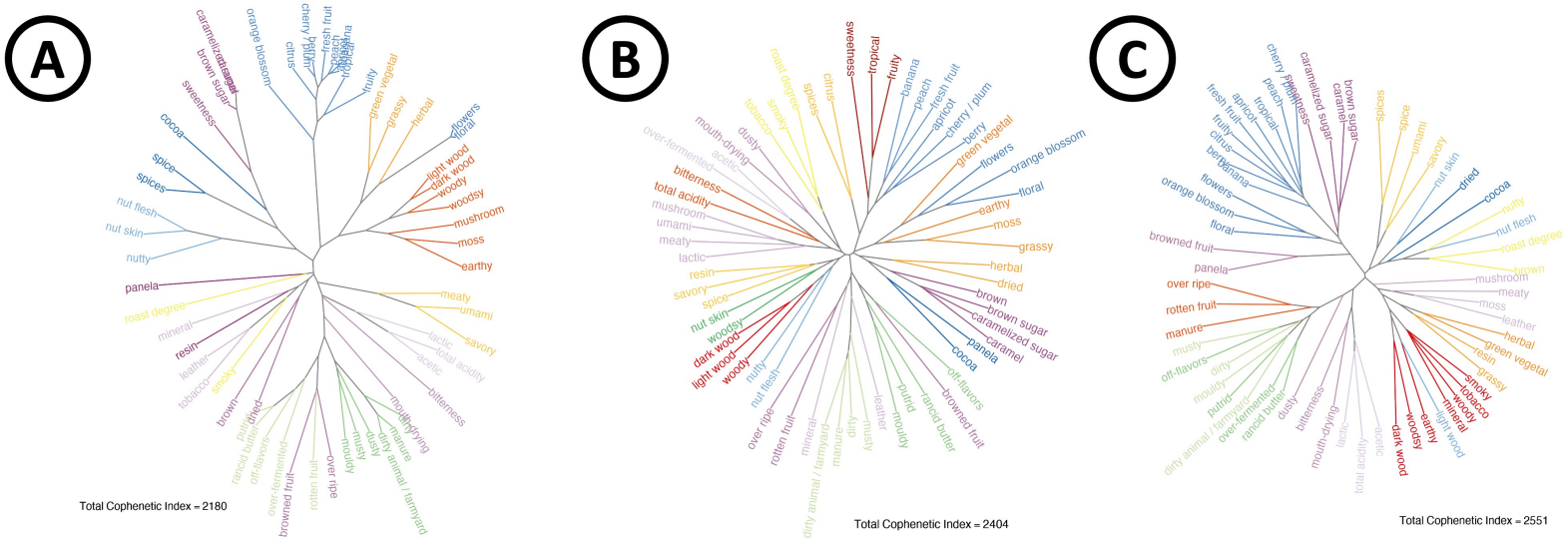
Figure 3

Figure 4

Figure 5

Figure 6

Figure 7


A
caramelized sugar
brown sugar
sweetness
cocoa
spice
spices
nut flesh
nut skin
nutty
panela
roast degree
mineral
resin
leather
tobacco
smoky
brown
dried
putrid
ranced butter
off-flavors
over-fermented
browned fruit
rotten fruit
over ripe
moldy
musty
dirty
dirty animal / farmyard
manure
mouth-drying
bitterness
total acidity
acetic
lactic
savory
umami
meaty
mushroom
moss
earthy
woody
dark wood
light wood
orange blossom
citrus
cherry / plum
fresh fruit
peach
banana
tropical
fruity
green vegetal
grassy
herbal
flowers
floral
Total Cophenetic Index = 2180
B
sweetness
tropical
fruity
banana
peach
fresh fruit
apricot
cherry / plum
berry
green vegetal
flowers
orange blossom
floral
earthy
moss
grassy
herbal
dried
brown
brown sugar
caramelized sugar
caramel
panela
cocoa
off-flavors
ranced butter
browned fruit
putrid
moldy
leather
musty
dirty
manure
mineral
dirty animal / farmyard
rotten fruit
over ripe
nut flesh
moldy
woody
dark wood
light wood
nut skin
spice
savory
resin
lactic
meaty
umami
mushroom
umami
total acidity
bitterness
mouth-drying
dusty
acetic
over-fermented
tobacco
smoky
spices
citrus
roast degree
Total Cophenetic Index = 2404
C
caramelized sugar
brown sugar
caramel
sweetness
cherry / plum
peach
tropical
apricot
fresh fruit
fruity
citrus
berry
banana
flowers
orange blossom
floral
browned fruit
panela
over ripe
rotten fruit
manure
musty
off-flavors
dirty
moldy
putrid
dirty animal / farmyard
over-fermented
ranced butter
dusty
bitterness
mouth-drying
lactic
total acidity
acetic
dark wood
woody
earthy
light wood
mineral
tobacco
smoky
resin
grassy
green vegetal
herbal
leather
meaty
mushroom
brown
roast degree
nut flesh
nutty
cocoa
dried
nut skin
savory
umami
spice
spices
Total Cophenetic Index = 2551