The Danger of Testing by Selecting Controlled Subsets, with Applications to Spoken-Word Recognition

David Liben-Nowell,[1*] Julia Strand,[2*] Alexa Sharp,[3,4]
Tom Wexler,[3,5] Kevin Woods[6]

[1]Department of Computer Science, Carleton College, Northfield, MN 55057
[2]Department of Psychology, Carleton College, Northfield, MN 55057
[3]Department of Computer Science, Oberlin College, Oberlin, OH 44074
[4]Google, Cambridge, MA 02139
[5]Verily Life Sciences, Cambridge, MA 02139
[6]Department of Mathematics, Oberlin College, Oberlin, OH 44074

[*]To whom correspondence should be addressed;
Email: dln@carleton.edu, jstrand@carleton.edu.

Acknowledgements: We are grateful to Mitchell Sommers and John Morton for generously providing the SWR data. Carleton College supported this work.

Abstract

When examining the effects of a continuous variable x on an outcome y, a researcher might choose to dichotomize on x, dividing the population into two sets—low x and high x—and testing whether these two subpopulations differ with respect to y. Dichotomization has long been known to incur a cost in statistical power, but there remain circumstances in which it is appealing: an experimenter might use it to control for confounding covariates through subset selection, by carefully choosing a subpopulation of Low and a corresponding subpopulation of High that are balanced with respect to a list of control variables, and then comparing the subpopulations' y values. This "divide, select, and test" approach is used in many papers throughout the psycholinguistics literature, and elsewhere. Here we show that, despite the apparent innocuousness, these methodological choices can lead to erroneous results, in two ways. First, if the balanced subsets of Low and High are selected in certain ways, it is possible to conclude a relationship between x and y not present in the full population. Specifically, we show that previously published conclusions drawn from this methodology—about the effect of a particular lexical property on spoken-word recognition—do not in fact appear to hold. Second, if the balanced subsets of Low and High are selected randomly, this methodology frequently fails to show a relationship between x and y that is present in the full population. Our work uncovers a new facet of an ongoing research effort: to identify and reveal the implicit freedoms of experimental design that can lead to false conclusions.

1 Introduction

There is growing concern in psychology and other disciplines that the scientific literature has a much higher rate of false positives than was previously assumed [56, 36]. This fear has grown based on the observation that many published findings fail to replicate [46]. The high false-positive rate is attributed, in part, to the tremendous flexibility that researchers have when making methodological and statistical decisions [5]. For example, researchers make choices throughout the experimental process about whether and how to exclude participants or observations, what covariates to include, how to combine or transform dependent variables, and when to terminate data collection [67]. These "researcher degrees of freedom" provide enough flexibility that, when used opportunistically, even impossible outcomes may be rendered statistically significant [56, 57].

In some subfields of psychology, experiment design includes deciding which stimuli to present to participants. Given that data collection requires time and other resources, and participants may become frustrated or withdraw from the experiment if testing is excessive, experimenters must make choices about which subset of survey questions, trial types, or stimulus items to include from a larger pool of possible items. Sometimes the choices of which items to include are dictated by prior work (e.g., a shortened form of a personality test that has been validated against a longer form [20]), but often a small subset of items may be selected with the implicit expectation that they represent the population from which they are drawn.

An assumption common to psychological research is that the findings of a particular study should generalize beyond the participants sampled. Concerns about this assumption have gained traction in the literature [30], and, more recently, there has been a push for researchers to explicitly state and justify the target population for the findings, thus defining the "constraints on generality" [58]. Although many researchers, if pressed, might agree that typical research participants (often college students) do not represent the general population, much less attention has been paid to whether the subsets of stimuli selected are representative of the broader population of stimuli from which they have been chosen.

1.1 Approaches to subset selection

When selecting multiple subsets of stimuli to assign to different experimental conditions, researchers often need to control for other relevant variables. For example, to carry out a study on gender stereotyping, Hettinger et al. [31] needed to identify two sets of household chores from a longer list—one set to assign to a male character in a story, one to a female character—so that the chosen sets matched on genderedness, pleasantness, difficulty, and time consumption. This approach is used widely in studies of word recognition, the focus of this paper, and has also been used in a variety of other psychological research, including the relationship between race and face perception [69], between attentional processing and obesity [12], and between emotion and memory [53], among others. Outside of psychology, this stimulus-selection approach has been used in applications ranging from echocardiographic interpretation [62] to deforestation [38].

Until recently, selecting matched subsets of items was typically achieved via manual selection by the experimenters themselves. To do so, the researchers laboriously select items that fit specified criteria (i.e., differing on an explanatory variable of interest, while being closely matched on a number of control variables)—presumably by starting from some rough-hewn item lists and iteratively improving their selections by adding and removing individual items to make the lists better matched in the control dimensions. The need to create matched subsets of items is widespread, though, and the manual process suffers from a number of problems: manual selection is tedious and painstaking work [19], precludes even the logical possibility of reporting every aspect of the reasoning in selection, may not result in well-matched stimuli [61], can be prone to error [4], and may introduce bias [23].

As such, several algorithmic approaches to generating matched subsets have been proposed recently, including MATCH [61], SOS [4], and BALI [18]. The underlying computational problem is provably difficult, and these algorithmic approaches vary in the ways—generally, which forms of randomization and heuristic approaches—that they use to handle that difficulty. MATCH finds a set of (yoked) pairs that are similar in control dimensions using backtracking and pruning; SOS ("stochastic optimization of stimuli") finds sets of items that are close in aggregate by starting from a random seed and making randomized local improving swaps using simulated annealing. There are also approaches based on genetic algorithms (BALI, "balancing lists"), as well as an algorithmic tool based on k-means clustering to give computational support for manual selection of items [28].

An alternate approach—also founded on the idea of selecting a carefully chosen subset of a large population, although here in a *post hoc* way—is based on the statistical technique of *matching* in observational studies [49, 6, 27, 21]. In this scenario, the goal is typically to infer the effect of an intervention in a population in settings where the assignment of individuals to the treatment group is chosen by some external decision-maker rather than being specified by the researcher; thus the allocation may be biased in any number of ways. As such, matching uses *post*-intervention subset selection to simulate a randomized controlled trial: a set of untreated controls is chosen from a large population of candidate untreated individuals, so that the selected subset matches the set of treated individuals with respect to the covariates. There are multiple approaches to selecting the matched control set, but *propensity score matching (PSM),* which aims to match the "propensity"—the probability of treatment conditioned on covariate values—is perhaps the most prominent. This approach applies, and is widely used, under reasonable assumptions about the way that individuals' treatment decisions were made and about the characteristics of the broader population, including the supposition that the population contains individuals who vary sufficiently on the measures of interest. (See [32] and [11] for more on the assumptions and implementation of PSM.) Indeed, an algorithmic tool for *ex ante* item selection based on PSM has recently been proposed [33].

1.2 The "Divide, Select, and Test" methodology

In both the experimental and observational methodologies just outlined, the researcher is seeking to assess the impact of a categorical variable x, typically representing population 1-vs.-population 2 membership or a treatment/no-treatment decision, on an outcome y. But a balanced-subset methodology is also sometimes used in conjunction with a "high–low split" when trying to understand how a continuously measured *explanatory variable* x predicts a *response variable* y, again while controlling for d different *control variables* $c_1, c_2, \ldots, c_d$. A high–low split is also sometimes called *dichotomization.* (In the examples described in the previous section, the researcher or some other decision-maker *assigns* the value of x to carefully chosen members of the population, or there are two discrete groups with different x values; here, the value of x is a continuously varying quantity that differs across members of the population.) In dichotomization, the population is divided into two sets, L (low x) and H (high x), and the y values of L and H are compared.

When dichotomization is combined with balanced-subset selection, sets $A \subseteq L$ and $B \subseteq H$—that is, a subset A of the low values L, and a subset B of the high values H—are selected so that two conditions hold:

(i) $|A| = |B|$ (that is, the sizes of A and B are identical); and
(ii) A and B match, on average, with respect to each of the control variables $c_1, c_2, \ldots, c_d$.

The values of y in A and B are then compared. This comparison is typically done using a t-test, which evaluates whether A and B differ significantly in their y values. We refer to this three-part methodology as "divide, select, and test" (DS&T): *divide* the population based on x into L and H; *select* sets $A \subseteq L$ and $B \subseteq H$ that match on $c_1, c_2, \ldots, c_d$; and *test* using a t-test whether A and B differ on y.

Under circumstances in which dichotomization is appropriate, existing algorithmic implementations [18, 4, 61, 33] are well-suited to efficiently selecting subsets while avoiding bias. However, dichotomization has well-known limitations, including a cost in statistical power [15, 25, 35, 51, 45, 34]. Here, we seek to assess the reliability and robustness of the DS&T approach by comparing it to other methods by which stimuli could be selected and effects tested.

1.3 "Divide, Select, and Test" in psycholinguistics

The DS&T methodology is used frequently in the psycholinguistics literature to support claims about how particular lexical properties affect the perception and production of spoken and written words. Specifically, in the context of spoken-word recognition (SWR) tasks, papers using DS&T have informed much of our understanding about lexical characteristics that make a word easier or harder for listeners to recognize. It has long been known that the greater *frequency* with which a word appears in natural language, the more quickly and accurately it is recognized [52, 17, 43]. Other research has explored effects based on the *competitors* of the word—that is, other words that are within a single phonemic insertion, deletion, or substitution of the word itself. For example, the competitors of *car* include *Carl* (an insertion); *are* (a deletion); and *bar, core, care,* and *call* (substitutions). DS&T studies first explored the most basic measure of competition on a word's recognizability, namely the word's *total number of competitors*: the more competitors a word has, the harder it is to recognize, even controlling for the word's frequency [43, 26].

More subtle metrics about a word's competitors have also been investigated, including the *clustering coefficient,* the fraction of pairs of the word's competitors that are themselves competitors of each other. For example, if *Carl, are, bar, core, care,* and *call* were the only six competitors of *car,* then *car* would have a clustering coefficient of $3/15 = 0.2$: of the 15 pairs of words that are competitors of *car,* the only three pairs that are themselves competitors are *Carl/call* (a deletion/insertion), *are/bar* (another deletion/insertion), and *core/care* (a substitution). (We treat clustering coefficient as undefined for any word with fewer than two competitors.) DS&T has also been used to show that high clustering coefficient is negatively associated with spoken- [13, 1] and written-word [68] recognition, even controlling for frequency and number of competitors.

In the present work, we concentrate on the effects of frequency, number of competitors, and clustering coefficient, but a large number of other lexical properties have been evaluated throughout the psycholinguistics literature. Other experiments using DS&T have concluded the presence of a significant effect from a host of other word properties, even controlling for previously known effects. These properties include the perceived subjective *familiarity* of the word [16]; the *phonotactic probability* of the word [the frequency with which a particular segment occurs in a given position in a word [66, 65]]; the *average frequency of occurrence of a word's competitors* [43]; the number of *onset competitors* [competitors that result from a substitution of the word's first phoneme [63]]; the competitors' *spread* [the number of phonemic positions in which a substitution creates another word [64]]; its *2-hop density* [the density of connections among competitors and their competitors [55]]; and the *isolation point* of the word [how many phonemes into the word one has to go before the word is uniquely identified [44]]. SWR provides a convenient venue for testing whether effects observed in small samples of items generalize to the larger population of items because data collection is sufficiently tedious that researchers tend to minimize the number of stimuli, but not so tedious that it is not possible to collect data on a much larger set of items.

2 Materials

SWR data were collected using standard methods. Stimuli were recorded by a native English speaker with a standard Midwestern American accent in a double-walled sound-attenuating chamber and leveled to total RMS in Adobe Audition. Participants were native English speakers who reported normal hearing and vision, recruited from the Washington University in St. Louis subject pool. Informed consent was obtained from participants and the research was approved by the Institutional Review Board where the SWR data were collected. Subsets of a list of 1120 monosyllabic words were presented to 94 participants (such that each word was presented to 30–32 participants), who attempted to identify the words by typing them. Word order was randomized. Stimuli were presented through headphones using E-Prime in six-talker babble at signal-to-noise ratio of -5. Both homophones and unambiguous nonwords whose obvious phonology matched the correct orthography (e.g., “turse” for “terse”) but no other deviations in spelling (e.g., pluralizations) were scored as correct. Starting from the correct/incorrect tags from the original dataset, we manually flipped a small number of correct/incorrect designations based on the identical-pronunciation criterion, changing a total of 362 participant–stimulus pairs out of 33 510, less than 1.1% of the data. (To ensure these changes did not systematically affect the outcomes, we also ran all of our analyses on the uncorrected data; the results were nearly identical to those reported here.) Of the 1120 words used as stimuli, 38 were not consonant-vowel-consonant words and 1 (“bass”) was pronounced differently than its form in the English Lexicon Project (ELP [8]), the dataset we use to calculate lexical characteristics. Thus, the analyses reported here were conducted on the remaining 1081 words, referred to here as the SWR1081 dataset. The dataset and all code necessary to run our analyses is available online through the Open Science Framework at `https://osf.io/x73dy/`.

In all analyses, including the integer linear program (ILP) in Figure 2(c), we use the z-scores of each numerical field (e.g., word frequency, number of competitors, clustering coefficient, accuracy) with respect to the full SWR1081 dataset. This choice puts all variables on comparable scales. However, for ease of interpretation, Figure 1 and Figure 2(b) show the raw numbers in the dataset.

3 Analyses & Results

In the present work, we show that DS&T-based analysis can lead to conclusions that are not well supported by data, including both the possibility of false positives (observing an effect through DS&T that is not present in the full population) and false negatives (frequently missing an effect through DS&T that is present in the full population). We also show that using either linear regression or linear mixed effects models on a subset of items provides greater statistical power than DS&T in SWR1081; we show similar results for linear regression on several synthetic datasets. (Generating a synthetic dataset suitable for analysis by linear mixed effects models requires more assumptions than we were willing to make in our generative process. Specifically, linear mixed effects models require participant-by-item data. Generating such synthetic data relies on a large number of parameters and assumptions about both participants and items, including the shape of distributions; many choices of parameters and assumptions would be consistent with the limited data that we would try to match.)

3.1 Demonstrating both positive and negative effects of the same variable

As just described, DS&T-based analysis has been used to argue that certain properties of words in the lexicon (with each of these properties serving as a candidate explanatory variable x) can predict human performance in recognizing those words (with recognition performance serving as the response variable y). Using the DS&T methodology with prior knowledge of the response variable, however, one can show contradictory results about the influence of an explanatory variable. Typically, of course, in a SWR experiment, the researcher would choose the sets A and B—the selected subpopulations of L and H, the low x and the high x segments of the population—*before* experimentally gathering the y values. One would want to choose these sets in advance because collecting data with human participants and their responses to lexical stimuli is resource intensive. And one would also want to select A and B in advance to ensure the moral equivalent of a double-blind study: it would be possible to put one's "thumb on the scale" if the y values of $L \cup H$ were known at the time of selection of A and B. [Having collected y values first would make the methodology much more similar to the observational setting of matching; for similar reasons, the matching literature counsels refraining from looking at outcome values when doing selection to avoid the risk of implicit bias in selecting which individuals to include in the dataset [50, 39].]

To illustrate this possibility of contradictory results, we use the SWR1081 dataset to ask how the selection of words affects the results that we observe. Following precisely the divide-and-select methodology (though peeking at the values of y), we are able to build two different pairs of balanced subsets $\{A_1, B_1\}$ and $\{A_2, B_2\}$, where $\{A_1, B_1\}$ demonstrates a strong positive effect of x on y and $\{A_2, B_2\}$ demonstrates a strong negative effect of x on y: that is, (i) the sets A_1 and B_1 are matched in all control dimensions (to within a specified tolerance, which we denote by δ) and $y(A_1)$ is much less than $y(B_1)$, whereas (ii) the sets A_2 and B_2 are also matched in all control dimensions but $y(A_2)$ is much greater than $y(B_2)$ (Figure 1).

3.2 Testing the effects of explanatory lexical variables using many different pairs of balanced subsets

We just showed that it is possible to find two pairs of extreme balanced subsets of SWR1081 showing highly positive or negative effects of an explanatory variable on a response variable. We now turn to generating *many* different pairs of balanced subsets through algorithmic means. This problem is a concrete, algorithmic task:

- We are given two populations L and H.[1] For each $x \in L \cup H$, we are given control-variable values $c_1(x), \ldots, c_d(x)$. We are also given a target set size k, and a tolerance $\delta > 0$.
- We must find sets $A \subseteq L$ and $B \subseteq H$, subject to two constraints:

  (i) $|A| = |B| = k$ (that is, both A and B have size equal to k); and

[1] In this paper, for a positive parameter $\rho \leq 0.5$, we construct L as the ρ-fraction of the population with the lowest values of x; similarly, H is the ρ-fraction of the population highest in x. Throughout this paper, we use $\rho = 0.5$, a median split. Note that using $\rho = 0.5$ means it is conceivable that words that have precisely the median value on the explanatory value could appear in both the L and H sets, although for SWR1081 and for the explanatory variables considered in this paper, this situation did not happen to occur.

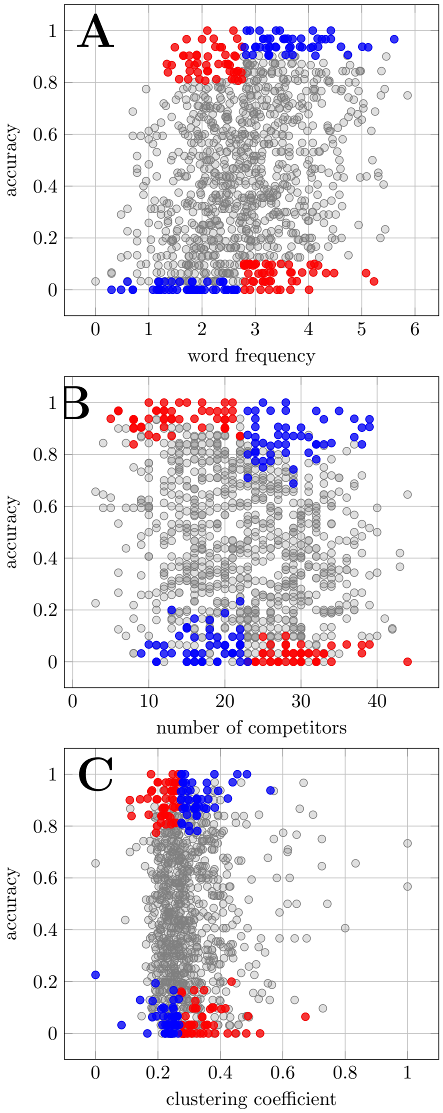


Figure 1: Maximizing the apparent positive and negative effect of an explanatory variable on SWR accuracy. Each panel shows all 1081 words in SWR1081, plotting for each word w an explanatory variable against the response variable *accuracy*, the fraction of participants who correctly identified this word when it was presented in a noisy environment. The explanatory variables are **(A)** the log frequency of w in a large corpus of natural text [10], **(B)** the number of competitors of w in the ELP lexicon [8], and **(C)** the clustering coefficient of w in the ELP lexicon. (Every word in SWR1081 has at least three competitors, so clustering coefficient is well defined.) Each panel identifies two pairs of 50-word subsets $\{A_1, B_1\}$ (blue; positive effect) and $\{A_2, B_2\}$ (red; negative effect). Each pair of color-matched subsets controls for all explanatory variables in previous panels: in (B), A_i and B_i's average frequency (in z-score) differ by less than $\delta = 0.05$, and likewise in (C) for both average frequency and number of competitors. Among all such δ-balanced 50-element subsets, the displayed subsets show the largest possible difference (positive and negative) in y for low-x and high-x words. (See Supplementary Materials for how these subsets are computed.)

(ii) for each control dimension i, the difference in A and B's average values in that control dimension is within δ:

$$\left| \frac{\sum_{x \in A} c_i(x)}{|A|} - \frac{\sum_{x \in B} c_i(x)}{|B|} \right| \leq \delta.$$

When there are a large number of control dimensions, selecting $A \subseteq L$ and $B \subseteq H$ is a tedious and difficult task (even if, unlike in the previous section, we do not try to push the response variable in either direction); thus we seek a general, systematic, and unbiased procedure to choose A and B.

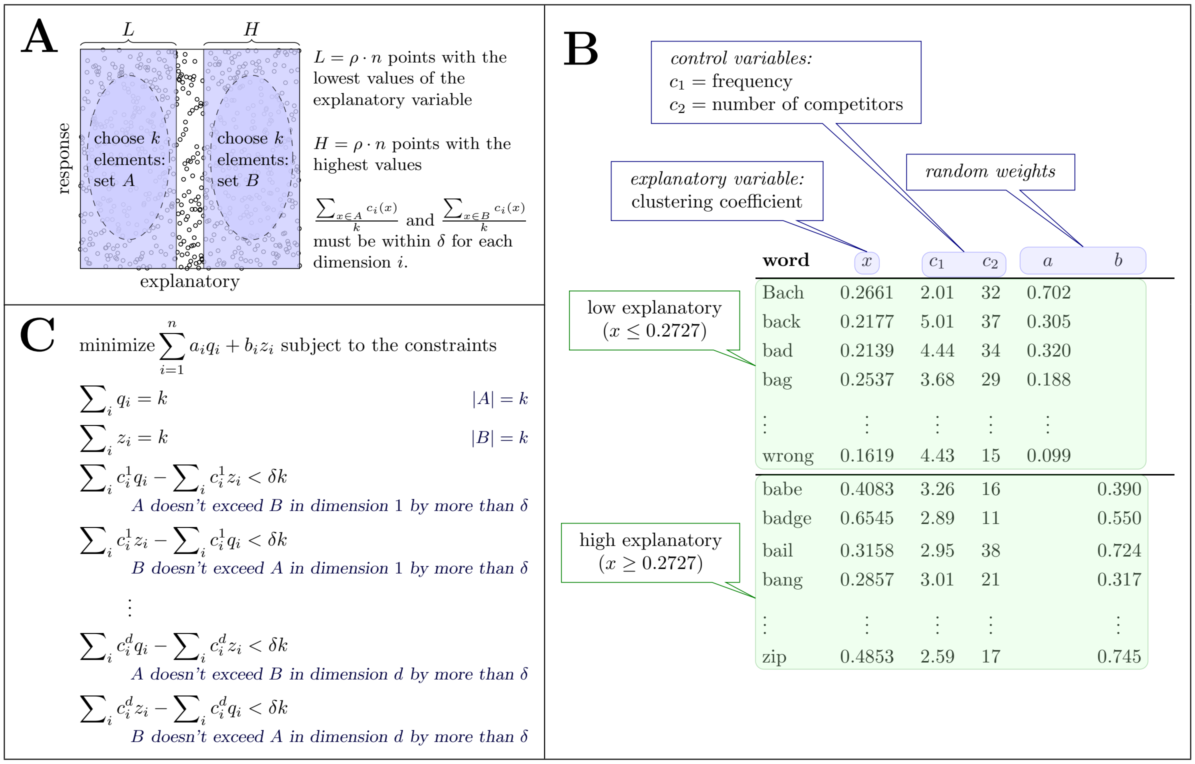
This problem is intractable in general—it is NP-hard to solve (see Supplementary Materials)—but for practical instances of reasonable size, this problem can be solved using an Integer Linear Program (ILP) and an off-the-shelf ILP solver called Gurobi [29]. The problem that we solve with our ILP is similar to the one solved in selection via the algorithmic approaches to subset selection detailed above, but we have chosen to design an algorithm to solve precisely the problem that corresponds to the DS&T methodology appearing regularly in the SWR literature that (i) optimally solves the item-selection problem and (ii) naturally allows the calculation of many pairs of balanced sets A and B.[2] We can achieve (ii)—that is, we can produce many different solutions for the same instance of the problem—by assigning randomly chosen weights to each element of $L \cup H$, and defining an ILP that selects, from among all sets satisfying the balance conditions, those sets A and B whose total weights are minimized (Figure 2). In this way, we are able to rapidly construct many different pairs of balanced sets A and B.

Papers on word recognition have used DS&T to claim effects on human recognition performance, using a single balanced pair of low/high word sets in each experiment: word frequency [high frequency corresponds to high recognizability [52, 17, 43]], number of competitors [many competitors corresponds to low recognizability [43, 26]], and clustering coefficient [high clustering corresponds to low recognizability [13, 1, 68]]. However, by sampling over many different balanced pairs of word sets, we see that the apparent strength of the effect on recognition accuracy varies dramatically across the chosen subsets.

Specifically, we ran 5000 DS&T experiments using the ILP in Figure 2 for the division/selection steps for these three lexical properties (Figure 3). The positive effect of high word frequency and the negative effect of a large number of competitors are largely evident and support prior research. (Frequency: 3632 of 5000 runs significant at $p < 0.05$; competitors: 2690 of 5000.)

However, in contrast to the results claimed in the literature, only in rare runs of the ILP does clustering coefficient show a significant effect on recognition accuracy when controlling for word frequency and number of competitors (Figure 3(c), 105 of 5000 runs significant at $p < 0.05$). Furthermore, even those rare runs that show a clustering-coefficient effect are split on the direction of effect (41 of the 105 show a negative effect; 64 of 105 show a positive effect).

[2] The fact that we optimally solve the item-selection problem (despite the problem's intractability) means that our ILP-based algorithm reliably produces the same output in every instantiation, though it comes at the cost of additional computation time compared to existing algorithmic approaches like SOS and MATCH [4, 61], which use randomization and heuristic optimization techniques. Note too that these existing tools also support a wider range of constraints and quality measures on the selected sets. For example, matching-based approaches—like MATCH or PSM—generally aim for individual-level matches for each treated individual, not just "on average" matches between the treated and untreated populations; our algorithm only seeks group-level similarity.



$L = \rho \cdot n$ points with the lowest values of the explanatory variable

$H = \rho \cdot n$ points with the highest values

$\frac{\sum_{x \in A} c_i(x)}{k}$ and $\frac{\sum_{x \in B} c_i(x)}{k}$ must be within δ for each dimension i.

control variables:
c_1 = frequency
c_2 = number of competitors

explanatory variable:
clustering coefficient

random weights

low explanatory ($x \leq 0.2727$)

high explanatory ($x \geq 0.2727$)

word	x	c_1	c_2	a	b
Bach	0.2661	2.01	32	0.702	
back	0.2177	5.01	37	0.305	
bad	0.2139	4.44	34	0.320	
bag	0.2537	3.68	29	0.188	
⋮	⋮	⋮	⋮	⋮	
wrong	0.1619	4.43	15	0.099	
babe	0.4083	3.26	16		0.390
badge	0.6545	2.89	11		0.550
bail	0.3158	2.95	38		0.724
bang	0.2857	3.01	21		0.317
⋮	⋮	⋮	⋮		⋮
zip	0.4853	2.59	17		0.745

minimize $\sum_{i=1}^{n} a_i q_i + b_i z_i$ subject to the constraints

$\sum_i q_i = k$ $\quad |A| = k$

$\sum_i z_i = k$ $\quad |B| = k$

$\sum_i c_i^1 q_i - \sum_i c_i^1 z_i < \delta k$

A doesn't exceed B in dimension 1 by more than δ

$\sum_i c_i^1 z_i - \sum_i c_i^1 q_i < \delta k$

B doesn't exceed A in dimension 1 by more than δ

$\vdots$

$\sum_i c_i^d q_i - \sum_i c_i^d z_i < \delta k$

A doesn't exceed B in dimension d by more than δ

$\sum_i c_i^d z_i - \sum_i c_i^d q_i < \delta k$

B doesn't exceed A in dimension d by more than δ

Figure 2: A schematic of the selection process, with parameters k (size of chosen subsets), ρ (fraction of data considered "high" or "low"), and δ (tolerance in control variables). **(A)** We must choose $2k$ of n given data points, in two equal-sized sets A and B, where A is chosen from among the $\rho \cdot n$ points with lowest explanatory variable values and B is chosen from among the $\rho \cdot n$ highest points. In every control dimension c_i, the elements of A and B are, on average, within δ. **(B)** A particular example of this input data in a SWR context, with data from the ELP lexicon [10, 8]. The weights a_i and b_i are chosen uniformly at random from $[0, 1]$. The desired solution is the lightest-weight pair of sets A and B (with respect to these particular a and b weights) that satisfies the control-dimension constraints. **(C)** The integer linear program (ILP) used to compute the solution. We define variables $q_i \in \{0, 1\}$ and $z_i \in \{0, 1\}$ indicating whether to include a point in A and B, respectively. Solving the ILP finds optimal values of q_i and z_i. Fresh random weights are chosen in each run of the algorithm.

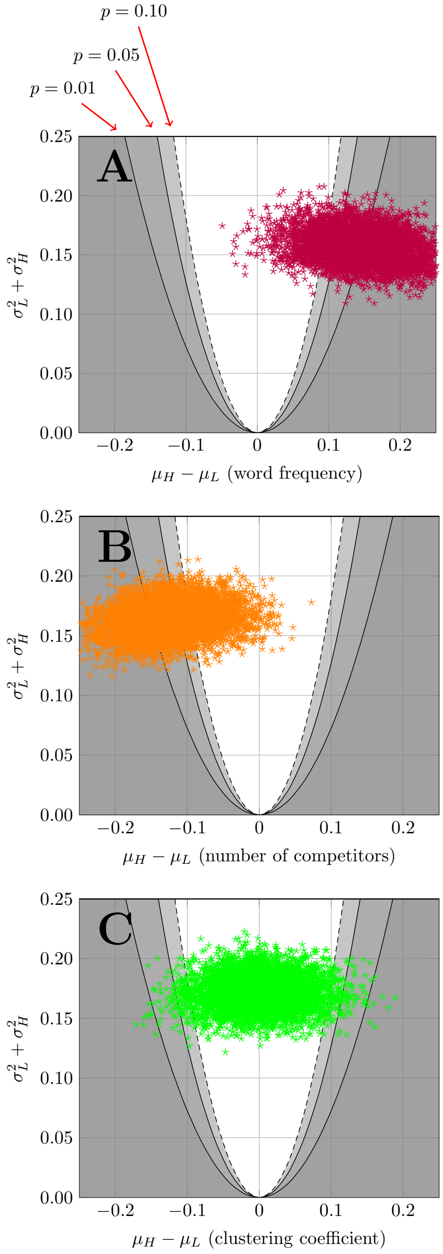


Figure 3: The result of 5000 runs of our ILP, with $k = 50$ words per subset, $\delta = 0.05$ tolerance for control variables, and $\rho = 0.5$ (dichotomizing on the median). Each point in each panel corresponds to a single run of the ILP to select sets A and B; the point plots the difference in mean recognition accuracy between A and B, vs. the sum of the variances of the recognition accuracies in A and B. The parabolas correspond to significance levels in a t-test on A vs. B. **(A)** The effect of frequency on recognition; 72.6% of these runs show that higher frequency is associated ($p < 0.05$) with more accurate recognition. **(B)** The effect of number of competitors; 53.8% of these runs show that having more competitors is associated ($p < 0.05$) with less accurate recognition. **(C)** The effect of clustering coefficient; 2.1% of these runs show an effect of clustering coefficient on recognition ($p < 0.05$), split between showing positive and negative effects. All experiments were controlled as in Figure 1. (See Figure S1 for the variant of this analysis that tests the effect of each variable while controlling for the other two.)

3.3 Comparing DS&T, linear regression, and linear mixed effects models

The literature includes at least two alternative approaches to the DS&T methodology. Researchers may instead opt to run correlations or linear regressions on continuously valued datasets [43, 59]. Using linear regressions differs from DS&T in both how stimuli are selected—by including words that vary continuously on the explanatory variable rather than binning high and low sets—and how the analysis is conducted—by statistically, rather than experimentally, controlling for the influence of control variables.

Another approach is to use linear mixed effects models (LMEMs), which provide a general, flexible approach to dealing with nested or hierarchical data (e.g., the fact that each word is identified by multiple participants). In these analyses, the LMEMs take raw, trial-level word-recognition correct/incorrect tags as input, rather than averaged accuracy values that collapse across participants. Like regressions, LMEMs can accommodate continuously valued data and statistically control for the influence of other variables. Much has been written about the benefits of the LMEM approach [37, 7, 14]; most notably for our project, an advantage of LMEMs over linear regressions is that LMEMs have access to information about variability at both the item and participant level, which may help reduce the error variance that can arise in large-scale studies [54].

To assess these different approaches, we compared the DS&T methodology (i.e., using the ILP with random weights to select A and B with sizes $|A| = |B| = k$ and using a t-test to look for a difference between A and B), denoted by **ILP‖*t*-test**, against four other methodologies. First, we considered two other ways of analyzing the sets produced by the ILP:

ILP‖lin-reg: ILP selection as just described, but using linear regression on the $2k$ elements of $A \cup B$ to test for an effect. (This methodology is appropriate only when $\rho = 0.5$, because an unpopulated central range of x values would violate the assumptions of linear regression.)

ILP‖LMEM: ILP selection as above, but using LMEM on the $2k$ elements of $A \cup B$ to test for an effect. In these models, participants and items were entered as random effects and the explanatory and control variables were entered as fixed effects. Given the large number of simulations that we ran, and the well-known problems with convergence for models that include them, by-participant random slopes for the lexical variables were not included. Omitting by-participant random slopes can increase the rate of false positives, but, as we shall see, our most interesting results are about the *low* positive rate associated with clustering coefficient. We used a logit linking function given that the explanatory variable (word-recognition accuracy) was binomial.

We tested each method's prediction of the effect of word frequency and number of competitors (Figure 4). We conducted the t-tests under the assumption of equal variance to make them more analogous with linear regression. In these experiments, **ILP‖*t*-test** had a higher likelihood (≈ 2 to 3 times) of failing to detect a significant effect than **ILP‖lin-reg** and a higher likelihood (again, ≈ 2 to 3 times) of failing to detect a significant effect than **ILP‖LMEM**. While the dichotomization literature has long recognized the gain in power of linear regression over t-tests for complete datasets [15, 25, 35, 51, 45, 34], here we are testing effects on carefully selected subpopulations; the results in Figure 4 suggest that a t-test remains far more likely than linear regression or LMEM to miss a true effect, even for intentionally chosen balanced subsets.

We then compared the use of linear regression and LMEM on sets selected via ILP to sets selected in a different way: via pure randomization—i.e., choosing the same number of elements

		fraction of runs with no significant effect				
explanatory variable	control variables	**ILP‖*t*-test**	**ILP‖lin-reg**	**ILP‖LMEM**	**uniform‖lin-reg**	**uniform‖LMEM**
frequency	(none)	0.274	0.123	0.085	0.127	0.090
competitors	frequency	0.462	0.265	0.227	0.277	0.247
clustering coefficient	frequency, competitors	0.979	0.965	0.959	0.962	0.958
clustering coefficient	frequency, competitors, subjective familiarity, spread of the neighborhood, number of neighbors formed in each phoneme position, neighborhood frequency, phonotactic probability [13]	0.975	0.963	0.948	0.960	0.943
clustering coefficient	frequency, competitors, subjective familiarity, frequency-weighted number of competitors, phonotactic probability, word length [1]	0.975	0.965	0.955	0.961	0.945

Figure 4: Comparison of power of various testing methodologies. We considered two explanatory variables with significant effect on recognition accuracy in SWR1081, as determined by linear regression and LMEM applied to the entire dataset ($p < 0.001$ for both frequency and competitors). We measured how often five different testing methodologies failed to detect the correct effect (at the $p = 0.05$ level) in subpopulations of 100 total words, over 5000 trials. ILP selection follows Figure 2; uniform selection chooses the same number of elements by uniform random sampling. We tested for a relationship using either a t-test comparing the low and high sets' response-variable values, via linear regression (in either case controlling for the listed control variables), or LMEM. The first two rows show settings in which there is a true effect (measured on the whole dataset); here, linear regression and LMEM correctly detect an effect more frequently than t-tests. When used with linear regression or LMEM, ILP performed slightly better than uniform sampling. For contrast, the last three rows show a setting in which there is no apparent relationship on the whole dataset ($p > 0.5$ using both linear regression and LMEM), where all three methodologies showed no effect $> 94\%$ of the time. The last two of these rows perform the clustering coefficient analysis while controlling for a much larger list of variables, following the methodology of [13] and [1]. Note that we did not directly attempt to correct for multicollinearity among variables; however, given the close similarity of the analyses in the last three rows of the table, which correspond to very different settings of control variables, and the fact that all of the ILP analyses (which control for covariates via selection rather than only statistically) are consistent with the uniform analyses, multicollinearity is not likely to substantially affect the results.

uniformly at random from the full population. (The t-test does not apply for the uniform selection mechanism, as it does not select two distinct populations.) Uniform selection yields two further possible analyses:

uniform‖lin-reg: we use uniform random sampling to select $2k$ elements from the entire population, and then use linear regression as above to test for an effect.

uniform‖LMEM: we use uniform random sampling to select $2k$ elements from the entire population, and then use LMEM as above to test for an effect.

Comparing **uniform‖lin-reg** and **uniform‖LMEM** to their ILP-generated counterparts, we see only a small difference in power: the ILP selection strategy has a mild advantage in the rate of runs with significant effects (for both frequency and number of competitors, both **ILP‖lin-reg** and **ILP‖LMEM** outperform their uniform counterparts). (See Figure 4 and also Figure S1.)

Using both linear regression and LMEM on the entire SWR1081 dataset suggests that, in keeping with Figure 3, word frequency and number of competitors do matter in word recognition ($p < 0.001$ for both variables and both analyses). We also tested the effect of clustering coefficient on recognizability, which does not have a significant effect in the full population ($p > 0.5$ for both linear regression and LMEM). When controlling for frequency and number of competitors, all five of our analysis methods do not find a relationship between clustering coefficient and recognition accuracy in the vast majority of runs (for each analysis method, fewer than 6% of runs found a relationship at the $p = 0.05$ level). To be consistent with previous published work that has shown significant effects of clustering coefficient on SWR [1, 13], we also ran the same analyses with a larger set of control variables to match the conditions of the previous studies (note that these studies used lexicons other than ELP, so our values for degree and clustering coefficient are close but not numerically identical to theirs); again, over 94% of runs fail to identify a significant relationship between clustering coefficient and accuracy.

ILP‖*t*-test was less likely than the other approaches to detect significant effects present in the full population (frequency and number of competitors). But the lower power of **ILP‖*t*-test** cannot be fully attributed to it being a conservative approach; when testing the effect on accuracy of clustering coefficient—which has no apparent explanatory effect on the response variable—the ILP-based selection of balanced sets still yields false positives $>2\%$ of the time, suggesting that such sets can be generated accidentally.

3.4 DS&T versus uniform selection on synthetic data

We have considered several methodologies for testing the effect of number of competitors (x) on word-recognition accuracy (y) while controlling for word frequency (c) in the SWR1081 dataset. To ensure that the issues with the DS&T methodology were not specific to the particular psycholinguistic dataset that we considered, we also attempted to replicate our results in a number of synthetic datasets. These synthetic datasets were designed to match the SWR1081 dataset in size and in relationships among these three variables—but with data that are drawn precisely from a multivariate normal distribution.

Specifically, using the covariances from the SWR1081 dataset (shown in the first row of Figure 5), we generated 10 synthetic datasets as follows: we generate $n = 1081$ data points, where each generated point is drawn from a multivariate normal distribution with all means equal to 0 and whose covariance matrix matches that of the z-scores of SWR1081. (Note that the covariance of

	covariances			fraction of runs with no significant effect		
dataset	x and y	y and c	x and c	**ILP**$|$$t$**-test**	**ILP**$\|$**lin-reg**	**uniform**$\|$**lin-reg**
SWR1081	−0.18	0.30	0.18	0.462	0.265	0.277
synthetic$_1$	−0.26	0.28	0.16	0.348	0.101	0.091
synthetic$_2$	−0.24	0.25	0.20	0.292	0.101	0.117
synthetic$_3$	−0.23	0.29	0.20	0.351	0.093	0.101
synthetic$_4$	−0.22	0.28	0.18	0.337	0.166	0.182
synthetic$_5$	−0.19	0.32	0.24	0.331	0.142	0.151
synthetic$_6$	−0.18	0.36	0.15	0.557	0.245	0.254
synthetic$_7$	−0.18	0.28	0.19	0.544	0.308	0.322
synthetic$_8$	−0.16	0.28	0.22	0.579	0.311	0.318
synthetic$_9$	−0.15	0.31	0.18	0.711	0.377	0.406
synthetic$_{10}$	−0.12	0.32	0.20	0.779	0.486	0.494

Figure 5: Analysis of SWR1081 and ten synthetic datasets generated to have (approximately) the same covariance matrix. The ten synthetic datasets are sorted in decreasing order of the strength of the relationship between x and y.

the *generated* synthetic datasets does not match SWR1081 precisely, because the synthetic datasets by definition are randomly constructed and do not precisely achieve their expected values.) We then ran the same t-test and regression analyses as in Figure 4 on all ten synthetic lexicons. Given the similarity of the results generated by linear regression and LMEM and the hard-to-justify assumptions necessary for generating synthetic trial-level data that accurately represents both item- and participant-level variability, we limited the analysis of the synthetic data to t-tests and linear regression.

The results are shown in Figure 5. We see the same broad patterns in the synthetic datasets as we do in SWR1081. First, **ILP**$\|$$t$**-test** has a higher likelihood of failing to detect a significant effect than **ILP**$\|$**lin-reg** or **uniform**$\|$**lin-reg**—generally by a factor of ≈2–3. Second, **ILP**$\|$**lin-reg** and **uniform**$\|$**lin-reg** have broadly similar false-negative rates. (There again appears to be a slight benefit for **ILP**$\|$**lin-reg** over **uniform**$\|$**lin-reg**, but the difference is modest.)

While the relative differences among the testing methodologies are fairly consistent across datasets, the raw values of the probability of detection of an effect varies across the ten synthetic datasets. The difference tracks the magnitude of the relationship between x and y; unsurprisingly, weaker correlations in the full population are more likely to be missed in the selected subsets.

4 Discussion

Taken together, our results show that DS&T increases the likelihood of false negatives without reducing the false-positive rate. DS&T also notably lacks transparency in how word sets are generated and matched: the key step of selecting which words from the lexicon to study is sublimated, and the principles by which selections were made are typically left opaque when the research is published. These choices have the potential to qualitatively affect the conclusions of a study, and thus serve as another researcher degree of freedom. Similar concerns have been raised about studies using matching, particularly PSM [39, 41, 3]. Unlike in matching-based studies about the effect of an intervention, though, we have a different option available: simply leave the continuous variable continuous. At a time when others are calling for larger participant sample sizes [2], we also recommend using larger sets of stimuli that more completely represent the population of stimuli—and analyzing the results using continuous statistics.

Continuous statistical methods—using linear regression or LMEMs on a randomly selected subset of stimuli—provides greater power and transparency of process than DS&T. Of course, these methods may come with their own challenges including how to deal with multicollinearity among predictor variables [22], how to make decisions about which potential covariates to include, and, in the case of LMEM, how to specify a random effects structure [9]. There may also be circumstances in which continuous statistics are not available or relevant: in many medical and policy-based studies, groups are truly categorical (e.g., control and experimental). In such cases, existing algorithms [18, 4, 33, 61] can be viewed as an alternative to dynamic allocation or matching techniques to assign individuals to treatment groups. Even here, though, care must be taken with statistical tests deployed in studies using dynamic allocation [48].

There appears to be a slight benefit in the true-positive rate of **ILP‖lin-reg** over **uniform‖lin-reg**: the ILP-based methodology modestly outperforms the uniform approach in SWR1081 and in 9 of the 10 synthetic datasets. **ILP‖LMEM** also outperforms **uniform‖LMEM** in SWR1081. This modest improvement may derive from the fact that the ILP necessarily selects a subpopulation that has a good spread of x values, whereas a uniform sample may have only a narrow swath of elements. [Note that ensuring this kind of variation is a feature that the SOS algorithm can support directly, by preferring subsets that have higher entropy in a given variable [4].]

Despite the intuitive nature of DS&T—all three of its components (dichotomization, controlling for covariates using subset selection, and t-tests) are seemingly innocuous—their combination not only weakens statistical power but also fails to eliminate the risk of false positives. Concretely, the previously published conclusions about the effect of clustering coefficient on spoken-word recognition are not supported by our analysis; we do not see evidence that clustering coefficient plays any significant role in recognition accuracy. Given the current replication crisis in psychology [46, 42], these results indicate that a certain attractive statistical approach in fact can lead to erroneous conclusions or suggest an unwarranted degree of confidence. DS&T approaches remain common, but compelling alternatives are appearing in the form of large-scale mega-studies on both written- [8] and spoken-word [60] recognition. The flexibility afforded by DS&T in choosing which data points to study allows a researcher to analyze a subpopulation that may be atypical; conclusions about that subpopulation do not validly imply anything about the population as a whole.

References

[1] Nicholas Altieri, Thomas Gruenenfelder, and David B. Pisoni. Clustering coefficients of lexical neighborhoods: Does neighborhood structure matter in spoken word recognition? *Mental Lexicon*, 5(1):1–21, 2010.

[2] Samantha F Anderson and Scott E Maxwell. Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3):305–324, May 2017.

[3] Kevin Arceneaux, Alan S Gerber, and Donald P Green. A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociological Methods & Research*, 39(2):256–282, 2010.

[4] Blair C Armstrong, Christine E Watson, and David C Plaut. SOS! an algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*, 44(3):675–705, September 2012.

[5] Jens B Asendorpf, Mark Conner, Filip De Fruyt, Jan De Houwer, Jaap J A Denissen, Klaus Fiedler, Susann Fiedler, David C Funder, Reinhold Kliegl, Brian A Nosek, Marco Perugini, Brent W Roberts, Manfred Schmitt, Marcel A G van Aken, Hannelore Weber, and Jelte M Wicherts. Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2):108–119, March 2013.

[6] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

[7] R H Baayen, D J Davidson, and D M Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, November 2008.

[8] D. A. Balota, M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and R. Treiman. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459, 2007.

[9] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), April 2013.

[10] Marc Brysbaert and Boris New. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990, 2009.

[11] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementaion of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.

[12] Megan A. Carters, Elizabeth Rieger, and Jason Bell. Reduced inhibition of return to food images in obese individuals. *PLOS ONE*, 10(9):1–20, September 2015.

[13] Kit Ying Chan and Michael S. Vitevitch. The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, 35(6):1934–1949, November 2009.

[14] Herbert H Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359, August 1973.

[15] Jacob Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253, 1983.

[16] Cynthia M. Connine, John Mullennix, Eve Shernoff, and Jennifer Yelen. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6):1084–1096, 1990.

[17] Cynthia M. Connine, Debra Titone, and Jian Wang. Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1):81–94, 1993.

[18] Christophe Coupé. BALI: A tool to build experimental materials in psycholinguistics. In *Architectures and Mechanisms of Language Processing*, 2011.

[19] Anne Cutler. Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10:65–70, 1981.

[20] M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. The mini-IPIP scales: tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, 18(2):192–203, June 2006.

[21] Marcello D'Orazio, Marco Di Zio, and Mauro Scanu. *Statistical Matching: Theory and Practice (Wiley Series in Survey Methodology)*. John Wiley & Sons, 2006.

[22] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1):92–107, 1967.

[23] K Forster. The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28(7):1109–1115, 2000.

[24] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

[25] Andrew Gelman and David K. Park. Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, 63(1):1–8, 2009.

[26] S. D. Goldinger, P. A. Luce, and D. B. Pisoni. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28:501–518, 1989.

[27] Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[28] Marc Guasch, Juan Haro, and Roger Boada. Clustering words to match conditions: An algorithm for stimuli selection in factorial designs. *Psicológica*, 38(1):111–131, 2017.

[29] Gurobi Optimization, Inc. Gurobi optimizer reference manual, 2015.

[30] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, June 2010.

[31] Vanessa E. Hettinger, Derek M. Hutchinson, and Jennifer K. Bosson. Influence of professional status on perceptions of romantic relationship dynamics. *Psychology of Men & Masculinity*, 15(4):470–480, October 2014.

[32] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[33] Stefan Huber, Julia F. Dietrich, Benjamin Nagengast, and Korbinian Moeller. Using propensity score matching to construct experimental stimuli. *Behavior Research Methods*, 49(3):1107–1119, June 2017.

[34] Dawn Iacobucci, Steven S. Posavac, Frank R. Kardes, Matthew J. Schneider, and Deidre L. Popovich. The median split: Robust, refined, and revived. *J. Consumer Psychology*, 25(4):690–704, 2015.

[35] Dawn Iacobucci, Steven S. Posavac, Frank R. Kardes, Matthew J. Schneider, and Deidre L. Popovich. Toward a more nuanced understanding of the statistical properties of a median split. *J. Consumer Psychology*, 25(4):652–665, 2015.

[36] John P A Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, August 2005.

[37] T Florian Jaeger. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446, November 2008.

[38] Seema Jayachandran, Joost de Laat, Eric F. Lambin, Charlotte Y. Stanton, Robin Audy, and Nancy E. Thomas. Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science*, 357(6348):267–273, 2017.

[39] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. Working paper, 2016.

[40] Jon Kleinberg and Eva Tardos. *Algorithm Design*. Addison-Wesley, 2005.

[41] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, pages 604–620, 1986.

[42] Eric Loken and Andrew Gelman. Measurement error and the replication crisis. *Science*, 355(6325):584–585, 2017.

[43] P. A. Luce and D. B. Pisoni. Recognizing spoken words: The Neighborhood Activation Model. *Ear & Hearing*, 19(1):1–36, 1998.

[44] William Marslen-Wilson and Lorraine Komisarjevsky Tyler. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71, 1980.

[45] Gary H. McClelland, John G. Lynch Jr., Julie R. Irwin, Stephen A. Spiller, and Gavan J. Fitzsimons. Median splits, type II errors, and false-positive consumer psychology: Don't fight the power. *J. Consumer Psychology*, 25(4):679–689, 2015.

[46] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

[47] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Inc., 1982.

[48] GR Pond. Statistical issues in the use of dynamic allocation methods for balancing baseline covariates. *British Journal of Cancer*, 104(11):1711–1715, 2011.

[49] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[50] Donald B Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, pages 808–840, 2008.

[51] Derek D. Rucker, Blakeley B. McShane, and Kristopher J. Preacher. A researcher's guide to regression, discretization, and median splits of continuous variables. *J. Consumer Psychology*, 25(4):666–678, 2015.

[52] H. B. Savin. Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, 35(2):200–206, 1963.

[53] Katherine Schmidt, Pooja Patnaik, and Elizabeth A Kensinger. Emotion's influence on memory for spatial and temporal context. *Cognition and Emotion*, 25(2):229–243, February 2011.

[54] D Sibley, C Kello, and M Seidenberg. Error, error everywhere: A look at megastudies of word reading. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.

[55] Cynthia S. Q. Siew. The influence of 2-hop network density on spoken word recognition. *Psychonomic Bulletin & Review*, pages 496–502, 2016.

[56] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[57] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive citations. *Perspectives on Psychological Science*, 2017.

[58] Daniel J Simons, Yuichi Shoda, and D Stephen Lindsay. Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6):1123–1128, November 2017.

[59] Julia F Strand and Mitchell S Sommers. Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *Journal of the Acoustical Society of America*, 130(3):1663–1672, September 2011.

[60] Benjamin V Tucker, Daniel Brenner, D Kyle Danielson, Matthew C Kelley, Filip Nenadić, and Michelle Sims. The massive auditory lexical decision (MALD) database. *Behavior Research Methods*, June 2018.

[61] Maarten Van Casteren and Matthew H Davis. Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, 39(4):973–978, November 2007.

[62] A. Varga, E. Picano, C. Dodi, A. Barbieri, L. Pratali, and O. Gaddi. Madness and method in stress echo reading. *European Heart Journal*, 20(17):1271–1275, 1999.

[63] Michael S. Vitevitch. Influence of onset density on spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):270–278, 2002.

[64] Michael S. Vitevitch. The spread of the phonological neighborhood influences spoken word recognition. *Memory and Cognition*, 35(1):166–175, 2007.

[65] Michael S. Vitevitch and Paul A. Luce. When words compete: Levels of processing in perception of spoken words. *Psychological Science*, 9(4):325–329, 1998.

[66] Michael S. Vitevitch, Paul A. Luce, David B. Pisoni, and Edward T. Auer. Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1):306–311, 1999.

[67] Jelte Wicherts, Coosje Veldkamp, Hilde Augusteijn, Marjan Bakker, Robbie van Aert, and Marcel van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7:1832, 2016.

[68] Mark Yates. How the clustering of phonological neighbors affects visual word recognition. *J. Experimental Psychology: Learning, Memory, & Cognition*, 39(5):1649–1656, 2013.

[69] Leslie Zebrowitz, Benjamin White, and Kristin Wieneke. Mere exposure and racial prejudice: Exposure to other-race faces increases liking for strangers of that race. *Social Cognition*, 26(3):259–275, 2008.

5 Supplementary Materials

5.1 Descriptive Data

	mean	standard deviation	range
accuracy	0.45	0.29	0–1
frequency	2.78	1.01	0–5.86
number of competitors	22.52	7.89	3–44
clustering coefficient	0.29	0.1	0–1

5.2 Integer Linear Programs to choose subsets of words

In the main text, we describe an algorithm to identify two sets A and B that are different with respect to an explanatory variable x (A comes from the part of the population with "low" x, and B from the "high" x subset) and such that A and B are balanced with respect to a given list of control variables. In Figure 2, we describe how we compute the balanced subsets A and B, using an integer linear program (ILP). For an introduction to integer linear programming, see [47]. In Figure 2, we use the following ILP, for randomly chosen weights a and b:

$$\text{minimize} \sum_{i=1}^{n} a_i q_i + b_i z_i \text{ subject to the constraints that}$$

$$\begin{aligned}
&\sum\nolimits_i q_i = k && |A| = k \\
&\sum\nolimits_i z_i = k && |B| = k \\
&\sum\nolimits_i c_i^1 q_i - \sum\nolimits_i c_i^1 z_i < \delta k && \textit{A doesn't exceed B in dimension 1 by more than } \delta \\
&\sum\nolimits_i c_i^1 z_i - \sum\nolimits_i c_i^1 q_i < \delta k && \textit{B doesn't exceed A in dimension 1 by more than } \delta \\
&\qquad\vdots \\
&\sum\nolimits_i c_i^d q_i - \sum\nolimits_i c_i^d z_i < \delta k && \textit{A doesn't exceed B in dimension d by more than } \delta \\
&\sum\nolimits_i c_i^d z_i - \sum\nolimits_i c_i^d q_i < \delta k && \textit{B doesn't exceed A in dimension d by more than } \delta
\end{aligned}$$

Here the idea is that, after generating random weights a_i for each "low" word and b_i for each "high" word, we select the balanced sets of low and high words that are lightest with respect to the chosen weights. The weights a and b describe the "cost" of selecting particular words; by randomly choosing those weights differently from run to run of our ILP algorithm, different words have high cost in different runs of the algorithm, so the balanced pairs of subsets we compute consequently differ across the algorithm's runs.

In Figure 1, we carefully choose balanced sets to maximize the apparent effect of the explanatory variable by using the response variable as the guide to choosing sets, instead of randomly chosen weights: we seek the balanced sets of low and high words that are *most different with respect to the response variable.* To do so, we use a very similar ILP, but with a different objective function:

$$\text{maximize} \sum_{i=1}^{n} r_i q_i - r_i' z_i$$

where

$$r_i = \begin{cases} x(\text{word } i) & \text{if word } i \text{ is in the low set} \\ 0 & \text{otherwise} \end{cases} \qquad r_i' = \begin{cases} x(\text{word } i) & \text{if word } i \text{ is in the high set} \\ 0 & \text{otherwise,} \end{cases}$$

where $x(\text{word } i)$ denotes the response-variable value for word i. The remainder of the calculation is exactly as described in Figure 2.

5.3 Computational complexity of finding balanced subsets

We claim that finding balanced sets A and B is an intractable problem, in general. Here is a precise statement and outline of a proof, using a reduction from SubsetSum, a standard NP-complete problem [40, 24]. Thus we would not expect to identify an efficient algorithm to solve the balanced subset problem; hence, the Integer Linear Program is an appropriate approach to solving the problem.

The specific algorithmic problem that we wish to solve (as described in the main text) is the following:

Definition 1. *The* BalancedSubset *problem is defined as follows.*

> **Input:** *two sets $A \subseteq \mathbb{R}^d$ and $B \subseteq \mathbb{R}^d$ of d-dimensional vectors, a positive integer $k \in \mathbb{Z}$, and a tolerance $\delta \geq 0$.*
>
> **Output:** *do there exist subsets $A' \subseteq A$ and $B' \subseteq B$, with $|A| = |B| = k$, such that, for every dimension $i \in \{1, 2, \ldots, d\}$,*
>
> $$\left| \frac{\sum_{x \in A'} c_i(x)}{k} - \frac{\sum_{x \in B'} c_i(x)}{k} \right| \leq \delta?$$

To demonstrate the hardness of the general BalancedSubset problem, we will prove the hardness of a special case of it. Specifically, we consider the EqualHalves problem, which is the special case of BalancedSubset in which:

- $d = 1$: there is only one control dimension.
- $\delta = 0$: the tolerance is zero (so we have to find subsets that match exactly in that one dimension).
- $c_1(x) > 0$ for all x: the values of all points in that one control dimension are strictly positive.
- $|A| = |B| = 2k$: the given sets are identical in cardinality, precisely twice that of the desired subsets.

Here is the formal definition of EqualHalves:

Definition 2. *The* EqualHalves *problem is defined as follows.*

> **Input:** *two sets of positive integers A, B with $|A| = |B| = n = 2k$. (We permit duplicates in A and B.)*
>
> **Output:** *do there exist subsets $A' \subset A$ and $B' \subset B$, both with size k and with equal sums?*

We will prove the hardness of EqualHalves via reduction from SubsetSum; from this fact, we conclude the hardness of its generalization BalancedSubset.

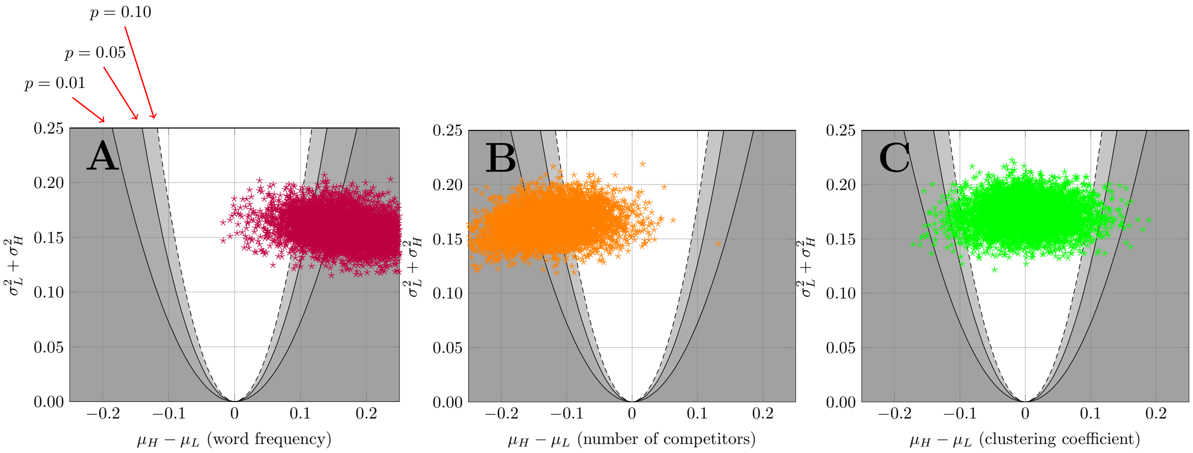
Theorem 3. *The* EqualHalves *problem is* NP-*Complete.*

Proof. Via reduction from SubsetSum. An instance of the SubsetSum problem consists of a set of positive integers $X = \{x_1, x_2, \ldots, x_m\}$ and a target sum W. The goal is to determine whether there is an $X' \subseteq X$ whose sum is W. Without loss of generality, we can assume that $\sum_{x \in X} > W$. SubsetSum is well known to be an NP-complete problem [40, 24]. (We permit duplicates in X, which doesn't affect the hardness of the problem.)

Given such an instance $\langle X, W \rangle$ of SubsetSum, construct an EqualHalves instance as follows. Define A to be the union of X and $m - 2$ zeroes. Define B to contain W and $2m - 3$ zeroes. We claim that $\langle X, W \rangle$ is a yes-instance of SubsetSum if and only if $\langle A, B \rangle$ is a yes-instance of EqualHalves.

($\Longrightarrow$) Suppose A' and B' is a solution to $\langle A, B \rangle$. The set A' must have a positive sum because strictly fewer than half of the elements of A are zero, and thus B' must contain W. Therefore the sum of elements in A' is W. Removing any zeroes from A' yields a subset of X whose sum is W.

($\Longleftarrow$) Suppose X' is a solution to $\langle X, W \rangle$. Generate A' by adding zeroes to X' until $|A'| = m - 1$. Let B' be W plus $m - 2$ zeroes. Both sets have size $m - 1$ and sum W. □



D

		fraction of runs with no significant effect				
explanatory variable	control variables	**ILP‖*t*-test**	**ILP‖lin-reg**	**ILP‖LMEM**	**uniform‖lin-reg**	**uniform‖LMEM**
frequency	competitors, clustering coefficient	0.168	0.048	0.030	0.058	0.034
competitors	frequency, clustering coefficient	0.459	0.276	0.234	0.318	0.281
clustering coefficient	frequency, competitors	0.979	0.965	0.959	0.962	0.958

Figure S1: The result of 5000 runs of our ILP, with $k = 50$ words per subset, $\delta = 0.05$ tolerance for control variables, and $\rho = 0.5$ (dichotomizing on the median). All points shown in panels **(A)**, **(B)**, and **(C)** are as in Figure 3 from the main text, except that *here each panel controls for the explanatory variables shown in the other two panels.* [In the main text, we analyze the effect of (A) word frequency on accuracy; (B) the number of competitors on accuracy (controlling for frequency); and (C) clustering coefficient on accuracy (controlling for frequency and number of competitors). Here, we analyze the effect of (A) word frequency on accuracy (controlling for number of competitors and clustering coefficient); (B) the number of competitors on accuracy (controlling for frequency and clustering coefficient); and (C) clustering coefficient on accuracy (controlling for frequency and number of competitors). The results are qualitatively identical to the results in the main text.] Panels (A), (B), and (C) are the analogue of Figure 3, and Panel **(D)** is the analogue of Figure 4.