Rich stimulus sampling for between-subjects designs improves model selection

Michael E. Young · James J. Cole · Steven C. Sutherland

Published online: 15 July 2011
© Psychonomic Society, Inc. 2011

Abstract The choice of stimulus values to test in any experiment is a critical component of good experimental design. This study examines the consequences of random and systematic sampling of data values for the identification of functional relationships in experimental settings. Using Monte Carlo simulation, uniform random sampling was compared with systematic sampling of two, three, four, or N equally spaced values along a single stimulus dimension. Selection of the correct generating function (a logistic or a linear model) was improved with each increase in the number of levels sampled, with N equally spaced values and random stimulus sampling performing similarly. These improvements came at a small cost in the precision of the parameter estimates for the generating function.

Keywords Stimulus sampling · Model selection · Parameter estimation · Nonlinear · Experimental design

Rich stimulus sampling for between-subjects designs improves model selection

Psychological investigations frequently focus on research that examines evidence for a categorical distinction between conditions or groups (e.g., depressed vs. not depressed). Throughout psychology's history, however, some behavioral scientists have also focused on identifying the nature of the relationships among continuous variables such as (1) whether the relationship between time and recall is an exponential or power function (Rubin & Wenzel, 1996), (2) whether delay

discounting is exponential, hyperbolic, or hyperboloid (McKerchar et al., 2009), and (3) whether the reaction time distribution is Weibull, ex-Gaussian, or gamma (Rouder, Lu, Speckman, Sun, & Jiang, 2005; Van Zandt, 2000). Furthermore, many categorical distinctions really reflect either a choice of two values along a continuum or the collapsing of disparate values along a continuum using median splits or similar approaches (the latter with oft-discussed shortcomings; Irwin & McClelland, 2003; Maxwell & Delaney, 1993).

The focus of the present research was to investigate the consequences of two alternative methods of choosing data values for the identification of functional relationships. The intent was to move beyond categorical distinctions between small and large stimuli, short and long delays, or low, medium, and high spatial frequencies to more richly sample the full range of stimulus dimensions. This broader sampling might improve the scientist's ability to identify the form of the relationship between independent variables (IVs) and dependent variables (DVs). To accomplish this goal, we propose that researchers should consider either choosing their predictor values randomly from the dimension or equally spacing them along the continuum.

Representative design

The notion of random stimulus sampling harkens back to the earliest writings of Egon Brunswik (1947, 1955). He suggested that just as subjects are randomly sampled from a larger population, so should stimuli be randomly sampled from the environment. This *representative design* serves two purposes: It ensures that the stimuli being tested are representative of the environments in which the behaviors of interest actually occur, and it improves generalization to the population of stimuli due to the sound statistical

M. E. Young (✉) · J. J. Cole · S. C. Sutherland
Department of Psychology, Southern Illinois University,
281 LSII, Mailcode 6502, Carbondale, IL 62901, USA
e-mail: meyoung@siu.edu

inferences possible with random sampling. Unfortunately, representative design is rarely pursued. Dhami, Hertwig, and Hoffrage (2004) examined studies conducted inside and outside the Brunswikian tradition; only one third of studies involving real cases randomly sampled stimuli from the environment, and random sampling of hypothetical cases was nonexistent. Instead, stimulus sampling most often resembles nonprobability sampling techniques such as purposive or convenience sampling, in which a handful of values along a dimension are chosen for testing. When we surveyed the 2004 issues of a leading experimental psychology journal, we discovered that 64% of continuous IVs involved sampling only two or three levels of the variable and only 16% sampled five or more levels. An identical survey of this journal 5 years later (2009) revealed that 78% of the continuous IVs involved sampling two or three levels and only 12% sampled five or more. Thus, there is no indication of a trend toward sampling more levels of continuous IVs in order to better obtain a richer sampling of the relationship between IVs and DVs.

Embedded within Brunswik's (1947, 1955) notion of representative design is the more modest proposal that stimulus sampling should be random. To illustrate the benefit of randomly sampling within a relevant range, consider the example of a neuroscientist who has tested three conditions involving a control (0-mg/l dose) and two experimental groups (50 and 500 mg/l). The scientist discovers that a 50-mg/l dose improves recovery but a 500-mg/l dose does not. A natural question concerns the shape of this relationship; clearly, it appears to be a relationship in which some of the drug is beneficial but too much is not. What level is most beneficial? Without a long series of follow-ups involving the choice of various dosages in each study, the relationship and, thus, the most beneficial level of the drug are unknown.

For situations like this one, it is clear that a richer sampling of the stimulus range is necessary. Although increasing the number of dosages sampled would help, scientists are hesitant to do so, because the commonly used analytical approaches often involve a series of planned or post hoc pairwise comparisons and undersampling any particular level would weaken these statistical inferences. If we shift the goal of a study from detecting pairwise differences to identifying the best mathematical function that describes the relationship between the predictors and the DV, concerns about undersampling any particular value of the continuous predictor would be misplaced. Given that a richer sampling of a stimulus range is desired, we considered the notion that each subject (between subjects) or experience (within subjects) could involve a different value from the range of values likely to be of interest and these values could be chosen randomly or at equal intervals to span the range. The present project provides the first test

of the benefits and costs of rich stimulus sampling by focusing on the simplest design, one involving a single IV manipulated between subjects.

One of the reasons why richer stimulus sampling may not have received serious consideration in the past is the analytical challenge of this type of data. Systematic sampling with replications is conducive to the variations on analysis of variance (ANOVA) commonly used in the behavioral sciences. Rich sampling without replication necessitates an approach similar to regression in which one or more continuous IVs are used to predict a continuous DV. However, regression assumes linearity, and the presence of nonlinearities offers its own set of challenges (the need for transformations or more complex approaches, such as polynomial or kernel regression); combining nonlinearities with the presence of variables that vary between or within subjects complicates matters even further, and many researchers may be unfamiliar with methods like multilevel or mixed effects modeling to do repeated measures linear or logistic regression and their nonlinear analogs (e.g., Gelman & Hill, 2006; Pinheiro & Bates, 2004). Despite these challenges, we believe that the benefits may make the effort worthwhile.

Theoretical issues in model selection and parameter estimation

Optimal experimental design involves choosing the optimal set of IV values, either for the purpose of identifying the best model among a palette of options (the goal of *model selection*) or for the purpose of producing the most accurate parameter estimates (the goal of *parameter estimation*). There is considerable attention to these issues in the statistics and engineering literatures (e.g., Atkinson & Donev, 1992; Box & Hill, 1967; Ford, Titterton, & Kitsos, 1989; Hill, Hunter, & Wichern, 1968; Steinberg & Hunter, 1984). However, there is very little coverage of the topic in the behavioral sciences, as evidenced by the small number of articles in psychological journals on choosing the best values along a continuum in the pursuit of either of these modeling goals. Although Myung, Pitt, and colleagues (Cavagnaro, Myung, Pitt, & Kujala, 2010; Cavagnaro, Pitt, & Myung, 2011; Myung & Pitt, 2009) recently have been actively developing methods of choosing a set of IV values in order to improve selection among known models, there are only a handful of other psychology publications involving general design approaches to optimal parameter estimation (Berger, 1994; Berger, King, & Wong, 2000; Passos & Berger, 2004; Vermeulen, Goos, & Vandebroek, 2008), with each receiving no more than a handful of citations (ranging from 0 to 8 in the Social Science Citation Index). One exception is the common use of adaptive methods in the area of psychophysics, where stimulus levels

are dynamically chosen on the basis of unfolding performance in order to estimate the threshold or slope of an ogival psychometric function (Leek, 2001).

These prior approaches focus on identifying an optimal set of values to test in either single-factor or factorial designs. For the goal of parameter estimation, the theoretical justification for the selection process rests on the assumption that the correct model has been chosen (Atkinson & Donev, 1992). For the goal of model selection, the scientist must compare a set of models that are already formalized, and the models often vary in terms of their emphasis on the mechanistic accuracy of the models (Box & Hill, 1967). The challenge of meeting both goals simultaneously has been addressed (e.g., Hill et al., 1968) but introduces the complex issue of the need for a process that emphasizes model selection early in a research program and parameter estimation later, once greater confidence in the choice of model has emerged.

The present proposal posits that rich stimulus sampling may provide an easy and efficient method of achieving both goals, especially in the earlier stages of theoretical development when formal mechanistic models do not yet exist. In the absence of extant models, a rich sampling technique may help map out the general shape of the relationship among the IVs and DVs and do so with sufficient accuracy of parameter estimates to provide utility. As models become more formalized later in the development of a scientific subdomain, there will need to be a greater emphasis on the types of formal value selection methods discussed elsewhere (Cavagnaro et al., 2010; Cavagnaro et al., 2011; Myung & Pitt, 2009). But the field may be slow to reach that stage if insufficient sampling of stimulus dimensions persists in the field of experimental psychology.

To investigate the impact of rich stimulus sampling on model selection and parameter estimation, our laboratory began with a series of Monte Carlo simulations involving the simplest possible experimental design. Each simulated subject received one level of one IV, and each subject generated one associated value of the DV. There were no other factors that varied between or within subjects. The goal was to identify the functional relationship between the IV and DV and the best-fitting parameter values for that function. In [Experiment 1](#), the data were generated on the basis of a noisy logistic relationship, but each simulated experiment involved different sample sizes, effect sizes, sampled range of the stimulus dimension, and stimulus sampling methods. In [Experiment 2](#), the data were generated on the basis of a noisy linear relationship to ensure that nothing was lost for those situations in which the relationship was truly linear.

Although the utility of the proposed method will importantly hinge on extending the present analyses to the

type of designs more commonly used in the behavioral sciences (involving multiple predictors or within-subjects independent variables), we believed that it was necessary to start with a simple design in order to identify the basic strengths or limitations of rich stimulus sampling without the complicating factor of additional complexity. Our simulations also focused on relatively large effect sizes. When the effect size is small, there is often insufficient data for the purposes of identifying any nonlinearities in a relationship, unless the sample size is very large or the nonlinearity is caused by floor or ceiling effects (e.g., in percentage or proportion data).

We used nonlinear curve fitting techniques based on specific functions, rather than transformations to linearity followed by linear analyses (a common approach) or generalized linear modeling and related methods (e.g., logistic regression). The rationale for this choice was to use a method that would generalize to any mathematical relation, not just those relations commonly present in current statistical packages (e.g., through generalized linear modeling's link functions) or achievable through standard transformations.

Experiment 1

Method

Procedure Four factors were manipulated in the creation of our hypothetical experiments. Of primary focus was the number of IV levels used; data generated either were systematically sampled from two, three, four, or N levels or were randomly (uniformly) sampled from a 100-point range. To simulate situations in which the researcher samples from a range that is suboptimal for ascertaining the true relationship, data sampled from the middle of the functional relationship, where there was maximal sensitivity of the DV to changes in the IV, were compared with data sampled from a high range, where sensitivity might be less optimal. Because changes in the variance of IV distributions affect parameter estimates (Cohen & Cohen, 1983), systematically sampled levels were chosen such that the population variance, σ^2 , of the IV in all cases would be approximately 833, thus matching the population variance of a continuous uniform distribution with a 100-point range. [Table 1](#) lists the specific values or range used for each sampling design. The remaining factors were two aspects of a data set commonly known to affect power: sample size and effect size. Sample sizes used were 12, 24, 48, or 60 simulated subjects; in cases of systematic sampling, values were split evenly between all sampled levels (e.g., for the three-level condition, a sample size of 12 would involve four values for each level). Population

Table 1 Sampled values of the independent variable (values for N -level sampling were rounded to the nearest integer for display only)

Sampling Method	Sampling Range	
	Complete	High
Two-level	22, 78	47, 103
Three-level	15, 50, 85	40, 75, 110
Four-level	11, 37, 63, 89	36, 62, 88, 114
N -level (example for sample size of 12)	6, 14, 22, 30, 38, 46, 54, 62, 70, 78, 86, 94	31, 39, 47, 55, 63, 71, 79, 87, 95, 103, 111, 119
Random	0 to 100	25 to 125

The N levels for N -level sampling in the complete condition were chosen by equally spacing the values between 5.96 and 100 minus 5.96 for $N = 12$, 3.05 and 96.95 for $N = 24$, 1.54 and 98.46 for $N = 48$, and 1.24 and 98.76 for $N = 60$. This method ensured that the variance was 833.33 for all sample sizes.

effect sizes (Cohen's f) of .58, .82, and .94 were tested. One thousand data sets were generated for each of the 120 experiment scenarios (5 sampling methods $\times$ 2 ranges sampled $\times$ 4 sample sizes $\times$ 3 effect sizes).

The true relationship in the simulated population was a logistic function of the form

$$y_{true} = \frac{1}{1 + e^{a(x-b)}}, \quad (1)$$

where a , the gradient, was assigned values of .025, .050, and .097 to produce effect sizes of .58, .82, and .94, respectively, and b , the point of maximal inflection (PMI), was held constant at 50. This form was chosen because of its common presence in psychology. Each simulated subject produced only one outcome value. In order to simulate variability in the data without producing illegal values for a sigmoid (i.e., y -values beyond its asymptotes), a logit transform was first performed on the generated values to linearize them. Gaussian error ($SD = 1.0$) was added to this transformed DV (Eq. 2), and the result was reverse-logit transformed back into a logistic (Eq. 3):

$$y_{logit} = \ln\left(\frac{y_{true}}{1 - y_{true}}\right) + \varepsilon \quad (2)$$

$$y = \frac{e^{y_{logit}}}{1 + e^{y_{logit}}}. \quad (3)$$

This method produces a noisy logistic curve with the heteroscedasticity and non-Gaussian error that characterize a sigmoid, thus avoiding illegal values. Figure 1 shows examples of data generated using $N = 48$ for each sampling method and effect size used. The N -level method is not shown, because the result looks very similar to that for random sampling with large sample sizes.

Using regression methods in the R statistical environment (*lm* for linear and *nls* for nonlinear; available at www.r-project.org), each data set was fit using two-parameter linear, exponential, and logistic models (Eqs. 4, 5, and 6, respectively) as plausible models that a researcher might have used to represent the relationship:

$$\hat{y} = a + bx \quad (4)$$

$$\hat{y} = ae^{-bx} \quad (5)$$

$$\hat{y} = \frac{1}{1 + e^{a(x-b)}}. \quad (6)$$

Although *nls* assumes homoscedasticity, the violation of this assumption inherent in these data sets was modeled using the *weights = varPower()* parameter option, which allows *nls* to model heterogeneous variance.

Outcome measures: Assessing model selection and parameter estimation The Akaike information criterion (AIC) was chosen as the evaluative measure of each model for each data set because it gauges not only goodness of fit, but also generalizability, since it punishes models that are more flexible (although it is worth noting that because all three models being compared here use the same number of degrees of freedom, the full benefits of using the AIC are not realized in these simulated experiments). AICs were calculated using the standard form:

$$AIC = -2LL + 2k, \quad (7)$$

where LL is the log likelihood of the fit and k is the number of degrees of freedom.

Fig. 1 Examples of data plots ($N = 48$) for two-level (first row), three-level (second row), four-level (third row), and random (fourth row) sampling methods at effect sizes of .58 (first column), .82 (second

column), and .94 (third column). The N -level sampling produces figures very similar to those for random sampling with this large sample size

Lower AIC values indicate better (and more generalizable) fits. However, rather than simply labeling the fit with the lowest AIC as the chosen model for a given data set and, thereby, losing any information about the similarities of said fits, the degree of evidence in favor of the correct model was derived using Akaike weights (Eqs. 8 and 9) (Wagenmakers & Farrell, 2004). The Akaike weight spans the 0 to 1 interval and is taken as a measure of discriminability between the fits, with .33 indicating that the three AICs were all very similar, values above .33 denoting evidence in favor of the correct logistic model and values below .33 signifying more evidence in favor of one or both of the alternatives. The following equations were used to compute the Akaike weight (what we will call the discriminability score for maximum explanatory clarity) for each of the three alternative generating functions:

$$\Delta_i AIC = AIC_i - \min AIC \quad (8)$$

$$\text{Discriminability}_i = \frac{e^{-\frac{1}{2}\Delta_i AIC}}{\sum_j e^{-\frac{1}{2}\Delta_j AIC}}, \quad (9)$$

where the negative difference between each model's AIC and the best AIC served as evidence in favor of each model.

We also analyzed the best-fitting parameter estimates for the logistic model for each data set to determine each method's ability to identify the correct parameter value for each model. Error was assessed by computing the absolute value of the difference between the true value and the estimated value. The PMI (b in Eq. 1) of the true function was always 50, and the true gradient (a in Eq. 1) varied as a function of effect size (gradient = .025, .050, or .097, respectively; see Fig. 1). The absolute value of the parameter errors required a logarithmic transform to produce a normal distribution of the DV for the ANOVA.

Results

The results are summarized in Fig. 2 and are collapsed across the conditions in which the sampling range was symmetrical about either an x -value of 50 or an x -value of 75 (see Table 1), because the differences due to this manipulation were very small and not particularly informative. With regard to the ability to identify the correct generating function using discriminability scores (left graph), (1) discriminability improved as the sample size increased and as the nonlinear nature of the logistic grew stronger (i.e., as effect size increased); (2) sampling only two levels of the predictor provided no ability to identify the correct function, an unsurprising result; (3) discriminability steadily increased as more levels were systematically sampled, increased even further when N equally spaced values or random sampling was used, and the benefits of increasing the number of values sampled were greater as the effect size and sample size increased; and (4) equal spacing was slightly superior to random sampling for small samples, but the two methods were very similar for larger samples.

These observations were confirmed using an ANOVA to compare discriminability values as a function of our IVs. Because of the large number of simulations and, hence, high statistical power, all main effects and interactions were statistically significant ($ps < .01$). For simplicity of presentation, we will focus on those terms involving variations in the sampling method; results are summarized as effect sizes in Table 2. The main effect of sampling method, $F(4, 119880) = 13,926.7$, reflected the increase in discriminability with increases in the levels sampled (.33, .46, .54, .59, and .57 for two, three, four, and N levels and random, respectively). The method $\times$ sample size interaction, $F(12, 119880) = 199.6$, reflects the stronger benefit for N -level and random sampling as the sample size increased (see Fig. 2a; planned contrasts revealed that N -level was superior to random only for the 12 and 24

Fig. 2 Discriminability and parameter errors as a function of key independent variables for Experiment 1

Table 2 Effect sizes for the main effect of sampling method and the two-way interactions involving sampling method for the analyses of discriminability scores and parameter errors

Term	Experiment 1–Cohen's <i>f</i>			Experiment 2–Cohen's <i>f</i>		
	Disc	PMI Error	Gradient Error	Disc	Intercept Error	Slope Error
Method	.68	.05	.06	.23	.06	.07
Method × sample size	.14	.01	.05	.05	n.s.	n.s.
Method × range	.06	.02	.03	.13	.04	n.s.
Method × effect size	.64	.03	.05	.37	n.s.	n.s.

“n.s.” indicates an effect size that was not significantly different from zero.

sample sizes). Although discriminability was lower when we sampled only the upper end of the stimulus range, a significant method × range interaction, $F(4, 119880) = 120.6$, reflects a smaller effect of sampling range for the two-level method (of little interest because neither method could identify the correct function) and for the four-level sampling method; all differences were very small, .03 or less. For the final two-way interaction, method × effect size, the differences among the sampling methods were much larger for the most nonlinear logistic functions than for the weakest and nearly linear relationships, $F(8, 119880) = 6,210.2$ (see Fig. 1 for examples of these relationships and Fig. 2a for the results). *N*-level sampling produced slightly higher discrimination scores than did random sampling for the two strongest effect sizes: .51 vs. .50 for the .82 effect size and .90 vs. .88 for the .94 effect size. Although the three- and four-way interactions were also significant, the effects were much weaker and represented issues of small and less interesting quantitative differences, rather than qualitative ones. Details are available from the first author.

The second and third graphs of Fig. 2 reveal the obvious result that larger samples generate smaller parameter errors. Figure 2b (PMI errors) shows that a sharper gradient (generated by bigger effect sizes) makes it easier to identify the correct PMI, an unsurprising result given the nature of logistic functions where the location of the PMI is easier to see when the change in inflection is more abrupt. Figure 2c (gradient errors), however, reveals that bigger effect sizes are associated with larger errors in estimates of the gradient. This finding is also not surprising, because very similar curves generate very different gradient values at the upper end of the gradient scale but very similar gradient values at the lower end of the scale.

More important, the two-level systematic sampling approach tended to generate the smallest errors for both parameters, presumably because there are more samples of the two values tested, and the random sampling approach tended to generate the largest errors (with the three-, four-, and *N*-level approaches often similar to the random sampling approach). This result must be qualified by noting that with one exception, the increase in parameter error was of little practical significance (0%–5%). The exception was when there was a very small sample of the function that had

the sharpest gradient (upper right portion of Fig. 2c). Here, there was a sizable increase in parameter error with each step away from two-level systematic sampling, with an approximately 100% increase in the error of the gradient estimate for the random sampling approach relative to the two-level approach for the smallest sample size ($N = 12$) and a 40%–50% increase in error for the next smallest sample size ($N = 24$). Smoothed density functions for the discriminability scores and the estimates of the 0.097 gradient are shown in Fig. 3 for some critical conditions. The *N*-level and random sampling approaches created a strong shift toward better model selection for the most nonlinear functions (effect size = .94), but with an increase in the upper tail of the distribution of estimated gradients (there was no such effect on PMI estimates; not shown). The *N*-level and random sampling density curves are indistinguishable for both discriminability scores and gradient error, except for the 24-sample estimates of the gradient, where random sampling generated more variable gradient estimates.

All approaches tended to underestimate the gradient value by approximately 15% ($Mdns = .021, .043$, and $.085$ for true values of $.025, .050$, and $.097$). This result is almost certainly a product of our ignoring the non-Gaussian nature of the error for a logistic function. Figure 1 highlights the fact that the error distribution is skewed away from the extremes of the DV and, thus, biases the analysis toward lower gradient estimates.

These observations were confirmed using ANOVAs of $\log(\text{error})$ for each parameter as a function of our IVs. Unlike the results for discriminability, not all effects were significant, but we will still focus on those terms involving variations in the sampling method. Each parameter will be considered separately.

For the PMI parameter, the main effect of sampling method, $F(4, 119880) = 69.4$, reflected lower error for the two-level approach and higher error for the random sampling approach as evaluated using a Tukey's HSD test (2.38, 2.67, 2.66, 2.70, and 2.82, for two, three, four, and *N* levels, and random, respectively; $SE = 0.02$). The method × sample size interaction, $F(12, 119880) = 2.0$, was evidenced in a slightly greater sensitivity to sample size as the number of levels sampled increased (see Fig. 2b). The method × sampling

Fig. 3 Top: Probability densities for the largest effect size for discriminability scores. Bottom: Estimated gradients for the largest effect size conditions ($f = 0.94$; true gradient = .097) for Experiment 1.

Sample size increases from left to right ($n = 24, 48$, and 60 are shown), and the five densities in each figure are for the five sampling methods. The random (R) and N lines are largely indistinguishable

range interaction, $F(4, 119880) = 15.6$, was evidenced in a small unusual result in which three-level sampling was unaffected by sampling range, whereas the others had higher error when the range was shifted to the right. For the final two-way interaction, method $\times$ effect size, the differences among the sampling methods were slightly larger for the most nonlinear logistic functions than for the weakest and nearly linear relationship, $F(8, 119880) = 14.2$ (see Fig. 2b). The three- and four-way interactions either did not reach statistical significance or produced small quantitative differences of little practical significance ($F_s < 2.6$). Details are available from the first author.

For the gradient parameter, the main effect of sampling method, $F(4, 119880) = 121.7$, reflected lower error when fewer levels were sampled, as evaluated using a Tukey's HSD test (0.0059, 0.0063, 0.0067, 0.0068, and 0.0069 for two, three, four, N, and random, respectively). The method $\times$ sample size interaction, $F(12, 119880) = 30.6$, reflects a much larger increase in error for the $n = 12$ samples as the number of levels sampled increased than for the larger samples (see Fig. 2c). The method $\times$ sampling range interaction, $F(4, 119880) = 26.9$, revealed slightly more parameter error (about 10%–15% more) for the centralized

range than when the range was shifted upward, but not for the three-level sampling, which was unaffected by the shift. For the final two-way interaction, method $\times$ effect size, the differences among the sampling methods were much larger for the most nonlinear logistic functions but quite small for the weakest and nearly linear relationship, $F(8, 119880) = 42.4$ (see Fig. 2c). A critical three-way interaction, method $\times$ sample size $\times$ effect size, $F(24, 119880) = 6.6$, is clearly illustrated in Fig. 2c; gradient error was much higher as the number of levels sampled increased for the most nonlinear relationship when the sample size was very small. Further details are available from the first author.

Discussion

Our simulations of a logistic relationship suggest that random sampling of the IV can produce substantial improvements in the accuracy of model selection, especially for highly nonlinear relations (see Figs. 2 and 3). The nature of random sampling creates the possibility of producing highly nonrepresentative samples if the sample size is too small. This small-sampling problem was especially problematic for the most nonlinear relationships (effect

size of .97) when estimating the logistic gradient, as shown in Fig. 2c. Here, the gradient is shown to be highly sensitive to unusual values in the sampling distribution. However, the possibility of an unusual sample was not the entire reason behind the poor gradient estimation of random sampling for small samples, because the gradient errors systematically increased as the number of levels sampled was increased. It is likely that increasing the number of levels sampled leads to unnecessary sampling of the extremes of the function (the left and right ends of the logistic), which does not help inform the estimates of the gradient. However, sampling of the extremes is not truly wasted, because these values help to identify the logistic nature of the relationship (see Figs. 1 and 2a). Because of this outcome, we recommend that any researcher considering the use of random stimulus sampling for a single, between-subjects stimulus variable sample at least 24 levels of the variable.

Although these results were encouraging, we needed to verify that the random stimulus sampling approach would not have negative consequences when the data were indeed linear. It is possible that stimulus sampling might be more likely to falsely identify nonlinear relationships in general. Thus, Experiment 2 used a linear generating function, while still attempting to fit linear, logistic, and exponential functions.

Experiment 2

Method

In Experiment 2, the same values for all IVs were used as in Experiment 1; however, the true relationship was a linear function of the form

$$y_{true} = a + bx, \quad (9)$$

where a , the intercept with the y -axis, was held constant at 1.0 and b , the slope, was given values of $-.0080$, $-.0045$,

and $-.0021$ to achieve the desired effect sizes. Unlike in Experiment 1, illegal values were not a concern here, so Gaussian error ($SD = .086019$ to maintain the specified effect sizes) was added directly to this raw score to produce a noisy linear relationship that conformed to the assumptions of simple regression. As in Experiment 1, 1,000 data sets were generated and fit with Eqs. 4, 5, and 6. The resulting AICs from these fits were again used to calculate a discriminability score for each set according to Eqs. 8 and 9. Data regarding the error of the parameter estimates were also collected and transformed, using the method described for Experiment 1. Example data for a random sample of 60 are shown in Fig. 4; the best-fit linear, exponential, and logistic functions are superimposed.

Results

The results are summarized in Fig. 5 and are collapsed across the conditions in which the sampling was symmetrical about either an x -value of 50 or an x -value of 75. With regard to the ability to identify the correct generating function using discriminability scores (left graph), (1) discriminability was much weaker for Experiment 2, as compared with Experiment 1, evidence that the chosen nonlinear functions can effectively mimic linear functions under some conditions (see Fig. 4 for an example); (2) discriminability among the functions was much stronger for the highest sloped line, where there also was a steady increase as a function of the sample size and number of levels sampled (see also Fig. 4); and (3) there was a paradoxical result in which smaller samples generated better function discriminability for the shallowest linear relationship. This oddity was due to small samples creating problems with fitting the logistic and, thus, generating much weaker model fits; thus, only the linear and exponential functions were viable candidates for many of these small-sample, shallow-slope runs.

Fig. 4 Examples of data plots ($N = 60$) for the random sampling method at effect sizes of .58 (first column), .82 (second column), and .94 (third column). The best-fit linear, exponential, and logistic functions are superimposed

Fig. 5 Discriminability and parameter errors as a function of key independent variables for Experiment 2

These observations were confirmed using an ANOVA of discriminability values as a function of our IVs. Because of the large number of simulations and, hence, high statistical power, all main effects and interactions were statistically significant ($ps < .01$). For simplicity of presentation, we again will focus on those terms involving variations in the sampling method; results are summarized as effect sizes in Table 2. The main effect of sampling method, $F(4, 119880) = 1670.1$, reflected the small increase in discriminability with an increase in levels sampled (.36, .43, .43, .43, and .44 for two, three, four, N , and random, respectively). The method $\times$ sample size interaction, $F(12, 119880) = 24.4$, reflected very small and idiosyncratic differences. Discriminability was higher when we sampled only the upper end of the stimulus range [method $\times$ range, $F(4, 119880) = 538.8$], but this interaction was largely produced by no effect of sampling range for the two-level method (of little interest, because neither method could identify the correct function). For the final two-way interaction, method $\times$ effect size, the differences among the sampling methods was larger for the steepest linear relationship than for the shallowest one, $F(8, 119880) = 2042.4$ (see Figs. 4 and 5a). Although the three- and four-way interactions were also significant, the effects were relatively small and represented issues of less theoretical interest. Details are available from the first author.

The second and third graphs of Fig. 5 reveal the obvious result that larger samples generate smaller parameter errors and that sampling more levels generated a small increase in parameter error. The linear function well estimated the true intercept and slope without bias and, thus, did not show the skew present in Experiment 1 that was caused by the non-Gaussian error distributions (cf. Figs. 3 and 6).

These observations were confirmed using ANOVAs of $\log(\text{error})$ for each parameter as a function of our IVs. Unlike the results for discriminability, not all effects were significant, but we will still focus on those terms involving variations in the sampling method. Each parameter will be considered separately.

For the intercept parameter, the main effect of sampling method, $F(4, 119880) = 101.9$, reflected lower error for the two-level approach and higher error for the random sampling approach, as evaluated using a Tukey's HSD test (.020, .022, .023, .024, and .024 for two, three, four, N , and random, respectively; $SE = 0.007$). The only significant interaction involving sampling method was the method $\times$ sampling range interaction, $F(4, 119880) = 41.3$, which revealed an increase in intercept error when the range was shifted to the right, but more so for the two-level method (29% higher) than for the others (13%–15% higher).

For the slope parameter, the main effect of sampling method, $F(4, 119880) = 168.7$, reflected higher error when more levels were sampled (0.00030, 0.00032, 0.00032, 0.00034, and .00034 for two, three, four, N , and random, respectively). The remaining interactions either were not significant or generated even smaller quantitative differences than those for the sampling method main effect. Figure 5c accurately reflects the very small magnitude of any effects of sampling method on slope estimation for our linear functions.

Smoothed density functions for the discriminability scores and the estimates of the intercept for the steepest function are shown in Fig. 6. For the larger sample sizes, the N -level and random sampling density curves are indistinguishable. The rich sampling approaches created a clear shift toward better model selection for these conditions at the cost of a small increase in variability in the intercept parameter estimates (there was no such effect on slope estimates; not shown).

Discussion

When the data conformed to a linear relationship, the benefits of random stimulus sampling were less evident for model selection, because our nonlinear functions could mimic linear relationships in the range tested. Indeed, we specifically avoided making the task too easy by not

Fig. 6 Top: Probability densities for largest effect size for discriminability scores. Bottom: Estimated intercepts for the largest effect size conditions ($f = 0.94$) for Experiment 2. Sample size increases from left

to right ($n = 24, 48$, and 60 are shown), and the five densities in each figure are for the five sampling methods. The random (R) and N lines are largely indistinguishable

allowing the linear generating function to produce values much beyond the range of a logistic (below 0 or above 1) or exponential (below 0) function. Our assumption was that any astute researcher should be able to eliminate functions that could not fit their data due to the restricted y -ranges possible with some functions. Importantly, when the slope of the line was steepest, the random sampling approach did improve the identification of the correct linear generating function, with a small cost in intercept estimation and no apparent cost in slope estimation.

General discussion

In order to identify the true relationship between continuous IVs and DVs, sampling only two to three levels of the IV is insufficient unless one already knows the nature of the function and is merely attempting to identify the best-fitting parameter values of that function. Our simulations of a simple between-subjects design suggest that sampling more levels results in better identification of the function at the sacrifice of precision in estimating the parameters of that function and that random stimulus sampling provides benefits equivalent to N -level sampling for larger sample

sizes. Although we tested only a small number of functions, we hope that the results will generalize to many other functional relationships, although the ability to discriminate among very similar functions (e.g., power and exponential) would likely require much larger samples.

The simulations suggest that the sacrifice in parameter precision is quite small relative to the large gain in model selection accuracy, at least for the models tested (see Table 2). Although small, any loss of precision in parameter estimates may make it more difficult to identify group or condition differences in those parameters (e.g., the gradient of sensitivity may be larger for some conditions than for others). If the wrong function has been identified, however, this loss of precision may be moot. An attempt to compare conditions on the basis of the wrong functional form would produce significant problems that would overwhelm any loss in precision. In future simulation work, we will examine the trade-off between correct model selection and precise parameter estimation when a categorical IV is incorporated into the design.

Given the occasional small differences between random and N -level sampling and given that those differences favor the latter, why would an investigator consider random stimulus sampling rather than N -level sampling? First,

using N -level sampling restricts the investigator to predetermined sample sizes and to spacing the values accordingly. If one chose to run a few more subjects or needed to drop a few due to subject inattention or experimenter error, some values along the nicely spaced continuum would be systematically oversampled or missing. Furthermore, N -level sampling would still require randomly ordering the presentation of the N values across time as each subject enters the laboratory, to ensure that sampling level is not confounded with time. Thus, as long as investigators are obtaining a sufficiently large sample, random stimulus sampling provides the same advantage as random subject sampling (generalizability to a larger population), while also offering the same benefits and minor costs of N -level sampling.

Analysis challenges

Random stimulus sampling and N -level sampling will create some analysis challenges for many research scientists. For example, given that these sampling approaches may produce no replications, the ability to test statistical assumptions (such as homogeneity and normality in the error distribution) becomes more complicated but not untenable. For instance, one method of testing for heteroscedasticity in regression involves identifying any relationship between the absolute value of the residuals and the predicted value of y (Cohen & Cohen, 1983). Likewise, non-Gaussian error distributions are often revealed by a simple frequency histogram of the residuals. Thus, these complications in the use of random sampling of the predictor values are not insurmountable but require greater comfort with the many flavors of regression, rather than the pro forma use of an ANOVA.

Another set of challenges involves visualizing the nature of the relationship between a richly sampled continuous variable and a DV. For simple designs like the one used here, spline fits can provide a graphical representation of the function. As design complexity increases, multidimensional spline fitting (e.g., generalized additive modeling), neural networks, and related techniques can be used as visualization tools. However, visualization must be followed by choosing candidate functions for subsequent model selection. Although the exponential family of functions are familiar to some researchers, along with methods for choosing among them (e.g., Shull, 1991), there exists a considerable palette of options from which to choose. A mathematical or statistical consultant may be necessary to identify the most promising candidates.

Once a set of candidate functions is identified, a final statistical challenge must be addressed: nonlinear fitting of the function. In some cases, a nonlinear relationship can be captured using linear techniques that include transformations of the IVs or DVs, although care must be taken to ensure that

other properties of the data (e.g., homogeneity of variance) are not adversely affected by the transformation. In other cases, a nonlinear fit can be achieved by specifying nonnormal error distributions, using methods like generalized linear modeling. Finally, it may be necessary to learn nonlinear curve fitting techniques to best estimate the parameter values of the proposed functions.

Extending our results

A necessary next step in our research program is to determine how our results generalize to more complex designs, especially those involving repeated measures. Although there are natural limits to how many levels of each IV can be tested for an individual subject, modern statistical techniques such as multilevel modeling (Gelman & Hill, 2006; Pinheiro & Bates, 2004; Young, Clark, Goffus, & Hoane, 2009) can effectively aggregate at both the individual and group levels in order to increase statistical power for model comparison and parameter estimation. These techniques also ensure that each subject's contribution to identifying functions and estimating parameters will be appropriately impacted by the quality of the subject's data, both its variability and its completeness.

The suitability of random stimulus sampling depends on the limitations inherent in one's research paradigm. For example, scientists who work with animals often are limited to running a small number of subjects, which suggests that randomly sampling a between-subjects variable in these studies would be problematic. However, laboratory animals sometimes can be tested for many trials per day for many weeks, thus making it easier to sample a large number of values for a within-subjects variable. The converse is often true for scientists working with human subjects, where large samples of people are easily obtained but each subject is tested for a relatively short period of time. Thus, random sampling of a between-subjects variable may be more valid than random sampling of a within-subjects variable. Addressing these questions, however, will require an additional set of simulations for mixed (between–within) designs.

Although the present experiments examined only uniform random stimulus sampling, the method is readily modified to oversample some ranges (e.g., through Gaussian random sampling) or to systematically sample key values along the continuum, especially at an endpoint, while randomly sampling other values. For example, judgment and decision-making researchers know that “free” appears to be a special value that prompts categorically different behavior (Shampanier, Mazar, & Ariely, 2007), and random sampling may never sample this particular value along the continuum. Thus, a researcher could opt to sample the no-cost option for 20% of the subjects but use

random sampling for the remaining subjects. Without further simulations showing the impact of using other types of random sampling, however, investigators should not automatically assume that our results necessarily translate to nonuniform random sampling methods.

Finally, we must acknowledge that a disturbingly large portion of the data collected in psychology is so noisy that it may prove untenable to identify any deviations from a linear relationship. This variability may be one of the reasons that the field has not been previously motivated to encourage the use of nonlinear functional forms. Common exceptions lie in those domains in which greater control can be established or behavior is more consistent (e.g., in the field of perception) or obvious asymptotic behavior is observed that creates clear deviations from linearity.

Transforming behavioral science

Although some areas of psychology have made significant progress toward identifying the nature of the functional relationship between IVs and DVs, the vast majority of psychology's domains have not. Thus, it is difficult to argue that parameter estimation is the paramount problem to be solved; rather, most psychology experiments should be better designed to identify the functional forms that might be in play. Achieving this goal will present its own analytical challenges, because behavioral scientists will need to be better trained to use multidimensional spline fitting, nonlinear modeling, generalized linear modeling, and related approaches in order to capture the full range of possible functional relationships that one may encounter. Rather than a daunting task, we believe that such an approach will inject excitement into the field and allow for leaps forward in the development of theory. Theories that predict only that two conditions will differ will never be as compelling as those that can predict the precise nature of the functional relationship among the variables under investigation (Roberts & Pashler, 2000).

References

- Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. New York: Oxford University Press.
- Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, 19, 43–56.
- Berger, M. P. F., King, C. Y. J., & Wong, W. K. (2000). Minimax D-optimal designs for item response theory models. *Psychometrika*, 65, 377–390.
- Box, G. E. P., & Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9, 57–71.
- Brunswick, E. (1947). *Systematic and representative design of psychological experiments; with results in physical and social perception*. Berkeley: University of California Press.
- Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.
- Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22, 887–905.
- Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review*, 18, 204–210.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.
- Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130, 959–988.
- Ford, I., Titterton, D. M., & Kitsos, C. P. (1989). Recent advances in nonlinear design. *Technometrics*, 31, 49–60.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Hill, W. J., Hunter, W. G., & Wichern, D. W. (1968). A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics*, 10, 145–160.
- Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40, 366–371.
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Attention, Perception, & Psychophysics*, 63, 1279–1292.
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113, 181–190.
- McKerchar, T. L., Green, L., Myerson, J., Pickford, T. S., Hill, J. C., & Stout, S. C. (2009). A comparison of four models of delay discounting in humans. *Behavioural Processes*, 81, 256–259.
- Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116, 499–518.
- Passos, V. L., & Berger, M. P. F. (2004). Maximin calibration designs for the nominal response model: An empirical evaluation. *Applied Psychological Measurement*, 28, 72–87.
- Pinheiro, J. C., & Bates, D. M. (2004). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358–367.
- Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, 12, 195–223.
- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103, 734–760.
- Shampanier, K., Mazar, N., & Ariely, D. (2007). Zero as a special price: The true value of free products. *Marketing Science*, 26, 742–757.
- Shull, R. L. (1991). Mathematical description of operant behavior: An introduction. In I. H. Iversen & K. A. Lattal (Eds.), *Experimental analysis of behavior* (Vol. 2, pp. 243–282). New York: Elsevier.
- Steinberg, D. M., & Hunter, W. G. (1984). Experimental design: Review and comment. *Technometrics*, 26, 71–97.
- Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, 7, 424–465.
- Vermeulen, B., Goos, P., & Vandebroek, M. (2008). Models and optimal designs for conjoint choice experiments including a no-choice option. *International Journal of Research in Marketing*, 25, 94–103.
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11, 192–196.
- Young, M. E., Clark, M. H., Goffus, A., & Hoane, M. R. (2009). Mixed effects modeling of Morris water maze data: Advantages and cautionary notes. *Learning and Motivation*, 40, 160–177.