THE MARKETING JOURNAL OF THE INSTITUTE FOR OPERATIONS RESEARCH AND THE MANAGEMENT SCIENCES

informs.

MARKETING SCIENCE

Self-Signaling and Prosocial Behavior: A Cause Marketing Experiment
Jean-Pierre Dubé, Xueming Luo, Zheng Fang
161

Signaling Virtue: Charitable Behavior Under Consumer Elective Pricing
Minah H. Jung, Leif D. Nelson, Uri Gneezy, Ayelet Gneezy
187

A Cross-Cohort Changepoint Model for Customer-Base Analysis
Arun Gopalakrishnan, Eric T. Bradlow, Peter S. Fader
195

Is Advance Selling Desirable with Competition?
Gérard P. Cachon, Pnina Feldman
214

Can Retail Sales Volatility be Curbed Through Marketing Actions?
Mercedes Esteban-Bravo, Jose M. Vidal-Sanz, Gökhan Yildirim
232

The Effects of Product Line Breadth: Evidence from the Automotive Industry
Antonio Moreno, Christian Terwiesch
254

Entry of Copycats of Luxury Brands
Sarah Yini Gao, Wei Shi Lim, Christopher S. Tang
272

"Ten Million Readers Can't be Wrong!," or Can They?
On the Role of Information About Adoption Stock in New Product Trial
Coby Morvinski, On Amir, Eitan Muller
290

Could Good Intentions Backfire? An Empirical Analysis of the Bank Deposit Insurance
Weining Bao, Jian Ni
301

Focus on Authors
320

2016 Guest Editors, Guest Associate Editors, and Ad Hoc Reviewers
324

Volume 36 • Number 2 • March–April 2017

Marketing Science

When Random Assignment Is Not Enough: Accounting for Item Selectivity in Experimental Research

Fred M. Feinberg, Linda Court Salisbury, Yuanping Ying

MARKETING SCIENCE
Vol. 35, No. 6, November–December 2016, pp. 976–994
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

https://doi.org/10.1287/mksc.2016.0991

When Random Assignment Is Not Enough: Accounting for Item Selectivity in Experimental Research

Fred M. Feinberg
Ross School of Business and Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, feinf@umich.edu

Linda Court Salisbury
Carroll School of Management, Boston College, Chestnut Hill, Massachusetts 02467, salisbli@bc.edu

Yuanping Ying
Yum! Brands, Plano, Texas 75024, yuanping.ying@yum.com

Experimental methods are critical tools in marketing, psychology, and economics to isolate the effects of key variables from vagaries intrinsic to field data. As such, they are often considered exempt from the sort of sample selectivity artifacts widely documented in empirical research, in part because participants are randomly assigned to experimental conditions. To conserve time and resources, experiments often focus on items participants have chosen or are familiar with, for example, postchoice satisfaction ratings, certain free recall tasks, or specifying consideration sets preceding brand choice. When consumer input even partially influences the items about which researchers request subsequent data, the potential for *item selectivity* arises. In such situations, analyses are contingent on both the choice context(s) of the experiment and the alternatives participants elect to evaluate, potentially leading to substantial item selectivity overall and to differing degrees across conditions. We examine situations in which a nonignorable "choose one of many" (polytomous) selection process limits which items offer up subsequent information, and develop methods to allow substantive results to pertain to the full set of items, not only those selected. The framework is illustrated via two experiments in which participants choose and then evaluate a frequently purchased consumer good as well as data first examined by Ratner et al. [Ratner RK, Kahn BE, Kahneman D (1999) Choosing less-preferred experiences for the sake of variety. *J. Consumer Res.* 26(1):1–15]. Results indicate substantial item selectivity that, when corrected for, can lead to markedly different interpretations of focal variable effects, such as large effect size changes and even sign reversal. Moreover, failing to flexibly account for item selectivity *across experimental conditions*, even in well-designed experimental settings, can lead to inaccurate substantive inferences about consumers' evaluative criteria. We further demonstrate robustness to theoretically driven (but not overtly misspecified) selection rules and provide researchers with a simple, "two-step" exploratory procedure akin to a "control function" approach—involving just one additional variable added to standard models—to determine whether and to what degree item selectivity may be affecting their substantive results.

Data, as supplemental material, are available at https://doi.org/10.1287/mksc.2016.0991.

Keywords: choice models; consumer behavior; decision making; econometric models; sample selection; Heckman model; Markov chain Monte Carlo; hierarchical Bayes; variety seeking; assortment size
History: Received: December 13, 2014; accepted: February 1, 2016; Preyas Desai served as the editor-in-chief and Eric Bradlow served as associate editor for this article. Published online in *Articles in Advance* August 16, 2016.

Introduction

Field research faces the formidable task of making valid inferences from samples that are often far from random. People who fill out surveys, firms that are publicly traded, and industries characterized by heavy advertising are seldom representative of the larger groups that marketers, economists, and social scientists wish to understand. Making such inferences relies on statistical methods that correct for a sample's being *selected* in some manner. Such methods date back to the classic articles of Tobin (1958) and Heckman (1979), and have become core tools in cognate disciplines (e.g., Winship and Mare 1992) as well as economics proper (Heckman 1990, Puhani 2000).

A similar path was forged in quantitative marketing, where selectivity on which consumers participate or provide key information has been addressed in a wide variety of settings, including models accounting for panel attrition (Danaher 2002) and underreporting (Yang et al. 2010), assessing long-run promotion effects using recency, frequency, and monetary (RFM) variables (Anderson and Simester 2004), of online banking decisions conditional on having signed up and logged in (Lambrecht et al. 2011), for correlations in incidence and strength of online product opinions (Moe and Schweidel 2012, Ying et al. 2006) or ad impressions and visit behavior in online display advertising (Braun and Moe 2013), to interrelate content creation and purchase decisions (Albuquerque et al. 2012), and culminating in Wachtel and Otter's (2013) comprehensive framework to account for multiple waves of selectivity (e.g., scoring and targeting) enacted deliberately by marketers.

Two distinct sorts of selectivity, one common and widely addressed, the other the subject of the present article, are often unwittingly conflated. The common form is where *participants* (consumers or respondents) self-select into "conditions," that is, those that buy a specific car, see a target film, or subscribe to a particular website. Random assignment (of participants) into experimental settings is often taken as sufficient to control for selectivity artifacts that can arise. While in many cases it does, in others it does not. A second, distinct type of selectivity can and does occur, in both field and experimental data, one that random assignment alone cannot fully overcome. It occurs when researchers observe feedback—ratings, evaluations, or any sort of contingent information—about *items* whose selection is even partly influenced by consumer input. In this case, then the potential for *item selectivity* arises. For example, if participants are asked to evaluate items with which they are most familiar (i.e., items compete with one another on familiarity for inclusion), then selectivity takes place on characteristics those items share—characteristics that can never be fully captured by researchers—potentially altering the estimated effects of experimental manipulations. Moreover, the *degree* of selectivity exhibited by a consumer can vary, depending on the manipulation used in the condition into which she has randomly been assigned. The point we hope to make is that random assignment alone does not ensure that experimental results are free from selectivity artifacts, so long as the information provided by consumers is in any way contingent on overt (i.e., explicitly required by the experiment) or covert (allowed in the experiment) choices they have made. In our empirical illustrations, we have limited our purview for comparability purposes to one specific but widely used paradigm, where (randomly assigned to conditions) consumers make a series of choices and then provide feedback on each. Yet the general point applies to any experiment where the participants are not required to provide an equal degree of feedback on all items used in the study, or where the items evaluated are themselves not selected randomly by the researcher. We find striking evidence of selectivity artifacts in our empirical applications, including some large changes in key effect sizes and even effect sign reversal.

The method we introduce later falls under the rubric of a *nonignorable missing data* mechanism: evaluative data are "missing" on items participants did not select. A review of such methods is beyond the scope of this article and have been treated extensively for survey research by, for example, Rubin (2004, Chapter 6). Specifically, the framework developed by Heckman (1979) can be viewed as "bivariate normal stochastic censoring" (see Little and Rubin 2002, p. 322; Enders 2010, p. 293), where values on one latent variable ("selection" utility) determine the likelihood of observing an outcome ("prediction," e.g., postconsumption satisfaction).

A hallmark of nonignorable selectivity artifacts is culling the set based on anything related to the dependent variable *over and above* what is attributable to (available) covariates. As described nontechnically in the review article of Schafer and Graham (2002, p. 151), nonignorability stems from the data being "missing not at random" (MNAR; contrasted with "missing at random" (MAR), which is generally "ignorable"). Specifically, given covariates X and dependent variable Y, "under MAR, there could be a relationship between missingness and Y induced by their mutual relationships to X, but there must be no residual relationship between them once X is taken into account. Under MNAR, some residual dependence between missingness and Y remains after accounting for X."[1] This is rarely if ever the case in behavioral research: specifically, which *items* a subject chooses cannot be presumed unrelated to her eventual evaluation of those items, even after available covariate effects have been regressed out. This will be true regardless of whether subjects are randomly assigned to conditions. In our experiments, this lack of relation is provably false: in each data setting, there is significant and meaningful residual explanatory power in formally "linking" the selection and evaluation processes. Zanutto and Bradlow (2006) provide an extended discussion of ignorability in marketing applications in the context of "data pruning"—removing alternatives, observations, households, etc., from the data—including an example of parameter bias based on (nonignorable) selection of top brands. Using their framework, Andrews and Currim (2005) conduct a detailed simulation study for scanner data applications, verifying substantial parameter bias and suggesting "best practices" for empirical research, but do not provide a statistical approach for correcting for selectivity artifacts, our main goal here.

The classic Heckman (1979) model allows a researcher to use a *nonrandomly selected sample of individuals* to make inferences about the entire population.

[1] For additional information regarding selection (of participants) on observables versus unobservables, see Bronnenberg et al. (2010, p. 1006).

By way of analogy, our framework uses information about *items consumers select from some set* to make inferences about *all* items in that set. This distinction has been noted in educational testing: Wainer et al. (1994), for example, pointed out that if students can at least partly select which questions to answer on a multiquestion exam, the unanswered ones are *nonignorable nonresponses*; a student's success on answered questions cannot be simply extrapolated to the others.[2] Wainer and Thissen (1994, p. 159), in a general review of "examinee choice" in testing, highlight the difficulty of comparing performance on self-selected activities, asking, "Is your French better than my calculus?" That is, what students have *not* selected (or avoided) is nonignorable. We stress here and subsequently that if a researcher wishes to make inferences about *selected items only*—e.g., how price affects ratings of *only* items you purchased, as opposed to all items available—the present modeling framework may be of limited benefit. Yet, when the missing data (evaluations of items consumers did *not* choose) are nonignorable, our modeling framework allows researchers to establish, measure, and correct for parametric estimation artifacts stemming from item selectivity in the wider scope of understanding choice processes, consumer evaluations, and the full array of available products.

In the Heckman (1979) framework, sample inclusion is binary: something is either "in or out," evaluated for inclusion *irrespective of other available items.* This is justified when items (or individuals) are considered on their own merits alone, e.g., having a positive net profit for being sent a catalog is not affected by how many other customers "make the cut" (Bult and Wansbeek 1995). In most marketing contexts, however, items *do* compete for inclusion. Consider choosing an entrée from a menu. Whether the restaurant is of high quality (many entrées are appealing) or poor, we do not choose multiple items in the first case or zero in the second, but one in each. We stress here that item selectivity can occur even when "selection" is not overt: for example, participants provide more *informative* evaluations for items with which they are knowledgeable or fervent.[3] Such selection or screening processes are frequently a critical, if unheralded, part of experimental research in marketing and decision theory. For example, research participants are often instructed to choose from a set of alternatives before reporting relevant measures about the chosen item(s), such as satisfaction (e.g., Diehl and Poynor 2010, Gu et al. 2013, Litt and Tormala 2010, Ratner et al. 1999) or willingness to pay (e.g., Carmon et al. 2003, Pham and Chang 2010). These diverse experimental settings share one crucial point of commonality: consumer input determines which items "survive" to be reported on.

Our main goals in this article, therefore, are to introduce and explore the notion of item selectivity; to develop models more directly applicable to the type of "choose one from many" item selectivity problems typically encountered in marketing, choice theory, and empirical behavioral research; to extend these models to allow for varying degrees of selectivity across experimental conditions; to provide "exact" Bayesian methods for their estimation; to provide a simple, two-step approximation to these methods to enable exploratory data analysis (EDA) for detecting item selectivity; to present varied empirical evidence of their importance; and to demonstrate a sufficient degree of substantive robustness to the specification of the selection mechanism.

We start by briefly reviewing standard selectivity models and then showing how to extend and estimate them. The importance of carefully modeling selectivity is then demonstrated in three data settings: one a new analysis of data examined in Ratner et al. (1999) and two involving postchoice satisfaction for a frequently purchased consumer good. In the online appendix (available as supplemental material at https://doi.org/10.1287/mksc.2016.0991), we use a RESET-based testing procedure (Peters 2000) to rule out omitted regressor bias for our results, demonstrate the importance of accounting for observed preference heterogeneity in our model specification, and test for evidence of error covariance (finding none). We also develop an approximate method for behavioral researchers engaged in EDA, requiring only the addition of one simple new covariate to their analyses.

Binary Selection and Extension to Polytomies

Heckman's (1979) original model for selectivity can be written as a two-equation system

$$Y_s = X_s\beta_s + \varepsilon_s, \tag{1}$$

$$Y_p = X_p\beta_p + \sigma\varepsilon_p \quad \text{if } Y_s > 0, \tag{2}$$

$$(\varepsilon_s, \varepsilon_p) \sim N[0, 0, 1, 1, \rho]. \tag{3}$$

The outcome or "prediction" variable, Y_p, is observed only in cases where the (binary) selection variable, Y_s,

[2] This line of work addresses how *which items one selects* (an observable) is informative about overall capability (test performance), with methodology focused on rescoring methods, via item response theory, to correct for selectivity stemming from test designs that allow examinees to choose the question(s) they answer. Here, we focus on how the *unmodeled* portion of selectivity (a latent residual) can improve predictions (of postchoice evaluation). We later assess the informativeness of this unmodeled part of selectivity.

[3] Bradlow and Zaslavsky (1999) provide one of the first examinations of this phenomenon, using hierarchical Bayesian techniques. In their data, "no answer" on a satisfaction survey suggested customers who may have been relatively uninformed about product features, and so entailed a nonignorable "item nonresponse" model.

Table 1 **Data Structure and Notation for Two Respondents under Item Selectivity**

Respondent number (r)	Number of alternatives (k_r)	Selection: polytomous (Y_s)	Selection covariates (X_s)	Prediction: interval[a] (Y_p)	Prediction covariates (X_p)
1	2	1	$X_{s,1,[1:2]}$	9	$X_{p,1,[1:2]}$
1	2	0	$X_{s,1,[2:2]}$		
2	3	1	$X_{s,2,[1:3]}$	6	$X_{p,2,[1:3]}$
2	3	0	$X_{s,2,[2:3]}$		
2	3	0	$X_{s,2,[3:3]}$		

[a] The dependent measure in the prediction submodel, Y_p, is interval in our applications, although extensions to ordinal and other dependent variable (DV) types are possible with relatively minor adjustments (e.g., Ying et al. 2006).

is positive. Selectivity is absent when the correlation, ρ, is near 0. Given regressors $\{X_s, X_p\}$, the system (1)–(3) is ordinarily estimated by maximum likelihood techniques, or "two-step" estimation approaches, which can be inaccurate for large $|\rho|$ (Puhani 2000) and sensitive to specification of $\{X_s, X_p\}$.

Notation, Model Likelihood, and Estimation

Let us first consider data for two specific respondents, whose choice sets may differ in composition, size, experimental manipulation, or otherwise, as shown in Table 1.

We denote selection and prediction estimates for respondent r as $V_{s,r,[i:k_r]} = X_{s,r,[i:k_r]}\beta_s$ and $V_{p,r,[i:k_r]} = X_{p,r,[i:k_r]}\beta_p$, respectively. The r subscript can be suppressed for clarity, and it is convenient to number the item chosen (i.e., selected) by each respondent as $[1\colon k_r]$ or simply as 1. We therefore observe second-stage values only for $i = 1$, that is, associated with deterministic utility portion $V_{p,[1:k_r]}$ or $V_{p,1}$.

Given the joint error distribution, $(\varepsilon_s, \varepsilon_p) \sim N[0, 0, 1, 1, \rho]$, the joint density for a particular observation (that is, suppressing k_r) is[4]

$$P\big[V_{s,1} + \varepsilon_{s,1} > \{V_{s,i} + \varepsilon_{s,i}\}_{i>1} \text{ and } Y_p = V_{p,1} + \sigma\varepsilon_{p,1}\big]. \tag{4}$$

This yields a "mixed" likelihood, where selection is a discrete probability mass function and prediction a continuous probability density function, for which σ is an estimated dispersion parameter. If $\{\varepsilon_{s,i}\}$ are multinormal with zero mean and identity covariance,[5] (4) can be evaluated by isolating $\varepsilon_{s,1}$, decomposing $\varepsilon_{p,1} = \rho\varepsilon_{s,1} + \bar{\rho}z$, for z a standard normal draw and $\bar{\rho}^2 = 1 - \rho^2$. We can therefore rewrite (4)

$$P\big[\varepsilon_{s,1} > \{(V_{s,i} - V_{s,1}) + \varepsilon_{s,i}\}_{i>1} \quad \text{and} \quad Y_p = V_{p,1} + \sigma(\rho\varepsilon_{s,1} + \bar{\rho}z)\big].$$

This can be further simplified by fixing $\theta = \varepsilon_{s,1}$, so that

$$P\left[\theta > \{(V_{s,i} - V_{s,1}) + \varepsilon_{s,i}\}_{i>1} \text{ and } z = \frac{(Y_p - V_{p,1}) - \sigma\rho\theta}{\sigma\bar{\rho}}\right]$$

cleaves into two probabilistic statements: one about $\{\varepsilon_{s,i}\}$ and one about z, all of which are standard normal by construction. We can therefore simply integrate across θ to complete the likelihood statement

$$\int_{\theta\in R} \phi(\theta)\left(\prod_{i>1} \Phi[\theta - (V_{s,i} - V_{s,1})]\right) \cdot \phi\left[\frac{(Y_p - V_{p,1}) - \sigma\rho\theta}{\sigma\bar{\rho}}\right] d\theta. \tag{5}$$

When $\rho = 0$, the portion of the integrand in (5) from the prediction model reduces to $\phi[(Y_p - V_{p,1})/\sigma]$; this no longer depends on $\theta = \varepsilon_{s,1}$ and can therefore come outside the integral. In other words, when selection is "ignorable," the likelihood factors into its selection and prediction component parts.

We estimate all parameters in the likelihood built up from (5)—σ, ρ, and the coefficients within V_s and V_p—across observations and participants, using both standard classical (e.g., gradient search, quadrature) and Bayesian (Markov chain Monte Carlo (MCMC)) methods, which agreed closely in all cases. Because some of our key tests will involve bounded parameters like ρ, which cannot have a limiting normal or t density, we report Bayesian highest density regions (HDRs) for such quantities. Bayesian estimates are based on a burn-in of 20,000 iterations, inference on an additional 20,000, and convergence checked by trace plots and typical diagnostics (e.g., Gelman–Rubin, Geweke) using multiple chains. Model comparisons (and significance levels) are carried out for nested models and parametric restrictions via likelihood ratio tests, and for nonnested models using Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002). We report posterior means for all parameters using highly diffuse conjugate priors for all parameters but ρ, which has a flat prior.

Our data are typical of many behavioral studies, where the number of observations per participant

[4] We will eventually allow ρ to vary by experimental condition, but leave it unsubscripted here. For conciseness, we use $x \geq \{y_i\}$ to mean $x \geq \max_i\{y_i\}$.

[5] We shall test this empirically for our data, finding support in all three experiments (see the online appendix).

can be far fewer than potential explanatory covariates, particularly when interaction effects are incorporated. Because we have few choices per participant (only 3 each in Studies 2 and 3), particularly so compared with potentially heterogeneous coefficients (e.g., over 20 in Study 3), as per Andrews et al. (2008) we do not attempt to recover "unobserved" heterogeneity.[6] Instead, as we stress throughout and test for, it is critical to account for "observed" heterogeneity, which in all three data sets takes the form of "prior ratings" (and related measures like self-stated "favorite") for each item by each participant; results will consistently show that these prior ratings are by far the statistically strongest effects of all measured.

We note that Bayesian estimation of the model can be slow, and searching the model space in this way, which we take up later, may not be feasible for behavioral researchers focused on theory testing or simply exploring their data. To that end, we develop a "two-step" approximation method that allows for such exploration and involves adding a new covariate in the *prediction* model, $X = \Phi^{-1}(\exp(-p_1))$, where p_1 is the (logit or probit) probability for the chosen item from the *selection* model. This merely involves running the selection and prediction models separately, in order, using standard software; details appear in the appendix along with derivations and a comparison of results versus the "correct" estimates from the empirical applications in the online appendix.

Empirical Applications

We illustrate the importance of accounting for item selectivity by applying models accounting for it to data from three studies. A critical commonality across the three studies is the random assignment of *participants* to experimental conditions, so any observed selectivity artifacts cannot stem from this assignment alone. However, in all experimental conditions, participants will choose items, which they will subsequently evaluate. If differences in experimental conditions thereby influence selection, *the existence and degree of item selectivity can systematically vary across conditions*. For example, characteristics of the choice set from which research participants choose—such as the number of available alternatives or the relative attractiveness of alternatives—likely influence selection and, consequently, degree of item selectivity. We examine effects of choice set composition on item selectivity in our applications, allowing for the possibility that item selectivity varies across conditions.

Table 2 Overview of Data Set Structures for Three Applications

	Study 1 (RKK)	Study 2	Study 3
Set size (No. of items)	3 or 6	6 or 12	6
No. of choices	10	3	3
Individualized sets	Yes	No	Yes
Stimuli	Songs	Snacks	Snacks
Time between choices	1 minute	<1 minute or 1 week	<1 minute or 1 week

Summary of the Three Studies

Our first study is a reanalysis of data from Ratner et al. (1999, Experiment 5), which suggested that when the number of available alternatives increases, consumers choose more varied sequences of items "for the sake of variety" rather than choosing items that are more preferred a priori. Other researchers have associated larger assortments with higher product expectations and lower evaluations of chosen items (Iyengar and Lepper 2000, Broniarczyk 2008, Diehl and Poynor 2010). We test for item selectivity and explore whether the degree of selectivity varies by choice set size condition. Study 2 explores a similar question about differing selectivity by choice set size, but for a different well-known and well-documented choice context—the choice task typically employed to examine the so-called diversification bias (Simonson 1990, Read and Loewenstein 1995). Study 3 explores whether varying attractiveness of options in the choice set, without increasing the number of alternatives, can also impact the degree of item selectivity across conditions. We present each participant with a choice set of the same size, individualized to contain her most- and least-favored items, and manipulate the relative attractiveness, the "bunchiness" of the internal items (as defined subsequently). A summary of the commonalities and differences across the studies is shown in Table 2.

Model Comparison and Theory-Based Specification of the Selection Process

For each of our three data sets, we will refer to a particular model as "best." This designation is based on a near-exhaustive search of the model space that included all possible combinations of available covariates theoretically relevant to the variety-seeking contexts we examine: a priori item rating, choice lag, choice frequency, time since last chosen, a binary indicator of which item is most preferred a priori, and indicator variables representing the experimental choice set conditions in each study.[7] For the prediction submodels, which are linear regressions, we

[6] Model convergence was poor (particularly so for Studies 2 and 3) for the traditional random-coefficients specification, and in some cases Bayesian measures of model fit yielded implausible values. Mean effects, however, were broadly consistent with the "observed heterogeneity only" effects reported.

[7] We did an exhaustive search to ensure that reported substantive conclusions were truly based on the "best" model for each empirical application. Researchers in consumer behavior and behavioral decision theory rarely require an exhaustive search; however, when

run all possible combinations of covariates; for the selection submodels, we begin with a standard stepwise probit, run both "forward" and "backward," as well as Least Angle Regression (LARS) procedures (Efron et al. 2004). When the best-fitting prediction and selection submodels are determined, we estimate the conjoined (full) model including selectivity and all "nearby" models (i.e., adding in/subtracting out covariates one at a time) with the crucial error correlation parameter(s) (values of ρ in each condition) freely estimated; that is, the resulting "best" model is such that (1) all its covariates are significant, (2) no excluded covariates are significant, and (3) if a higher-order interaction appears, all lower-order interactions (significant or not) for those same covariates do as well, to allow for appropriate interpretation of higher-order effects. (As such, for example, we do not include all possible two-way interaction effects, only those identified as significantly improving fit.) Before any models are estimated, we standardize all covariates, except for binary variables, which are mean centered, to aid in the estimation and interpretation of models with interaction terms. In a later section, we examine the robustness of our estimated effects to the selection submodel specification.

In comparing models in any of our data settings, we stress one point: the prediction submodel does not itself change. Rather, our main concern is how the presence of and type of error correlation (ρ) codified by the selection model affect deductions (from the prediction model); that is, our focus is on the proper specification, and informativeness, of the selection model. Across the three data sets, we will address two questions primarily: Is there evidence of item selectivity *overall*? Is the *degree* of item selectivity substantially different across experimental conditions? We also examine whether, and how, estimated effects are impacted by two common procedures: failing to account for selectivity at all and specifying a theory-driven account of the selection process (as opposed to searching for the "best" model). Findings indicate that substantive results can be strongly altered by the former (failing to account for selectivity), but are moderately (but not "infinitely") robust to the latter.

The key primary check in all three studies will be on the significance of ρ, the error correlation. Although the literature on selectivity offers many discussions of the "meaning" of ρ, it cautions researchers against pinning specific interpretations onto it, as it can arise from several sources (including omitted regressors, which we test for in the online appendix). It is, however, uncontroversial regarding the *effects* of omitting this correlation: when the selection process and evaluation process are based on similar underlying criteria *that cannot be fully accounted for via covariates*—such as finding superior options for purchase—we would expect a positive value of ρ. This is what we find in each of our studies, and we also document substantively relevant coefficient sign changes.

The goal across our studies is to go beyond a substantive examination of these data per se, toward demonstrating the need to statistically account for selectivity by experimental condition when the experimental design setup allows. Because we present three studies, descriptions of the experimental procedures and data are deliberately concise, allowing more emphasis on the role of item selectivity in the results proper. Full details for all studies are available from the authors.

Study 1: Reexamination of Ratner, Kahn, and Kahneman Data

A number of studies (e.g., Simonson 1990) suggest that consumers can behave as if varied sequences—of products or other experiences—were intrinsically superior to repetitive ones. Ratner et al. (1999; henceforth, RKK) concluded that more varied sequences of popular songs resulted in *diminished* enjoyment during consumption (even though the authors ruled out satiation with top-ranked songs). RKK's analysis made use of established statistical methods and, because their article dealt with many topics substantively unrelated to selectivity, no Heckman-type corrections were applied (this would not have been possible in any case because their analysis was performed at the aggregate level). By contrast, we conduct our analyses at the individual item level, and so we can specify a model that allows for item selectivity and compare the analysis results with and without constraining the error correlation parameter(s), ρ, to be zero.[8]

RKK (Experiment 5) studied the effects of the number of available alternatives on the selection and ratings of popular songs using a two-cell, within-subjects experimental design.[9] Participants first provided a priori ratings, on a 100-point scale, for 12 popular songs that were presented to them through a computer program. Idiosyncratic preference rankings were constructed based on these initial ratings. Participants

specific hypotheses require testing, the proposed model need be run only for each hypothesized relationship. They can also first avail of the "two-step" procedure detailed in the appendix.

[8] RKK performed a repeated measures analysis of covariance, with *average* real-time satisfaction ratings and choice set size predicting *average* retrospective evaluations of participants' chosen sequences. They found a significant effect of set size, with participants in the larger set condition reporting global evaluations greater than their satisfaction ratings. Our analyses are conducted at the *individual item* level and thus cannot be directly compared to the RKK results.

[9] We thank the authors for allowing us to reanalyze their data. Forty-eight study participants are included in our analysis.

were next presented with either a set of three items or a set of six items. They chose, listened to, and rated the song of their choice, also on a 100-point scale; this was done 10 times. For an additional 10 occasions, participants were presented with the other-sized choice set and completed similar choice and rating tasks. The {small; large} pair of choice sets from which each participant chose was constructed using prior preference rankings and randomly assigned: either ranks $\{(1,3,6);(2,4,7,10,11,12)\}$ or ranks $\{(2,4,7);(1,3,6,8,9,10)\}$. Because these data involve two phases—choice (of which song one listens to) from sets of varied sizes and ratings (of satisfaction or liking of the chosen song)—it is well suited to the models developed here and to evaluating the substantive implications of choice-based selectivity.

Our re-examination is narrowly focused on whether *limiting analysis only to items chosen by participants* (selection) affects estimates related to satisfaction with one's choice (prediction). To this end, we predict posterior ratings (*Rating*) using prior ratings (*Prior Rating*), how frequently an item was chosen in past occasions (*Frequency*), whether the song was chosen last time (*Choice Lag*), and the size of the choice set (*Set Size*, either three or six songs), while accounting for selectivity effects arising from the choice process by accounting for its *Prior Rating*, how often it was chosen (*Frequency*), and again whether the song was chosen last time (*Choice Lag*).[10] These regressors were drawn from among those examined by RKK, and the *Frequency* and *Choice Lag* covariates were also included because they have long been examined in the variety-seeking literature (e.g., Van Trijp et al. 1996). Model estimates are given in Table 3.

We do not engage in a substantive reinterpretation of the very rich RKK data from this experiment, nor link it to the numerous companion studies in that article. We can, however, consider model implications strictly from the vantage point of item selectivity, which are indicated for these data, as we discuss next. In a later section, we examine substantive findings in terms of robustness to the specification of the selection submodel.

Results

We discuss three models, in order of their appearance in Table 3: (1) assuming no selectivity (restricting $\rho = 0$); (2) allowing for an identical degree of selectivity across conditions (free ρ); and (3) allowing selectivity to vary across choice set conditions ($\rho_{SetSize}$). In this way, one can "decompose" the influences of the various modeling constructs systematically. Because all parameters are estimated via Bayesian techniques, we assess significance via HDRs for posteriors. Thus, "significant" denotes zero lies outside a specific HDR, usually at the 0.05 level, although for convenience we list traditional means and standard errors.

Item Selectivity. The model estimates reveal an intriguing and, to our knowledge, novel pattern of selection effects across conditions. Allowing for selection, but assuming ρ equal across (choice set size) conditions, yields a ρ estimate that is not significantly different from zero ($\hat{\rho} = 0.043$, HDR $= [-0.289, 0.332]$). This would appear to suggest there are no selection effects for these data. However, allowing ρ to vary by condition reveals significant selectivity, but only for the larger choice set condition: $\hat{\rho}_{Small} = -0.125$, HDR $= [-0.404, 0.138]$; $\hat{\rho}_{Large} = 0.571$, HDR $= [0.353, 0.727]$. Thus, degree of selectivity increases with choice set size; later, we revisit this finding in light of analogous ones from Studies 2 and 3.

Estimated Effects. Importantly, we also find that allowing for selectivity by condition impacts estimated effect sizes. Substantively, the pattern of effects for the selection model is of lesser interest, as coefficients across the various models (in Table 3) are very close; HDRs overlap to a degree that render them statistically indistinguishable. It is worth noting, however, that the estimated negative effects of *Choice Lag* and its interaction with *Prior Rating* (e.g., $\hat{\beta}_{S,\,Lag} = -0.481$, $\hat{\beta}_{S,\,Lag\times PrRate} = -0.226$, multiple ρ) are consistent with previous research examining variety seeking and preference (under)weighting in repeated sequential choice (RKK; Simonson 1990).

A very different pattern emerges across the prediction models. The most striking result is that the effects of *Set Size* are *vastly* different across models. When there is no selection ($\rho = 0$) or equal selection across choice set size conditions (common ρ), the effects of *Set Size* are significantly positive and not statistically distinguishable ($\hat{\beta}_{P,\,SetSize} = 0.126$, no ρ; $\hat{\beta}_{P,\,SetSize} = 0.119$, common ρ); that is, if one analyzed only the *Rating* (i.e., prediction) data, choice set size could be confidently claimed as positively affecting evaluation. However, when selectivity is accounted for across set size conditions ($\rho_{SetSize}$), we see that *Set Size* has a strongly *negative* effect ($\hat{\beta}_{P,\,SetSize} = -0.322$, $p < 0.003$, multiple ρ). This is consistent with extant literature demonstrating that consumers tend to be less satisfied with items chosen from relatively larger assortments (e.g., Iyengar and Lepper 2000, Diehl and Poynor 2010). A posteriori, in comparing a choice set of size 3 to one of size 6 (as RKK did), *chosen items are rated about one-third standard deviation lower, on average, when chosen from the larger set*, net of other covariate effects. The valence of an important main effect

[10] Including the same covariates in selection and prediction is permissible, particularly so when, as here, theory suggests doing so. An additional covariate representing the highest ranked item in each set (*Favorite*) was also tested in the selection and prediction submodels; it was not significant and is not discussed further.

Table 3 Study 1 (RKK Data) Model Comparisons: Posterior Means for Selection, Prediction, ρ, and σ

	Parameter estimates (std. err.)		
Model	No ρ	Common ρ	Multiple ρ
Selectivity/ρ	$\rho = 0$	Free ρ	$\rho_{SetSize}$
Selection model			
Prior Rating	**0.440 (0.039)**	**0.439 (0.039)**	**0.432 (0.038)**
Choice Lag	**−0.496 (0.078)**	**−0.495 (0.079)**	**−0.481 (0.077)**
Frequency	**0.190 (0.063)**	**0.186 (0.066)**	**0.168 (0.061)**
Choice Lag × *Prior Rating*	**−0.242 (0.068)**	**−0.242 (0.068)**	**−0.226 (0.067)**
Choice Lag × *Frequency*	**0.762 (0.098)**	**0.761 (0.101)**	**0.791 (0.096)**
Prediction model			
Intercept	0.033 (0.021)	0.009 (0.091)	**−0.129 (0.054)**
Prior Rating	**0.713 (0.021)**	**0.717 (0.025)**	**0.747 (0.023)**
Choice Lag	**0.175 (0.058)**	**0.173 (0.060)**	**0.155 (0.058)**
Frequency	**0.134 (0.027)**	**0.138 (0.032)**	**0.167 (0.029)**
Set Size	**0.126 (0.042)**	**0.119 (0.048)**	**−0.322 (0.112)**
Frequency × *Prior Rating*	**−0.127 (0.021)**	**−0.128 (0.022)**	**−0.147 (0.022)**
Frequency × *Set Size*	0.026 (0.043)	0.027 (0.043)	**0.099 (0.047)**
Sigma, σ	**0.621 (0.014)**	**0.625 (0.015)**	**0.651 (0.100)**
ρ		0.043 [−0.289, 0.332]	
ρ_{Small}, small set size			−0.125 [−0.404, 0.138]
ρ_{Large}, large set size			**0.571 [0.353, 0.727]**
Number of parameters	13	14	15
DIC	4,252.867	4,254.891	4,243.864
Log likelihood	−2,113.411	−2,113.344	−2,106.836
Likelihood ratio tests: p-values			
Common ρ vs. no ρ		0.714	
Multiple ρ vs. no ρ			0.001
Multiple ρ vs. common ρ			0.000
Add *Frequency* × *Set Size* in selection ($\text{df}_{\text{diff}} = 1$; LL = −2,106.836; DIC = 4,248.798)			0.987
Add *Favorite* in selection and prediction ($\text{df}_{\text{diff}} = 2$; LL = −2,105.106; DIC = 4,244.383)			0.177
Add *Favorite* and all possible two-way *Set Size* interactions in selection and prediction ($\text{df}_{\text{diff}} = 9$; LL = −2,101.782; DIC = 4,251.811)			0.342

Notes. Bold denotes statistical significance. Numbers in brackets represent the 95% Bayesian HDR.

is therefore reversed when the *Ratings* data are analyzed in the absence of an associated model for choice that not only allows for selectivity, but that also does not restrict selectivity to be fixed across experimental conditions.

The interaction between *Set Size* and how frequently an item has been chosen (*Frequency*) also has strongly differing effects. When selectivity is not accounted for by condition, the interaction effect is not significantly different from zero ($\hat{\beta}_{P,\,Freq\times SetSize} = 0.026$, $p > 0.27$, $\rho = 0$). However, when selectivity is allowed to vary across conditions ($\rho_{SetSize}$), the interaction between *Frequency* and *Set Size* becomes larger and significant ($\hat{\beta}_{P,\,Freq\times SetSize} = 0.099$, $p < 0.05$). In other words, frequently selecting the same item from a larger set is a stronger indicator of enjoyment, compared to repeated choice in a smaller set. The model without selection ($\rho = 0$) does not suggest this more nuanced insight into repeated choice.

Model Fit. Finally, one is left with the question of which model represents the data best, which can be assessed via both Bayesian and classical metrics. DIC speaks clearly for multiple ρ, the values of which, for {no ρ, common ρ, and multiple ρ}, are {4,253, 4,255, 4,244}.[11] Likelihood ratio tests corroborate the DIC comparison, while allowing statistical tests for nested models (like those in Table 3): the model allowing selection to vary by choice set size (multiple ρ) offers a better fit for the data than both the model with no selection (no ρ; difference in log likelihood (LL_{diff}) = 6.58, df = 2, $p < 0.002$) and the model restricting selection to be equal across set size conditions (common ρ; $\text{LL}_{\text{diff}} = 6.51$, df = 1, $p < 0.001$). The models with no selection versus equal selection across set size conditions exhibit no difference in fit (no ρ versus common ρ; $\text{LL}_{\text{diff}} = 0.07$, df = 1, $p > 0.7$); this is consistent with the nonsignificance of ρ when it is

[11] Table 3 (bottom) summarizes whether other potentially relevant covariates had been excluded from the focal model: adding a *Frequency* × *Set Size* interaction to the selection model is nonsignificant ($p > 0.98$), as are adding a *Favorite* item indicator main effect ($p > 0.15$) and all possible two-way interactions in selection and prediction ($p > 0.3$).

restricted to be constant across experimental conditions. The slightly less parsimonious model thus more than compensates for its additional complexity, and allowing the correlation between selection and prediction to differ across conditions not only improves fit, but affects interpretation of focal substantive effects.

Study 2: Choice Set Size

The previous analysis, based on data collected in a classic prior experimental study, demonstrated the potential for selection effect strength to vary across experimental conditions with differing choice set sizes. Study 2 was designed and conducted with an explicit goal in mind: to examine whether the same potential exists when the time between each item selection is much greater than that examined in Study 1. To do so, we examine another well-known repeated choice phenomenon widely documented in the marketing and psychology literatures. Numerous prior studies have observed that people choosing multiple items at once now, to consume later, tend to choose a more varied set of items than if they had chosen the items one at a time, just before consuming each one (e.g., Simonson 1990, Read and Loewenstein 1995). While prior research has focused on the impact of these two choice modes on variety seeking, we will instead focus our analysis on the influence of the size of the available choice set on evaluation of chosen items and potential selectivity artifacts.

Participants chose three snacks, either from a set of 6 snacks or from a set of 12. Half the participants chose all three snacks at once ("simultaneous choice"); the remaining participants chose each snack one at a time across three choice occasions ("sequential choice"). A 2 (simultaneous choice versus sequential choice) × 2 (small choice set versus large choice set) between-subjects design was employed.

Snacks included well-known brands of crackers, chips, candy bars, cookies, and nuts. The *small set* condition included six snack options, and the *large set* condition included those six snacks plus six more. The small set condition stimuli and task replicate experiments found in Simonson (1990) and Read and Loewenstein (1995). The six additional snacks in the large set were chosen to mirror the six snacks in the small choice set, in terms of both product attributes and market share, so as not to increase perceptions of attribute variety or general product desirability. One hundred four undergraduate students participated in the study to earn course credit in an introductory marketing course.

The study was composed of four sessions spaced one week apart. In Session 1, participants' prior preferences were measured; participants rated how much they liked each snack using an 11-point Likert scale (1 = dislike very much, 11 = like very much). Participants also ranked the snacks from their most favorite to their least favorite. The choice tasks took place during Sessions 2, 3, and 4; we refer to these as Choice Weeks 1, 2, and 3, respectively. Participants in the *sequential* condition chose and ate one snack in Choice Week 1, chose and ate a second snack during Week 2, and chose and ate a third snack in Week 3. Participants in the *simultaneous* condition selected all three snacks in Choice Week 1, designating the first snack to be eaten in Choice Week 1, the second snack to be eaten in Choice Week 2, and the third snack in Choice Week 3. Immediately after participants ate each of their chosen snacks, they rated how much they liked the snack using an 11-point Likert scale (1 = dislike very much, 11 = like very much).

The snack evaluation rating measured immediately after a participant ate her chosen snack is the dependent variable, and it is observed only for the single item chosen for that time period. As in Study 1, we considered potential regressors from published research analyses examining this particular task (Simonson 1990) and variety seeking in general, as well as a covariate to represent our choice set size manipulation. Regressors for the selection model (for which item is chosen) are *Prior Rating* (the a priori item rating), *Favorite* (whether the item was designated the favorite; a priori rank equals one), *Choice Lag* (whether the item had been chosen in the prior time period), *Choice Lag* × *Favorite* interaction, and *Choice Lag* × *SEQ*, where *SEQ* represents the choice mode manipulation (equals one for sequential choice, zero for simultaneous choice). The regressors for the prediction model (for the single brand chosen) include *Prior Rating*, *Favorite*, *Choice Lag*, and *Choice Lag* × *Favorite* interaction as well as *Set Size* (equals one for the large set, zero for the small set).

Results

Item Selectivity. We find a pattern of selection effects mirroring those found in Study 1, again revealing evidence that ignoring item selectivity risks the possibility of drawing incorrect substantive conclusions. Table 4 summarizes model estimation results in a manner similar to Study 1, featuring three candidate models that differ in how flexibly they allow for selection: no selectivity ($\rho = 0$), selectivity common across conditions (free ρ), and selectivity varying across choice set conditions ($\rho_{SetSize}$). The pattern of results in Table 4 shows strong evidence of selectivity in the large choice set condition ($\hat{\rho}_{Large} = 0.568$, HDR = [0.280, 0.787]; multiple ρ), but not in the small choice set condition ($\hat{\rho}_{Small} = -0.273$, HDR = [−0.731, 0.433]; multiple ρ). This is consistent with the results in Study 1, and the estimated values of ρ_{Small} and ρ_{Large} are remarkably similar across the two studies. The model allowing for selectivity, but restricting ρ to be constant

Table 4 Study 2 Model Comparisons: Posterior Means for Selection, Prediction, ρ, and σ

	Parameter estimates (std. err.)		
Model	No ρ	Common ρ	Multiple ρ
Selectivity/ρ	$\rho = 0$	Free ρ	$\rho_{SetSize}$
Selection model			
Prior Rating	**0.516 (0.054)**	**0.518 (0.053)**	**0.527 (0.054)**
Favorite	**0.327 (0.111)**	**0.322 (0.110)**	**0.315 (0.109)**
Choice Lag	−0.226 (0.140)	−0.243 (0.138)	**−0.289 (0.138)**
Choice Lag × *SEQ*	**0.872 (0.270)**	**0.918 (0.263)**	**0.970 (0.267)**
Choice Lag × *Favorite*	**−0.862 (0.308)**	**−0.855 (0.304)**	**−0.792 (0.299)**
Prediction model			
Intercept	0.012 (0.050)	**−0.419 (0.199)**	−0.185 (0.187)
Prior Rating	**0.485 (0.058)**	**0.554 (0.065)**	**0.493 (0.064)**
Favorite	0.193 (0.120)	**0.340 (0.140)**	**0.273 (0.134)**
Choice Lag	0.272 (0.154)	0.276 (0.157)	0.286 (0.153)
Set Size	0.034 (0.103)	−0.065 (0.112)	**−0.881 (0.322)**
Choice Lag × *Favorite*	**−0.675 (0.297)**	**−0.808 (0.313)**	**−0.711 (0.310)**
Sigma, σ	**0.845 (0.036)**	**0.895 (0.055)**	**0.901 (0.048)**
ρ		**0.413 [0.055, 0.703]**	
ρ_{Small}, small set size			−0.273 [−0.731, 0.433]
ρ_{Large}, large set size			**0.568 [0.280, 0.787]**
Number of parameters	12	13	14
DIC	1,655.320	1,652.189	1,645.995
Log likelihood	−815.604	−812.930	−808.703
Likelihood ratio tests: p-values			
Common ρ vs. no ρ		0.021	
Multiple ρ vs. no ρ			0.001
Multiple ρ vs. common ρ			0.004
Add *Prior Rating* × *Set Size* to prediction ($df_{diff} = 1$; LL $= -807.926$; DIC $= 1{,}646.697$)			0.212
Remove *Set Size* from prediction ($df_{diff} = 1$; LL $= -812.956$; DIC $= 1{,}651.920$)			0.004

Notes. Bold denotes statistical significance. Numbers in brackets represent the 95% Bayesian HDR.

across conditions (common ρ), yields an estimated ρ value that is significantly positive ($\hat{\rho} = 0.413$, HDR $= [0.055, 0.703]$). Thus, selection effects are clearly evident in the data; however, presuming that *degree* of selectivity is the same across all choice set conditions would be erroneous in this case, just as in Study 1.

Estimated Effects. A key substantive implication of failing to appropriately account for selectivity is that, just as in Study 1, the estimated effect of choice set size on evaluation changes dramatically. The estimated effects of *Set Size* are not distinguishable from zero when there is no selection ($\hat{\beta}_{P,\,SetSize} = 0.034$, $p > 0.37$; no ρ) or equal selection across choice set size conditions ($\hat{\beta}_{P,\,SetSize} = -0.065$, $p > 0.28$; common ρ). However, when selectivity *is* accounted for across set size conditions, choice set size has a very large negative effect ($\hat{\beta}_{P,\,SetSize} = -0.881$, $p < 0.004$; multiple ρ). Again we find that when a model allowing varying degrees of selectivity across choice set conditions is employed, the results reveal that participants tend to be less satisfied with items chosen from the larger choice set. This pattern of results is remarkably concordant with Study 1 (RKK), even though that experiment was run by other researchers, using a within-subjects design and different stimuli, and with many more repetitions.

Differences in selectivity also reveal their substantive importance when comparing the estimated *Favorite* effect strengths between the $\rho = 0$ prediction submodel (i.e., no selection effects) and its more flexible variants. "Common" selectivity yields $\hat{\beta}_{P,\,Fav} = 0.340$ ($p < 0.008$; common ρ), a significant effect; presuming there is no selectivity yields $\hat{\beta}_{P,\,Fav} = 0.193$ ($p > 0.05$; $\rho = 0$), a nonsignificant value about half the size. Allowing selectivity to differ across conditions also yields a positive effect of most-favored status, $\hat{\beta}_{P,\,Fav} = 0.273$ ($p < 0.03$; multiple ρ). Thus, allowing for item selectivity reveals a crucial role of the favorite option in evaluation: the most-favored option gets a "boost" in evaluation, over and above that accounted for by its higher prior rating. Note that a significant negative interaction between *Favorite* and *Choice Lag* is observed in all three models—$\hat{\beta}_{P,\,Lag\times Fav} = -0.675$ ($p < 0.02$) when $\rho = 0$, $\hat{\beta}_{P,\,Lag\times Fav} = -0.808$ ($p < 0.005$) when ρ is unrestricted, and $\hat{\beta}_{P,\,Lag\times Fav} = -0.711$ ($p < 0.02$) when ρ is allowed to vary across set size conditions—suggesting that, in the case where the favorite was chosen in the prior period, the favorite item's evaluation is discounted. Thus, for these data, a modeling approach that allows for selectivity reveals that the favorite option almost always "gets a boost" in evaluation (except in the case when it was chosen last time).

Model Fit. In addition to the substantive insights gained from allowing for selectivity, we find better model fits for both models with unrestricted ρ, as measured using DIC: 1,655, 1,652, and 1,646 for no ρ, common ρ, and multiple ρ, respectively. This evidence is bolstered by likelihood ratio tests: the model with free common ρ offers a better fit than one restricting ρ to zero ($\text{LL}_{\text{diff}} = 2.67$, $\text{df} = 1$, $p < 0.03$), and the model allowing selectivity to vary across choice set conditions fits the data better than both one restricting ρ to be constant across conditions ($\text{LL}_{\text{diff}} = 4.23$, $\text{df} = 1$, $p < 0.005$) and the model restricting ρ to be zero ($\text{LL}_{\text{diff}} = 6.90$, $\text{df} = 2$, $p < 0.002$).[12] Overall, analysis of these data provides evidence that appropriately accounting for selectivity adds substantially to both model fit and interpretation of effects.

Study 3: Choice Set "Bunchiness"

The prior two studies demonstrated that degree of item selectivity can vary with the number of alternatives in a choice set. This study assesses whether selection effects can vary across choice set conditions even when the number of selection alternatives stays constant. We explore this question with the same choice task employed in Study 2, and we examine the impact of varying the relative attractiveness of items in the choice set on degree of selectivity and the value of ρ.

"Bunchiness" refers to the degree to which a choice set contains items that are perceived to be equally attractive to one another (i.e., the extent to which items are "bunched" together), from the decision maker's perspective. For example, a choice set comprising six equally attractive items or one with six equally unattractive items would be high in bunchiness.[13] Note that bunchiness is a *characteristic of the choice set itself*, rather than of any one item in the set. We expect that bunchiness will have a negative relationship with degree of selectivity (and ρ) because, for experimental procedures in which participants choose the items they evaluate, items from bunchy choice sets have similar probabilities of sample inclusion, so analyses of bunchy sets may be less prone to selection artifacts. In other words, everything in a bunch is similar in overall attractiveness, so there is not much to be gained (or lost) by the consumer's choosing one over another, suggesting at best modest item selectivity.

One hundred twenty-six participants followed a four-week procedure analogous to that in Study 2. They were asked to rate and rank 12 snacks in Week 1 (the same as those in the Study 2 large set condition). The number of available items was held constant across conditions, at six, and we manipulated choice set bunchiness. To this end, the items available in the choice set varied across three bunchiness conditions based on the idiosyncratic rankings provided by each participant: *bunchy attractive* (ranks 1, 2, 3, 4, 5, and 12), *bunchy unattractive* (ranks 1, 8, 9, 10, 11, and 12), and *not bunchy* (ranks 1, 4, 6, 8, 10, and 12). We include the bunchy-*unattractive* condition in the experimental design to assess whether any potential impact of bunchiness is conditional on the bunched alternatives being perceived as (more) attractive; we will allow separate measures of selectivity for all three conditions to determine whether, empirically, ρ estimates for *bunchy attractive* and *bunchy unattractive* are close in magnitude. Note that all choice sets include both the most-favored item (rank = 1) and the least-favored item (rank = 12), so that the range of relative attractiveness of all items in the set is consistent across conditions. The three conditions are reminiscent of the experimental design in Wang et al. (1995), where three pairs of exam questions were varied on both mean difficulty and difficulty difference. By asking test takers to indicate a preference in each pair, but still answer all six questions (so there is no "missing" evaluation data), they were able to verify that post hoc correction for selectivity is critical.

We include two binary variables to represent bunchiness: *Bunchy Attractive* (equals 1 for the bunchy-attractive condition, 0 for not-bunchy and bunchy-unattractive conditions) and *Bunchy Unattractive* (1 for bunchy-unattractive, 0 for not-bunchy and bunchy-attractive conditions). The selection submodel regressors are a priori rating (*Prior Rating*), a priori most-favored item (*Favorite*), whether an item was chosen last time (*Choice Lag*), choice lag interacted with sequential/simultaneous choice mode (*Choice Lag* × *SEQ*), choice lag interacted with prior rating (*Choice Lag* × *Prior Rating*), a priori rating interacted with choice mode (*Prior Rating* × *SEQ*), and most-favored item interacted with the two choice set condition indicator variables (*Bunchy* {*Attractive*, *Unattractive*} × *Favorite*). Prediction submodel regressors are *Prior Rating*, *Favorite*, *Choice Lag*, *SEQ*, *Bunchy* {*Attractive*, *Unattractive*} (i.e., two main effects), *Prior Rating* × *SEQ*,

[12] Additional tests of model fit (Table 4, bottom) demonstrate that adding *Prior Rating* × *Set Size* to the prediction model is nonsignificant ($p > 0.2$), while removing *Set Size* from the selection model significantly reduces model fit ($p < 0.01$).

[13] We operationalize bunchiness here based on how attractive the items are to individuals; there may be other ways, including perceived variety in a particular set (e.g., it is possible to like all items in a set, while also finding them to have important differences). However, because we focus on the degree to which respondents may be indifferent between two randomly chosen items in the set, attractiveness seems an appropriate criterion.

Table 5 **Study 3 Model Comparisons: Posterior Means for Selection, Prediction, ρ, and σ**

	Parameter estimates (std. err.)		
Model	No ρ	Common ρ	Multiple ρ
Selectivity/ρ	$\rho = 0$	Free ρ	$\rho_{Bunchiness}$
Selection model			
Prior Rating	**0.393 (0.066)**	**0.450 (0.069)**	**0.454 (0.065)**
Favorite	**0.617 (0.120)**	**0.539 (0.124)**	**0.517 (0.118)**
Choice Lag	0.024 (0.121)	−0.017 (0.121)	−0.003 (0.117)
Prior Rating × *SEQ*	−0.174 (0.091)	**−0.215 (0.093)**	**−0.196 (0.092)**
Choice Lag × *SEQ*	**0.915 (0.241)**	**0.922 (0.235)**	**0.883 (0.233)**
Choice Lag × *Prior Rating*	**−0.275 (0.121)**	**−0.255 (0.117)**	**−0.266 (0.116)**
Bunchy Attractive × *Favorite*	**−0.577 (0.206)**	**−0.537 (0.206)**	**−0.497 (0.206)**
Bunchy Unattractive × *Favorite*	−0.071 (0.208)	−0.101 (0.208)	−0.075 (0.207)
Prediction model			
Intercept	−0.102 (0.048)	**−0.563 (0.151)**	**−0.607 (0.135)**
Prior Rating	**0.472 (0.057)**	**0.531 (0.060)**	**0.546 (0.060)**
Favorite	**0.315 (0.121)**	**0.601 (0.152)**	**0.640 (0.153)**
Choice Lag	−0.070 (0.102)	−0.047 (0.107)	−0.062 (0.108)
SEQ	−0.087 (0.085)	−0.086 (0.085)	−0.075 (0.084)
Bunchy Attractive	−0.170 (0.110)	**−0.248 (0.114)**	0.073 (0.204)
Bunchy Unattractive	**−0.407 (0.115)**	**−0.390 (0.113)**	−0.093 (0.202)
Prior Rating × *SEQ*	**−0.184 (0.087)**	**−0.220 (0.090)**	**−0.240 (0.089)**
Choice Lag × *SEQ*	0.279 (0.202)	**0.488 (0.224)**	**0.536 (0.222)**
Bunchy Attractive × *Favorite*	**−0.721 (0.247)**	**−0.905 (0.262)**	**−1.093 (0.259)**
Bunchy Unattractive × *Favorite*	0.554 (0.310)	**0.704 (0.320)**	0.504 (0.362)
Bunchy Attractive × *Prior Rating*	**0.368 (0.139)**	**0.350 (0.137)**	**0.288 (0.136)**
Bunchy Unattractive × *Prior Rating*	−0.151 (0.140)	−0.195 (0.140)	−0.256 (0.138)
Sigma, σ	**0.802 (0.030)**	**0.869 (0.050)**	**0.891 (0.047)**
ρ		**0.543 [0.190, 0.788]**	
$\rho_{BunchyAttr}$, bunchy attractive set			**0.450 [0.028, 0.740]**
$\rho_{BunchyUnattr}$, bunchy unattractive set			**0.470 [0.040, 0.823]**
$\rho_{NotBunchy}$, not bunchy set			**0.835 [0.578, 0.959]**
Number of parameters	22	23	25
DIC	1,892.286	1,884.578	1,882.726
Log likelihood	−924.025	−919.016	−916.022
Likelihood ratio tests: *p*-values			
Common ρ vs. no ρ		0.002	
Multiple ρ vs. no ρ			0.001
Multiple ρ vs. common ρ			0.050

Notes. Bold denotes statistical significance. Adding *Bunchy* {*Attractive*, *Unattractive*} × *Prior Rating* to the selection model is nonsignificant ($p > 0.2$; df = 2; LL = −914.46; DIC = 1,884.9). Numbers in brackets represent the 95% Bayesian HDR.

Choice Lag × *SEQ*, *Bunchy* {*Attractive*, *Unattractive*} × *Prior Rating*, and *Bunchy* {*Attractive*, *Unattractive*} × *Favorite*. As in the two previous studies, we standardize all variables, except binary (dummy) variables, which are mean centered.

Results

Item Selectivity. We again find strong evidence of selectivity, this time for all choice set conditions, with the value of ρ systematically varying with bunchiness. Table 5 presents estimation results for three models differing in how flexibly selectivity is modeled: no selection ($\rho = 0$), restricting selectivity to be constant across choice set conditions (common ρ), and allowing selectivity to differ across choice set conditions (multiple ρ; $\rho_{Bunchiness}$). Restricting ρ to be constant across conditions yields an estimated ρ value that is significantly positive ($\hat{\rho} = 0.543$, HDR = [0.190, 0.788]; common ρ). Allowing ρ to differ across conditions ($\rho_{Bunchiness}$) reveals that degree of selectivity varies with bunchiness: the not-bunchy choice set produced the greatest degree of selectivity ($\hat{\rho}_{NotBunchy} = 0.835$, HDR = [0.578, 0.959]), while the bunchy-attractive and bunchy-unattractive choice sets generated lower, and nearly identical, degrees of selectivity ($\hat{\rho}_{BunchyAttr} = 0.450$, HDR = [0.028, 0.740]; $\hat{\rho}_{BunchyUnattr} = 0.470$, HDR = [0.040, 0.823]). Consistent with both Studies 1 and 2, we find clear evidence of selection effects in the data, but they vary in degree across choice set conditions, in this case, without varying choice set size; decreasing the degree to which available options are perceived as equally attractive (i.e., a "less bunchy" choice set) increases degree of item selectivity.

Estimated Effects. Comparing the prediction submodels with increasing flexibility of accounting for selection reveals striking differences in several of the prediction covariates. First, consistent with our finding in Study 2, the estimated effect of *Favorite* doubles in size when selectivity is accounted for in the model ($\hat{\beta}_{p,\,Fav} = 0.315$, $p < 0.005$ when $\rho = 0$; $\hat{\beta}_{p,\,Fav} = 0.601$, $p < 0.001$ with common ρ; $\hat{\beta}_{p,\,Fav} = 0.640$, $p < 0.001$ with multiple ρ). Similarly, the negative interaction effect of *Bunchy Attractive* × *Favorite* without selectivity ($\hat{\beta}_{P,\,BunchyAttr \times Fav} = -0.721$, $p < 0.003$, $\rho = 0$) grows increasingly negative as selectivity is introduced into the model ($\hat{\beta}_{P,\,BunchyAttr \times Fav} = -0.905$, $p < 0.001$, common ρ) and then allowed to vary across conditions ($\hat{\beta}_{P,\,BunchyAttr \times Fav} = -1.093$, $p < 0.001$, multiple ρ). The combined result of these two covariate estimate changes is that failing to account for selectivity across conditions would lead a researcher to underestimate the size of the postchoice evaluation "boost" for most-favored options in the bunchy-unattractive and not-bunchy set conditions.

Second, when selectivity is restricted to be zero ($\rho = 0$) or common across choice set conditions (free ρ), *Bunchy Unattractive* has a negative effect on evaluation ($\hat{\beta}_{P,\,BunchyUnattr} = -0.407$, $p < 0.001$ when $\rho = 0$; $\hat{\beta}_{P,\,BunchyUnattr} = -0.390$, $p < 0.001$ with common ρ). However, when ρ is allowed to vary across conditions, the estimated effect of *Bunchy Unattractive* shrinks dramatically to nonsignificance ($\hat{\beta}_{P,\,BunchyUnattr} = -0.093$, $p > 0.32$ with multiple ρ). Similar to Studies 1 and 2, we find a stark change in the estimated effect of choice set condition on evaluation, but with one critical twist: in Studies 1 and 2, modeling selectivity (across conditions) led to enhanced effects or overt sign reversals, but here, an effect that would otherwise appear significant recedes to nonsignificance. This highlights that modeling selectivity does not always lead to a particular conclusion—e.g., effects get stronger or reverse—but rather depends on the nature of the data and best-fitting model.

Third, we find that the estimated effect of *Choice Lag* × *SEQ* without selection ($\hat{\beta}_{P,\,Lag \times SEQ} = 0.279$, $p > 0.08$; $\rho = 0$) nearly doubles and becomes statistically significant when selectivity is accounted for in the model ($\hat{\beta}_{P,\,Lag \times SEQ} = 0.488$, $p < 0.02$ with common ρ; $\hat{\beta}_{P,\,Lag \times SEQ} = 0.536$, $p < 0.008$ with multiple ρ). Thus, for these data, failing to account for selectivity would lead the analyst to erroneously conclude that inertial choices have no distinct effect on evaluation, when they do, even after accounting for prior preference rating (*Prior Rating*).

Model Fit. Finally, we assess model fit and find that it consistently improves as selection is more flexibly accounted for in the model. Model fit, as measured using DIC, improves when ρ is assumed constant across choice set conditions (common ρ; 1,885 versus 1,892), and improves further when ρ is allowed to vary across choice set conditions (multiple ρ; 1,883 versus 1,885). Likelihood ratio tests also indicate that model fit improves: presuming common ρ improves fit versus restricting ρ to zero ($\text{LL}_{\text{diff}} = 5.01$, $\text{df} = 1$, $p < 0.003$), allowing ρ to vary across choice set conditions improves fit versus presuming zero ρ ($\text{LL}_{\text{diff}} = 8.00$, $\text{df} = 2$, $p < 0.002$), and allowing ρ to vary across choice set conditions marginally improves model fit versus presuming common ρ ($\text{LL}_{\text{diff}} = 2.99$, $\text{df} = 1$, $p \approx 0.05$).[14] In conclusion, the findings from this study offer further support for the importance of accounting for selectivity by experimental condition (when warranted) as well as its potential impact on both model fit and the substantive interpretation of effects.

Common Findings Across Studies 1–3 and Using the Model in Practice

We find evidence for and substantive implications of item selectivity across all three studies. First, across the board, the key footprint of item selectivity—a significant value of ρ in at least one condition—is observed. Second, assuming the degree of selectivity is the same in all conditions is strongly rejected: in Studies 1 and 2, assuming ρ is constant across conditions leads to poorer fit and substantively altered results, and even in Study 3, when all conditions show significant positive ρ, there is evidence of ρ being higher in the "not-bunchy" condition. Last, in all cases, allowing for item selectivity alters the substantive interpretation of the study in question, strikingly so in the case of *Set Size* for Studies 1 and 2, where the selectivity-based model indicates very significantly diminished evaluations in the larger set. By contrast, the results of Studies 2 and 3 show enhanced effects of the most-favored option on evaluation when selectivity is fully accounted for in the analysis.

One element uniting our results is that the conditions showing the greatest selectivity in all three studies are those with arguably the greatest "informativeness." For example, the larger sets in Studies 1 and 2 contain more items, and therefore more information on which consumers can base their judgments; note that in RKK, there is also a greater variation in the judged quality of the items available. In Study 3, the nonbunchy condition is the one lacking a large

[14] Additional tests confirmed that including *Bunchy* {*Attractive*, *Unattractive*} × *Prior Rating* in the selection model does not improve fit ($p > 0.20$), while removing any of the two-way interactions between *Bunchy* {*Attractive*, *Unattractive*} and *Prior Rating* or *Favorite* significantly reduces model fit (all $p < 0.001$).

"lump" of similarly rated items, and therefore presents a more diverse collection of item differences to judge in selection.

Having detected substantive artifacts associated with item selectivity in each of the three studies, we feel justified in advocating its being explicitly accounted for by behavioral researchers. However, we recognize two practical impediments to doing so: the complex and time-consuming nature of estimation (for even one model) and the need to explore a very large model space for most applications. We address these in turn, providing practical solutions explored further in the appendix and the online appendix.

Robustness to Selection Model Specification

As mentioned earlier, researchers often have provisional theories they wish to test, as opposed to engaging in an automated, exhaustive search of the model space. Because this article is about carefully modeling selection, one might reasonably question whether the results here are "robust" (versus "brittle") with regard to the specification of the selection model. In one obvious sense, no: in each case, failing to allow for selectivity fit the data worse and altered a substantively important measurement. However, what if researchers had used theory to guide them and prespecified the selection model, rather than searching widely for the best-fitting one? For each study, we examine this issue, as well as examine what happens if a covariate of *known* substantive importance—*Prior Rating* in the selection model—is deliberately left out.

The choice of *available* covariates in each of the three studies was guided by prior/anticipated findings, although in each case there were additional covariates collected as controls, confluence with other research, etc., that were not directly theoretically implicated; this is especially so for the large number of potential interaction terms. For each data set, we examine what a substantively oriented researcher might posit—based solely on prior findings—to be a strongly plausible selection model, to assess robustness (of findings in the prediction model) to not searching the entire model space. We also assess how far this can be "stretched" by deliberately leaving out a common covariate of known importance to selection: *Prior Rating*. Table A.2 summarizes our approach and key results, as described here. (Full estimation results for all models appear in the online appendix.)

Study 1 (RKK). While alternating choices (e.g., variety) and repeated consumption of an item have each been examined separately as drivers of varied choice behavior, their interaction (i.e., *Choice Lag* × *Frequency*) has not. Doing so (model M1a) entails a significant reduction in model fit ($\text{LL}_{\text{diff}} = 35.1$, $\text{df}_{\text{diff}} = 1$, $p < 0.001$) *but no substantive difference in estimates of* ρ (*by condition*) *or prediction covariate effects*. RKK further ruled out satiation as an explanation for their findings (in their Experiment 5); model M1b excludes *Frequency* with similar results: far worse overall fit ($\text{LL}_{\text{diff}} = 56.4$, $\text{df}_{\text{diff}} = 2$, $p < 0.001$; model M1b), but no substantive differences in ρ (by condition) or covariates. The same is *not* true if *Prior Rating* is excluded: model fit is far worse ($\text{LL}_{\text{diff}} = 80.084$, $\text{df}_{\text{diff}} = 1$, $p < 0.001$; model M1c), but there *are* substantive differences in both ρ ($\rho_{Small} = -0.448$ is significantly negative) and covariate effects (most notably, *Set Size* erroneously doubles). We next replicate this discussion for Studies 2 and 3.

Study 2. Prior diversification bias theory (Read and Loewenstein 1995) suggests consumers' a priori favorite item may influence choice, but does not suggest it interacts with switching behavior. Removing *Choice Lag* × *Favorite* from the selection model degrades fit ($\text{LL}_{\text{diff}} = 3.511$, $\text{df}_{\text{diff}} = 1$, $p < 0.01$; M2a), but entails no substantive differences in ρ by condition, nor in substantive prediction covariate effects. Removing *Prior Rating*, however, causes both loss of fit ($\text{LL}_{\text{diff}} = 69.6$, $\text{df}_{\text{diff}} = 1$, $p < 0.001$; M2b) and two important substantive changes in the prediction model: *Set Size* and *Favorite* both become nonsignificant (and *Favorite* becomes three times more important in the selection model, standing in for *Prior Ratings*).

Study 3. Although Wang et al. (1995) examined a type of "bunchiness" for question pairs in scholastic testing, no prior research has focused on it as a choice predictor. So it is natural to consider dropping bunchiness condition dummies, which we include as interaction effects with the favorite item in the selection model. Doing so leads to worse fit ($\text{LL}_{\text{diff}} = 3.157$, $\text{df}_{\text{diff}} = 2$, $p < 0.05$), with neither substantive differences in ρ (by bunchiness condition) nor in prediction coefficients. The same is true for further dropping *Favorite*: poorer fit ($\text{L}_{\text{diff}} = 11.388$, $\text{df}_{\text{diff}} = 3$, $p < 0.001$) and similar values of both ρ and prediction coefficients. However, dropping *Prior Rating* ($\text{LL}_{\text{diff}} = 49.4$, $\text{df}_{\text{diff}} = 1$, $p < 0.001$) causes ρ to recede to nonsignificance in both the *Bunchy Attractive* and *Unattractive* conditions, and the theoretically implicated *Choice Lag* × *SEQ* interaction to become nonsignificant as well.

In conclusion, *all three data sets tell an identical story*: substituting a theoretically driven selection model specification for the "best" one arrived by exhaustive search did degrade overall model fit, but had no real substantive implications for focal effects (ρ by condition, prediction coefficients). This robustness was not boundless: doing the same with *Prior Rating* led not only to losses in overall fit, but important substantive differences in ρ values *and* focal prediction coefficients. We conclude that although researchers need not always search the entire model space as we have here, due diligence in specifying covariates included in the selection model is likely to secure similar substantive conclusions in the prediction model.

Two-Step Approach for EDA

Estimating the model using Bayesian tools can be slow, and of course requires a familiarity with interpreting MCMC output. One might reasonably ask, "How can I determine whether my data require this additional step?" We describe (in the appendix) and both derive and test (in Tables B1–B3 in the online appendix) the performance of a simple, "two-step" procedure that broadly replicates the results of the "exact" Bayesian analysis in this article. This two-step procedure has much in common with the "control function" approach described by Petrin and Train (2010), in which a researcher wishes to estimate a choice model including endogenous covariates and so first regresses these on other available covariates (including instruments), and then includes the *residuals* from that regression among the covariates in the choice model. This is practicable whenever residuals in the first-stage model (for the endogenous regressors) can be directly observed, as in ordinary least squares (OLS). In our set-up and applications, the first-stage model is (discrete) choice, and residuals from a choice model can only be estimated, not observed.

The gist of the procedure is running the selection (choice) model—which can be multinomial logit or probit—retaining the probabilities for the selected items, $\{p_1\}$, and then creating a new covariate (for each experimental condition), $X = \Phi^{-1}[\exp(-p_1)]$. This is simply added to the prediction model along with all other covariates, and then $\hat{\rho} = b/\sqrt{b^2 + \text{MSE}/\sigma_X^2}$ is calculated for each condition, where b is the estimated coefficient of X. Table A.1 summarizes the results of doing so for all three studies versus the "correct" estimates. The magnitude and significance of $\hat{\rho}$ by condition are quite close to the correct values, with the exception of the large set size in RKK. Notably, all of the "common ρ" cases are well within a 95% confidence interval (CI) of their correct values, although the procedure can produce CIs somewhat smaller than the "correct" Bayesian method.

Quantifying the Informativeness of the Selection Model

It is reasonable to ask, to what degree does carefully accounting for correlations "help" in forecasting? That is, how *informative* is the selection model? When $\rho = 0$, this is simple: the selection likelihood is just over half the total (57.4% in RKK; 57.2% in Study 2; 51.8% in Study 3). When $\rho \neq 0$, this is complicated by the fact that the likelihood in (5) cannot be neatly cleaved into selection and prediction portions, but rather is integrated over the latent error distribution. However, standard output provides a direct answer: the residual error variance is found to be $\hat{\sigma}^2(1-\hat{\rho}^2)$. Summing these for each observation in each condition provides a clear indication of the "informativeness" of the selection model, as a percentage reduction in mean squared error (MSE) ($\hat{\sigma}^2$). These vary quite a bit by study and whether ρ is common or multiple: 0.0% (RKK, common), 9.0% (RKK, multiple), 7.0% (Study 2, common), 8.9% (Study 2, multiple), 17.2% (Study 3, common), 23.1% (Study 3, multiple). The poor informativeness of the RKK common ρ stems from the erroneous conclusion that $\rho \approx 0$ when it was forced to be identical across conditions.

A simpler (though only approximate) measure is to follow the two-step procedure above and look directly at the MSE reduction when the "residual" from the selection model is entered into the prediction model: respectively, 0.0% and 0.6% (RKK), 2.5% and 7.8% (Study 2), 6.5% and 9.3% (Study 3). These are proportionally consistent with the exact proportions above, and suggest that *even when the selection model coefficients are not optimally adjusted through joint estimation*, the reduction in MSE due to the "informativeness" of the selection model can be substantial and readily assessed via ordinary regression.

Conclusions and Potential Extensions

Model frameworks developed by Heckman (1979), Tobin (1958), and others address analysis of field data with selectivity not amenable to researcher control. Although similar corrections have been applied in field data studies in marketing, their use in laboratory experiments, which typically offer the luxury of random assignment of participants, has been tacitly seen as less pressing, or perhaps nonexistent. It is also possible that the "in or out" selectivity of the classic Heckman (1979) model may have limited its applicability in choice-based behavioral research.

Our intent was to demonstrate that *item selectivity* can arise in a broad class of decision problems in consumer and psychological research, even when participants are randomly assigned to conditions, and to show how to account for it in a fairly general setting. Three studies—one a reanalysis of an influential, classic data set and two theory-driven experiments designed specifically for this purpose—converged on similar conclusions, namely, that selectivity effects can be substantial even in controlled, randomized laboratory studies; ignoring selectivity can alter focal substantive results, including effect sizes and even sign reversals; and allowing for different *degrees* of selectivity across experimental conditions can be crucial. We also show that substantive results are broadly (but boundedly) robust to a reasonable variety of selection model specifications, and provide researchers with a readily applied, "two-step" method to determine whether selectivity corrections may impact their results.

Although we have not reported on them in this article, we have successfully extended the model to different types of selection and prediction outcomes, including binary and "pick-*k*-of-*n*" selection and both ordinal (i.e., on a discrete, ordered scale) and discrete choice prediction (i.e., we observe only what was finally chosen, not any evaluation of it).[15] A common example of such extensions is discussed by Wachtel and Otter (2013), where only some consumers receive a catalog, some of whom purchase, and then some of *those* repurchase, with selectivity at each stage. The information we receive from consumers differs by how far they are in this winnowing process, so that item selectivity can result. Researchers ignoring the individual-specific selection phase(s) may be led astray in gauging market drivers: for example, price may influence which items are eliminated early on, and may thus appear relatively unimportant if only later stages of the purchase process are analyzed.

We can envision several fruitful extensions of the basic methodology. For example, "selection" in our model requires full knowledge of available options and item covariates, which are rarely available in field data and require foresight in experimental settings. Although we underscored the importance of observed heterogeneity, the incorporation of (unobserved) parametric heterogeneity when there are few observations per consumer (relative to the number of estimated coefficients) remains a challenge in choice modeling. It is possible, for example, that such unaccounted for unobserved heterogeneity can mask, attenuate, or falsely imply selectivity artifacts. Studies 2 and 3, with just three choices per respondent, were too "shallow" to investigate this, but the RKK data did allow a preliminary investigation. Using hierarchical Bayes techniques and estimating such a heterogeneous model for all reported coefficients yielded strongly overlapping 95% HDRs for both "common ρ" and "multiple ρ" models,[16] but obviously these limited findings cannot claim to generalize to other settings.

Among other extensions for practical model applications, one might similarly wish to rule out omitted regressors as a formal part of the model, without post hoc RESET-based procedures. Similarly, we presented robustness checks on the specific form of the selection model, but this required prior theory to determine reasonable alternatives; future work might explore an automated or disintermediated method to explore only "theoretically guided" selectivity specifications. Avoiding exhaustive search (e.g., stepwise or LARS methods on the selection and prediction submodels separately) to determine the best regressor set for the "full" conjoined model would enhance usability for practical applications; the covariate space for the conjoined model can be vast, especially when, as in our applications, interaction effects are considered. This was merely when linearity was presumed; allowing nonlinearities, in the form of power or other transforms of selection or prediction covariates, would expand the model space dramatically further. We view these as primarily issues of implementation and processing speed, and to be prohibitive mainly when researchers require a full search of the model space. The methods used to estimate the "full" model are especially attractive when multiple selection models have been run—based on theory or prior findings—which thereby allow for greater predictive stability via Bayesian model averaging.

Researchers often employ models (without selectivity) to test specific relationships between variables, as opposed to prediction per se; as such, they may be erroneously finding support (or lack thereof) for key hypotheses because estimates critical to testing them are not corrected for item selectivity. The models presented here can be readily estimated using a variety of available software platforms with modest run times, and can be first explored using the derived two-step approximation. As such, behavioral researchers could apply them "out of the box" to determine whether item selectivity substantively altered conclusions when initial consumer choice, screening, or input was a critical feature of their experiments.

Supplemental Material

Supplemental material to this paper is available at https://doi.org/10.1287/mksc.2016.0991.

Acknowledgments

The authors wish to thank Greg Allenby, Christie Brown, Pierre Chandon, Clint Cummins, Terry Elrod, Gabor Kezdi, Mike Palazzolo, and Carolyn Yoon for their assistance and suggestions.

Appendix. Simple Test and "Two-Step" Estimation Procedure for Multinomial Selectivity

Recall that the researcher wishes to estimate the system (1)–(3). One way to do so is to estimate the residuals, ε_s, *for the chosen items* and introduce them as regressors in (2). When ε_s are (unconditionally) normal, there is no closed-form solution, when conditioned on their being for the chosen items, for their expectation, median, or mode. It is

[15] Using Bayesian estimation tools, any of these outcome types can be back-sampled from to draw the "underlying" latent interval-scaled variable(s) that gave rise to it, rendering the remaining estimation equivalent (or nearly so) to a standard Heckman or a seemingly unrelated regression.

[16] Values for homogeneous versus heterogeneous, respectively, are as follows: common ρ, $(-0.289, 0.331)$ versus $(-0.029, 0.491)$; multiple ρ, small set size, $(-0.404, 0.138)$ versus $(-0.213, 0.425)$; multiple ρ, large set size, $(0.353, 0.727)$ versus $(0.042, 0.598)$.

possible instead to estimate (1) as a multinomial logit model, in which case the (unconditional) distribution of $\{\varepsilon_s\}$ is i.i.d. Gumbel, with mode 0 and mean the Euler–Mascheroni constant, $\gamma \approx 0.5772$. In this case, we can ask about the mean, median, or modal value of ε_s *for the selected item*. In the online appendix, we show that these have closed-form solutions, since the conditional density for error of the chosen item is again Gumbel, with location parameter $\ln(p_1)$, where the "chosen" item is denoted by subscript 1. We further show that the researcher can avail of any of the following as a new covariate, "X," in the prediction equation:

$$\begin{aligned} &\text{Mean:} && \Phi^{-1}[\exp(-p_1\exp(-\gamma))]; \\ &\text{Median:} && \Phi^{-1}[\exp(-p_1\ln(2))]; \\ &\text{Mode:} && \Phi^{-1}[\exp(-p_1)]. \end{aligned}$$

These new potential covariates (compared empirically below) are available from any program estimating the (conditional) logit model, since they are simple functions of the predicted probabilities (p_1) for the chosen items. We also note that, because X depends *only* on predicted probabilities, it is also possible to use a multinomial *probit* model to compute p_1; results doing so were extremely similar to those using the logit for selection, as shown in detail in the online appendix. For all three data sets, conclusions about "common ρ" were identical (to the fully Bayesian approach) using the two-step method; they were nearly so when ρ varied by condition, failing only for the large set size in RKK.

Estimating ρ

To estimate ρ, we augment the original prediction model with the new variable, X, obtained as above from the selection model

$$Y_p = X_p\beta_p + X\beta + \varepsilon_{new}.$$

The key insight is that ρ is just the correlation between the error in the selection model, represented by the new variable (X), and ε_p, which has been decomposed into $X\beta + \varepsilon_{new}$; that is

$$\rho \approx \operatorname{corr}(X, X\beta + \varepsilon_{new}).$$

This is simple to calculate, because X and ε_{new} are uncorrelated, since they are in the same regression. In the online appendix, we show that

$$\rho \approx \frac{b}{\sqrt{b^2 + \mathrm{MSE}/\sigma_X^2}},$$

where b is the estimated coefficient for X, σ_X^2 is the variance of X, and MSE is the estimated mean square error of the regression. Note that this ensures that ρ lies in $(-1, 1)$ and that $b \approx 0$ leads to $\rho \approx 0$.

Significance levels for ρ can be gauged by those for b; CIs can be computed using the upper and lower estimates for b itself. When ρ is estimated separately by condition—one of the main points of this article—one can simply put in separate "X" variables for each condition, obtain separate estimates of b, and thereby ρ, being sure to use the MSE for the overall regression. Because b and MSE are not necessarily uncorrelated, one can run a Bayesian regression instead of a classical one; the quantity $b/\sqrt{b^2 + \mathrm{MSE}/\sigma_X^2}$ is calculated for each draw of b and MSE, giving a *distribution* for ρ requiring no asymptotic assumptions, from which tests and CIs (highest density regions) arise.

We compare the accuracy of these methods—using the residual mean, median, and mode; logit versus probit selection; and classical versus Bayesian regression—in the online appendix for all three data sets in the article. The results are summarized in Table A.1. We find the proposed method to be reasonably accurate and to work (marginally) best by using the following for inference: the mean residual, probit selection, and Bayesian regression.

To summarize, the entire procedure, which can be used in an "exploratory" manner, entails four straightforward steps:

(1) Run a multinomial logit or probit selection model; retain the estimated probabilities for the chosen items, p_1.

(2) Compute the new covariate, X, using the mean, median, or mode (e.g., for the mode, $\Phi^{-1}[\exp(-p_1)]$). This should be done separately for each condition; e.g., three conditions requires three separate "X" variables, placed in the next step simultaneously.

Table A.1 **Bayesian Estimation of the Full Model vs. Two-Step Approaches**

			Mean of selected item residual		Median of selected item residual		Mode of selected item residual	
			Probit selection		Probit selection		Probit selection	
Study	Condition	Correct Rho	OLS	Bayes	OLS	Bayes	OLS	Bayes
1	All	0.043	0.022	0.014	0.022	0.023	0.020	0.020
(RKK)	Small set size	−0.125	−0.068	−0.063	−0.069	−0.073	−0.070	−0.074
	Large set size	0.571	0.117	0.140	0.117	0.146	0.117	0.119
2	All	0.413	0.325	0.321	0.326	0.328	0.327	0.328
	Small set size	−0.273	0.077	0.065	0.078	0.076	0.080	0.086
	Large set size	0.568	0.467	0.457	0.468	0.456	0.470	0.466
3	All	0.543	0.630	0.654	0.634	0.613	0.640	0.610
	Bunchy unattr.	0.470	0.322	0.354	0.328	0.386	0.339	0.313
	Not bunchy	0.835	0.757	0.756	0.758	0.762	0.760	0.671
	Bunchy attr.	0.450	0.591	0.598	0.593	0.578	0.595	0.618

Table A.2 Auxiliary Model Comparisons: Selection Model Specifications Tested and Posterior Means for ρ

Selection model	Focal model: Multiple ρ	Alternative selection model specifications tested		
Study 1 (RKK)		**M1a**	**M1b**	**M1c**
Prior Rating	✓	✓	✓	Omitted
Choice Lag	✓	✓	✓	✓
Frequency	✓	✓	Omitted	✓
Choice Lag × *Prior Rating*	✓	✓	✓	Omitted
Choice Lag × *Frequency*	✓	Omitted	Omitted	✓
ρ_{Small}, small set size	−0.125 [−0.404, 0.138]	−0.149 [−0.419, 0.134]	−0.058 [−0.355, 0.253]	**−0.448 [−0.739, −0.051]**
ρ_{Large}, large set size	**0.571 [0.353, 0.727]**	**0.456 [0.167, 0.661]**	**0.489 [0.102, 0.686]**	**0.679 [0.477, 0.811]**
Number of parameters	15	14	13	13
Log likelihood	−2,106.836	−2,141.98	−2,163.24	−2,186.92
LR test: *p*-value vs. focal model		<0.001	<0.001	<0.001
Study 2		**M2a**	**M2b**	
Prior Rating	✓	✓	Omitted	
Favorite	✓	✓	✓	
Choice Lag	✓	✓	✓	
Choice Lag × *SEQ*	✓	✓	✓	
Choice Lag × *Favorite*	✓	Omitted	✓	
ρ_{Small}, small set size	−0.273 [−0.731, 0.433]	−0.364 [−0.799, 0.276]	−0.320 [−0.748, 0.175]	
ρ_{Large}, large set size	**0.568 [0.280, 0.787]**	**0.573 [0.249, 0.786]**	0.346 [−0.152, 0.682]	
Number of parameters	14	13	13	
Log likelihood	−808.703	−812.214	−878.644	
LR test: *p*-value vs. focal model		0.008	<0.001	
Study 3		**M3a**	**M3b**	**M3c**
Prior Rating	✓	✓	✓	Omitted
Favorite	✓	✓	Omitted	✓
Choice Lag	✓	✓	✓	✓
Prior Rating × *SEQ*	✓	✓	✓	Omitted
Choice Lag × *SEQ*	✓	✓	✓	✓
Choice Lag × *Prior Rating*	✓	✓	✓	Omitted
Bunchy Attractive × *Favorite*	✓	Omitted	Omitted	✓
Bunchy Unattractive × *Favorite*	✓	Omitted	Omitted	✓
$\rho_{BunchyAttr}$, bunchy attractive set	**0.450 [0.028, 0.740]**	**0.465 [0.052, 0.747]**	**0.528 [0.156, 0.786]**	−0.039 [−0.355, 0.354]
$\rho_{BunchyUnattr}$, bunchy unattractive set	**0.470 [0.040, 0.823]**	**0.474 [0.043, 0.824]**	**0.500 [0.052, 0.835]**	0.236 [−0.179, 0.635]
$\rho_{NotBunchy}$, not bunchy set	**0.835 [0.578, 0.959]**	**0.854 [0.633, 0.962]**	**0.893 [0.727, 0.962]**	**0.565 [0.238, 0.806]**
Number of parameters	25	23	22	22
Log likelihood	−916.022	−919.179	−927.410	−965.452
LR test: *p*-value vs. focal model		0.046	<0.001	<0.001

Notes. Bold denotes statistical significance. Numbers in brackets represent the 95% Bayesian HDR.

(3) Run the prediction model, including X: $Y_p = X_p\beta_p + X\beta + \varepsilon_{new}$.

(4) Compute $\hat{\rho} = b/\sqrt{b^2 + \mathrm{MSE}/\sigma_X^2}$, either once (if the regression was classical) or at each draw of the sampler (if it was Bayesian), with CIs computed accordingly.

We summarize the Bayesian and OLS procedures for a probit model on all three data sets; complete estimation results, including logit and CIs, appear in the online appendix.

References

Albuquerque P, Pavlidis P, Chatow U, Chen KY, Jamal Z (2012) Evaluating promotional activities in an online two-sided market of user-generated content. *Marketing Sci.* 31(3):406–432.

Anderson ET, Simester DI (2004) Long-run effects of promotion depth on new versus established customers: Three field studies. *Marketing Sci.* 23(1):4–20.

Andrews RL, Currim IS (2005) An experimental investigation of scanner data preparation strategies for consumer choice models. *Internat. J. Res. Marketing* 22(3):319–331.

Andrews RL, Ainslie A, Currim IS (2008) On the recoverability of choice behaviors with random coefficients choice models in the context of limited data. *Management Sci.* 54(1):83–99.

Bradlow ET, Zaslavsky AM (1999) A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *J. Amer. Statist. Assoc.* 94(445): 43–52.

Braun M, Moe WW (2013) Online display advertising: Modeling the effects of multiple creatives and individual impression histories. *Marketing Sci.* 32(5):753–767.

Broniarczyk SM (2008) Product assortment. Haugtvedt CP, Herr PM, Kardes FR, eds. *Handbook of Consumer Psychology* (Laurence Erlbaum Associates, New York), 755–779.

Bronnenberg BJ, Dubé JP, Mela CF (2010) Do digital video recorders influence sales? *J. Marketing Res.* 47(6):998–1010.

Bult JR, Wansbeek T (1995) Optimal selection for direct mail. *Marketing Sci.* 14(4):378–394.

Carmon Z, Wertenbroch K, Zeelenberg M (2003) Option attachment: When deliberating makes choosing feel like losing. *J. Consumer Res.* 30(1):15–29.

Danaher PJ (2002) Optimal pricing of new subscription services: Analysis of a market experiment. *Marketing Sci.* 21(2):119–138.

Diehl K, Poynor C (2010) Great expectations?! Assortment size, expectations and satisfaction. *J. Marketing Res.* 47(2):312–322.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann. Statist.* 32(2):407–499.

Enders CK (2010) *Applied Missing Data Analysis* (Guilford Publications, New York).

Gu Y, Botti S, Faro D (2013) Turning the page: The impact of choice closure on satisfaction. *J. Consumer Res.* 40(2):268–283.

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Heckman JJ (1990) Varieties of selection bias. *Amer. Econom. Rev.* 80(2):313–318.

Iyengar S, Lepper M (2000) When choice is demotivating: Can one desire too much of a good thing? *J. Personality Soc. Psych.* 79(6):995–1006.

Lambrecht A, Seim K, Tucker C (2011) Stuck in the adoption funnel: The effect of interruptions in the adoption process on usage. *Marketing Sci.* 30(2):355–367.

Litt A, Tormala ZL (2010) Fragile enhancement of attitudes and intentions following difficult decisions. *J. Consumer Res.* 37(4):584–598.

Little RJA, Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Moe WW, Schweidel DA (2012) Online product opinions: Incidence, evaluation, and evolution. *Marketing Sci.* 31(3):372–386.

Peters S (2000) On the use of the RESET test in microeconometric models. *Appl. Econom. Lett.* 7(6):361–365.

Petrin A, Train K (2010) A control function approach to endogeneity in consumer choice models. *J. Marketing Res.* 47(1):3–13.

Pham MT, Chang HH (2010) Regulatory focus, regulatory fit, and the search and consideration of choice alternatives. *J. Consumer Res.* 37(4):626–640.

Puhani PA (2000) The Heckman correction for sample selection and its critique. *J. Econom. Surveys* 14(1):53–68.

Ratner RK, Kahn BE, Kahneman D (1999) Choosing less-preferred experiences for the sake of variety. *J. Consumer Res.* 26(1):1–15.

Read D, Loewenstein G (1995) Diversification bias: Explaining the discrepancy in variety seeking between combined and separated choices. *J. Exp. Psych.: Appl.* 1(1):34–49.

Rubin DB (2004) *Multiple Imputation for Nonresponse in Surveys*, Vol. 81 (John Wiley & Sons, New York).

Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. *Psych. Methods* 7(2):147–177.

Simonson I (1990) The effect of purchase quantity and timing on variety-seeking behavior. *J. Marketing Res.* 27(2):150–162.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc., Ser. B* 64(4):583–639.

Tobin J (1958) Estimation of relationships among limited dependent variables. *Econometrica* 26(1):24–36.

Van Trijp HC, Hoyer WD, Inman JJ (1996) Why switch? Product category-level explanations for true variety-seeking behavior. *J. Marketing Res.* 33(3):281–292.

Wachtel S, Otter T (2013) Successive sample selection and its relevance for management decisions. *Marketing Sci.* 32(1):170–185.

Wainer H, Thissen D (1994) On examinee choice in educational testing. *ETS Res. Report Ser.* 64(1):159–195.

Wainer H, Wang XB, Thissen D (1994) How well can we compare scores on test forms that are constructed by examinees choice? *J. Educational Measurement* 31(3):183–199.

Wang XB, Wainer H, Thissen D (1995) On the viability of some untestable assumptions in equating exams that allow examinee choice. *App. Measurement Ed.* 8(3):211–225.

Winship C, Mare RD (1992) Models for sample selection bias. *Ann. Rev. Soc.* 18:327–350.

Yang S, Zhao Y, Dhar R (2010) Modeling the underreporting bias in panel survey data. *Marketing Sci.* 29(3):525–539.

Ying Y, Feinberg F, Wedel M (2006) Leveraging missing ratings to improve online recommendation systems. *J. Marketing Res.* 43(3):355–365.

Zanutto EL, Bradlow ET (2006) Data pruning in consumer choice models. *Quant. Marketing Econom.* 4(3):267–287.