Using Multiple Methods to Improve Validity

David R. Just, Cornell University, djust@cornell.edu
Jie Jiao, Cornell University, jj546@cornell.edu

Invited Paper prepared for presentation at the 2024 Agricultural & Applied Economics Association Annual Meeting, New Orleans, LA; July 28-30, 2024

Using Multiple Methods to Improve Validity

David R. Just, Cornell University

djust@cornell.edu

Jie Jiao, Cornell University

jj546@cornell.edu

I. Introduction

Behavioral research, particularly in the context of providing evidence for policy, suffers from a unique and acute challenge of demonstrating real world relevance while establishing unassailable causality (Byrne & Just, 2020). We refer to this as the *behavioral policy challenge*. Natural tension arises in selecting a research methodology that maximizes both internal and external validity. Laboratory experiments are effective at establishing causality but face challenges in generalizing findings beyond their original settings. The uses of naturally occurred field or market data can offer a high degree of external validity but are inadequate in controlling for confounding factors due to selection biases and unobserved heterogeneity. Intermediate approaches, which attempt to balance these trade-offs, also have their own limitations.

Internal validity is essential for evaluating behavioral effects, as it establishes whether changes in the dependent variable (behavior) are caused by the independent variable (intervention or treatment) and not by other potential confounders. This is particularly important in behavioral research, where maintaining internal validity enables researchers to assert causality with confidence. Given the strong theoretical

tradition and practice presuming rational economic behavior, there is a high bar of evidence required to demonstrate behavioral effects that are either unrelated to rational considerations or in violation of them. Without strong internal validity, such results will often be dismissed as a result of misspecification, endogeneity or statistical outlier. High internal validity ensures that research findings can be properly understood, thereby enhancing the credibility and reliability of the scientific conclusions. Still, the applicability of behavioral findings for policy-making may be inappropriate if they are derived from studies conducted in highly controlled lab environments (Levitt & List, 2007a,b).

External validity is about the relevance and applicability of research findings. The utilization of naturally occurring field data can offer a significant advantage in understanding the magnitude of behavioral effects relative to other uncontrolled factors, yielding high external validity. However, because these data are uncontrolled, any observed relationships are likely subject to endogeneity bias. Consequently, the level of internal validity that can be achieved through econometric estimation is typically quite low (Learner, 1983). Even randomized field trials can compromise the internal validity often seen as guaranteed by randomization (McMillan, 2019).

Given the well known tradeoffs in external and internal validity (Roe & Just, 2009), using multiple methods has been suggested as the best way to overcome the joint need for external and internal validity in the case of behavioral policy (Roe & Just, 2009; Just & Byrne, 2020). We propose that the set of conditions under which such multiple methods will be effective in establishing increased validity is much narrower than is

often acknowledged. Given the challenges of behavioral work, the conditions under which multiple methods could jointly establish internal and external validity for a behavioral phenomenon is substantively limited and are likely violated specifically due to the same conditions that afflict internal and external validity in behavioral work generally. Without strong theoretical assumptions regarding the data generating process, establishing behavioral phenomena may be reliant on the use of field experiments, which can unfortunately face substantial restrictions in practice (Just & Gabrielyan, 2018; Just & Byrne, 2020).

In this paper we argue that while multiple methods are likely to be necessary, practical elements in research are such that a large class of multiple methods approaches are likely to fall short. Intuitively, a random assignment experiment accompanied by analysis of naturally occurring field data could not jointly establish causality and relevance unless uncontrolled factors can be excluded from the field data and the lab data is generated under sufficiently realistic conditions to instill confidence.

We demonstrate this point both theoretically and empirically. Our empirical example draws on two previously published studies, one a field experiment, the other an analysis of secondary transaction data. Both studies appear to demonstrate that students in a lunch setting who use debit cards to purchase food are more likely to select less healthy options. On the surface this may suggest that debit accounts are creating a public health issue in schools. However, a careful description of the conditions under which both datasets were collected raises substantial doubt about

any such conclusions. Differences in the restrictions on choice and novelty of choice environments may overwhelm the primary behavioral effect in question in the laboratory, while selection effects may affect interpretations of the secondary analysis. In such a case, the original weaknesses of each approach can only be bridged by the beliefs of the researcher regarding the statistical properties of the two studies.

II. The Behavioral Policy Challenge

The use of behavioral economic research to inform policy requires a unique demonstration of both internal and external validity. Internal validity is required to ensure the behavioral effect is actually caused by the intervention in question. External validity ensures that the effect when implemented in a natural setting will be of a magnitude that is policy relevant. We refer to this requirement for both internal and external validity as "the behavioral policy challenge." We argue that the behavioral policy challenge is unique to behavior economic research.

Behavioral economics examines the effects of psychological, social, cognitive, and emotional factors on economic decisions. Economic theory has a strong foundation in rational choice theory—which famously ignores purely psychological, social, cognitive, and emotional factors (Thaler, 2000; Urbina & Ruiz-Villaverde, 2019). The notion of rational choice is widely applied in policy circles even outside of economics (Neimun & Stambough, 1998). Thus, demonstrating a behavioral effect is a unique challenge in establishing internal validity within the economics literature and more broadly. A study demonstrating an effect that is inconsistent with the rational choice

model is rightly subject to added scrutiny as often subtle uncontrolled factors may make rational behavior appear irrational.

The rational model of choice is generally the first explanation for behavior employed by both behavioral and neoclassical economists alike (Just, 2014). Theoretical frameworks such as those proposed in economic models can help elucidate mechanisms behind observed relationships—providing us some explanation for why or how a stimulus may cause some specific behavior. Knowing the "why" of behavior provides deeper insights into the underlying economic processes and lends confidence to the robustness of the observed relationship under variations in context. Such confidence may be of significant value to a policymaker who likely wishes to implement policies in wider circumstances than those that prevail in a specific study.

Compared to economic models, behavioral phenomena face added scrutiny regarding the identification of causal effects. Behavioral economics focuses on the demonstrating the existence and importance of behavioral anomalies, which by definition are behaviors that violate the accepted economic models of behavior. Behavioral anomalies inherently face a burden of proof that the accepted model does not apply. It would be particularly difficult for an apparent anomaly to meet this burden if using non-experimental data. For example, consider the canonical endowment effect, in which ownership of an object increases an individual's valuation of the object (Kahneman et al., 1991). It would be difficult to argue based on observational data because of endogenous selection in ownership and valuation. The fact that those who own a particular object value it more than those who don't

may arise simply because those who place high value on the object are more likely to purchase it. Without random assignment of ownership, it would be impossible to establish the direction of causality between valuation and ownership. Similar arguments arise for many behavioral anomalies.

By randomly assigning participants to different groups, researchers can ensure that any pre-existing differences between these groups are distributed equally and randomly. This process eliminates selection bias, and makes a clear case for internal validity. Because randomization with large samples ensures observations in all treatments have similar distributions of potential confounds, it reduces the risk that other observed and unobserved factors might confound the results. This ensures that the systematic difference between groups is directly attributable to the intervention itself. This process enhances the credibility of the research findings by facilitating replication of the study, further validating the findings.

Random assignment at the individual level is often most easily achieved in laboratory settings (Shadish et al., 2002). In these experiments, researchers have full control over the environment and variables which reinforces the strength of internal validity. By necessity, however, controlled laboratory experiments present decisions that abstract away from reality. Decisions are often simplified, reducing the number of options available or simplifying the types of incentives offered. Moreover, participants enter a laboratory experiment by choice understanding that their behavior is being observed for research purposes. All of this suggests that behavior in the laboratory may not mirror behavior in less controlled settings.

The logistics of conducting experiments are also less complex in a lab compared to field settings. Random assignment may not be possible at an individual level in the field due to environmental complexity. Randomizing over time, geographic or institutional units introduces potential confounds that are not randomized away. Moreover, not every field setting will be amenable to every potential intervention (Just & Gabrielyan, 2020).

For policy considerations, it is essential not only to identify what is causal but also to determine that the size of effects are relevant when implemented in a policy setting—a question of external validity, To argue for policy relevance, we must establish that a behavioral phenomenon not only occurs regularly but also at a rate and magnitude that elicit the desired policy outcome despite the noise and confounding factors occurring in the field. Making this argument faces, perhaps, just as steep of a climb as establishing internal validity.

External validity is difficult to establish in a lab. Human behavior may deviate from the rational norm due to simplifying behaviors, calculation errors (Thaler, 2016). Such errors may be more likely when facing the smaller incentives one faces in a lab, or when facing a novel and unfamiliar choice context. Experiments tend to simplify choice contexts in order to clarify internal validity. These simplifications are exactly the factors that undermine external validity. In general it is likely too costly and too chaotic to replicate complex real-world behaviors within laboratory settings. These threats to external validity in a laboratory setting include (Roe & Just, 2009):

Artificial setting: Labs often create a controlled, simplified environment that does not accurately mimic the complexity of real-world situations. This can limit the applicability of the findings to actual environments where variables are more dynamic and less controllable.

Small incentives: Incentives used in lab experiments are typically smaller and may not effectively simulate real-world stakes. This discrepancy can affect participants' motivation and behavior, making it challenging to translate results to contexts where decisions have significant consequences.

Simplified choices: Laboratory studies often reduce the complexity of decision-making processes to a few controlled choices. While this simplification helps isolate specific variables, it also strips away the complexity typical in real-life situations, potentially oversimplifying the real life decision process.

Novel effects: Participants in lab experiments might experience novelty effects, where their behavior is influenced by the unfamiliarity of the experimental setting rather than the variables being tested. This can skew results, making it difficult to determine if the same behavior would occur outside the lab.

Alternatively, establishing external validity often requires the use of secondary data or conducting field experiments. Data collected in natural decision contexts allow researchers to confirm that their findings are applicable in real-world settings beyond the controlled conditions of the initial study. But sometimes even field experiments face challenge, in that the data collected usually still contain uncontrolled factors that

can obscure causal relationships. The same issues that undermine the use of non-random assignment data to establish causality also undermines attribution of the effects even if a behavioral pattern is known to exist.

Behavioral policy challenge requires methods that somehow could demonstrate real world relevance while establishing unassailable causality. Laboratory experiments and analysis of secondary data represent two ends of an evidence possibility frontier, with each offering validity in one dimension (internal or external) at the expense of the other. Field and natural experiments have gained wide popularity in recent decades because these methods relax the inherent tension between uncontrolled field and controlled laboratory data collection approaches (Roe & Just, 2009).

Field experiments maintain the advantage of establishing external validity due to minimal interference with the natural context. In 2010, the UK government established the Behavioral Insights Team (BIT) to improve public policy and government services by applying concepts from behavioral economics. Since then, many governments have formed their own nudge units, with a goal of using behavioral experimentation at a grand scale to inform policy. Such groups overcome the tradeoff between internal and external validity through large-scale field experiments often conducted at a national level. Much of their work is focused on finding “nudges” (Thaler & Sunstein, 2009) that improve the effectiveness of policy. "Nudges" are defined as "changes in the decision context, or, more precisely, changes in the choice architecture that alter people's behavior in predictable ways."

A field experiment that appears to have both internal and external validity arises in the school lunch context. This experiment tests different types of nudges to reduce the consumption of added sugar. The experiment aimed to enhance the impact of the National School Lunch Program by encouraging healthier food choices and improving nutritional intake. This experiment was highly efficient as it evaluated the effects of various nudges within a single study. The result shows that prompts alone increased the proportion of white milk chosen from 20% to 30%, whereas adding health or taste messaging to the prompt does not seem to be effective (Lai et al., 2019). Impressively, prompts seem to be about as effective as classical incentives (a glow bracelet) as found in a similar field experiment (List & Samek, 2017).

Another field experiment that appears to have both internal and external validity is showcased in a tax compliance experiment conducted by the Behavioral Insights Team (Behavioral Insights Team 2016, see Update Report 2015-16). They examine the impact of various types of messages included in letters sent to taxpayers, such as social norms ("9 out of 10 people in your area pay their taxes on time") and public goods messages (a letter emphasizing the provision of public goods, highlighting how taxes are used to fund essential services), on the timely payment of taxes. Taxpayers were randomly assigned to different treatment groups, ensuring that any differences in tax payment behavior could be attributed to the content of the letters rather than other factors. The experiment found that the social norm message significantly increased the rate of timely tax payments compared to the control group (Hallsworth et al., 2017). The randomized controlled trial design ensured internal validity, while the

real-world application, diverse participant pool, and scalability of the intervention contributed to its external validity. The success of this nudge has been replicated in various contexts, further supporting its robustness.

Some of the large field experiments conducted by the behavioral nudge units have demonstrated the great need for external validity. Some behavioral-based interventions that have been well established in experimental laboratory setting experienced surprising failures in the field. Government Nudge units have been using behavioral science principles to design and implement different labeling systems to encourage healthier food choices among consumers. A controlled experiment on the traffic light labeling system (color coded label indicating healthiness of food) conducted in Australia showed that the TL system was the most effective in assisting consumers to identify healthier foods (Kelly et al., 2009). For the purpose of the experiment, two-dimensional mock packages for three different product categories were created. A total of 790 participants were recruited from three different socioeconomic groups based on quota allocation. However, another 2009 field study using supermarket point-of-sales data from a major UK retailer operating a chain of over 1,000 supermarket stores provides evidence that the introduction of traffic light labels did not substantially influence supermarket sales of ready meals and sandwiches. The study examined the sales changes of certain product categories in the 8-week period surrounding the introduction of the labels (Sacks et al., 2009).

Efficacy aside, such experiments are not always possible to run at scale due to the ethics and resources involved. Large scale experiments of this nature are only possible

with government resources as they generally involve hundreds of thousands of participants or even millions. Without substantial prior evidence that the intervention cannot lead to unexpected results, it is not ethical to experiment on such a large group of unwitting participants. Some nudges may also have heterogeneous treatment effects, including the potential to backfire (Banerjee et al., 2022b, Maier et al., 2022, Mertens et al., 2022). For example, a nudge may demotivate those who are already motivated to engage in a positive behavior, especially in the short term. Nudges may also interact with other incentives already extant in the environment. Besides resource and scalability constraints, these nudge experiments face the same limitations as field experiments in general. The topic areas and subject matters available are often more limited than the laboratory environments that originally produced the potential intervention (Roe & Just, 2009). Notably, many large scale field experiments can only be run as policy is implemented. This precludes the use of such evidence to support the initial implementation.

Where large field experiments are not feasible, Roe and Just, among others, suggest the use of multiple methods to overcome the *behavior policy challenge.* By layering laboratory studies with studies using field data, together these studies in aggregate, it is argued, can establish both internal and external validity. Mixed methods may engage several separate studies and involve both experimental and non-experimental data. The potential of using such mixed methods has been offered many times as the solution, but little work has been done to examine how such layering of studies adds to the validity of any single result.

III. A Theory of Corroboration

Consider two independent studies which we will refer to as A and B. In each study we observe the dependent variable, $Y_i \in \mathbb{R}$ with $i \in \{A, B\}$, and the state of the independent variable $X_i \in \mathbb{R}$, but do not observe some related independent variable, $Z_i \in \mathbb{R}$, with $Y_i = \beta_{i,X} X_i + Z_i$ and $E(Z_i) = 0$. Thus, the coefficients that define the relationship between the dependent and independent variables may differ by study (e.g., a lab versus a field setting). Moreover, let A represent a controlled experiment so that X_A is chosen independent of Z_A. Thus realizations of Y_A are drawn from a conditional distribution which can be represented as $f_{(Y|X)}(Y|X)$. However, B is not controlled, and thus the distribution observed is the joint distribution $f(Y, X)$.

When can we use study A to learn about the sign of $\beta_{B,X}$ and study B to learn about the magnitude of $\beta_{B,X}$. Let $g(\beta_{A,X}, \beta_{B,X})$ represent the experimenter's prior regarding the relationship between the two coefficients, which implies a conditional distribution $g_{B|A}(\beta_{B,X}|\beta_{A,X})$, as well as marginal distributions $g_A(\beta_{A,X}), g_B(\beta_{B,X})$. So long as $g(\beta_{A,X}, \beta_{B,X}) \neq g_A(\beta_{A,X}) g_B(\beta_{B,X})$, estimating $\beta_{A,X}$ using experimental data will result in an updated marginal distribution $\tilde{g}_A(\beta_{A,X}|Y_A, X_A)$, implying an updated posterior for $\beta_{B,X}$ given by $\tilde{g}_B(\beta_{B,X}|Y_A, X_A) = g_{B|A}(\beta_{B,X}|\beta_{A,X}) \tilde{g}_A(\beta_{A,X}|Y_A, X_A) / f(Y_A, X_A)$. Thus, if the coefficients between the two experiments are not a priori regarded as independent, then we can make inference regarding the sign of the coefficient in study B from observations in study A. The degree to which we learn will be determined by the strength of relationship embedded in the prior $g(\beta_{A,X}, \beta_{B,X})$. This could be undermined by a laboratory experiment that

is too artificial or which imposes restrictions on choice that do not appear in the natural settings. These are common criticisms of laboratory experiments. Nevertheless, this does not present a very high bar for the laboratory experiment to provide value.

This leaves open the second part of our question. In this case, we may know that there is an independent causal effect of X on Y, but we may not know the size of this effect relative to other background unobservables in a setting where X cannot be exogenously controlled. Here it may be useful to consider a specific common parameterization for illustrative purposes. Suppose that $(X_B, Z_B) \sim N\left(\begin{bmatrix}\mu_X \\ 0\end{bmatrix}, \begin{bmatrix}\sigma_X^2 & \sigma_{XZ} \\ \sigma_{XZ} & \sigma_Z^2\end{bmatrix}\right)$. Then $Y_B|X_B \sim N\left(\beta_B X_B + E(Z_B|X_B), \sigma^2_{Z_B|X_B}\right)$ The additional value of observations in study B will depend heavily on the prior imposed on the relationship between Z_B and X_B, and specifically the mean of Z_B conditioned on X_B. If this expectation can be ascertained precisely a priori, then additional observations in study B clearly identify β_B, providing valuable information about the magnitude of the effect. But in this case, the entire exercise of study A was superfluous—we already knew how to eliminate the endogeneity bias in a study that would yield externally valid estimates. Alternatively, if we do not know $E(Z_B|X_B)$, estimation of β_B is infeasible only using observations of X_B and Y_B. Montes-Rojas and Galvao (2014) demonstrate how prior information regarding the conditional mean can allow for Bayesian estimation in the absence of instruments, with the requirement being that we can create an informed prior about the relationship between the endogenous observed and unobserved variables. If we have any

knowledge of that relationship, then additional observations will allow us to update our beliefs regarding the size of parameter β_B.

From this simplified discussion, we can glean a few important points about using multiple methods to cover both internal and external validity.

1. Meaningful causation can only be established if the laboratory experiment is a sufficient analog of field based decisions. Restrictions on the choice set, novel environments or obtrusive observation may cause the relationships to be unique to the laboratory, undermining the case for internal validity.
2. Adding external validity through field data requires some information regarding the magnitude of endogenous relationships that may bias traditional estimates. Instrumental variables, when feasible, provides one way of obtaining this information. Field experiments provide another in which the endogenous variable can be manipulated, though with some limitations (see Roe & Just, 2009; Just & Gabrielyan, 2018).

In both cases, it is key that the prior beliefs regarding the underlying distributions be made explicit. By glossing over these prior beliefs, researchers risk misleading readers into unwittingly believing in results that may only be suggestive of a case for both internal and external validity, or may not add weight to the case at all.

IV. An Example

Behavioral studies on debit and credit card spending show that debit card payment systems induce frivolous purchases and greater overall spending by adults and college

students (Lo & Harvey, 2011). The theory is that using a card obscures the pain of payment, leading to less restraint when faced with a potential purchase. This has led some to suspect that cashless systems within a school lunch environment could lead to additional purchases on snacks and desserts, increasing the risk of overeating and poor nutrition.[1]

In the case of prepaid debit cards, a similar effect may prevail as payment has occurred in the past and one has perhaps already internalized any pain associated with diverting money to the debit account. This effect may be enhanced in the context of a school lunch where the purchaser (a child) is unlikely to be the one who is funding the card (the parent). Prepayment functions as a form of commitment mechanism, with the consumer knowing they need to expend the money on the account eventually (Just et al. 2008). Compared with their counterparts who pay cash for lunch, students using debit cards tend to discount the costs of food and be less sensitive to price variations. This can lead to less thoughtful food choices and overspending. Unrestricted prepayment may also increase people's sensitivity to environmental factors by reducing their overall cognitive engagement and encouraging impulse buying.

Consider first an experiment performed at a Cornell cafeteria (Just et al. 2008). While conducted in a field setting, participants had a reduced number of choices and the incentives offered them were artificial in nature. Thus, this may be regarded as

[1] The examples used in this section are selected both for the keen fit to the point of the paper, and for the fact that the authors of the papers are unlikely to take umbrage at the criticism. This is not intended as a comment on the credibility of any of the authors or the veracity of these studies individually.

being very close to a laboratory experiment. The experiment is designed to test how payment options and the timing of food selection affect participants' food choices. Payment options varied by treatment with a CASH condition in which participants received $20 in cash that could be used to purchase any item on the menu, an UNRESTRICTED debit condition in which participants received $10 in cash and $10 on a debit card that could be used to purchase any food item on the menu, and a RESTRICTED debit condition in which participants received $10 in cash that could be used to purchase any items on the menu, and $10 on a debit card that could only be used to purchase items on the menu designated as healthy. For the purpose of this paper, we focus on two of the three treatments: CASH and UNRESTRICTED. Outcome variables of interest are food choices, calories consumed, nutritional intake, and expenditures. Participants were recruited from the undergraduate population of Cornell campus for convenience, with the majority being freshman business students. Any money left unspent on the debit card was returned to the individual two weeks after completing the study.

Each participant was randomly assigned to one of the three payment treatments, and a total of 323 observations were collected, with 95 in the cash group and 109 in the unrestricted card group. Items available in a nearby food outlet were offered for sale, though the menu was relatively limited including bacon cheeseburgers, chicken breast sandwiches, turkey sandwiches, chicken fingers, French fries, baked potato

chips, salad, macaroni and cheese, peaches, brownies, skim milk, full calorie soda and bottled water.[2]

Results show that individuals using an unrestricted debit card are significantly more likely to purchase a brownie and a soda (less nutritious food) but less likely to buy skim milk and similarly priced healthful side items and desserts (more nutritious food) than those using cash. Given the random assignment, the experiment provides evidence that prepaid cards cause more indulgent purchases.

In a study using secondary data analysis (Just & Wansink, 2013) authors compare food purchases at schools with that only allow debit purchases to those at schools with both debit and cash systems. Data are drawn from the School Nutritional Dietary Assessment Study III recall data collected from a nationally representative sample of 285 public schools within 94 school districts. The field data involves 1,036 students in grades 1-12, among which 725 attend debit-only schools and 311 attend schools that allow either debit or cash. Summary statistics for survey participants show that age, gender, and BMI percentage are not significantly different between the two payment groups.

Results show that students in debit and cash schools purchase more fresh fruits and vegetables and consume fewer total calories. The conclusion is that payment systems

[2] While randomized, some analyses use propensity score matching to control for potential confounding factors such as gender, weight, body mass index (BMI), and hours since the last meal to isolate the impact of payment treatment.

with cash options have a lower purchase incidence of less healthy foods and a higher purchase incidence of more healthy foods, indicating more conscious food choices.

A summary of research results appears in Table 1. In both studies, it seemed that the use of prepaid debt cards was associated with more calories dense items and higher total calories.

Table 1. A Comparison of Debit versus Cash in Two Studies

	Study A		Study B	
Research method	Experiment		Field Data	
Results:	Cash	Unrestricted card	Debit and Cash	Debit only
Calories	644.37(275) $^{+}$	692.14(306.64) $^{+}$	721[a]	752[a]
Calories from more nutritious foods	248.88(198.27) $^{**++}$	192.36(222.97) $^{**++}$	343	311
Calories from less nutritious foods	397.43(346.19) $^{**++}$	502.01(377.42) $^{**++}$	378[a]	441[a]
Expenditures	$6.53(2.26)	$6.33(1.96)	NA	NA

$^{*, **}$ Mean of cash treatment and unrestricted card treatment differ by10, 5 percent

$^{+, ++}$ Difference are significant at 10- and 5-percent level after using the Bonferroni corrected p-value

[a] Payment Mechanism differences at $P<0.05$

On the surface, both studies suggest that the use of debit cards might create a public health issue, having both a causal impact and one that is of a size that is relevant at a policy level. However, if we consider these two examples in light of the theory from the previous section, we might conclude that this is something of a mirage. In particular, study A can only inform study B to the extent that the impact effects are

correlated, which is likely connected to the level of fidelity to realistic choices in the experiment. Here it is notable how the choice incentives differ:

- In study A, money not spent will be returned to the individual after 2 weeks. In study B, money not spent will be saved for future purchases and if returned would go to the parent.
- In study A, participants were in a designated study area when eating and completing a survey and may have been in a hurry to go to leave. In study B, students would have had a fixed amount of time to eat before the next period.
- Students in the study A were restricted and could only spend at most $20 (and only $10 on a card). Students in B spending cash might be limited by the amount of cash their parent gave them that day, while those spending on a debit account would likely be unlimited in their spending.
- The menu in study A was intentionally divided into healthy and less healthy items, with a limited number of items to ensure feasibility. Study B included schools with a wide selection of foods and some with smaller selections. All schools in study B would offer items meeting the school lunch guidelines and many would avoid offering soda specifically, while almost all would offer some options for desserts (cookies, ice cream, etc.) beyond what was available in A.

- Participants in study A were college students who chose to participate in a study in exchange for a cash payment. Participants in B were K-12 students whose parents responded to a survey.

Given the very different choice options, context and perhaps most importantly the difference in financial incentives, it is not at all clear that we would expect a high correlation of the debit effect across these two contexts. Indeed, these may be nearly independent. There are many reasons one might consider the laboratory experiment is not a sufficient analog of field-based choice in this case. If this is the case, any causal effect found in A is not necessarily a sign that we could presume causality in B.

Consider now the potential for learning about the size of the impact from study B. This is a question of whether there is some endogenous selection that might cause a bias in the data. It may seem unlikely that children would select into schools with particular payment options based on their preferences for high calories foods. However, two mechanisms still remain for influencing the results. First, schools may select their payment methods while considering the types of offerings and the revenue they need to fund their activities. For example, a school that is particularly strapped for budget may decide one way to increase their budget is to sell more cookies and candy, and they may recognize this is more likely to sell if students can spend as much as they want per day. Additionally, in the schools where both cash and cards are allowed, parents have a choice as to whether they fund a debit account or send their child with a restricted amount of cash per day. A parent who worried their child was particularly prone to overconsumption may opt to use cash. Each of these may suggest

a different mechanism for why we would see the differences in behavior in B that is not connected to pain of payment.

Income also plays some role in school selection of payment methods. As we could see from Table 1, in debit-only schools, 27% of households have an annual income of less than $30K, whereas in debit-or-cash schools, this number is 36%. In contrast, the higher income group (annual income ranging from $40K to $60K) accounts for 22% in debit-only schools and for 14% in debit-or-cash schools. The statistics are not significantly different for other income groups. Family income could also affect student's food choices by the formation of eating habits. In this case, it is not clear that either of these studies provides substantive additional policy value to the other. Rather, each must be evaluated independently and could only provide weak support for consideration of school debit policies.

V. Conclusion

This manuscript makes a plea for additional thought about how evidence from multiple studies can be additive in support of policy. A simple accumulation of studies with similar sounding results does not imply a stronger argument for policy. Rather, it is important to take into consideration the assumptions that are necessary in order to learn about field effects from laboratory or other studies. While it may seem as though we are arguing for field experiments as the only sufficient study mechanism, in truth we believe there are reasonable criteria in which multiple

methods can bolster a policy case. In order to achieve this, it is important to consider the specific conditions we outline in this paper.

Particularly in establishing behavioral phenomena, the absence of an instrumental variable leads researchers often to turn to field experiments. In some cases, studying and establishing behavioral phenomenon may be entirely reliant on the use field experiments. Nevertheless, such field experiments are not always feasible. Thus we can expect many potentially important results will need to rely on a combination of laboratory data and field data. Overcoming the behavioral policy problem will require a much more thoughtful approach to the marriage of multiple studies.

References

Banerjee, Sanchayan; Matteo M. Galizzi; Peter John; and Susana Mourato. 2023. "Immediate backfire? Nudging sustainable food choices and psychological reactance." *Food Quality and Preference* 109: 104923.

Byrne, Anne T.; and Just, David R. 2019. "Food consumer trends: food experience, pleasure, and policy in the United States: Emerging trends to watch for in American food consumption for researchers, policymakers, and consumers." In *Food and Experiential Marketing*: Routledge: 38-56.

Caputo, Vincenzina; and Just, David R. 2022. "Chapter 92 - The economics of food related policies: Considering public health and malnutrition." In *Handbook of Agricultural Economics*, Elsevier, 6: 5117-5200

Halpern, David; and Sanders, Michael. 2016. "Nudging by Government: Progress, Impact, & Lessons Learned." *Behavioral Science & Policy*, 2(2), 53-65.

Hallsworth, Michael; List, John A.; Metcalfe, Robert D.; and Vlaev, Ivo. 2017. "The behavior list as tax collector: Using natural field experiments to enhance tax compliance." *Journal of Public Economics*, 148, 14-31.

Hummel, Dennis; and Maedche, Alexander. 2019. “How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies.” *Journal of Behavioral and Experimental Economics*, 80: 47-58.

Just, David R.. 2014. *Introduction to behavioral economics: noneconomic factors that shape economic decisions*. Hoboken, NJ: Wiley and Sons.

Just, David R.; and Byrne, Anne T.. 2020. “Evidence-based policy and food consumer behaviour: How empirical challenges shape the evidence.” *European Review of Agricultural Economics*, 47(1): 348-370.

Just, David R.; and Gabrielyan, Gnel. 2018. “Influencing the food choices of SNAP consumers: Lessons from economics, psychology and marketing.” *Food Policy*, 79: 309-317.

Just, David R.; and Wansink, Brian. 2013. “School lunch debit card payment systems are associated with lower Nutrition and higher calories.” *The Obesity Society*, 22(1): 24-26

Just, David R.; and Wansink, Brian; Mancino, Lisa; and Guthrie, Joanne F.. 2008. “Behavioral Economic Concepts To Encourage Healthy Eating in School Cafeterias: Experiments and Lessons From College Students.” *Research in Agricultural & Applied Economics*, Economic Research Report (68).

Kahneman, Daniel; Knetsch, Jack L.; and Thaler, Richard H..1991. “Anomalies: The endowment effect, loss aversion, and status quo bias.” *Journal of Economic perspectives*, 5(1): 193-206.

Kelly, Bridget; Hughes, Clare; Chapman, Kathy; Louie, Jimmy Chun-Yu; Dixon, Helen; Crawford, Jennifer; King, Lesley; Daube, Mike and Slevin, Terry. 2009. “Consumer testing of the acceptability and effectiveness of front-of-pack food labelling systems for the Australian grocery market.” *Health Promotion International*, 24(2): 120–129

Lai, Chien-Yu; List, John A; and Samek, Anya. (2019). ”Got Milk? Using Nudges to Reduce Consumption of Added Sugar.” *American Journal of Agricultural Economics*, 102(1): 154-168.

Leamer, Edward E..1983. “Let's take the con out of econometrics.” *The American Economic Review*, 73(1): 31-43.

Levitt, Steven D.; and List, John A.. 2007a. “What do laboratory experiments measuring social preferences reveal about the real world?” *Journal of Economic perspectives*, 21(2):153-174.

Levitt, Steven D.; and List, John A.. 2007b. “On the generalizability of lab behaviour to the field” *Canadian Journal of Economics/Revue canadienne d'économique*, 40(2): 347-370.

List, John A.; and Samek, Anya. 2017. “A Field Experiment on the Impact of Incentives on Milk Choice in the Lunchroom.” *Public Finance Review*, 45(1): 44-67.

Lo, Hui-Yi; and Harvey, Nigel. 2011. “Shopping without pain: Compulsive buying and the effects of credit card availability in Europe and the Far East.” *Journal of Economic Psychology*, 32:79-92.

McMillan, James H.. 2019. “Randomized field trials and internal validity: Not so fast my friend.” *Practical Assessment, Research, and Evaluation*, 12(1):15.

Montes-Rojas, Gabriel; and Galvao, Antonio F.. 2014. “Bayesian endogeneity bias modeling.” *Economics letters*, 122(1): 36-39.

Neimun, Max; and Stambough, Stephen J.. 1998. “Rational choice theory and the evaluation of public policy.” *Policy Studies Journal*, 26(3): 449-465.

Raghubir, Priya; and Srivastava, Joydeep. 2008. “Monopoly money: The effect of payment coupling and form on spending behavior.”*Journal of Experimental Psychology: Applied*, 14(3): 213–225.

Roe, Brian E.; and Just, David R.. 2009. “Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data.” *American Journal of Agricultural Economics*, 91:1266-1271.

Sacks, Gary; Rayner, Mike; and Swinburn, Boyd. 2009. “Impact of front-of-pack ‘traffic-light’ nutrition labelling on consumer food purchases in the UK.” *Health Promotion International*, 24(4): 344–352

Shadish, William R.; Cook, Thomas D.; and Campbell, Donald T.. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Thaler, Richard H.. 2000. “From homo economicus to homo sapiens.” *Journal of economic perspectives*, 14(1): 133-141.

Thaler, Richard H.. 2016. “Behavioral Economics: Past, Present, and Future.” *American Economic Review*, 106 (7): 1577-1600.

Urbina, Dante A.; and Ruiz-Villaverde, Alberto. 2019. “A critical review of homo economicus from five approaches.” *American Journal of Economics and Sociology*, 78(1): 63-93.