15 The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects – Judd Kessler and Lise Vesterlund

<1>Introduction

Laboratory experiments are used to address a wide variety of questions within economics, including whether behavior is consistent with the predictions and assumptions of theory and how various mechanisms and institutions affect the behavior of economic agents (see Roth 1987 for an overview). The experimental laboratory has become an integral part of the field of economics and a productive dialog now exists between theory, laboratory experiments, and field studies.

Recently, however, a set of papers by Levitt and List (2006, 2007a, 2007b, 2008) has questioned what we can learn from laboratory studies. At the center of their critique is the statement that "the critical assumption underlying the interpretation of data from lab experiments is that the insights gained can be extrapolated to the world beyond" (Levitt and List 2007a, p. 153) and the subsequent argument that there are "many reasons to suspect that these laboratory findings might fail to generalize to real markets" (Levitt and List 2008, p. 909), suggesting that the critical assumption about extrapolation may not hold. Specifically, the authors highlight five factors that differ between the laboratory and "the world beyond."[1] They recognize that external validity is also a concern for field experiments and for naturally occurring data; however, their statement that "field experiments avoid many of the important obstacles to generalizability faced by lab experiments" (Levitt and List 2008, p. 910) has caused many to interpret their papers as an attempt to discredit laboratory experiments and to rank field experiments as a superior methodology.

The papers by Levitt and List have caused quite a stir both inside and outside of the

field of experimental economics. A common response in defense of laboratory experiments has been to counter attack field experiments, arguing that field experiments suffer from the same charges levied at laboratory experiments, namely a lack of external validity.[2]

In his reply to Levitt and List, Camerer (2012) moves beyond the generalizability of field experiments and systematically addresses the five factors that Levitt and List (2007a) argue reduce the generalizability of laboratory studies. Camerer (2012) notes that the features of the lab that differ from the field make less of a difference on behavior than Levitt and List (2007a) suggest. He comments that when concerns about one of these factors arise, lab studies can be altered to better mirror an external environment of interest.[3] In addition, he compares the strengths and weaknesses of data collected in the lab and field, essentially arguing that laboratory experiments are more easily replicated whereas field experiments are less obtrusive. However, Camerer also makes the striking argument that considering external validity is distracting for a large class of laboratory studies. While he states that external validity is crucial for studies that aim to inform policy, he argues that it is not necessary for studies aiming to understand general principles. Referring to the former as the policy view and the latter as the science view, Camerer (2012) argues that since experiments conducted under the science view do not aim to forecast behavior in a particular external target setting we should not care whether these laboratory results generalize to the field.

The papers by Levitt and List and the reply by Camerer (2012) contribute to what many view as an overdue debate on the contribution of laboratory experiments to economics. Unfortunately, much of the debate has been aimed at a straw-man version of external validity. While the debate has centered on the extent to which the *quantitative* results are externally valid, we will argue that for most laboratory studies it is only relevant to ask whether the *qualitative* results are externally valid. A quantitative result refers to the precise magnitude or parameter estimate of an effect and a qualitative result refers to the direction or the sign of the estimated effect.[4] Interestingly, among the authors

on both sides of the debate there is much less (and possibly no) disagreement on the extent to which the qualitative results of a laboratory study are externally valid.

In Section 2, we explain why for most laboratory studies it is only relevant whether the *qualitative* or directional results of the study are externally valid. In Section 3, we argue that laboratory studies are conducted to identify general principles of behavior and therefore promise to generalize. In Section 4, we examine whether laboratory experiments live up to this promise. We discuss the extent to which qualitative results persist outside of the lab, and how we should respond when they do not. We will avoid the debate on whether the concerns about external validity are more or less warranted in laboratory or field environments. We do not see this debate as being productive as it presupposes that the methodologies are in competition. We conclude the paper by arguing, as many others do, that the lab and field methodologies are highly complementary and that both provide important insights to our understanding of economics.[5]

<1>Quantitative versus qualitative external validity

In the debate about whether laboratory studies are "generalizable" or "externally valid," these terms are often not explicitly defined. Indeed, formal definitions of external validity vary substantially. Some definitions of external validity simply require that the *qualitative* relationship between two variables hold across similar environments. For example, Guala (2002, p. 262) states: "an experimental result is internally valid, if the experimenter attributes the production of an effect B to the factor ... A, and A really is the ... cause of B in the experimental set-up E ... [The experimental result] is externally valid ... if A causes B not only in E, but also in a set of other circumstances of interest F, G, H, etc." More demanding definitions of external validity additionally require that the *quantitative* effect of a one unit change in A on B identified in one set-up hold in other comparable settings.

Levitt and List (2007a) describe concerns about laboratory experiments meeting the

higher standard. In their conclusion, they accept that laboratory experiments meet the first definition of external validity, and argue that (emphasis added) "lab experiments that focus on *qualitative* insights can provide a crucial first understanding and suggest underlying mechanisms that might be at work when certain data patterns are observed" (p. 170-171).[6] However they argue that laboratory experiments fail to meet the higher standard, questioning whether (emphasis added) "the experimental findings are *equally* descriptive of the world at large" (p. 158). More directly, Levitt and List (2007b, p. 351) write "even for those experiments that are affected by our concerns, it is likely that the qualitative findings of the lab are generalizable, even when the quantitative magnitudes are not." In responding to Levitt and List, the subsequent debate has centered on the extent to which *quantitative* lab findings are externally valid.

This focus on quantitative external validity is misplaced for many (if not most) experimental studies, however, as the emphasis in laboratory studies is to identify the direction rather than the precise magnitude of an effect. Indeed, the non-parametric statistical methods commonly used to infer significance rely solely on qualitative differences. Few experimental economists would argue that the precise magnitude of the difference between two laboratory treatments is indicative of the precise magnitude one would expect to see in the field or even in other laboratory studies in which important characteristics of the environment have changed.[7] For example, the revenue difference between an English auction and a first-price sealed bid auction in the lab is not thought to be indicative of the quantitative difference one would find between any other set of English and first-price sealed bid auctions. Similarly, despite the clear objective of finding externally valid results, the experiments that tested various designs of FCC spectrum auctions were not aiming to identify magnitudes that would generalize to the field. Instead, they were run with the expectation that the directional effects and general principles of behavior identified in lab auctions also would be present in the field (e.g. Ledyard, Porter and Rangel 1997; Plott 1997).

The emphasis on qualitative results is in part explained by the fact that all theoretical and empirical models require simplifying assumptions. In constructing these models, we eliminate any factors that we think are not central. A consequence of abstracting away from environments of interest is that we likely fail to capture the precise magnitude of the effect we expect to see in those environments.[8]

Since most experimental studies focus on directional effects, the debate about external validity should center on qualitative rather than quantitative predictions. Falk and Heckman (2009) introduce a framework that we can use to conceptualize the difference between the two types of external validity. Considering a relationship between an outcome Y and a number of variables as defined by $Y=f(X_1, X_2, ...X_N)$, which they refer to as an all-causes model since it "captures all possible causes of Y", they note that the causal effect of X_1 on Y is the effect of varying X_1 holding fixed $X^* = (X_2, ...X_N)$.[9] Following on Levitt and List, Falk and Heckman also focus on the conditions under which the quantitative findings of the laboratory are externally valid. They show that the substantial requirements for quantitative external validity are that f is separable in X_1 and Y is linear in X_1. Notice, however, that the requirements securing external validity of the qualitative effects are weaker. For the qualitative results to be externally valid, we simply require Y to be monotonic in X_1 and for changes in X^* to not reverse the relationship of X_1 on Y.

In the all-causes model, the concerns about external validity raised by Levitt and List are concerns that in the laboratory the magnitude of X_1 and the level at which X^* is held fixed do not correspond to environments outside of the lab. If, in contrast to the current debate, the concern is whether the qualitative effect generalizes, then the differences between the laboratory and the field are only relevant if they are thought to reverse the relationship of X_1 on Y.

Take, for example, the winner's curse. Early experimental demonstrations of the winner's curse, using student subjects, found that increasing the number of bidders increased seller revenue while providing public information about the value of the item for

auction decreased seller revenue (Kagel and Levin 1986). Since experience may influence a bidder's understanding of the incomplete information problem at the core of the winner's curse, the quantitative effect of increasing the number of bidders or the effect of increasing public information may differ across subject pools. However, independent of the subject pool, we expect that increasing the number of bidders will increase the number of individuals who fail to understand the winner's curse thus increasing revenue. And we expect that providing public information will mitigate the effect of incomplete information on the bids of anyone who had previously failed to recognize the winner's curse thus decreasing the seller's revenue. The magnitude of these comparative statics could be different between students and, say, oil company executives, but we expect the qualitative results to be the same.[10]

<1>Do laboratory studies promise generalizability?

In the debate about the external validity of laboratory experiments there has been disagreement about when external validity is important. Levitt and List state "the critical assumption underlying the interpretation of data from lab experiments is that the insights gained can be extrapolated to the world beyond." Schram (2005, abstract) has a more moderate statement noting "External validity is relatively more important for experiments searching for empirical regularities than for theory-testing experiments." Camerer (2012) takes this argument one step further and argues that external validity is important for experiments conducted from a policy view, but not for experiments conducted from a scientific view.[11] Camerer (2012) states "since the goal is to understand general principles, whether the 'lab generalizes to the field' ... is distracting, difficult to know ... and is no more useful than asking whether 'the field generalizes to the lab'." (p. 3).[12]

While there may be disagreement on whether there is a promise for quantitative results of an experiment to be externally valid, we do not think there can be much disagreement on the extent to which the qualitative results promise external validity. Since laboratory

experiments are meant to uncover general principles of behavior, it is difficult to see how a concern for external validity is not warranted. Even without a particular external target in mind, the general rules that govern behavior in the experimental environment must apply in other environments with similar characteristics.[13] Surely it is a minority of experimental studies that examine environments that have no counterpart outside of the study and for which we would not expect that the "insights gained can be extrapolated to the world beyond." While laboratory studies may not promise quantitative external validity they do promise qualitative external validity. The question of interest is whether they live up to this promise.

<1>Do laboratory results inform us about the world outside the lab?

Over the course of the debate, authors have suggested two conditions under which we can extrapolate from the laboratory to other environments of interest. Falk and Heckman (2009) summarize the two conditions: "When the exact question being addressed and the population being studied are mirrored in an experiment, the information from it can be clear and informative. Otherwise, to transport experimental findings to new populations or new environments requires a model." Camerer (2012) also highlights that extrapolation is warranted either when the population and environment examined in the laboratory mirrors an environment of interest or when one uses previous studies to account for differences between the lab and field (implying one has an underlying model in mind). Camerer (2012) states "parallelism does not require that students in a lab setting designed to resemble foreign exchange behave in the same way as professional foreign exchange traders on trading floors. ... The maintained assumption of parallelism simply asserts that if those differences could be held constant (or controlled for econometrically), behavior in the lab and the trading floor would be the same." Levitt and List (2007b, p. 364) also stress the value of a model in extrapolating experimental results when they write "even in cases where lab results are believed to have little generalizability, some number [a laboratory

estimate] is better than no number, provided the proper theoretical model is used to make inference."[14]

While mirroring a particular environment of interest or using a model for inference are both appealing, it is important to recognize that these conditions are very stringent. It is difficult to envision a laboratory study that fully mirrors the circumstances of the external environment of interest and it is unrealistic to think that we can find a model that allows us to predict how differences between the lab and the field will interact with any comparative static that we observe in the lab. If these conditions were necessary for external validity, then laboratory studies would provide limited insight about behavior outside the lab. Fortunately, neither of these conditions is necessary for the qualitative results of a lab experiment to extrapolate to other environments of interest. As noted earlier, the qualitative effects will be externally valid if the observed relationship is monotonic and does not change direction when changing the level of variables seen in the field relative to those in the lab.

In a laboratory experiment, subjects are presented with incentives that are meant to capture the central features of the environment in which the economic decisions are usually made. The experimenter has in mind a model that assumes that the laboratory environment does not differ from a comparable field environment on a dimension that would change the sign of the comparative static.[15] Provided the experimental model is correct, the qualitative results should generalize. What does that mean in practice? What can we conclude about behavior outside the laboratory when we reject, or when we fail to reject, a directional hypothesis in the laboratory?

Suppose that laboratory results reject the hypothesis that a variable affects behavior in a certain way. To what extent does this finding allow us to draw inference on the role the manipulated variable will have outside of the laboratory? If in the very controlled laboratory setting we reject our hypothesis, then it is unlikely that the manipulated variable will affect behavior in a more complicated external environment.[16]

What if instead we fail to reject a hypothesis in the lab? Does that imply that the hypothesis is likely to find support in field settings with similar characteristics? Schram (2005, p. 232) argues that "After a theoretical design, a test [of a new airplane] in a wind tunnel is the stage of laboratory experimentation. If it does not 'crash' in this experiment, the plane is not immediately used for the transport of passengers, however. One will typically conduct further tests in the wind tunnel under extreme circumstances. In addition, further testing including 'real' flights without passengers will be conducted." Thus finding lab evidence consistent with a theory will typically lead to repeated investigations of the result, and ideally these will be done under various stress-test conditions in the lab and in the field. Absent these stress tests, however, is it reasonable to expect the documented comparative static in the lab to also hold in the field? The answer may depend on the strength of our prior, but identifying a comparative static in the lab certainly increases our posterior belief that the comparative static will be found in the field. Since the lab is thought to investigate general principles of behavior, we expect these principles to hold both inside and outside of the laboratory.

As with any finding, however, caution is needed to generate predictions in different settings. For example, in documenting statistical discrimination against women in the sports-card market, List (2004) does not claim that women always will be charged a price that is a specific magnitude greater than that for men, or that women always will be charged a higher price, or that there always will be statistical discrimination, but rather that when there are grounds for statistical discrimination against a particular group the market is likely to respond in a predictable way. For example, in a study on taxi fare negotiations in Lima, Peru, Castillo et al. (2009) confirm this prediction by showing that statistical discrimination leads to inferior outcomes for men since they have a greater willingness to pay for taxi rides than do women.

So what should be done if we identify a comparative static in the lab but fail to find evidence of the comparative static outside of the lab?[17] When designing an experiment, the

experimenter assumes the lab setting captures the important characteristics of environments of interest and that the qualitative result will hold outside the lab. Failure to replicate a lab finding in the field may result from the experimenter's model failing to capture central features of the decision environment outside the lab. That is, it is not a question of the laboratory failing to identify general principles of behavior, but rather a question of the laboratory model not capturing the external environment of interest. This is akin to when a result that holds true in a model is not observed in the world. In these cases, we infer that the model has an assumption that does not hold or that the model has abstracted away from something important. Consequently, failure to replicate an experimental finding should cause us to revisit the question at hand, as it may be an indication that the laboratory and field environments were different on a dimension that plays an important role in driving the comparative static results.

For example, theoretical studies by Goeree et al. (2005) and Engers and McManus (2007) along with lab studies by Orzen (2008) and Schram and Onderstal (2009) all demonstrated that all-pay charity auctions generated higher revenue than other fundraising mechanisms, while subsequent field studies contradicted this comparative static. Carpenter et al. (2006) and Onderstal et al. (2010) both found that contributions fell under the all-pay auction. Interestingly, the field studies also demonstrated why this discrepancy may have occurred. While the theory and laboratory experiments had assumed full participation, the field studies found that potential donors will opt out of participating in the all-pay auction. Thus the inconsistencies between the lab and field resulted from an incorrect (and restrictive) assumption of full participation in the auction.[18] By ignoring the importance of participation, the initial laboratory model did not capture an essential feature of the external environment of interest.

When results of a laboratory study are not observed in certain field settings, it is of interest to determine which assumption in the laboratory has failed to hold true. The fact that certain laboratory environments may fail to capture the central features of the

decision environment outside the lab is raised in Kessler (2010), which highlights a distinction between *methodological differences* and *strategic and informational differences*. *Methodological differences* are differences between environments inside and outside the lab that result from laboratory methodology. Factors highlighted by Levitt and List (2007a) like scrutiny and the voluntary participation of the actors are methodological differences, since they systematically differ inside and outside the lab. *Strategic and informational differences* are differences in information, incentives, actions, etc. that might vary from one environment to another and can be manipulated by an experimenter.

While it is tempting to conclude that inconsistencies between lab and field studies result from methodological differences, care should be given to determine whether instead strategic and informational differences are driving the results. Kessler (2010) aims to explain why gift exchange is more commonly seen in laboratory than field experiments. Using laboratory experiments, he shows that differences in the relative wealth of the firm, the efficiency of worker effort, and the action space available to the worker (strategic and informational differences, not methodological ones) contribute significantly to the differences in results between the laboratory studies and the field studies. Another example is the lab and field differences of Dutch and sealed bid auctions. While laboratory studies by Cox, Roberson and Smith (1982, 1983) find that the revenue in sealed bid auctions dominates that in Dutch auctions, a field study by Lucking-Reiley (1999) finds the reverse revenue ranking. While these results initially were ascribed to methodological differences between lab and field, a subsequent laboratory study by Katok and Kwasnica (2008) shows that *strategic and informational differences* can explain the divergent results. Specifically, they note that the clock speed in Lucking-Reiley was much slower than that in Cox et al., and they show that revenue in the Dutch auction is significantly lower than in the sealed bid auction at fast clock speeds, whereas the reverse holds at slow clock speeds.[19] As the initial study failed to account for the effect of clock speed on the revenue ranking, the model was misspecified and the results seen at fast clock speeds did not generalize to

environments with slow clock speeds.

Notice that in these examples the failure of results to generalize between laboratory and field was *not* a failure of laboratory methodology but rather evidence that the laboratory experiment and field experiment differed on an important feature of the decision environment. By identifying which features of the decision environment are causing the differential results (and which are not) we hone our model of behavior.[20]

<1>Conclusion

Economic research aims to inform us of how markets work and how economic agents interact. Principles of economic behavior are expected to apply outside of the unique environment in which they are identified. The expectation and promise of economic research is that the uncovered principles of behavior are general and therefore externally valid. However that promise does not imply that the magnitude of an estimated effect applies generally. In many cases, including many experimental economics studies, the expectation is simply that the qualitative or directional results are generalizable. The simplifying assumptions used to secure internal validity imply that the magnitude of the observed effect will likely differ from the magnitudes in other environments. Interestingly, there appears to be broad agreement that the qualitative results seen in the laboratory are externally valid. To our knowledge, there is no evidence suggesting that the lab-field differences discussed in the ongoing debate reverse directional effects identified in the lab.

In emphasizing the importance of qualitative results, we have ignored the studies that appear to estimate preference parameters in the laboratory. The objective of some of these studies is to derive comparative statics, whereas others emphasize the parameter estimates themselves – and while some of these parameter estimates may be thought to be scale free and generalizable, others are context dependent and therefore unlikely to generalize.

When authors use preference parameters to generate comparative statics, they often do so with the expectation that the comparative statics, rather than the estimated preference

parameters, will generalize. For example, while Andreoni and Vesterlund (2001) estimate male and female demand functions for giving in the laboratory, they solely emphasize the surprising comparative static result that women are less sensitive than men to the price of giving, and it is this comparative static that they subsequently try to extrapolate. They first note that Andreoni, Brown and Rischall (2003) find the same gender difference in price sensitivity when examining how annual giving responds to an individual's marginal tax rate. Then, using data on tipping by Conlin, O'Donoghue, and Lynn (2003), they find that tipping by men is more sensitive to the cost of tipping than it is for women. Thus, despite generating demand estimates for giving, Andreoni and Vesterlund (2001) do not examine whether the quantitative results generalize, instead they use the qualitative results to predict behavior outside the laboratory. This emphasis on comparative statics is also seen in some studies on individual risk preferences, time preferences, and preferences over payoffs to others, which aim to identify general principles such as loss aversion, probability weighting, present bias, and inequality aversion.

While we may expect the comparative statics derived from preference parameter estimates to generalize, it is questionable whether the estimates themselves will generalize. For example, while lab and field studies on other-regarding preferences help identify the general characteristics of behavior that result from such preferences, they are unlikely to identify the magnitude of such effects across domains. Considering the amount of work professional fundraisers put into soliciting funds, it is clear that other-regarding behavior depends greatly on context. One act of charity by an individual cannot predict all his other charitable acts; instead, each charitable act has specific characteristics. Hoping an estimated preference for giving can be extrapolated to all other environments is similar to hoping that we can predict a consumer's demand for all goods from an estimate on demand for one good.

Perhaps because other-regarding preferences are so complex, however, it would be particularly costly to dismiss a research methodology from shedding light on the

phenomenon. Indeed, research from both lab and field experiments have played a significant role in improving our understanding of what triggers giving. As noted earlier, field experiments helped us understand behavior in all-pay charity auctions. Lab experiments have played an important role in helping us understand charitable giving by providing a controlled environment that enables us to identify which mechanisms may be driving behavior.

For example, field studies have repeatedly shown that contributions in many settings can be impacted by information about the contributions made by previous donors, see for example, List and Lucking-Reiley (2002), Croson and Shang (2008), Frey and Meier (2004) and Soetevent (2005). While these studies demonstrate that individuals respond to the contributions of others, they provide little information on which mechanisms may be driving the result. One hypothesis is that information about the contributions of others may provide guidance when there is uncertainty about the quality of the product provided by the organization (e.g. Vesterlund 2003; Andreoni 2006). While one easily can show theoretically that sequential giving can generate an increase in donation, signaling is a difficult behavioral task and it may be questioned whether donors will be able to exploit their ability to signal quality. Unfortunately, the signaling model is not easily tested in the field, as it is hard to isolate changes in charity quality. However it is not difficult to conduct such a study in the laboratory, and indeed a substantial attraction of the lab is that one can easily contrast competing hypotheses. Potters, Sefton and Vesterlund (2005, 2007) investigate sequential giving both with and without uncertainty about the quality of the public good. They find that sequential contributions increase giving when there is uncertainty about the quality of a public good, but not when the quality of the public good is known. Thus, behavior is consistent with individuals seeing large initial contributions as a signal of the charity's quality. This result corresponds with field evidence that information on past contributions has a greater impact on new donors (Croson and Shang, 2008) and a greater impact when contributing to projects for which the donor has less

information on quality (Soetevent, 2005).

Lab and field experiments each add unique and complementary insights to our understanding of economic behavior. Discussions aiming to secure a relative ranking of the two methodologies are both unwarranted and unproductive. Instead, methodological discussions should highlight the ways in which laboratory and field experiments are complements. And ideally, those discussions will spark new research that takes advantage of their combined strengths.

<1> Notes

The authors thank George Lowenstein, Jack Ochs, Alvin Roth and Tim Salmon for their helpful and thoughtful comments, and we thank Guillaume Frechette and Andrew Schotter for inviting us to write this comment.

1. The five factors they discuss are: the level of scrutiny, the lack of anonymity, the context, the stakes, and the population.

2. In an echo of the attacks on laboratory experiments, critics have argued that certain markets studied in the field may differ substantially, and thus provide limited insights about, other markets of interest (not coincidentally, a common example has been the sports-card market studied by List in List (2006)). In addition, proponents of laboratory studies have argued that field experiments also lack internal validity as limitations on control in the field make it more difficult to identify causal relationships. Finally, some have raised concerns about the difficulty of replicating field experiments.

3. Camerer (2012, p. 3) writes: "We then consider which typical features of lab experiments might threaten generalizability. Are those features a necessary part of all lab experiments? Except for obtrusive observation in the lab (which is inherent in human subjects protection), the answer is 'No!'. The special features of lab experiments which might limit generalizability can be relaxed if necessary to more closely match particular field settings."

4. We thank George Loewenstein for pointing out that these definitions differ substantially from their common uses, in which qualitative refers to things that cannot be measured quantitatively. We use these definitions as they are commonly used in the debate see e.g. Levitt and List (2007a).

5. Roth (2008) notes "Lab and field experiments are complements not only with each other, but also with other kinds of empirical and theoretical work." Falk and Heckman (2009) write "Field data, survey data, and experiments, both lab and field, as well as standard econometric methods, can all improve the state of knowledge in the social sciences." In their Palgrave entry on field experiments, Reiley and List (2007) write "the various empirical approaches should be thought of as strong complements, and combining insights from each of the methodologies will permit economists to develop a deeper understanding of our science." Levitt and List (2007b, p. 364) point to the complementarities in stating "we believe that the sharp dichotomy sometimes drawn between lab experiments and data generated in natural settings is a false one.... Each approach has a different set of strengths and weaknesses, and thus a combination of the two is likely to provide more insight than either in isolation." As discussed below, Kessler (2010) highlights a specific way in which laboratory and field results are complements in the production of knowledge.

6. Levitt and List also note that "lab experiments can suggest underlying mechanisms that might be at work when certain data patterns are observed and provide insights into what can happen in other related settings" (2007b, p. 363).

7. While many field experiments are written up to emphasize the magnitude of an estimated effect, it is presumably not the intention of the authors that the level of this magnitude is expected to generalize to other environments. For example, List and Lucking-Reiley (2002) identify a nearly six-fold increase in contributions when they increase seed money for a fundraising goal from 10% to 67%. Few would expect this result to generalize to a six-fold increase in all other charitable campaigns. Presumably the

authors do not report this result in their abstract to suggest that it is quantitatively generalizable, but instead report the result: to demonstrate the strength of the effect, to compare it to the strength of the other results in their paper, and to suggest that it is of economic significance.

8. We would only describe quantitative relationships with our models if all the factors we assumed away were irrelevant for the magnitude of the examined effect.

9. Note that in many experimental studies X_1 is a binary variable indicating different market mechanisms or institutions.

10. Dyer, Kagel and Levin (1989) find that professionals also are subject to the winner's curse. See Frechette (2012) for a review of studies comparing the behavior of students and professionals. Out of 13 studies that allow comparison of professionals and students in standard laboratory games, he finds only one example where the behavior by professionals is closer to what is predicted by economic theory.

11. Camerer (2012, p. 7) states that: "If the purpose of an experiment is to supply a policy-relevant answer to a particular external... setting, then it is certainly valid to ask about how well the experimental setting resembles the target external setting. But this is rarely the case in experimental economics".

12. The many experimental studies on various FCC auction mechanisms demonstrate that policy makers and practitioners are often deeply interested in qualitative (as well as quantitative) effects.

13. For example, Plott (1982) argues that the markets examined in the lab also are real markets and therefore that the general principles of economics demonstrated in the lab should also hold in other markets.

14. Levitt and List (2007b) argue that a model is required to predict outside of the laboratory. "Our approach to assess the properties of the situation is to explore, both theoretically and empirically, how individual behavior changes across judiciously chosen levels of these factors, as moderated by both the task and the agent type. Until this bridge

is built between the lab and the field, any argument concerning behavioral consistency might be considered premature" (p. 363). They also note that the demands on this model are rather substantial, "unless considerable changes are made in the manner in which we conduct lab experiments, our model highlights that the relevant factors will rarely converge across the lab and many field settings....what is necessary are a model and a set of empirical estimates to inform us of when and where we should expect lab behavior to be similar to a particular field environment and, alternatively, when we should expect large differences" (p. 364).

15. If a laboratory experiment were expected to generate a result that was specific to the lab (i.e. rather than a result that identified a general principle) such that the sign of the result might change outside the lab, we contend that the experimenter should not have bothered to run the experiment in the first place.

16. For example, Schram (2005, p. 231) writes: "The bottom line is that there is no reason to believe that a general theory that is rejected in the laboratory would work well in the world outside of the laboratory." Of course this does not mean that the theory being tested is wrong, it just means is that it is not a good approximation of actual behavior.

17. Of course, some studies are conducted in the laboratory precisely because they cannot be conducted in the field. For example, it is difficult to see how a signaling experiment along the lines of Cooper et al. (1997a,b) could be conducted in the field.

18. See also Ivanova-Stenzel and Salmon (2008) for a further illustration that endogenous entry may influence the revenue rankings in auctions. Interestingly, Corazzini et al. (2010) show a similar decrease in participation in the lab when participants in the all-pay public good auction are given heterogenous endowments.

19. Katok and Kwasnica (2008, p. 346) note: "The Cox et al. (1982) study used clocks that descended between 0.75% and 2% of their maximum value every second; the Lucking-Reiley (1999) field study used a clock that decreased approximately 5% per day... Since slower auctions impose higher monitoring and opportunity costs on bidders and are

generally less exciting, the slow clock may cause the bidders to end the auction early." We thank Tim Salmon for suggesting this example.

20. In fact, if factors like scrutiny, decision context, or characteristics of the actors interact importantly with a comparative static in a way that we do not expect, the fact that we did not expect the interaction means our model is misspecified. In particular, it means we have left out an important interaction that will be important to include in the model to make predictions. For example, if only women (or only students) were to respond to the incentive of lowered prices, then a model of demand that does not account for gender (or student status) would fail to explain or predict the effect of prices on behavior.

<1>References

Andreoni, J. 2006. Leadership giving in charitable fund-raising. *Journal of Public Economic Theory* 8: 1-22.

Andreoni, J., E. Brown, and I. Rischall. 2003. Charitable giving by married couples: Who decides and why does it matter? *Journal of Human Resources* XXXVIII(1): 111-133.

Andreoni, J. and L. Vesterlund. 2001. Which is the fair sex? gender differences in altruism. *Quarterly Journal of Economics* 116(1): 293-312

Camerer, C. 2012. The Promise of Lab-Field Generalizability in Experimental Economics: A Reply to Levitt and List (2007). *The Methods of Modern Experimental Economics*. Oxford University Press.

Carpenter, J., A. Daniere, and L. Takahashi. 2006. Space, trust, and communal action: Results from field experiments in southeast asia. *Journal of Regional Science* 46(4): 681-705.

Carpenter, J., J. Holmes, and P. H. Matthews. 2008. Charity auctions: a field experiment. *Economic Journal* 118: 92-113.

Castillo, M., M. Torero, and L. Vesterlund. August 2009. A field experiment on bargaining. Working Paper.

Charles. 1997. Laboratory experimental testbeds: Application to the pcs auction. *Journal of Economics & Management Strategy* 6: 605-638.

Conlin, M., M. Lynn, and T. O'Donoghue. 2003. The norm of restaurant tipping. *Journal of Economic Behavior & Organization* 52(3): 297-321.

Cooper, D. J., S. Garvin, and J. H. Kagel. 1997a. Adaptive learning vs. equilibrium refinements in an entry limit pricing game. *Economic Journal* 107(442): 553-575.

Cooper, D. J., S. Garvin, and J. H. Kagel. 1997b. Signalling and adaptive learning in an entry limit pricing game. *RAND Journal of Economics* 28(4): 662-683.

Corazzini, L., M. Faravelli, and L. Stanca. 2010. A Prize to Give For: an Experiment on Public Good Funding Mechanisms. *Economic Journal*.

Cox, J. C., B. Roberson, and V. L. Smith. 1982. Theory and behavior of single object auctions. *Research in Experimental Economics* 2: 1-43.

Cox, J. C., V. L. Smith, and J. M. Walker. 1983. A Test that Discriminates Between Two Models of the Dutch-First Auction Non-Isomorphism. *Journal of Economic Behavior and Organization* 4: 205-219.

Croson, R. and J. Y. Shang. 2008. The impact of downward social information on contribution decisions. *Experimental Economics* 11(3): 221-233.

Dyer, D., J. H. Kagel, and D. Levin. 1989. Resolving uncertainty about the number of bidders in independent private- value auctions: An experimental analysis. *RAND Journal of Economics* 20(2): 268-279.

Engers, M. and B. McManus. 2007. Charity auctions. *International Economic Review* 48(3): 953-994.

Falk, A. and J. J. Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326: 535-538.

Forsythe, R., Horowitz, J. L., Savin, N. E., 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* 6: 347-369.

Fréchette, G. R. 2012. Laboratory experiments: Professionals versus students. *The*

Methods of Modern Experimental Economics. Oxford University Press.

Frey, B. S. and S. Meier. 2004. Social comparisons and pro-social behavior: Testing 'conditional cooperation' in a field experiment. *American Economic Review* 94: 1717-1722.

Goeree, J. K., E. Maasland, S. Onderstal, and J. L. Turner. 2005. How (not) to raise money. *Journal of Political Economy* 113(4): 897-918.

Guala, F. 2002. On the scope of experiments in economics: comments on Siakantaris. *Cambridge Journal of Economics* 26(2): 261-267.

Ivanova-Stenzel, R. and T. C. Salmon. 2008. Revenue equivalence revisited. *Games and Economic Behavior* 64(1): 171-192.

Kagel, J. H. and D. Levin. 1986. The winner's curse and public information in common value auctions. *American Economic Review* 76(5): 894-920.

Katok, E. and A. Kwasnica. 2008. Time is money: The effect of clock speed on seller's revenue in dutch auctions. *Experimental Economics* 11: 344-357.

Kessler, J. B. May 2010. Strategic and informational environment: Addressing the lab-field debate with laboratory gift exchange experiments. Working Paper.

Ledyard, J. O., D. Porter, and A. Rangel. 1997. Experiments testing multiobject allocation mechanisms. *Journal of Economics & Management Strategy* 6: 639-675.

Levitt, S. D. and J. A. List. June 2006. What do laboratory experiments tell us about the real world? Working Paper.

Levitt, S. D. and J. A. List. 2007a. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives* 21: 153-174.

Levitt, S. D. and J. A. List. 2007b. Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics* 40: 347-370.

Levitt, S. D. and J. A. List. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review* 53: 1-18.

List, J. A. 2004. Substitutability, experience, and the value disparity: evidence from the marketplace. *Journal of Environmental Economics and Management* 47(3): 486-509.

List, J. A. 2006. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy* 114: 1.

List, J. A. and D. Lucking-Reiley. 2002. The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign. *Journal of Political Economy* 110(1): 215-233.

List, J. A., Reiley, D. 2008. Field experiments. In *The New Palgrave Dictionary of Economics*, eds. S. N. Durlauf and L. E. Blume. Palgrave Macmillan Publishing.

Lucking-Reiley, D. 1999. Using field experiments to test equivalence between auction formats: Magic on the internet. *American Economic Review* 89(5): 1063-1080.

Onderstal, S., A. J. H. C. Schram, and A. R. Soetevent. 2010. Bidding to Give in the Field: Door-to-Door Fundraisers had it Right from the Start. Working paper.

Orzen, H. October 2008. Fundraising through competition: Evidence from the lab. Discussion Paper 2008-11, CeDEx, University of Nottingham.

Plott, C. R. 1982. Industrial organization theory and experimental economics. *Journal of Economic Literature* 20(4), 1485-1527.

Potters, J., M. Sefton, and L. Vesterlund. 2005. After you-endogenous sequencing in voluntary contribution games. *Journal of Public Economics* 89(8): 1399-1419.

Potters, J., M. Sefton, and L. Vesterlund. 2007. Leading-by-example and signaling in voluntary contribution games: an experimental study. *Economic Theory* 33(1): 169-182.

Roth, A. E. 1987. Laboratory experimentation in economics, and its relation to economic theory. In *Scientific Inquiry in Philosophical Perspective*. University Press of America.

Schram, A. J., Onderstal, S., May 2009. Bidding to give: An experimental comparison of auctions for charity. *International Economic Review* 50(2): 431-457.

Schram, A. J. H. C. 2005. Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology* 12(2): 225-237.

Soetevent, A. R. 2005. Anonymity in giving in a natural context – a field experiment in

30 churches. *Journal of Public Economics* 89(11-12): 2301-2323.

Vesterlund, L. 2003. The informational value of sequential fundraising. *Journal of Public Economics* 87: 627-657.