Validating List Experiment Estimates

Against an Incentive-Compatible Behavioral Measure

John A. List

University of Chicago, ANU, and NBER

First draft: Spring 1993

This draft: Winter 2026

Abstract

List Experiments are widely used across the social sciences to measure sensitive attitudes and behaviors, yet no prior study has validated their estimates against an incentive-compatible behavioral measure. I conduct a field experiment with 400 subjects at a sports card show, combining List Experiment treatments for willingness to pay, one for wolf reintroduction in Yellowstone Park, one for a graded sports card, with a Vickrey second-price auction that provides a real-money benchmark. The List Experiment estimates 26% would pay $50 for the card, compared to 22% who bid at least that amount in the auction; this difference is not statistically significant. These results provide the first criterion validity test of a List Experiment and suggest the method holds promise as a parsimonious alternative to conventional stated preference approaches in settings where survey space constraints preclude standard bias-mitigation interventions.

Keywords: CVM, benefit cost, hypothetical bias, List Experiment

JEL: C83, C93, Q51.

Acknowledgements

This study is a revised version of a previous study titled "Valuing Non-Marketed Goods and Services Using a List Field Experiment." The work was originally written for a class project in Ralph d'Arge's Environmental Economics course in the Spring of 1993 at the University of Wyoming. The paper received a strong A, so I figured I should revise it for a research publication. I tried to maintain as much of the original data analysis and writing as possible. Yet, the results remain unpublished to date. Excellent comments from two reporters, Ian Batemen, Marije Schaafsma, and Christian Vossler improved the study.
For correspondence: jlist@uchicago.edu

The core aim of environmental policy is straightforward: safeguard human well-being and protect the ecosystems crucial to our survival. With benefit-cost analyses now a staple at the federal level and increasingly at the state level, there's been a significant push toward developing robust, credible methods for estimating the benefits and costs of environmental programs. While policymakers have become adept at quantifying the explicit costs of enhanced environmental protection, they currently rely on a variety of methods to measure the economic value of environmental goods and services. Among these, the Contingent Valuation Method (CVM) stands out as particularly contentious. This method, which allows researchers to gauge the total value of the commodity in question, has faced scrutiny, most notably in the critical review by Diamond and Hausman (1994).

Chief among these concerns is the question of whether hypothetical bias is inherently present in Contingent Valuation Method (CVM) responses. Emerging evidence from recent studies highlights notable differences between responses to real and hypothetical valuation questions. To tackle this issue, researchers have employed both ex ante and ex post methods to mitigate hypothetical bias (see, for example, Cummings and Taylor, 1999; List, 2001; Hofler and List, 2004; Carson et al., 2014; Loomis, 2014; and for meta-analyses, see List and Gallet, 2001 and Murphy et al., 2005).

This study diverges from existing literature by employing a List Experiment, a technique that offers a distinct approach to mitigating response bias. In brief, participants are randomly sorted into treatment or control groups, where the control group receives a fixed list of n innocuous statements, while the treatment group receives an n+1 list that includes a single sensitive statement (e.g., a valuation statement) added to the same n innocuous items. Respondents report only the total number of statements they endorse, not which specific ones. Comparison of the mean counts between treatment and control groups then identifies the proportion of respondents who endorse the sensitive item, without any individual's response to that item being directly observable. Subsequent research in political science and other social sciences has extended this simple difference-in-means comparison to multivariate regression frameworks that control respondent characteristics (Imai, 2011; Blair and Imai, 2012).

The List Experiment potentially mitigates biased responding through at least two channels. First, by embedding the sensitive valuation item within a list of non-sensitive statements, the technique diffuses respondent attention across multiple items rather than focusing it on a single valuation question. This is analogous to how discrete choice experiments employ multiple attributes to reduce the salience of any one dimension, thereby attenuating strategic or socially desirable responses. Second, the control items in the list typically involve factual statements about actual behaviors (e.g., recycling habits, past charitable donations), which may establish a pattern of truthful responding that carries over to the embedded sensitive item. These features distinguish the List Experiment from other approaches to improving response quality, such as the randomized response technique (Coutts and Jann, 2011), which obscures individual responses through probabilistic mechanisms, and the dissonance-minimizing format (Blamey, Bennett, and

Morrison, 1999), which modifies the elicitation question to allow respondents to express qualified support. The List Experiment instead preserves the standard valuation question while concealing the individual's answer through aggregation.

The technique was pioneered by political scientists in the 1980s to measure racial prejudice and attitudes towards affirmative action. I draw my inspiration from Sniderman et al. (1992), which I believe was the first published List Experiment in political science.[1] Since then, there have been hundreds of studies in the social sciences using the List Experiment. Imai (2011) offers a comprehensive review of List Experiments in political science, discussing their statistical analysis and methodological considerations. Blair and Imai (2012) present a review of List Experiments in social science research, focusing on their application in measuring sensitive attitudes and behaviors. Treibich and Lépine (2020) provide an overview of List Experiments in health research, specifically examining their use in estimating the prevalence of sensitive health behaviors. Coutts and Jann (2011) review the use of List Experiments in sociology, discussing their effectiveness in reducing social desirability bias. Glynn (2013) offers a critical review of List Experiments across various social science disciplines, examining their strengths and limitations.

These studies collectively demonstrate the widespread use of List Experiments across the social sciences, including political science, sociology, and public health research. They highlight the method's effectiveness in measuring sensitive topics while also discussing its methodological challenges and potential improvements.

My study represents a novel application in the realm of economics experiments on non-market valuation. To the best of my knowledge, this is the first application of a List Experiment in the field to value non-marketed goods and services (indeed, when I conducted this field experiment I think it was the first application in the Economics literature). As discussed above, the List Field Experiment holds significant promise for benefit estimation within CVM, as it potentially mitigates social desirability bias by embedding the sensitive question within a list of non-sensitive items. This contribution addresses an important gap identified in the recent stated preference guidelines literature. Johnston et al. (2017, Recommendation 13) call for research into approaches that can identify and mitigate the effects of response biases, including social desirability effects, in stated preference studies. The List Experiment offers precisely such an approach.

My design is straightforward and follows the general approach of Sniderman et al. (1991). In the Spring of 1993, I conducted an experiment at a sports card show in Denver, CO that had

[1]Sniderman et al. (1992) measured racial attitudes using the National Race and Politics Survey. The control group received a list of three statements about things that made them angry (e.g., "the federal government increasing the tax on gasoline"). The treatment group received the same three statements plus a fourth: "a Black family moving next door." The difference in mean counts between groups provided an estimate of racial animus that did not require respondents to directly reveal a socially undesirable attitude.

two treatments: the sports card treatment and the Yellowstone Park treatment. For each treatment, I developed two sub-treatments: a control group list of four items and a treatment group list of five items. The treatment group list included all the items from the control list plus one additional item.

For the Yellowstone Park treatment, the added item was "I would pay a $100 one-time tax to reintroduce wolves into Yellowstone Park." For the sports card treatment, the added item was "I would pay $50 for a Ken Griffey Jr. 1989 Upper Deck PSA 10."[2] To explore external validity, I also conducted a Vickrey second-price auction for the Griffey Jr. sports card treatment subjects by holding a real auction for the card. I am unaware of such an incentive-compatible external validity test being done in the literature on List Experiments, either within economics or the broader social sciences.

The pilot results, drawn from 400 subjects, are encouraging. For the Yellowstone Park List treatment, the data suggest that 31% of respondents were willing to pay the $100 one-time tax to reintroduce wolves into Yellowstone Park. In the sports card treatment, 26% of subjects revealed that they would pay $50 for the Griffey card. Notably, in the auction for the Griffey Jr. sports card, approximately 22% of bidders bid at least $50. A test of proportions shows that these percentages are not significantly different at conventional levels. While these findings mark a step forward, they also present a challenge for future researchers to replicate these pilot results and to design List Experiments that can accurately recover true values in environments that have extraneous factors that can influence stated preferences.

I view the overarching set of results as important for policymakers, as they potentially provide a reliable method for valuing non-market goods and services, ensuring accurate and informed decisions about resource allocation and policy priorities. By understanding public support for environmental initiatives, such as the reintroduction of wolves to Yellowstone, policymakers can better gauge public preferences and justify expenditures. This approach therefore aids in designing policies that align with constituents' values and enhances the credibility and transparency of policy decisions, fostering greater public trust. Additionally, reliable valuation methods improve benefit-cost analyses, leading to better societal outcomes.

The remainder of the study proceeds as follows. The next Section describes the List field experimental design. Section III presents the experimental results. Section IV discusses the findings and their implications. Section V concludes.

II. Experimental Design

[2]PSA stands for Professional Sports Authenticator, a widely used third-party grading service for sports cards and memorabilia. A PSA 10, designated "Gem Mint," is the highest possible grade, indicating the card is in virtually perfect condition with sharp corners, full gloss, and precise centering.

I conducted the experiment at a sportscard show in Denver, CO in March 1993. For similar experiments conducted in this setting please see List and Shogren (1998) and List (2001, 2002, 2003), though this List field experiment pre-dated all others. The experiment followed a three- or four-step process. In Step 1, either I or my assistant approached a person entering the show and asked if they would like to participate in a study that would last no more than 10 minutes. If the individual agreed, in Step 2, I placed them into one of two treatments.

If they were placed in the Yellowstone Park treatment, then I further randomized if they were in treatment or control. For those in the control group, I asked one simple question: Please indicate how many (not which) of the following four statements apply to you:

> I recycle every week.
>
> I support environmental initiatives around Denver.
>
> In the past, I have donated to wildlife conservation causes, such as the NWF or the Sierra Club.
>
> I have tried to educate people, either through chatting or writing, on environmental issues in the past.

For those in the treatment group, I asked the same question, but I included a fifth statement:

> I would pay a $100 one-time tax to re-introduce wolves into Yellowstone Park

Since reintroduction of wolves into Yellowstone Park was a current local topic of interest for Coloradans, this fifth option needed little further description.

If the subject was placed into the sports card treatment, then I also further randomized them into treatment or control. For the control group, I included the following list:

> I own a baseball, or a baseball card, autographed by a current star player.
>
> I have traded sports cards with dealers at a card show in the last year.
>
> I have friends who are collectors.
>
> I have attended more than ten baseball card shows in the past year.

And, for the treatment group, I augmented that list by adding one more statement:

> I would pay $50 for a Ken Griffey 1989 Upper Deck PSA 10.

For subjects who were placed in the sports card treatment group, they proceeded to Step 3. In Step 3, I told the participant that he or she now had the chance to bid on the Griffey card from the survey. I then retrieved the card from under my dealer's table and the participant could pick up and visually examine the Griffey card, which was sealed with the PSA grade clearly marked on the cardholder. I further told the subjects that they would be bidding in a sealed bid second-

price auction and that for this type of auction it was best for them to bid their true value for this item because overbidding may cause them to pay too much, and underbidding decreases their odds of winning the item.

Finally, in Step 4 we debriefed the participant (Step 3 for non-auction subjects). The monitor explained that the actual bids would be ranked, and the participant would be contacted within 3 days after the show if he or she was the highest bidder. For those not in this treatment arm, they were thanked for their participation.

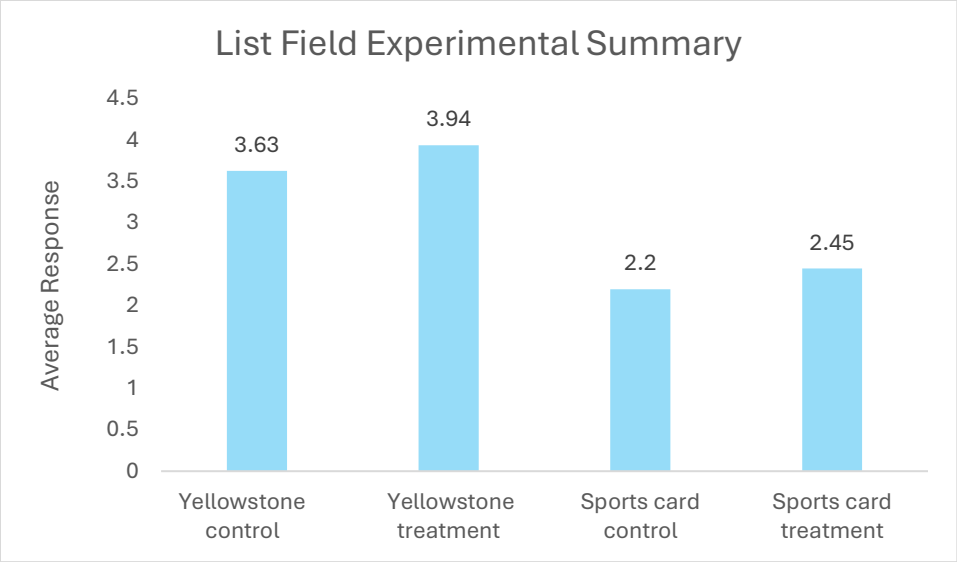
The treatments took approximately 18 hours to complete over two days. On the top of each hour the treatment was switched. I conducted each treatment until 100 subjects participated. Since no participant took part in more than one treatment, I had 400 subjects in total.

III. Experimental Results

Experimental results for List experiments are straightforward to present. Figure 1 shows the summary results for the average number of statements the subject agreed with in each treatment type. As Figure 1 reveals, the average number of statements the control group subjects agreed with in the Yellowstone Park treatment was 3.63 whereas the comparable figure for the treatment group was 3.94. The difference (3.94 - 3.63 = 0.31) suggests that approximately 31% of respondents would be willing to pay a $100 one-time tax for wolves to be re-introduced to Yellowstone Park.

For the sports card treatment, Figure 1 shows that the average number of statements the control group subjects agreed with was 2.20 whereas the comparable figure for the treatment group was 2.45. The difference (2.45 – 2.19 = 0.26) suggests that approximately 26% of respondents are willing to pay $50 for the Ken Griffey 1989 Upper Deck PSA 10. Interestingly, this figure compares well to the 22% of subjects who bid $50 or more for the Ken Griffey 1989 Upper Deck PSA 10 card in the second-price auction. When doing a test of proportions, I fail to reject the homogenous null of proportional equivalence at conventional levels.

Figure 1: Field Experimental Results



Note: Figures summarize results from the List Field Experiment.

IV. Discussion

The experimental results reported above provide initial evidence that the List Experiment can yield valuation estimates that are broadly consistent with incentive-compatible mechanisms. In this section, I discuss the interpretation of these results and the implications for future research.

A natural first question is how to interpret the comparison between the two treatments. The Yellowstone Park treatment yields an estimated 31% willing to pay the $100 one-time tax, while the sports card treatment yields 26% willing to pay $50. These figures are not directly comparable, as they involve different goods (a public environmental good versus a private collectible), different price points ($100 versus $50), and different populations of interest (though all subjects were sportscard show attendees). The more informative comparison is within the sports card treatment, where both a List Experiment estimate and an incentive-compatible benchmark are available.

The sports card results deserve careful attention. The List Experiment estimate (26%) exceeds the second-price auction estimate (22%) by 4 percentage points, though this difference is not statistically significant at conventional levels. This is an encouraging finding. The standard hypothetical bias literature typically finds that hypothetical valuations exceed real valuations by a factor of two or more (List and Gallet, 2001; Murphy et al., 2005; Loomis, 2014). The residual gap observed here, if it is real and not simply noise, is an order of magnitude smaller than what is typically documented. This suggests that the List Experiment substantially attenuates hypothetical bias, even if it may not eliminate it entirely.

It is also worth noting what the comparison between the List Experiment and the second-price auction means for the broader literature on List Experiments in the social sciences. To my knowledge, no prior List Experiment study has validated its estimates against an incentive-compatible behavioral measure. The close correspondence between the List Experiment and auction results therefore provides a novel form of criterion validity that strengthens the case for the technique's accuracy, not only in valuation contexts but potentially in other domains where List Experiments are used to measure sensitive attitudes and behaviors.

A legitimate concern, raised in the stated preference guidelines literature (Johnston et al., 2017), is how the List Experiment approach relates to current best practices for CVM surveys. Conventional CV practice emphasizes detailed scenario descriptions, explicit payment mechanisms, incentive compatibility, consequentiality, and supplementary interventions such as cheap talk scripts or certainty follow-ups. The List Experiment, by its nature, embeds the valuation question as a single brief statement within a list, making it difficult to provide the same level of scenario detail or to incorporate standard augmentations without compromising the camouflage that makes the technique effective.

I view this tension not as a fatal limitation but as a clarification of the approach's comparative advantage. There are important settings where conventional CV augmentations are infeasible or undesirable. National surveys that include a short valuation module typically impose strict space constraints; researchers may be allocated only a few questions and cannot include lengthy scenarios, cheap talk scripts, or batteries of follow-up questions to assess credibility and certainty. In these contexts, the List Experiment offers a parsimonious alternative that trades detailed scenario description for enhanced response privacy. Similarly, for contingent behavior questions commonly included in recreation demand surveys (e.g., "How many additional trips would you take if water quality improved from X to Y?"), the scenarios are inherently brief and the List format can be readily accommodated.

Several extensions of the basic design could enhance the method's utility. First, varying the bid amount across treatment cells would allow researchers to trace out a WTP distribution using nonparametric methods, rather than estimating the proportion willing to pay at a single price point. Second, the sensitive item could be framed as a referendum-style voting statement (e.g., "I would vote in favor of a ballot measure to..."), which aligns the List Experiment more closely with consequential referendum formats used in modern CV. Third, the List Experiment could be combined with the randomized response technique or other privacy-enhancing methods to provide additional layers of response protection. Fourth, advances in the statistical analysis of List Experiments (Blair and Imai, 2012; Glynn, 2013) enable multivariate regression approaches that can examine how WTP varies with respondent characteristics, moving beyond the simple difference-in-means comparison employed here.

V. Conclusion

The application of a List Field Experiment to address hypothetical bias in CVM holds significant promise for improving the accuracy of economic valuations for nonmarket commodities. Accurate valuations are critical for informed policymaking and resource allocation, particularly in environmental economics, where the true value of public goods like clean air, water, and natural landscapes must be estimated. By reducing the gap between stated intentions and actual behaviors, the List Experiment enhances the credibility and reliability of survey results, thereby leading to more effective and justified policy decisions that better reflect the public's preferences and willingness to pay for non-marketed goods and services. This, in turn, supports sustainable development and the efficient management of environmental resources.

The pilot results presented here are encouraging but should be viewed as a starting point. Future research should investigate the List Experiment's performance across different types of environmental goods, varying price levels, and larger samples. Within-study comparisons of the List Experiment against conventional CV, with and without standard bias-mitigation tools, would be particularly valuable for understanding the tradeoffs involved. I hope this paper stimulates such research and expands the methodological toolkit available to stated preference researchers.

References

Blamey, Russell K., Jeff W. Bennett, and Mark D. Morrison. 1999. "Yea-Saying in Contingent Valuation Surveys." Land Economics 75 (1): 126–141.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." Political Analysis 20 (1): 47–77.

Carson, Richard T., Theodore Groves, and John A. List. 2014. "Consequentiality: A Theoretical and Experimental Exploration of a Single Binary Choice." Journal of the Association of Environmental and Resource Economists 1 (1/2): 171–207.

Coutts, Elisabeth, and Ben Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)." Sociological Methods & Research 40 (1): 169–193.

Cummings, Ronald G., and Laura O. Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method." American Economic Review 89 (3): 649–665.

Diamond, Peter, and Jerry Hausman. 1994. "Contingent Valuation: Is Some Number Better Than No Number?" Journal of Economic Perspectives 8: 45–64.

Glynn, Adam N. 2013. “What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment.” Public Opinion Quarterly 77 (S1): 159–172.

Hofler, Richard A., and John A. List. 2004. “Valuation on the Frontier: Calibrating Actual and Hypothetical Statements of Value.” American Journal of Agricultural Economics 86 (1): 213–223.

Imai, Kosuke. 2011. “Multivariate Regression Analysis for the Item Count Technique.” Journal of the American Statistical Association 106 (494): 407–416.

Johnston, Robert J., Kevin J. Boyle, Wiktor Adamowicz, Jeff Bennett, Roy Brouwer, Trudy Ann Cameron, W. Michael Hanemann, Nick Hanley, Mandy Ryan, Riccardo Scarpa, Roger Tourangeau, and Christian A. Vossler. 2017. “Contemporary Guidance for Stated Preference Studies.” Journal of the Association of Environmental and Resource Economists 4 (2): 319–405.

List, John A. 2001. “Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards.” American Economic Review 91 (5): 1498–1507.

List, John A. 2002. “Preference Reversals of a Different Kind: The ‘More Is Less’ Phenomenon.” American Economic Review 92 (5): 1636–1643.

List, John A. 2003. “Does Market Experience Eliminate Market Anomalies?” Quarterly Journal of Economics 118 (1): 41–71.

List, John A., and Craig A. Gallet. 2001. “What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? Evidence from a Meta-Analysis.” Environmental and Resource Economics 20 (3): 241–254.

List, John A., and Jason F. Shogren. 1998. “Calibration of the Difference Between Actual and Hypothetical Valuations in a Field Experiment.” Journal of Economic Behavior & Organization 37 (2): 193–205.

Loomis, John B. 2014. “2013 WAEA Keynote Address: Strategies for Overcoming Hypothetical Bias in Stated Preference Surveys.” Journal of Agricultural and Resource Economics 39 (1): 1–13.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead. 2005. “A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation.” Environmental and Resource Economics 30 (3): 313–325.

Sniderman, P. M., P. E. Tetlock, and T. Piazza. 1992. Codebook for the 1991 National Race and Politics Survey. Survey Research Center, Berkeley, California.

Treibich, Carole, and Aurélia Lépine. 2020. “Estimating Misreporting in Sensitive Health Behaviours: Evidence from Condom Use of Female Sex Workers in Senegal.” PLoS ONE 15 (4): e0231670.