Willingness to Pay for an Agricultural Technology: An Economic Application of List Experiments

Bailey Peterson-Wilhelm, University of Saskatchewan, bailey27@ksu.edu
Benjamin Schwab, Kansas State University, benschwab@ksu.edu
Sara Burrone, Food and Agriculture Organization of the United Nations (FAO), Sara.Burrone@fao.org

Selected Paper Presentation prepared for presentation at the 2025 AAEA & WAEA Joint Annual Meeting
in Denver, CO; July 27-29, 2025

Willingness to Pay for an Agricultural Technology: An Economic Application of List Experiments

Bailey Peterson-Wilhelm[a]

Benjamin Schwab[a]

Sara Burrone[b]

Abstract

List experiments utilize indirect survey questions to reduce social desirability bias in measures of sensitive behaviors and sentiments. While often used to assess retrospective behavior or opinions of respondents, list experiments have not been widely applied to assessing "deep" parameters of economic models, such as willingness to pay. Common stated preference methods of estimating willingness to pay may be impacted by social desirability bias, particularly when a product has been provided to survey recipients for free. List experiments can uncover the share of respondents willing to pay a given price while reducing social desirability bias. Repeating the method at a variety of prices recovers a partial demand curve. This study discusses the conditions required to satisfy the list experiment validity assumptions and demonstrates the method in an e-extension platform randomized control trial in Sri Lanka. We show that the "no design effect" assumption for list experiments requires that the budget constraint for a household be nonbinding. Under conditions where that assumption is likely to hold, we find direct estimates overstate willingness to pay at low prices. Our findings suggest list experiments may provide a cheap method of more accurately assessing the typically large share of respondents unwilling to pay any non-zero-sum (extensive margin), but are less effective at reducing bias from exaggerated demand (intensive margin).

[a] Kansas State University

[b] Food and Agricultural Organization

1 Introduction

Producers' willingness-to-pay (WTP) for an agricultural technology in low- and middle-income countries is often of interest to academics, companies, governments, and nongovernmental organizations (NGOs). When a product has previously been given at a subsidized or zero price as part of a government or NGO intervention, WTP is relevant for decisions about the scale and longevity of the program. However, estimating the WTP may be a secondary or tertiary objective of a survey, with evaluating program objectives and outcomes taking precedence. In these events, as is often the case (Breidert, Hahsler, and Reutterer, 2006), the method to obtain WTP is largely dictated by the remaining time and resources.

Researchers have evaluated many strategies for measuring WTP (Breidert, Hahsler, and Reutterer, 2006; Miller et al., 2011). These methods can be broadly classified into revealed and stated preference methods (Breidert, Hahsler, and Reutterer, 2006). Revealed preference methods require actual purchase decisions in response to prices, and can be applied to market data or experimental designs conducted in the field or a laboratory setting. Stated preference methods, on the other hand, estimate WTP from consumer responses to hypothetical purchase or valuation decisions. These methods span a variety of techniques, including open-ended question formats, inferred valuation, and discrete choice analysis. Both types of WTP elicitation methods carry important trade-offs.

Revealed preferences methods, such as a Becker-DeGroot-Marschak (BDM) auctions, have been used to uncover WTP in the agricultural technology space (Berry, Fischer, and Guiteras, 2020; Channa et al., 2019; Fuller and Ricker-Gilbert, 2021; Schwab and Yu, 2024). Because these auctions require money to be provided in exchange for the good, the participants' bids should reflect their true demand. Thus, estimated WTP in such incentive compatible auctions should closely approximate true WTP. However, auction methods are intensive to deploy in surveys. They require additional time, training, and cognitive effort. If WTP is not the primary objective of data collection efforts, then revealed preference methods may not be feasible.

Stated preference questions are less intensive. They can be asked in a few questions appended to an existing evaluation survey. However, without the transfer of funds, stated WTP tends to yield inflated WTP estimates, which is commonly referred to as hypothetical bias (Champ, Moore, and Bishop, 2009; Murphy et al., 2005; Schmidt and Bijmolt, 2020). There are many causes of hypothetical bias that fall into two categories termed survey bias and social desirability bias by Entem et al. (2019). Survey bias includes elements that are related to the survey implementation itself. It can be circumvented by increasing the consequentiality of the question by, for example, adding a cheap talk script or certainty follow-up questions (Champ, Moore, and Bishop, 2009; Carlsson, Kataria, and Lampi, 2018; Entem et al., 2019).

Social desirability bias is driven by social pressures when the item of interest has normative expectations. Unlike survey bias, social desirability bias can influence responses in both stated and revealed preference contexts (List et al., 2004; Menapace and Raffaelli, 2020). The impact of social desirability bias on revealed preference methods is more severe, due to the increased scrutiny in an experimental setting (Levitt and List, 2007; Lusk and Norwood, 2009a), the costless nature of the signal (Carlsson, Daruvala, and Jaldell, 2010; Norwood and Lusk, 2011; Menapace and Raffaelli, 2020), and in some instances, the presence of the enumerator or experimenter (Leggett et al., 2003; List et al., 2004).

Most of the literature that examines social desirability bias in WTP is estimating the value of a public or environmental good, such as species at risk of extinction (Entem et al., 2019), animal welfare (Lusk and Norwood, 2010), water quality (Carlsson, Kataria, and Lampi, 2018), natural landscapes (Yadav, Van Rensburg, and Kelley, 2013), national parks (Leggett et al., 2003), and other related topics. There has been considerably less focus on social desirability bias influencing estimates of WTP for private goods, with Menapace and Raffaelli (2020) being a notable exception. Still, there are normative social pressures in the context of new, promoted agricultural technologies, not yet available in a market. The existence of the survey itself is a signal the product should be valued (Carlsson, Kataria, and Lampi, 2018; Zizzo, 2010). If a product has been promoted through training or subsidies, the signal is clear - this item should be valued at a positive price.

Outside of WTP, social desirability bias has been examined in social science literature, primarily examining behavior related to sensitive subjects (Cullen, 2023; Jouvin, 2023; Lépine et al., 2020; Lépine, Treibich, and D'Exelle, 2020; Rosenfeld, Imai, and Shapiro, 2016; Tadesse, Abate, and Zewdie, 2020). Respondents may be unwilling to honestly discuss sensitive subjects, and therefore may not honestly respond to direct survey questions on such topics. Their resulting responses may be misrepresented based on their perception of the socially correct answer. Four necessary conditions must be present for social desirability bias to influence responses: (1) a social referent, (2) the respondent assumes to know the referent's preferred response (3) the respondent must respond directly, and (4) some repercussion for not giving the preferred response (Blair, Coppock, and Moor, 2020).

Those four criteria are likely to be met when evaluating WTP for a promoted agricultural technology. An enumerator or the implementing agency would serve as the social referent; the person the respondent has in mind when responding. As noted above, reporting being unwilling to pay their own money for something they have been receiving free of charge may not be perceived as the ideal response on behalf of the referent. Directly responding to questions about WTP means the referent will certainly know the response. Finally, respondents may feel shame for being unwilling to enter the market or fear the service being discontinued if they reveal they do not want to pay for the product themselves.

The same criteria provide insight into mitigation strategies. If one criterion is disrupted, then social desirability bias should be eliminated. In the context of a survey, and especially a survey administered in person by an enumerator, a social referent cannot be removed. Similarly, the social expectation surrounding normative attitudes or actions can not be easily disrupted, as the term "norm" suggests. Removing personal feelings of shame, embarrassment, or fear of future actions is not feasible in a survey setting. In fact, including scripts aimed at decreasing social desirability bias, may inadvertently amplify it by drawing attention to the sensitivity of the item. Therefore, avoiding a direct response is the ideal mode of mitigating social desirability bias. Using indirect questioning should reduce the impact of social desirability bias.

Inferred valuation, a method proposed by Lusk and Norwood (2009a,b), is a commonly used indirect questioning method. People gain satisfaction from simply reporting a positive WTP for a good with normative characteristics, and so have an incentive to overreport WTP. However, a person does not gain satisfaction from overstating another person's WTP. Put another way, another's altruism does not improve my own self-image.[1] Therefore, inferred valuation asks a respondent to report their perception of the WTP of the average person.[2] The inherent assumption is that people use their own, unbiased preferences as a proxy for the preferences of others (Carlsson, Daruvala, and Jaldell, 2010). Inferred valuation has been shown to consistently uncover WTP and voting behavior closer to true outcomes when compared to traditional direct questioning methods (Entem et al., 2019; Lusk and Norwood, 2009b, 2010; Menapace and Raffaelli, 2020; Norwood and Lusk, 2011; Yadav, Van Rensburg, and Kelley, 2013). Even so, there are two instances where inferred valuation may be less effective. First, when the value placed on an item (or service) is highly heterogeneous, identifying the value placed by a random other person is challenging. Such a task would require a respondent to assume another person's values, which can often be biased (Menapace and Raffaelli, 2020). Second, inferred valuation is less effective when there are significant spill-over effects, also referred to as free-riding (Yadav, Van Rensburg, and Kelley, 2013).

These concerns will likely be particularly salient in a promoted agricultural technology program. Agricultural technologies will have variable returns related to farmers' production characteristics, so WTP is likely to be heterogeneous. More pressing is the nature of the spill-over effects. In the context of a program, the availability of a promoted technology is often randomized by geographic area, like farm organization. Since the randomization occurs for the full area, the threat of the technology being diminished is not just to the individual. Instead, the reduction in technology availability would influence the entire community. So, all individuals alike have the incentive to overstate WTP, even

[1] In fact, there is some evidence that there is satisfaction to be gained by understating another's WTP to elevate my own image in comparison (Menapace and Raffaelli, 2020; Yadav, Van Rensburg, and Kelley, 2013)

[2] Or the share willing to support a ballot measure.

when responding from the perception of their neighbors. In the context of a community-based program evaluation, inferred valuation should not eliminate the utility gained from signaling. Therefore, there is value in proposing a new method that does not hinge on the perception of others.

We look to the social science literature for an alternative method. Prior literature has used psychometric scales in surveys to identify *people* inherently prone to social desirability bias (Rawat et al., 2017; Reynolds, 1982). Norwood and Lusk (2011) cast doubt on this method. Alternatively, list experiments are an indirect question method that measures social desirability bias in normative behaviors. Yet, list experiments have not been widely applied to assessing "deep" parameters of economic models, such as WTP. When applied in an economic context, list experiments can uncover the share of respondents willing to pay a given price. Implementing this strategy at a variety of prices traces a partial demand curve, giving insight into consumers' WTP.

We are particularly interested in the ability to measure near-zero prices using stated preference methods. Prior studies that utilize auction methods have shown that demand for promoted health products is strongly discontinuous at zero. In Kenya, Cohen and Dupas (2010) demonstrate a 60 percent reduction in demand for insecticide-treated bednets when prices increase from zero to \$0.60. An experiment in Zambia testing demand for a new water purification product over a traditional domestic alternative at different subsidy levels found a similar phenomenon. Among those receiving the information treatment, take-up of the new product over the traditional fell from 50 percent to zero when the new price increased \$0.03 relative to the previous price(Ashraf, Jack, and Kamenica, 2013). In Japan, Iizuka and Shigeoka (2022) extremely small co-pays sharply reduce child visits to a physician. These studies have fueled debate on whether cost-sharing methods are appropriate to use when promoting beneficial products in low-income settings.

A recent line of literature also suggests that the initial provision of a good or service for a zero price depresses future demand (Fischer, van den Berg, and Mutengwa, 2015; Shukla, Pullabhotla, and Baylis, 2022). That phenomenon may arise because items or services given at a zero price are viewed fundamentally differently than those at small, but positive prices (Shampanier, Mazar, and Ariely, 2007).

Taken together, the literature suggests that price dynamics across the zero-price threshold comprise a key element for understanding demand for a novel product. However, understanding whether consumers value a product at any price is complicated by measurement concerns. The previous studies rely on incentive-compatible, revealed preference methods. Without a transfer of money, respondents may overstate WTP on the extensive margin to avoid the potential social repercussions of revealing that they place no value on the program's product. As a result, the risk of social desirability bias from stated preference methods may be especially high at trivially low prices.

Our primary objective is to evaluate the effectiveness of list experiments at uncovering WTP at a variety of prices. Using a list experiment as an alternative WTP estimation strategy capitalizes on the experimental benefits of revealed preference methods and the brevity benefits of stated preference methods. In addition, we discuss the considerations and limitations of using list experiments in an economic application. We demonstrate the effectiveness of this theoretical strategy in an empirical application to an e-extension platform in Sri Lanka.

The Food and Agriculture Organization of the United Nations (FAO) has implemented a randomized control trial (RCT) to assess the impact of the Smart Extension and Efficient Decision-making (SEED) Hub, which provides farmers in the treatment groups with free-of-charge, geo-localized, timely, and integrated advisories. This digital tool is expected to enhance farmers' risk management practices and agricultural productivity and strengthen resilience in the face of climate and market risks. Understanding the value participants place on the application is relevant for future program scaling and longevity.

While list experiments effectively reduce social desirability bias, the applicability of the method to demand elicitation is not universal. We show that this method is only suited for products and settings where the respondent's budget constraint is unlikely to bind. The necessary conditions for a valid list experiment are violated when the respondent's ability to purchase at a given price is limited by their income level. Consequently, list experiments generate more reliable estimates of demand at lower price levels.

In the Sri Lankan sample, where the product was provided free prior to the survey, we find stated willingness to pay at low price levels is overstated when elicited by direct questions. Compared to direct elicitation, the share of respondents willing to pay the lowest price level declines by 30 percent when measured using the list experiment. At higher prices, the list experiment appears to offer either no advantage, and may even perform worse at very high price levels.

Our study offers a novel application of the widely used list experiment method: estimating a partial demand curve. A list experiment can provide a light touch measurement option when the necessary conditions hold and strategic design choices are made.

The remainder of this paper will be structured as follows. We begin by describing the list experiment approach. Then we propose a conceptual framework before introducing the RCT study background and data. Finally, we will discuss our results before closing with our discussion of the potential use cases.

2 List Experiments

List experiments allow respondents to answer indirectly, and therefore anonymously. Instead of individually answering yes or no to each item in a list of statements, respondents provide one single aggregated count of their affirmative answers. List experiments can uncover the sample-level prevalence of the behavior or sentiment of interest (or WTP as we will demonstrate) while allowing each respondent to keep individual answers private.

In a standard list experiment, respondents are randomly assigned to either Group A or Group B, so that the groups are similar.[3] Both groups are read a list of items and asked how many are true for their household.[4] Vital to the success of the list experiment, respondents do not reveal which specific items are true, only the aggregate number. For Group B, the list includes items that are not sensitive but are topically similar to the sensitive item. Group A has a list of the same items as Group B, but with the addition of the sensitive item to be measured. The difference between the average number of affirmative responses in Group A and Group B is the sample prevalence of the sensitive behavior.

In our list experiment, each respondent is asked whether they are willing to pay for each of a list of items at price p. Adapting notation used by Tsai (2019), the share of respondents willing to pay the price p, is $P(S_p = 1)$. The observable individual response $Y_{i,p}$ to the list question is a function of the unobservable responses and the individual's group status T_i. For Group A $T_i = 1$ and $T_i = 0$ for Group B. The response for individual i at price p is $Y_{i,p} = S_{i,p}T_i + \sum_{j=1}^{J} R_{i,j,p}$. Here, $S_{i,p}$ is the unobservable response to the sensitive item at price p and $R_{i,j,p}$ is the unobservable response to the non-sensitive item j in the list of J items at price p. Using the observable responses, we estimate the following difference in means at each p in the set of prices $p \in \{150, 600, 1050, 1500\}$.

$$P(S_p = 1) = \left(\frac{\sum_{i=1}^{n} Y_{i,p}T_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} Y_{i,p}(1 - T_i)}{\sum_{i=1}^{n}(1 - T_i)} \right) \quad (1)$$

2.1 Assumptions

List experiment validity rests on three key assumptions: randomization, no liars, and no design effects. The literature contains ample discussion of best practices in the design and implementation of list experiments in order to satisfy these assumptions. However, that guidance focuses on identifying the sample prevalence of retroactive behaviors or opinions. As such, the standard recommendations may need to be reexamined for economic applications like assessing demand.

[3]Many studies refer to these as treatment and control groups. Since this experiment is nested within a randomized control trial, we use alternative language to avoid confusion. Group A refers to the group that answers a list of question that includes the sensitive item (i.e. the 'treatment group'), and group B refers to the group that does not see a sensitive item (i.e. the 'control group')

[4]Or in sentiment based list experiments, "how many do you agree with?".

2.1.1 Assumption 1: Randomization

The randomization assumption requires that Group A and Group B are similar. The difference in means framework hinges on Group A and Group B having on average equivalent responses to the non-sensitive items in the list. Standard practice is to check that observable characteristics are balanced across groups. The randomization assumption does not need to be adjusted in an economic context. While no formal test for randomization exists, a standard balance table gives assurances that the two groups are similar.

2.1.2 Assumption 2: No Liars

The second assumption necessitates that respondents do not falsely answer the question. Respondents may strategically give a false answer when reporting the true answer will reveal a socially undesirable response. This can occur when the true answer lies at the extremes of the list length. When that occurs, the list experiment fails to guarantee anonymity and thus becomes functionally the same as a direct question. A response at the high (ceiling effect) or low point (floor effect) reveals the sensitive response, which may prompt a respondent to give a strategically altered response. To avoid floor and ceiling effects, common practice is to include an item that is very likely to have an affirmative response and an item that is very unlikely to have an affirmative response in the list (Glynn, 2013).

The same practice can be applied to the WTP context by including both an item highly likely to be purchased and an item highly unlikely to be purchased at all proposed prices. Prevailing market prices can facilitate the choice of items to satisfy the criteria. To reduce the chance of floor effects, include an item with a prevailing value well above all of the proposed prices. Given the discounted price of the item, even those with small budgets or weak preferences would be likely to select the item. Similarly, including an item that is valued lower than the stated price reduces ceiling effects. Even those with large budgets and high idiosyncratic preferences are unlikely to report a desire to overpay for an item. Importantly, the non-sensitive items in the list should have well-established market values so that respondents have a clear reference for price. The sensitive item, on the other hand, may not have a clear market price so respondents must evaluate what they are willing to pay in the absence of a market reference.

As with the randomization assumption, no formal test for the no liars assumption exists. However, checking for floor and ceiling effects (i.e. responses of 0 and the maximum list length J) in Group B is a good indicator of the potential that the assumption may be violated. Even with the addition of prices, this practice remains a valid and important diagnostic check in an economic context.

2.1.3 Assumption 3: No Design Effects

Assuming no design effects requires that the inclusion of the sensitive item does not influence the responses to the non-sensitive items when the items are privately counted. If adding the sensitive item changes a respondent's answer to a non-sensitive item, then Group A and Group B are no longer comparable. If the inclusion of the sensitive item increases the non-sensitive response total, then the share willing to pay price p for the sensitive item is overstated. The share willing to pay price p for the sensitive item is understated if the sensitive item decreases the non-sensitive response total.

In a standard list experiment, common practice is to include items of a similar nature in the list so that the sensitive item does not stand out. While this is good practice in an economic context, it is not sufficient to ensure that the no design effects assumption holds. Two factors may also cause the no design effect assumption to be violated: (1) the sensitive item and any non-sensitive item are complements or substitutes and (2) the inclusion of the sensitive item changes the optimal purchase decision for non-sensitive items via a binding budget constraint.

When a good is a substitute or complement for the sensitive item, then the addition of the sensitive item may cause a respondent to change their response to non-sensitive items.[5] The direction of the bias would depend on whether the goods are substitutes or complements. When substitutes, the desire for the sensitive item would be underestimated as the switching between items would not be captured in aggregate. If the goods are complements, the share willing to pay would be overstated since increased demand for the non-sensitive complement item would be falsely attributed to the sensitive item. The impact of complements and substitutes can be mitigated with careful selection of the items in the list.

The binding budget constraint problem occurs when the likelihood that a respondent's hypothetical budget for the category of goods under study is exceeded differs based on group assignment. To motivate, in the scenario considered here, respondents aggregate the number of agricultural inputs they are willing to purchase at a given price. To simplify, assume a household has a predefined separable budget for agricultural input purchase. Respondents in the experiment are asked to consider their current budget constraint for such goods.[6]

The budget creates a two-step decision process. First, respondents must consider the total number of items at the given price that could be purchased within their budget. If the separable budget exceeds the total to purchase all items in the list, then the budget constraint is non-binding. In this case, in the second step, respondents report the number of items in the list where their own WTP exceeds the listed price. This scenario represents an interior solution. On the other hand, if the separable budget

[5] Non-sensitive items can be compliments or substitutes with each other, since both items appear in the list for Group A and Group B.

[6] Reminding participants of their budget constraint was originally recommended by Cummings and Taylor (1999). While the method was shown to be less effective in contingent valuation settings, we find evidence that in a list experiment context, the script is sufficient to remind participants of a binding budget constraint.

does not exceed the total cost of all items at the given price, then respondents must compare the items in the list relative to one another rather than individually against the given price. Importantly, this comparison of goods relative to one another is partially a function of the list length and, therefore, a function of the group assignment. This point is demonstrated by expanding on the list experiment notation used in Section 2.

$$Y_{i,p} = \begin{cases} T_i S_{i,p} + \sum_{j=1}^{J} R_{i,j,p} | B_i \geq (J + T_i)p \\ \lfloor \frac{B_i}{p} \rfloor | B_i < (J + T_i)p \end{cases} \tag{2}$$

The observable response $Y_{i,p}$ is now determined in part by the separable budget B_i. If the budget constraint is not binding, then the observable responses will continue to be a function of the unobservable responses to the sensitive and non-sensitive items. However, when the budget constraint is binding, the response is now a function of the budget, rather than the unobservable preferences. The design effects assumption is violated because the switching point between those two unobservable processes is potentially a function of the group assignment.

For example, if an individual could purchase two items at price p with their predetermined budget, then a respondent in Group B could select two of the three items while a respondent in Group A can still only select two but their list has four items. If both answered two, it would look like there was zero demand for the sensitive item. However, it could be that individuals in Group A were selecting the sensitive item instead of one of the non-sensitive items. The comparison of goods, due to a binding budget constraint, would lead to a no design effects assumption failure.

A binding budget constraint is a key limiting factor of using list experiments in an economic context. The next section explores a conceptual framework that demonstrates implications of a binding or non-binding budget constraint in a theoretical application.

3 Conceptual Framework

To illustrate the application of a list experiment to demand measurement, we build on the conceptual framework of WTP elicitation from surveys proposed by Lusk and Norwood (2009b) and Menapace and Raffaelli (2020). The key feature of a list experiment is the anonymity provided by the list, which eliminates sources of hypothetical bias related to social signalling that may be present when using a single direct response survey question. However, as noted above, the list of items makes a budget constraint more salient. In the following exposition, we assume the budget for agricultural technologies is separable from other household purchasing decisions. Following Lusk and Norwood (2009a,b), utility is derived from both market and nonmarket sources in our model. Here, we propose that utility is derived from normative behaviors and consumption derived from income maximization. Assuming the components are additive, the resulting framework is such that:

$$U = wM(S,H,E) + (1-w)V(I,G) \tag{3}$$

Where $M(S,H,E)$ is the utility gained from social behavior, which is a function of social signaling (S), honesty (H), and expenditure signaling (E). Utility can be derived by signaling pro-social behaviors ($\frac{\partial M}{\partial S} > 0$), giving honest responses ($\frac{\partial M}{\partial H} > 0$), and expenditure signaling ($\frac{\partial M}{\partial E} \geq 0$).While utility is strictly increasing in signaling and honesty, expenditure signaling has two distinct possibilities: $\frac{\partial M}{\partial E} = 0$ or $\frac{\partial M}{\partial E} > 0$, discussed further with Equation 6.[7]

$V(I,G)$ is a standard indirect utility function, where I is income and G is the level of consumption of the good of interest. In addition, w is the weight of importance for the moral and wealth maximizing components. It is worth noting that we assume the hypothetical budget is separable from other income categories, so the income I is distinct from the budget B. Since the stated choice setting is hypothetical, we set $w = 1$, so that $U = M(S,H,E)$ following Menapace and Raffaelli (2020).

We are interested in the unbiased stated WTP for the item of interest j^* (wtp_j^s). Within a list context, there are J other items of interest where $j \neq j^*$. This gives an aggregate stated WTP such that $WTP^S = \sum_{j=1}^{J} wtp_j^s$. Similarly, the true WTP is the sum of the WTP for J items, giving $WTP^T = \sum_{j=1}^{J} wtp_j^t$. We define true WTP to be equivalent to the WTP when a monetary transaction occurs. That is, we are not attempting to remove any social desirability bias from a real market interaction. When a direct question is asked about item j^*, then $WTP^S = \sum_{j=j^*} wtp_j^s = wtp_{j^*}^s$ and $WTP^T = \sum_{j=j^*} wtpts_j = wtp_{j^*}^t$.

Social signalling (S) is a function of the stated WTP ($wtp_{j^*}^s$). Direct and indirect questions have different signaling opportunities and therefore must be defined separately. In a direct question, individuals have the opportunity to boost moral behavior, and therefore utility, by overstating WTP. For simplicity, we assume $\alpha = 1$. The indirect question, in the form of a list experiment, does not allow for signaling.[8] The anonymous response prevents the respondent from revealing socially normative behaviors, so there is no incentive to alter responses. For the indirect question, the utility of signaling is zero since there is no way to link the signal to the individual. This is represented as such:

$$S(WPT^S) = \begin{cases} a + \alpha wtp_{j^*}^s \text{ if direct} \\ 0 \text{ if indirect} \end{cases} \tag{4}$$

[7]While not directly relevant here, there may be some circumstances where an individual gains utility from hiding wealth or appearing less well off.

[8]Recall, we argue community-level randomization in the context of program implementation allows for utility to be gained from signaling among inferred valuation indirect questions.

Honesty is a function of stated WTP (WTP^S) and true WTP (WTP^T). Honesty increases as the difference between the stated WTP and the true WTP decreases (Lusk and Norwood, 2009a). Individuals perceive honesty for each item individually. Meaning, aggregate honesty is not the total difference between stated WTP and true WTP, but the sum of the squared difference for each item. For continuity, we keep the functional form used previously by Menapace and Raffaelli (2020).

$$H(WTP^S, WTP^T) = -0.5 \sum_{j=1}^{J} (wtp_j^s - wtp_j^t)^2 \tag{5}$$

Lastly, the expenditure signal is similarly a function of stated WTP (WTP^S) and true WTP (WTP^T). In a list context, an individual has the opportunity to improve their own self-image by signaling their separable budget for agricultural technologies. Reporting more items signals a larger budget, which increases utility. A notable difference between the budget signal and honesty is that the budget signal occurs in aggregate. Therefore, it is the difference between the total stated WTP (WTP^S) and the total true WTP (WTP^T), not the individual differences as with honesty. We maintain the same functional form for simplicity, giving:

$$E(WTP^S, WTP^T) = 0.5(\sum_{j=1}^{J} wtp_j^s - \sum_{j=1}^{J} wtp_j^t)^2 \tag{6}$$

As noted in Section 2.1, a response to the list question is determined in two steps. The first step specifies if the budget constraint is binding or not. If the budget constraint does not bind, then the budget does not influence the second step. When the second step (determining WTP for each item) is unencumbered by the budget constraint, there is an interior solution and no expenditure signal such that $\frac{\partial M}{\partial E} = 0$. In this scenario, the assumption of no design effects is satisfied. When the budget constraint binds, the respondent's budget enters the the second step, a comparison that violates the design effects assumption and allows for expenditure signaling, so that $\frac{\partial M}{\partial E} > 0$.

We identify the utility maximizing stated WTP using the first order condition of the morality utility function with respect to $wtp_{j^*}^s$.

$$wtp_{j^*}^s = wtp_{j^*}^t + \frac{M_S}{M_H - M_E} + M_E \sum_{j \neq j^*}^{J} (wtp_j^s - wtp_j^t) \tag{7}$$

The first term captures the true WTP. The second term represents the utility trade-off between reporting normative behavior, honesty, and expenditure signaling. As expenditure signaling increases, honesty decreases. The final term is the utility gained from expenditure signaling. There are three potential stated WTP scenarios. First, a standard direct question in which case $M_S > 0$. In a hypothetical single-question, direct format, individuals are freely able to signal.[9] Individuals only consider the value of one item. Therefore, no other $j \neq j^*$ items appear and the third expenditure signaling term is zero. In this scenario stated WTP is:

$$wtp_{j^*}^{Direct} = \frac{M_S}{M_H - M_E} + wtp_{j^*}^t \tag{8}$$

In the second scenario, we consider an indirect question, so that $M_S = 0$, and a non-binding budget constraint. Since there is an interior solution, $M_E = 0$. For a household where the budget constraint is not binding, a household can report purchasing the items they value. Therefore, there is no utility to be gained from reporting the desire to purchase an item they do not value at the given price. This occurs because the budget is no longer salient when making decisions, so the marginal benefit of budget signaling goes to zero. In this scenario, the stated WTP collapses to the true WTP.

$$wtp_{j^*}^{Indirect,Non-Binding} = wtp_{j^*}^t \tag{9}$$

Finally, the third scenario is an indirect question with a binding budget constraint. The indirect question prevents signaling so that $M_S = 0$. However, there is an opportunity to increase utility by expenditure signaling ($M_E > 0$). Importantly, the budget signaling is related to the list length, which varies by group assignment as noted in 2.1. This again highlights the relevance of a binding budget constraint to the no design effects assumption. This gives:

$$wtp_{j^*}^{Indirect,Binding} = wtp_{j^*}^t + M_E \sum_{j \neq j^*}^{J} (wtp_j^s - wtp_j^t) \tag{10}$$

Comparing the three scenarios, there are important trade-offs to consider. An indirect approach will uncover the true WTP when the budget constraint is not binding, since there is no opportunity to signal pro-social or expenditure behavior. The budget constraint is less likely to bind, revealing an interior solution, when wtp^t is low or budgets are high. However, if the anticipated budget is low or the expected WTP is high, then the trade-off between the social signaling and honesty ratio and expenditure signaling must be evaluated. As the price increases, the probability of having an interior solution decreases, holding all else constant. In this case, a list question may perform worse relative to a direct question because respondents can signal exaggerated expenditure for an item category. The trade-off depends on which signal is stronger, which can vary by context. In situations with strong social expectations, indirect measure may be preferred. However, in scenarios with limited budgets or a highly valued item, then the likelihood the budget will bind and the size of the expenditure signal

[9]This is also likely the stated WTP for an inferred valuation method in a randomized program context.

may caution against using a list approach.

4 Background

In July 2023, the Food and Agriculture Organization of the United Nations (FAO) began a randomized control trial (RCT) aimed at improving Sri Lankan paddy farmers' access to information. Treated farmers received access to SEED Hub, an integrated digital extension service. SEED Hub was designed in conjunction with the FAO, the Sri Lankan Ministry of Agriculture, and research institutes to bring together weather, agricultural production, and market information into one comprehensive platform.

The motivation for the program was to reduce barriers to information and therefore optimize agricultural production and risk management strategies. Climate change and natural hazards jeopardize production, especially among Sri Lankan paddy farmers. Having access to more complete and reliable information may improve farmer decision-making in a way that improves production, reduces risk and loss, and ultimately improves farm household livelihoods.

A total of 220 farm organizations were selected, so that the sample was nationally representative of paddy farmers in terms of agro-ecological zones and irrigation schemes. Half of the farm organizations were designated as the control group. The remaining 110 farm organizations were then divided again into two treatment arms. The first treatment arm received the login information for the application, we refer to this group as the "Access Only" group. The second treated group received the login information and a training session on how to use the platform, which is the "Access + Training" group. For both groups, access to the SEED Hub platform was free.

In both treatment groups, farmers received the log-in information at a hub deployment meeting. The meetings took place at the farm organization before the Maha (primary) growing season. During this time, farm organizations in the "Access + Training" group were given the training. The training consisted of a lecture, demonstration, and time for questions. A typical training session took 2.5 hours.

4.1 Data

The data was collected from March 2024 to May 2024 in 10 regions in Sri Lanka by a team of enumerators using the digital Survey Solutions platform.[10] The measurement experiments were embedded in the endline survey. For which, respondents were randomly assigned to Group A or Group B, stratified by RCT group assignment and farmer organization. The RCT treatment groups included a total of 1,100 individuals.

[10]The regions are: Ampara, Anuradhapura, Badulla, Batticaloa, Galle, Kegalle, Kilinochchi, Kurunegala, Matara, Vavuniya

We use a single list experiment difference in means strategy to estimate the share of individuals willing to pay a given price for SEED Hub. While the improved efficiency associated with a double list experiment may be useful, it is not feasible in our setting. A double list experiment requires two list questions to uncover one prevalence. Each group alternatively has the list with the sensitive item and the non-key items in the list change.

We opted to implement single list experiments, despite the potential efficiency gains of a double list experiment for two reasons. First, the number of questions that would be required to uncover one prevalence was not feasible given the length of the existing survey. To maintain the sample size, each individual would need to answer six cognitively taxing questions. Doing so may jeopardize the data quality. More pressingly, in a double list experiment the individual would alternate between having SEED Hub in their list. At times, the non-key items would change but SEED Hub would remain the same. Doing so would draw attention to SEED Hub and increase the likelihood that respondents augment their answers, violating the no design effects validity assumption (discussed in Section 2.1). So, we maintain that a single list is the most appropriate estimation strategy.

The list experiment instructions and lists are outlined in Table 1. Many elements of the instructions do not change with the inclusion of prices. The instructions for the list experiment emphasize that respondents should not give individual answers for each item but should instead give the total number of responses. Even without the inclusion of prices, list experiments are more cognitively burdensome than comparable direct questions. To increase the accessibility of the questions, we ask respondents to count their affirmative statements on their fingers out of sight of the enumerator. Counting props or fingers is common practice in list experiments, given the varying levels of literacy and numeracy (Tadesse, Abate, and Zewdie, 2020; Jouvin, 2023).

There are two changes to the list experiment instructions for an economic application. First, we ask respondents to consider their current budget for purchasing farm tools, assets, or services. Since there is no transfer of funds for this question, hypothetical bias is a concern. Similar to a cheap talk statement, prompting respondents to think of their budget constraint should reduce the impact of hypothetical bias.

The second change made to the list experiment is the inclusion of the price. Each individual is asked the same question three times, with only the price varying. The price for each question was randomly drawn from a set of four prices without replacement. This design ensures each individual would see three unique prices in a random order.

The prices were selected based on responses in the midline survey for the same RCT. The midline and endline surveys consider four different types of agricultural information: weather, disaster, farming, and marketing. Weather data was the most commonly used SEED Hub function, as reported in the midline survey. In addition, farmers were asked if they had paid for any information services in any of the four categories. While the vast majority did not pay for information, among those that did pay the average monthly payment for weather information was 605 LRK with a standard deviation of 454. The average payments for the other sources of information were similar, ranging from 427 to 647 per month. Using the information from the midline, we selected 600 LRK, one standard deviation below the mean (150 LRK), and one and two standard deviations above the mean (1050 and 1500 LRK). All of these prices fall within the range of reported prices in the midline, with the highest price reporting being 3000 and the lowest 30 for one month. Selecting a near-zero price was of particular interest, given our interest in censoring in response to receiving the product freely in the past. For reference, a typical daily labor wage for men's agricultural work is 2,250 LRK, so 150 LRK would reflect less than one hour of agricultural labor.

The number of question iterations and prices were selected based on a balance of survey fatigue and statistical power. Since the measurement experiments were embedded in the endline survey of an RCT, the sample size was fixed at a maximum of 1,100. The list experiments can be burdensome, causing survey fatigue, so we limited the number of potential questions to three. Therefore, the number of potential prices we could include in our set was determined so that we could maintain statistical power to detect differences between direct and indirect methods. We used simulation to determine the necessary sample size per price level to detect (1) a WTP different from zero and (2) social desirability bias at a variety of potential levels of bias. Further discussion of this process is available in the Appendix C. Based on this work, we selected three questions with prices drawn from a set of four, leading to a maximum potential sample size of 885 for each price.

The indirect questions were administered after the Maha season production and crop sales modules. The prior questions would prime farmers to think about their production decisions, but not SEED Hub specifically. Three modules separate the indirect questions from the SEED Hub module, where the direct WTP questions appear. The direct SEED Hub questions appeared at the end of the information services module, which included questions about all sources of information used, including SEED Hub. The following direct question appeared four times for all respondents. *Would you be willing to pay **Price** Rs. for a one-month S.E.E.D. hub subscription, a mobile application that provides information on agronomic practices, market prices and weather forecasts, and related advisories?* Respondents were asked about all four prices in our set in a random order.

All respondents were asked both the list questions and the direct questions. There may be some concern that the list question may influence the direct response among individuals in Group A. Prior literature has suggested asking direct questions only of Group B. However, the distance between questions in the survey reduces the likelihood of the indirect response affecting the direct response.

Asking the question of all respondents increases statistical efficiency, and would therefore be preferable. We compare the share willing to pay each price using the direct measure by the list group and find no statistically meaningful differences (Table 2). Therefore, we use the full RCT treatment sample to construct both the indirect and direct estimates.

The prices for both the indirect and direct measures appeared in random order to avoid anchoring effects. However, in doing so, respondents may have given inconsistent answers. For example, indicating directly or indirectly, that they would be willing to pay 600 LRK but not 150 LRK. Such responses were removed from the estimation sample. For the indirect question, if a respondent gave a larger number of affirmative statements at a higher price than a lower price the individual was removed from all list questions. Similarly, if a respondent answered the direct question as yes for a higher price and no for a lower price, the respondent was removed from all questions. Given the cognitive burden of the indirect questions, the inconsistent response rate is much higher compared to the direct question (Table 3). However, notably, the number of inconsistent responses is balanced between list groups. Meaning, that the inconsistencies were not caused by the group assignment and so the randomization holds, which is relevant for the validity assumptions discussed in Section 2.1.

The high number of inconsistent responses raises concerns about internal validity. Households with inconsistent responses tend to have lower education, a larger total agricultural area, use more male hired labor, less household labor, and a larger harvest. This indicates that large-scale farmers may be underrepresented in our sample (Appendix Table 1). Prior literature in the consumer choice space has similarly found significant levels of inconsistent responses and has shown that including inconsistent responses can bias WTP (Colombo, Glenk, and Rocamora-Montiel, 2016; Sælensminde, 2002). Although we do not find that inconsistent statistically affects our results (Appendix Figure 1), we exclude them from our analysis nonetheless.

5 Results

5.1 Assumptions

As discussed in Section 2.1, a valid list estimation requires three assumptions to hold: (1) randomization, (2) no liars, and (3) no design effects. In this section, we discuss the validity of these assumptions in our empirical application. First, Table 4 checks for balance among observable household and agricultural production characteristics. The characteristics are balanced between the list groups, suggesting that the randomization assumption is likely to hold.

Next, we consider the no liars assumption. The advice outlined in Section 2.1 would suggest a high and low value item should be included. Urea is a valuable input, 10 kgs would be worth around 2,150

LRK at the subsidized price (3,400 LRK unsubsidized) (Centre, 2023). At any of the prices in our set, the urea would be a good deal. A half day of labor would be valued at around 1,125 LRK, so this item should have mixed selection. The low value item selected in the list (Table 1) is compost. Compost is generally available for free to all farmers (Centre, 2023), so any positive price would be unlikely to be accepted. Importantly, compost is widely available, consistently free, and frequently used. Even though SEED Hub is currently free, it is less widely used and does not have a clearly established market price, so it should not be viewed in the same way as compost.

Table 5 looks at the potential floor and ceiling effects. The first thing to note is the rate of misunderstood or declined responses. A respondent was recorded as misunderstood if they repeatedly tried to answer each individual item or if they gave a number outside of the allowable range for their group. The number of respondents that misunderstood is similar across list group assignment and price level, so that does not appear to be a threat to list experiment validity. We also see a relatively low frequency of zero and three responses in Group B, aside from price 150. In addition, the rates are comparable between Group A and Group B indicating that it is unlikely respondents are strategically changing their responses to avoid detection.

Price 150 does not seem to follow the same pattern as the other prices. Here, we see a higher rate of zero responses compared to other prices and in Group A relative to Group B. An answer of zero among Group A would reveal the socially undesirable response (not valuing SEED Hub), so the higher number of zeros among Group A compared to Group B does not suggest respondents are strategically altering their responses. However, if there were strategic altering of responses then that would lead to an overstatement of SEED Hub use by the indirect measure and an underestimate of social desirability bias, representing a lower bound. Comparing the frequency of zeros for price 150 compared to the higher prices is unexpected. However, the near-zero price could be such an undervaluation of the high value inputs that respondents are skeptical. Since this behavior seems to be relatively balanced between Group A and Group B, it should not bias the estimates of willingness to pay. Still, this suggests a smaller range of values may be preferred. In total, we do not find strong evidence that the no liars assumption is violated.

Finally, we proceed to the no design effects assumption which we anticipate being the most difficult to satisfy in an economic context. All of the items in the list are production inputs. There are no other information sources included in the list, so substitutes should not be a concern. Avoiding compliments is more difficult, as an increase in production information likely requires more of production inputs as well. However, the sources of information in available in SEED Hub (weather, production, and market prices) are more likely to influence planting, irrigation, and marketing decisions than fertilizer or labor decisions. Still, we would like to check for the possibility of design effects due to complementary goods.

Since the goods remain constant across all prices, if design effects fail universally, then it is a sign that the complementary nature of the items is a problem. On the other hand, if there is a price cut off when the design effects become present, then that indicates that a binding budget constraint is more likely to be the driver.

We use the Blair and Imai (2012) test for design effects. The test estimates the probability (π_{yt}) that a response will be given, where y is the number of non-key items and t indicates the sensitive response. Any negative probability indicates the design effect assumption is violated. The results of this test are in 6. The second part of the Blair and Imai (2012) tests for stochastic dominance, where $\#P$ is the Bonferroni-adjusted p-value. The results of this test are in Table 7.

The Blair and Imai (2012) test indicates in both Table 6 and 7 that the design effects are not universally violated. Instead, at a price of 600 per item the no design effects assumption is violated, indicating this is the price where the budget constraint is binding. To purchase all of the items in the list (three or four depending on the list group assignment) would be roughly equivalent to a day of labor. The negative probabilities are closer to zero at higher prices, indicating that as the maximum potential budget increases the impact of having one additional item is smaller. If the budget constraint would only allow an individual to purchase one good at the higher price the impact of having four items compared to three is smaller than at the lower prices where an individual may have been able to afford more items with their existing budget. This supports the suggestion that the budget constraint is the limiting factor, rather than the goods being complements.

Considering all three of the necessary conditions for a valid list, the list experiment at the price of 150 is likely to provide reliable responses. We show the results for the other prices, but they should be viewed skeptically. Given the existing literature on the misestimation of WTP near zero, the results for price 150 are of particular interest.

5.2 Willingness-to-Pay

We find a downward sloping demand curve with a steep drop off after 150 LKR per month when using the direct WTP questions (Figure 1). When asked directly, 63.6% of respondents indicated they would be willing to pay 150 LKR per month for the SEED Hub platform. Putting this number in context, 77.8% of people reported being aware of SEED Hub but only 30.6% indicated they had downloaded the app. The stark difference between those who had interacted with the application and those who were willing to pay a positive price for the application indicates that direct WTP is likely overstated. Overstatement of WTP at low prices aligns with our hypothesis that individuals may overstate their WTP when a product had previously been given for free and is associated with a program.

The indirect demand curve is initially downward sloping, with a slight upward slope at the high prices.[11] As theorized in Section 2.1 and demonstrated in Section 5.1, a binding budget constraint violates the no design effects assumption for prices 600 LKR and above. The assumptions required for valid estimates at those prices are unlikely to hold, so the results should be viewed with caution. Still, we see that at the higher prices when the budget constraint holds, the indirect estimate exceeds the direct measure estimate. This supports the theoretical conclusion that the expenditure signal has a larger impact than the morality signal, $\frac{M_S}{M_H - M_E} < M_E \sum_{j \neq j^*}^{J} (wtp_j^s - wtp_j^t)$, in this scenario. This finding suggests that, in this case, at higher prices, the direct measure would be preferred. Further supporting the theoretical model, at the low price of 150, where the budget constraint is far less likely to be binding, the indirect measure is preferred, because there is no opportunity to signal and there is no marginal utility to overstating the budget for unvalued items.

The estimate at a price of 150 should be unaffected by the design effect violations present at the prices. At this price, the demand estimate is lower than the direct measure. By the indirect estimate, only 47.4% of households would be willing to pay 150 LKR per month for SEED Hub. The 16.3 percentage point difference is marginally significant (10% significance level). The lack of precision owes to the lower efficiency of the indirect method and the decreased sample size from inconsistent responses.

In terms of magnitude, the 16.3 percentage points difference is a nearly 30% decrease in the share of farmers willing to pay 150 LKR. The difference in share willing to pay at low prices is consistent with the claim that social pressures lead to an overstatement of WTP, particularly when the product has been previously given freely. When alleviated of the pressure to directly reveal the proposed product does not have a positive economic value, respondents are more willing to reveal a WTP of zero. This finding suggests that an indirect question may be useful for uncovering more accurate measures of WTP when the budget constraint is not binding and researchers are concerned about biased WTP at the extensive margin.

We would expect those with more experience with the SEED Hub application to have a higher WTP but they may be subject to more social pressure as well. Having benefited from the application previously, downloaders of SEED Hub may feel the need to reciprocate and overstate their economic value of the product. To explore this possibility, we limit the sample to only those who reported downloading the application. Since downloading the application is potentially verifiable, respondents may be less likely to give biased responses to the download question than other self-reported use characteristics. When including only individuals who downloaded the application, we predictably see an increase in the share willing to pay 150 LKR per month in both the direct and indirect measures. The direct measure increases to 75.7% willing to pay while the indirect measure increases more marginally

[11] Despite concerns about internal validity, the results are similar with and without the inconsistent responses. See Appendix Figure 1

to 53.8% willing to pay. Among this sample, the 21.9 percentage point gap translates to a 33.8% overstatement of willingness to pay the lowest price.

Dividing the sample by those who downloaded the application raises two important points: (1) impact of program engagement and (2) subsetting the sample. First, we see that when we limit the sample to those with the most interaction with the service, and therefore the most interaction with the program, we see a slightly larger difference between measurement strategies. It seems that while those who have more experience with the program are more likely to have a positive WTP, they are also associated with more overstatement of WTP. This indicates that the level of intensity of the program or policy associated with giving a product may also be related to mismeasurement. If this is the case, we may also expect to see different levels of WTP by the RCT treatment arm.

In Figure 3, we see that the RCT treatment group that received access and training did have a higher WTP by both direct and indirect measures compared to just the group with access only. However, Figure 3 also highlights an important limitation of the indirect estimation strategy. Returning to the second point, subsetting the data greatly reduces precision. When exploring smaller subsets of the full sample, the direct measure continues to provide precise estimates, even if accuracy is in question. On the other hand, as sample size diminishes, the list experiment estimations grows increasingly noisy. These smaller subsamples make statistically distinguishing estimates difficult.

6 Conclusions

In this study, we discuss the potential for an economic application of a list experiment. A common use of list experiments is to estimate the behavior or sentiment of a sensitive subject. The indirect nature of the list experiments reduces the impact of social pressures on responses. The same effect is also desirable when estimating WTP. Applying list experiments to an economic context can uncover WTP without the social desirability bias and with a lighter touch than revealed preference methods require.

In a theoretical context, we demonstrate the benefits of an indirect question in reducing social desirability bias in stated WTP while highlighting the pitfalls of the method when individuals have a limited budget. Based on the conceptual model, we outlined the scenarios where a list experiment would and would not likely be useful. In addition, we discuss the ways a list experiment must be adjusted to be administered in an economic context. The randomization assumption can be satisfied without adjustment. Similarly, including a high and low value item increases the likelihood that the no liars assumption holds. However, we find that the no design effects assumption is more difficult to satisfy in an economic context. The items in the list should be topically similar, yet not complements or substitutes. More pressingly, the budget constraint should not be binding at any of the potential prices in the set, a concern highlighted in the conceptual framework. The range of prices should also be selected strategically. A wide range of prices makes selecting non-sensitive items more challenging, while a tighter range of prices uncovers a smaller part of the demand curve, representing an important trade-off.

This last consideration—avoiding a binding budget constraint—limits the general applicability of list experiments for WTP estimation. However, a key area where this method holds promise is in estimating WTP for products that have previously been provided for free or at subsidized prices. The discontinuous nature of zero and small positive prices makes a low price, where participants are deciding to enter the market or not, after a program intervention the most likely part of the demand curve to be biased. The indirect list method may provide a solution. Given the near-zero prices, budget constraints are unlikely to be binding so list experiments are a viable alternative. Our empirical application demonstrates that at low prices when the no design effects assumption is more likely to hold, direct WTP elicitation overstates WTP compared to the indirect list method.

To be implemented successfully, a list experiment should be piloted. The risk of not satisfying the validity assumptions is greater in an economic context than in a behavioral context, which itself suffers from a file-drawer problem (Blair, Coppock, and Moor, 2020). Meaning, that without strategic design, pre-testing, and piloting, there is a significant risk of having unusable responses. However, when done with care in appropriate circumstances, the economic application of a list experiment holds promise.

Our work is relevant for academics, NGOs, and governments that are interested in the future valuation of a product that is currently being given for free or at a substantially subsidized price. Understanding how recipients value a product given as a part of a program may be useful in decision-making about program expansion or extension and the future price of the product. However, future economic applications of a list experiment should be limited to lower value items or contexts ready for investment and the items in the list should be selected with caution. Ultimately, our findings suggest that economic applications of list experiments have the potential to reduce upwardly biased WTP found at the extensive margin.

Table 1: List Experiment Instructions and Question

List Instructions

Think of your current budget for purchasing farm tools, assets, or services. I will read you a list of four items or services, and I would like to know how many of these items or services you would purchase one or more of if each cost *Price*[a] Rs.? Do not tell me which you would purchase or the quantity, just the number of items you would purchase given the amount of money you have. Please put your hand behind your back. Lift a finger for every item that you would like to purchase at the price *Price* Rs. After I have read all the statements show me the number of fingers you have raised. That number should be the number of items you would like to purchase. Let's begin:

Group A	Group B
• 10 kg of urea	• 10 kg of urea
• Half day of hired labor	• Half day of hired labor
• **One month access to SEED Hub**[b]	• 1 kg of compost
• 1 kg of compost	

[a] Price was randomly selected from set of prices {150, 600, 1050, 1500}

[b] **One month subscription to market and weather information app** in the RCT control group

Table 2: Share Willing to Pay Each Price by List Group

Price	Group A	Group B	p-value
150	0.605	0.628	0.298
600	0.200	0.199	0.946
1050	0.061	0.052	0.394
1500	0.041	0.033	0.371
N	997	998	

Table 3: Inconsistent Responses By Question Type

Share Inconsistent	Group A	Group B	p-value
Direct	0.012	0.007	0.253
Indirect	0.291	0.292	0.994

Table 4: Randomization Balance Table

	Group A	Group B	p-value
Household size	3.961	4.041	0.290
Share with Male Head	0.501	0.499	0.837
Maximum Education	13.453	13.626	0.144
Number of Plots	3.592	3.520	0.368
Total Plot Area	4.492	4.424	0.891
Hired Male labor	5.870	5.993	0.804
Hired Female labor	0.443	0.410	0.719
HH male labor	1.505	1.468	0.878
HH female labor	0.506	0.507	0.991
Harvest (kg)	1415.692	1617.083	0.360
Largest plot	2.700	2.576	0.758

Table 5: No Liars Assumption Floor & Ceiling Frequency Check

	Price 150				Price 600			
	Group A		Group B		Group A		Group B	
Levels	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
0	106	12.71	155	17.67	30	3.8	25	3.17
1	217	26.02	244	27.82	273	34.56	396	50.19
2	185	22.18	233	26.57	319	40.38	268	33.97
3	165	19.78	188	21.44	95	12.03	59	7.48
4	106	12.71	-	-	31	3.92	-	-
Misunderstood	33	3.96	40	4.56	33	4.18	35	4.44
Declined	22	2.64	17	1.94	9	1.14	6	0.76
	Price 1050				**Price 1500**			
	Group A		Group B		Group A		Group B	
Levels	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
0	30	4.88	20	3.37	50	5.3	58	6.09
1	240	39.02	345	58.08	422	44.7	521	54.67
2	265	43.09	175	29.46	300	31.78	301	31.58
3	41	6.67	25	4.21	121	12.82	36	3.78
4	6	0.98	-	-	8	0.85	-	-
Misunderstood	25	4.07	26	4.38	32	3.39	34	3.57
Declined	8	1.3	3	0.51	11	1.17	3	0.31

Table 6: No Design Effect Assumption Test

	Price 150				**Price 600**			
Non-Key Value	π_{y1}	**Robust SE**	π_{y0}	**Robust SE**	π_{y1}	**Robust SE**	π_{y0}	**Robust SE**
0	0.24	0.13	0.13	0.02	-0.03	0.01	0.17	0.05
1	0.05	0.04	0.24	0.03	0.16	0.04	0.41	0.03
2	0.12	0.03	0.17	0.03	0.08	0.02	0.26	0.03
3	0.14	0.02	0.07	0.03	0.03	0.01	0.03	0.02
	Price 1050				**Price 1500**			
Non-Key Value	π_{y1}	**Robust SE**	π_{y0}	**Robust SE**	π_{y1}	**Robust SE**	π_{y0}	**Robust SE**
0	-0.02	0.02	0.06	0.01	-0.002	0.02	0.06	0.01
1	0.23	0.13	0.47	0.03	0.12	0.03	0.45	0.03
2	0.03	0.02	0.27	0.03	0.10	0.02	0.23	0.03
3	0.01	0.01	0.04	0.01	0.01	0.01	0.03	0.01

Table 7: No Design Effect Assumption Joint Test

	Test for design effects (with GMS)							
	Price 150				**Price 600**			
Ha:Pr<0	**K**	**Lambda**	**P>Lambda**	**#P>Lambda**	**K**	**Lambda**	**P>Lambda**	**#P>Lambda**
Pr(R,S=0)	0	0	1	1	0	0	1	1
Pr(R,S=1)	0	0	1	1	1	6.19	0.01	0.01
	Price 1050				**Price 1500**			
Ha:Pr<0	**K**	**Lambda**	**P>Lambda**	**#P>Lambda**	**K**	**Lambda**	**P>Lambda**	**#P>Lambda**
Pr(R,S=0)	0	0	1	1	0	0	1	1
Pr(R,S=1)	1	1.25	0.13	0.26	0	0.01	0.46	0.92

Figure 1: Partial Demand Curves for Seed Hub by Question Type

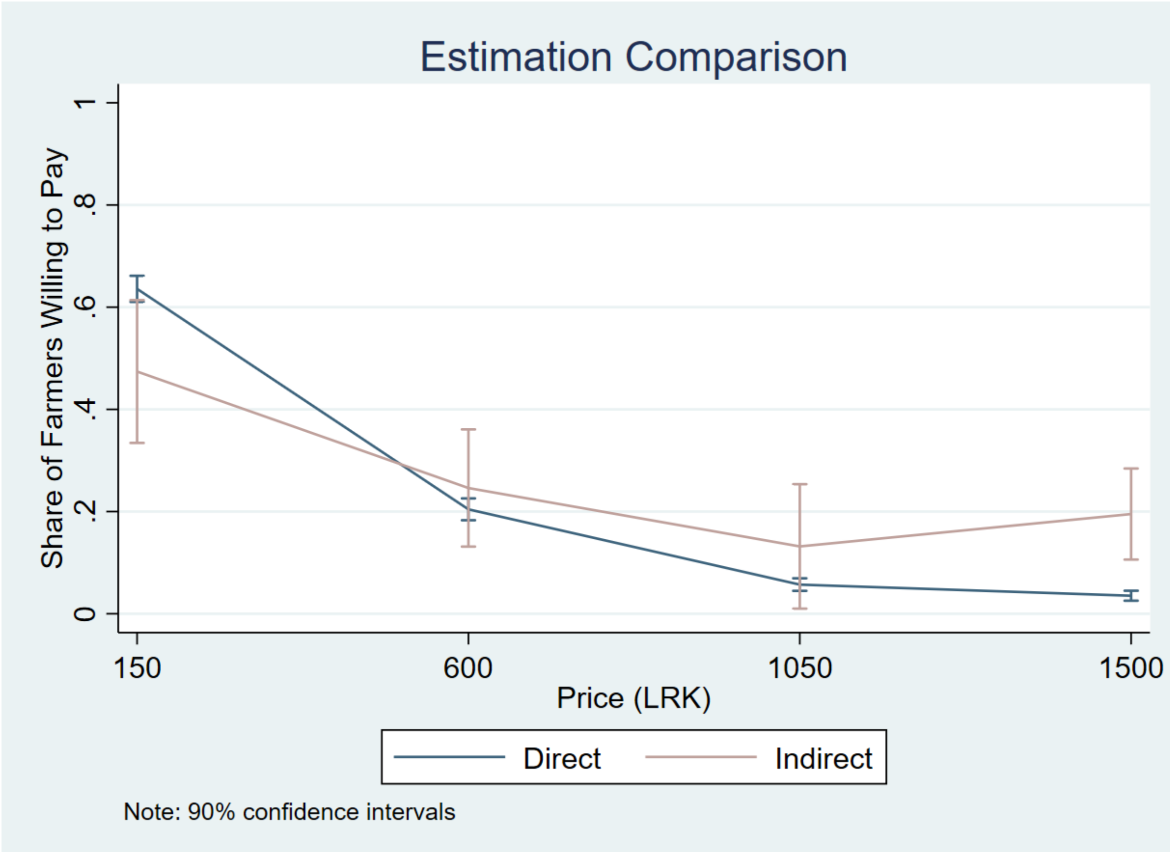


Figure 2: Partial Demand Curves for Seed Hub by Question Type Among Users

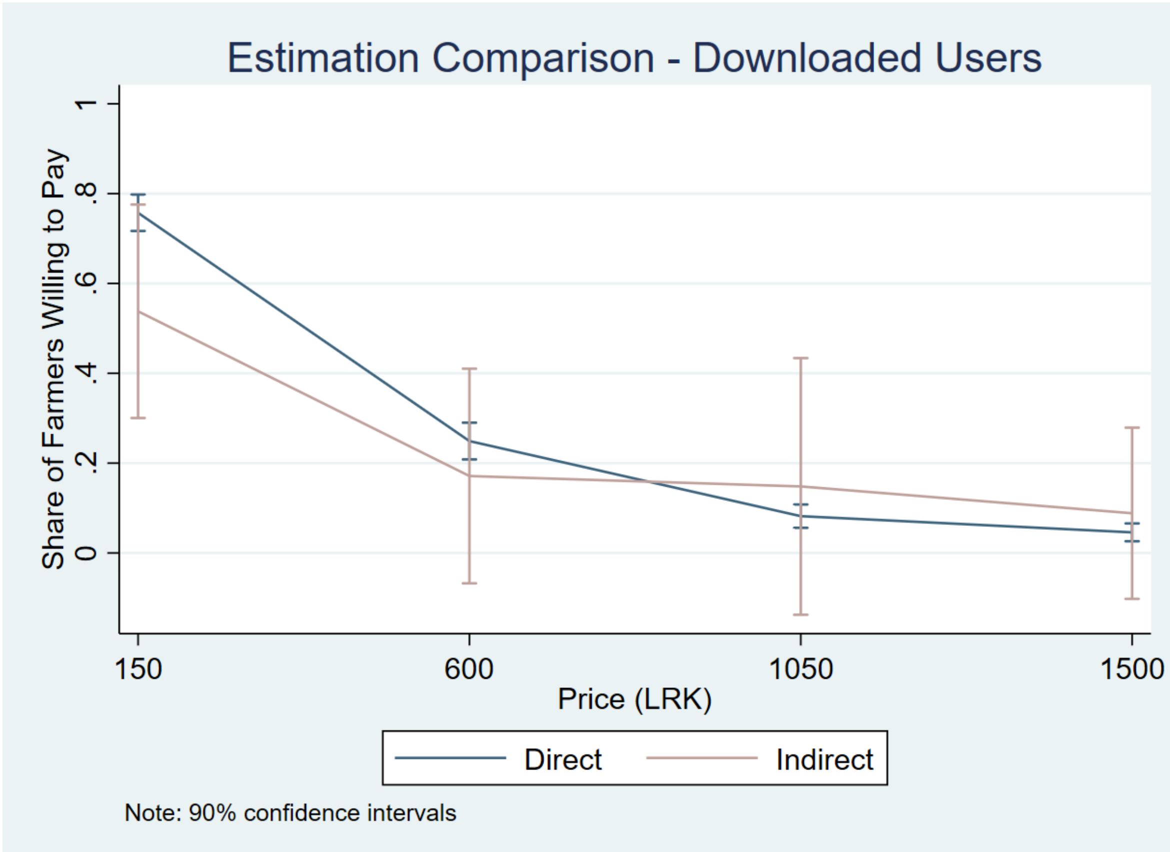
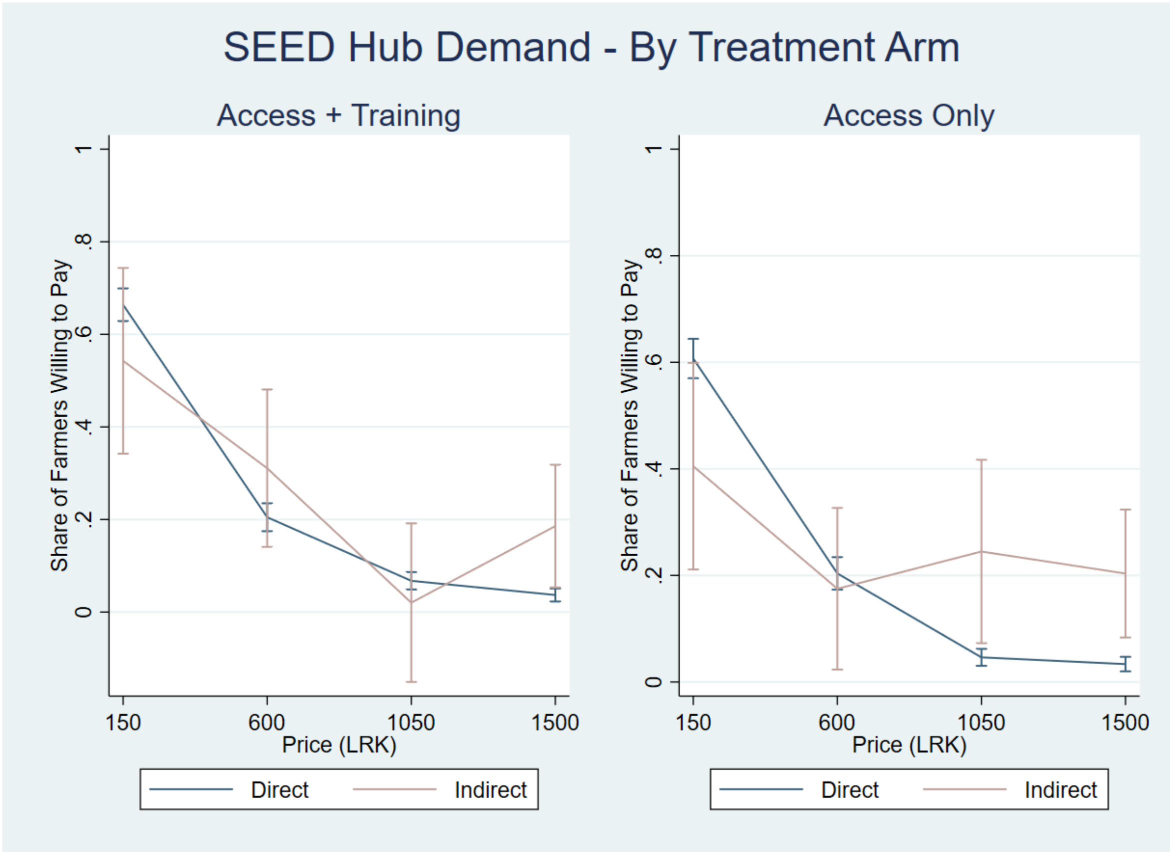


Figure 3: Partial Demand Curves for Seed Hub by Question Type and RCT Treatment Arm



References

Ashraf, N., B.K. Jack, and E. Kamenica. 2013. "Information and subsidies: Complements or substitutes?" *Journal of Economic Behavior & Organization* 88:133–139.

Berry, J., G. Fischer, and R. Guiteras. 2020. "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana." *Journal of Political Economy* 128:1436–1473, Publisher: The University of Chicago Press.

Blair, G., A. Coppock, and M. Moor. 2020. "When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114:1297–1315.

Blair, G., and K. Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20:47–77.

Breidert, C., M. Hahsler, and T. Reutterer. 2006. "A REVIEW OF METHODS FOR MEASURING WILLINGNESS-TO-PAY." *Innovative Marketing* 2.

Carlsson, F., D. Daruvala, and H. Jaldell. 2010. "Do you do what you say or do you do what you say others do?" *Journal of Choice Modelling* 3:113–133.

Carlsson, F., M. Kataria, and E. Lampi. 2018. "Demand effects in stated preference surveys." *Journal of Environmental Economics and Management* 90:294–302.

Centre, S.E.a.P. 2023. "Cost of Cultivation of Agricultural Crops." Working paper No. Volume 84, Department of Agriculture, Peradeniya, Dec.

Champ, P.A., R. Moore, and R.C. Bishop. 2009. "A Comparison of Approaches to Mitigate Hypothetical Bias." *Agricultural and Resource Economics Review* 38:166–180.

Channa, H., A.Z. Chen, P. Pina, J. Ricker-Gilbert, and D. Stein. 2019. "What drives smallholder farmers' willingness to pay for a new farm technology? Evidence from an experimental auction in Kenya." *Food Policy* 85:64–71.

Cohen, J., and P. Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment*." *The Quarterly Journal of Economics* 125:1–45.

Colombo, S., K. Glenk, and B. Rocamora-Montiel. 2016. "Analysis of choice inconsistencies in on-line choice experiments: impact on welfare measures." *European Review of Agricultural Economics* 43:271–302.

Cullen, C. 2023. "Method Matters: The Underreporting of Intimate Partner Violence." *The World Bank Economic Review* 37:49–73.

Cummings, R.G., and L.O. Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method." *American Economic Review* 89:649–665.

Entem, A., P. Lloyd-Smith, W..V.L. Adamowicz, and P.C. Boxall. 2019. "Using inferred valuation to quantify survey and social desirability bias in stated preference research." *American Journal of Agricultural Economics* 104:1224–1242, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajae.12268.

Fischer, K., J. van den Berg, and C. Mutengwa. 2015. "Is Bt maize effective in improving South African smallholder agriculture?" *South African Journal of Science* 111:1–2, Num Pages: 2 Place: Pretoria, South Africa Publisher: Academy of Science of South Africa Section: Commentary.

Fuller, A.J., and J. Ricker-Gilbert. 2021. "Estimating Demand for Third-party Quality Testing in Rural Grain Markets: Evidence from an Experimental Auction for Measuring Moisture Content in Kenya." *Journal of African Economies* 30:389–417.

Glynn, A.N. 2013. "What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77:159–172.

Iizuka, T., and H. Shigeoka. 2022. "Is Zero a Special Price? Evidence from Child Health Care." *American Economic Journal: Applied Economics* 14:381–410.

Jouvin, M. 2023. "Addressing Social Desirability Bias When Measuring Child Labor Use: An Application to Cocoa Farms in Côte d'Ivoire." *The World Bank Economic Review*, Sep., pp. lhad030.

Leggett, C.G., N.S. Kleckner, K.J. Boyle, J.W. Duffield, and R.C. Mitchell. 2003. "Social Desirability Bias in Contingent Valuation Surveys Administered through In-Person Interviews." *Land Economics* 79:561–575, Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Levitt, S.D., and J.A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21(2):153–174.

List, J.A., R.P. Berrens, A.K. Bohara, and J. Kerkvliet. 2004. "Examining the Role of Social Isolation on Stated Preferences." *American Economic Review* 94:741–752.

Lusk, J.L., and F.B. Norwood. 2009a. "Bridging the gap between laboratory experiments and naturally occurring markets: An inferred valuation method." *Journal of Environmental Economics and Management* 58(2):236–250.

—. 2010. "Direct Versus Indirect Questioning: An Application to the Well-Being of Farm Animals." *Social Indicators Research* 96:551–565.

—. 2009b. "An Inferred Valuation Method." *Land Economics* 85:500–514, Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Lépine, A., C. Treibich, and B. D'Exelle. 2020. "Nothing but the truth: Consistency and efficiency of the list experiment method for the measurement of sensitive health behaviours." *Social Science & Medicine (1982)* 266:113326.

Lépine, A., C. Treibich, C.T. Ndour, K. Gueye, and P. Vickerman. 2020. "HIV infection risk and condom use among sex workers in Senegal: evidence from the list experiment method." *Health Policy and Planning* 35:408–415.

Menapace, L., and R. Raffaelli. 2020. "Unraveling hypothetical bias in discrete choice experiments." *Journal of Economic Behavior & Organization* 176:416–430.

Miller, K.M., R. Hofstetter, H. Krohmer, and Z.J. Zhang. 2011. "How Should Consumers' Willingness to Pay be Measured? An Empirical Comparison of State-of-the-Art Approaches." *Journal of Marketing Research* 48:172–184, Publisher: SAGE Publications Inc.

Murphy, J.J., P.G. Allen, T.H. Stevens, and D. Weatherhead. 2005. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation.", pp. .

Norwood, F.B., and J.L. Lusk. 2011. "Social Desirability Bias in Real, Hypothetical, and Inferred Valuation Experiments." *American Journal of Agricultural Economics* 93:528–534, Publisher: [Agricultural & Applied Economics Association, Oxford University Press].

Rawat, R., P.H. Nguyen, L.M. Tran, N. Hajeebhoy, H.V. Nguyen, J. Baker, E.A. Frongillo, M.T. Ruel, and P. Menon. 2017. "Social Franchising and a Nationwide Mass Media Campaign Increased the Prevalence of Adequate Complementary Feeding in Vietnam: A Cluster-Randomized Program Evaluation1." *The Journal of Nutrition* 147:670–679.

Reynolds, W.M. 1982. "Development of reliable and valid short forms of the marlowe-crowne social desirability scale." *Journal of Clinical Psychology* 38:119–125, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-4679%28198201%2938%3A1%3C119%3A%3AAID-JCLP2270380118%3E3.0.CO%3B2-I.

Rosenfeld, B., K. Imai, and J.N. Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60:783–802.

Schmidt, J., and T.H.A. Bijmolt. 2020. "Accurately measuring willingness to pay for consumer goods: a meta-analysis of the hypothetical bias." *Journal of the Academy of Marketing Science* 48:499–518.

Schwab, B., and J. Yu. 2024. "Guaranteed storage? Risk and credit constraints in the demand for postharvest technology." *Oxford Economic Papers*, Jul., pp. gpae027.

Shampanier, K., N. Mazar, and D. Ariely. 2007. "Zero as a Special Price: The True Value of Free Products." *Marketing Science* 26:742–757, Publisher: INFORMS.

Shukla, P., H.K. Pullabhotla, and K. Baylis. 2022. "Trouble with zero: The limits of subsidizing technology adoption." *Journal of Development Economics* 158:102920.

Sælensminde, K. 2002. "The Impact of Choice Inconsistencies in Stated Choice Studies." *Environmental and Resource Economics* 23:403–420.

Tadesse, G., G.T. Abate, and T. Zewdie. 2020. "Biases in self-reported food insecurity measurement: A list experiment approach." *Food Policy* 92:101862.

Tsai, C.l. 2019. "Statistical analysis of the item-count technique using Stata." *The Stata Journal: Promoting communications on statistics and Stata* 19:390–434.

Yadav, L., T.M. Van Rensburg, and H. Kelley. 2013. "A Comparison Between the Conventional Stated Preference Technique and an Inferred Valuation Approach." *Journal of Agricultural Economics* 64:405–422.

Zizzo, D.J. 2010. "Experimenter demand effects in economic experiments." *Experimental Economics* 13:75–98.

Appendix

A Internal Validity

Table 1: Internal Validity Check

	Consistent	Inconsistent	p-value
Household size	4.072	4.010	0.528
Share with Male Head	0.505	0.494	0.403
Maximum Education	13.704	13.276	0.007
Number of Plots	3.629	2.877	0.000
Total Plot Area	5.028	7.266	0.003
Hired Male labor	5.770	10.281	0.000
Hired Female labor	0.460	0.158	0.006
HH male labor	1.165	0.557	0.013
HH female labor	0.393	0.063	0.002
Harvest (kg)	1655.168	4601.477	0.000
Largest plot	3.035	5.017	0.001
N	723	312	

B Power Calculations

Prior to survey implementation, we conducted power calculations to determine the appropriate number of questions and prices to include. The RCT treatment group had a total of 1,100 participants, making that our maximum potential sample size. Given the existing survey length and the complexity of the questions, the maximum number of list questions to be included was set at three per person. We then used the number of questions, the maximum potential sample, and the number of prices to calculate the sample size per price level for each scenario ($Sample_p = \frac{1100*Questions}{Prices}$). Table 2 summarizes the scenarios considered.

Next, we needed to identify how often we could detect (1) the share WTP different from zero at standard significance levels and (2) a statistically significant difference between the direct and indirect questioning. The first objective assumes a 20% true use rate of SEED Hub, which is drawn from the midline survey. Prevalence of the non-sensitive items in the list is assumed based on the mid-line survey and publicly available data sets. The level of bias in WTP is allowed to vary from 10% to 30%, based on social science literature levels of social desirability bias.

Based on the assumptions outlined, a data set was generated, objectives 1 and 2 were tested, and the p-value of the test was saved. The process was iterated 1,000 times. Finally, the share of the sample for each objective with $p < 0.05$ is reported in Table 3. Detect refers to the ability to detect social desirability bias and Difference refers to the ability to distinguish the indirect from the direct measure statistically. The final selection of three questions with four prices was selected based on the available results. Future applications of this process should include an additional variable to account for the possibility of inconsistent responses, which may further decrease sample size though the attrition need not be related to the price level.

Table 2: Power Calculation Key

Sample Size per price level	List Experiments	Prices
733	2	3
550	2	4
440	2	5
1100	3	3
825	3	4
660	3	5
550	3	6
471	3	7
413	3	8

Table 3: Power Calculation Results

Bias	Sample Size	Detect	Difference
0.1	367	70.80	17.50
0.15	367	71.40	32.10
0.2	367	72.20	52.00
0.25	367	74.70	71.60
0.3	367	71.70	86.90
0.1	412	75.20	19.30
0.15	412	76.10	36.40
0.2	412	77.50	54.80
0.25	412	74.60	76.10
0.3	412	77.20	86.90
0.1	440	81.30	17.40
0.15	440	78.80	37.70
0.2	440	79.70	60.40
0.25	440	79.40	78.50
0.3	440	80.30	90.70
0.1	471	81.70	20.60
0.15	471	81.60	41.40
0.2	471	82.10	62.70
0.25	471	79.60	83.20
0.3	471	82.70	93.40
0.1	550	87.30	23.40
0.15	550	87.90	43.00
0.2	550	86.10	69.20
0.25	550	86.50	86.80
0.3	550	85.40	97.10
0.1	550	87.10	23.50
0.15	550	88.40	44.10
0.2	550	86.70	69.20
0.25	550	87.30	87.00
0.3	550	86.30	95.30
0.1	660	92.70	27.60
0.15	660	91.50	52.20
0.2	660	92.60	76.20
0.25	660	92.70	90.30
0.3	660	93.20	98.20
0.1	733	95.10	25.20
0.15	733	96.30	57.40
0.2	733	94.40	80.30
0.25	733	94.00	92.80
0.3	733	94.40	99.30
0.1	825	96.60	32.70
0.15	825	96.60	61.40
0.2	825	96.20	86.30
0.25	825	96.00	96.50
0.3	825	97.90	99.90
0.1	1100	99.40	43.70
0.15	1100	99.40	75.90
0.2	1100	99.20	94.50
0.25	1100	99.80	98.90
0.3	1100	99.60	100.00

Detect is how often we could detect a statistically significant effect, given a 20% true prevalence of SEED hub usage. Difference is how often we could detect a statistically significant difference between list and direct questioning.

C Inconsistent Response

Figure 1: With and Without Inconsistent Responses