Learning from failures of protocol in cross-cultural research

Daniel J. Hruschka[a,1], Shirajum Munira[b], Khaleda Jesmin[b], Joseph Hackman[a], and Leonid Tiokhin[a]

[a]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85281; and [b]LAMB Project for Integrated Health and Development, Parbatipur 5250, Bangladesh

The many tools that social and behavioral scientists use to gather data from their fellow humans have, in most cases, been honed on a rarefied subset of humanity: highly educated participants with unique capacities, experiences, motivations, and social expectations. Through this honing process, researchers have developed protocols that extract information from these participants with great efficiency. However, as researchers reach out to broader populations, it is unclear whether these highly refined protocols are robust to cultural differences in skills, motivations, and expected modes of social interaction. In this paper, we illustrate the kinds of mismatches that can arise when using these highly refined protocols with nontypical populations by describing our experience translating an apparently simple social discounting protocol to work in rural Bangladesh. Multiple rounds of piloting and revision revealed a number of tacit assumptions about how participants should perceive, understand, and respond to key elements of the protocol. These included facility with numbers, letters, abstract number lines, and 2D geometric shapes, and the treatment of decisions as a series of isolated events. Through on-the-ground observation and a collaborative refinement process, we developed a protocol that worked both in Bangladesh and among US college students. More systematic study of the process of adapting common protocols to new contexts will provide valuable information about the range of skills, motivations, and modes of interaction that participants bring to studies as we develop a more diverse and inclusive social and behavioral science.

generalizability | diversity | cross-cultural | social discounting | Bangladesh

In 1932, the psychologist Rensis Likert (1) published his dissertation on a novel method for measuring attitudes. After giving university students printed statements about race relations, Likert asked them to check one of five options (strongly approve, approve, undecided, disapprove, and strongly disapprove) indicating how much they endorsed each of these statements. Likert then assigned numbers to these levels of approval and took an average across all statements. The simplicity of both the response format and construction of the scale soon spurred researchers to adopt elements of the technique to assess not only attitudes (Likert's original interest) but also subjective judgments along many dimensions, including likelihood, desirability, difficulty, and happiness (2). Today, after decades of testing and refinement on generations of participants, Likert's simple format has become a reliable mainstay of social and behavioral research.

Given its ubiquity in the social and behavioral sciences, one might guess that a five- or seven-item Likert format is a natural way of asking humans about their subjective judgments. However, in the rare cases when researchers have described their experience using Likert items outside of formally educated populations, they have been met with mixed success (3, 4). It turns out that respondents are often unable or unwilling to make a choice among graded levels, preferring instead to provide yes or no answers to either the entire statement or to each individual level (e.g., Always? Yes, Sometimes? Yes, Never? No) (3). These difficulties suggest that the participants who are most likely to find Likert items intuitive are those whom we typically study:

college students who have years of experience responding to Likert items.

A deeper problem with Likert items is that we currently have very little systematic knowledge or theory about when such a simple format will elicit useful information and when it will fail. In some cases, Likert items appear to generate meaningful data across diverse human populations (5, 6). In others, researchers report high rates of nonresponse or misunderstanding (3, 4). Except for analyses of extreme responding and acquiescence bias (7), only rare studies have examined what specific aspects of the task need to be revised in diverse settings to elicit meaningful responses (8, 9). With this lack of a systematic, comparative understanding of how humans interact with something as simple as a Likert item, it can be challenging to interpret and compare the results from diverse settings. It is perhaps for this reason that the longest running worldwide surveys of low- and middle-income countries (i.e., demographic and health surveys), which include populations varying dramatically in literacy, numeracy, and cultural background, often ask attitude questions with a yes/no, instead of Likert, format (10, 11).

Beyond the Likert Format

Likert's format for eliciting judgments is just one of many tools that social and behavioral scientists have invented to collect data from their fellow humans. Common tools include interview schedules, survey questions, hypothetical vignettes, number lines, 2D pictures, and abstract geometric forms to represent ineffable concepts. This impressive toolkit has been developed, refined, and passed down across generations as researchers have learned which techniques efficiently generate desired results, like the Likert format, and which ones do not (12). In many cases, these tools have evolved to capitalize on the skills, cognitive tendencies, and social expectations of the most typical study participants (i.e., college undergraduates, highly literate and numerate populations) (13–17). In turn, the success of many methods now depends on these same unique skills and proclivities, including the ability to read and write, to use a big vocabulary, and to quickly process large numbers and abstract geometric figures, as well as the willingness to focus on a few abstract details of a situation rather than the larger context of a problem, to entertain

[1]To whom correspondence should be addressed. Email: dhruschk@asu.edu.

hypothetical vignettes, and to give finely graded judgements (as in the Likert example).

In short, while protocols have been built to run smoothly and efficiently with certain rarefied populations, they can seriously break down in contexts where people have practiced different abilities, cultivated different motivations, and share different expectations about appropriate social interactions (13, 14, 18–22). A trivial example of a methodological breakdown would occur if a naive researcher conducted an interview in a language not spoken fluently by a study participant. Researchers are keenly aware of this problem and have devoted considerable attention to translating survey questions between languages and contexts to ensure that participants understand and respond to questions in roughly equivalent ways (23–25). Extensive work also deals with the effects of subtle changes in wording questions (26) and dialectical differences between participants assumed to speak the same language (27).

Even though it has received the most attention, language misuse is just the tip of the iceberg. Protocols also depend on tacit, unexamined assumptions about the following: (*i*) the meaning of concepts, symbols, and pictorial representations; (*ii*) normal ways of answering questions or responding to stimuli; and (*iii*) socially acceptable interactions among researchers and participants. As an example, US and European social surveys have tracked general trust in others for several decades with a question devised in 1948: "Generally speaking, would you say that most people could be trusted or that you need to be careful in dealing with people?" (28). The phrase "most people" in this question was intended to capture the trust placed in a broad set of unfamiliar others as opposed to one's closest friends and family. However, in many parts of world, the concept of most people has a meaning closer to family, neighbors, and friends. Without considering this conceptual difference, we would conclude that the average level of generalized trust in Australia and Thailand is equivalent, when, in fact, the two countries recently ranked fourth and 29th, respectively, in trust in unfamiliar others out of 51 countries (28).

Similar problems arise when using apparently natural graphic symbols to represent concepts (29) or pictorial designs to represent spatial arrangements (15, 30, 31). Indeed, if we turn to a literature where miscommunicating with symbols has life or death consequences (studies of safety, emergency, and traffic signs), a recurring finding is that there are few (if any) universally meaningful symbols and that correct interpretation usually depends on extended periods of learning and training (32, 33). However, tacit assumptions about the universality of certain visual stimuli can profoundly shape how we interpret responses to those stimuli. Consider a recent study that examined a standard pattern recognition tool for assessing cognitive abilities among 2,711 Zambian schoolchildren. A standard item consists of a sequence of 2D squares, circles, and triangles, with a missing space for which a child is asked to choose from several options to complete the pattern. When presented in this standard format, only 12.5% of Zambian schoolchildren correctly answered more than half of the questions. However, when the same patterns were presented to children with familiar 3D objects (e.g., toothpicks, stones, beans, beads), nearly threefold as many schoolchildren achieved this benchmark (34.9%) (14). Without the careful work of learning how schoolchildren viewed and interacted with "basic" objects, researchers naively applying the first standard tool would have dramatically underestimated the cognitive abilities of these Zambian schoolchildren.

Existing cross-cultural work on spatial and visual perception reveals many other opportunities for researchers to fail when using images as part of a protocol with new audiences. For example, researchers and respondents may not agree on the set of geometric transformations (e.g., rotation, translation, reflection, enlargement) that make two images "the same" (34). They may also draw from different bags of tricks for representing and perceiving spatial arrangements, such as the illusion of depth in a 2D space. The dominant technique in current European art traditions (i.e., perspective) relies on lines that converge at a vanishing point "in the distance." However, other traditions rely on different techniques, including oblique parallel lines, vertical stacking, changing hues, and obscuration with visual mists, to accomplish the same task (15, 35). As early European criticisms of "flat" Chinese landscapes suggest, it takes practice to be deceived by these techniques, and a naive audience viewing an unfamiliar technique may be immune to its effects (36). As such, trying to depict something as simple as a distant point might fail if the researcher uses the wrong bag of tricks for doing so.

In addition to response formats and stimuli, researchers' protocols rely on a uniquely stylized form of social interaction between researchers and respondents. In these protocols, researchers often embed tacit assumptions about normal or reasonable social behavior that may not conform to local norms of common or appropriate behavior. The anthropologist Charles Briggs (19) discovered this the hard way when elders in a northern New Mexico community were unwilling to take a subordinate role to a young investigator by submitting to questions and directions. Similarly, in a setting where one should avoid speaking in front of elders, children may be reluctant to respond to researchers' requests, not because they cannot respond but because they feel they should not (22). Researchers may also impose what they view to be natural social situations between participants, (e.g., face-to-face or one-on-one interaction between mothers or infants) to assess important concepts, such as attachment or sensitivity (37). In the many cultural settings where these imposed situations are not the dominant mode of interaction (e.g., in multiple caregiver settings), observations of behavior can pose deep interpretive challenges (37–40).

When such failures arise, they are more than just methodological flaws or artifacts (41, 42). They also represent weaknesses in researchers' (often tacit) theories of how participants should interact, think, perceive, and respond (43). When tacit assumptions about appropriate interactions are wrong, participants can refuse to respond (3, 19) or give agnostic responses (44). Researchers may also inadvertently interact with only participants who have the interest and ability to fully engage with the protocols (19) or even discard or ignore "abnormal" or "inconsistent" responses from their analyses (3). More broadly, incorrect tacit expectations about reasonable ways of perceiving, thinking, and responding also risk leading researchers to mistake unorthodox responses as deficit or deficiency (18, 45).

Despite the promise of learning from these failures, there are few published examples (outside language translation or extreme responding on ordinal scales) that describe the challenges researchers must overcome when meaningfully adapting protocols to new, cross-cultural contexts. To illustrate the pervasive failures that arise from these tacit assumptions (and how they can reveal weaknesses in these assumptions), we document the challenges we faced when adapting a common tool to assess social discounting in rural Bangladesh. We also describe the collaborative process that allowed us to develop tools to deal with these challenges. Our case study illustrates how adapting protocols to new settings can reveal mismatches between the hidden baggage that we bring with us and the demands of the local setting, as we work toward a more inclusive social and behavioral science.

Adapting a Common Protocol: Social Discounting

Social discounting, as defined in psychology, is the tendency to bear greater costs to benefit socially close others (46). Investigators across a range of disciplines, including psychology, economics, sociology, and anthropology, have consistently documented this behavioral bias, and its apparent pervasiveness has spurred considerable

theoretical work on its nature and origins. One line of theorizing suggests that social discounting results from a "merging" of neural representations of the self with socially close others in a way that partially confounds costs and benefits to self versus others (47). Others have identified a hyperbolic relationship between social distance and willingness to sacrifice, and have therefore hypothesized an underlying connection to hyperbolic discounting of delayed or probabilistic returns (48). The apparent regularity and pervasiveness of social discounting have even led one set of researchers to call it the "inverse distance law of giving" (49) and others to identify its underlying neural basis (50).

A growing database of more than 50 studies, including a preregistered direct replication, has reliably documented social discounting (51). However, current studies rely almost exclusively on highly educated populations in the United States and Europe, with the addition of some highly educated populations in China (52), India (53), and Singapore (54). Cross-cultural research with other populations suggests that the effect of social distance on generosity may be less important than other factors in some cultural settings. For example, Miller and Bersoff (55) found that among respondents from Mysore, India, a variable commonly related to social closeness, liking, had little effect on Indian respondents' perceived moral responsibility to help others, whereas it did among US participants. Miller and Bersoff (55) point out that this is consistent with a cultural view that behavior in relationships should depend on the formal duties of the relationship rather than on individual feelings toward another person. More broadly, this may reflect a different perspective on the self and the importance of personal feelings and goals in guiding behavior toward others (56). If this is true, then we should expect feelings of social closeness to poorly predict helping or sacrifice in these settings.

To have any hope of adequately testing this and other hypotheses about cross-population differences in social discounting, we must first develop protocols that capture the key variables, closeness and the willingness to sacrifice, in ways that are both locally meaningful and comparable across sites. For this reason, for the past 4 y, our team has been attempting to adapt the standard protocol for social discounting to work in low-literacy, low-numeracy farming communities in northwestern Bangladesh.

The Standard Protocol. The standard protocol for assessing social discounting in psychology and behavioral economics has been honed nearly exclusively on college students in the United States, Europe, and China. In one of its earliest forms, the protocol involved a paper-and-pencil task that first asked respondents to:

> "Imagine a list of 100 people closest to you in the world, ranging from your dearest friend or relative at position 1 to a mere acquaintance at 100. The person at 1 would be someone you know well and is your closest friend or relative. The person at 100 might be someone you recognize and encounter, but perhaps you may not even know [his or her] name."

The respondent is then asked to imagine a person at a specific location on that list (e.g., 1, 5, 10, 20, 50, 100) and is instructed to:

> "Make a series of judgments based on your preferences. On each line, you will be asked if you would prefer to receive an amount of money for yourself versus an amount of money for yourself and the person listed."

The respondent is then presented with a sequence of choices in the following format and asked to circle A or B for his or her choice on each line:

A. $85 for you alone. B. $75 for the ___ person on the list
A. $75 for you alone. B. $75 for the ___ person on the list
A. $65 for you alone. B. $75 for the ___ person on the list
A. $5 for you alone. B. $75 for the ___ person on the list

To assess how much a respondent would sacrifice for a given partner, the researcher identifies the point in the sequence of questions where the respondent switches from response A (some amount for you alone) to response B ($75 for the other person). The "amount foregone to give someone $75" is then calculated as the average amount foregone between the responses where the respondent crossed from A to B. Those rare cases (usually not more than 10% of participants) that have more than one crossover point are conventionally excluded from analyses (46, 51, 57). With an estimate of the amount foregone to give someone $75, a researcher can then examine how this varies across partners at varying social distances.

As one might expect, adapting this protocol to farming communities in rural Bangladesh required many modifications (*SI Appendix*). Some were apparent from the outset. For example, most respondents were not sufficiently literate to read instructions and complete a paper-and-pencil task. Other challenges required several rounds of piloting and collaborative discussion to identify and resolve. For example, it was necessary to create a hands-on protocol using concrete and visually intuitive artifacts that represented real, as opposed to hypothetical, payoffs from decisions. We did this by allowing respondents to literally hold the two choices on two slips of paper. We then asked them to place their preferred choice in a small "lottery" bucket, from which we randomly pulled one slip to pay out at the end of the experiment. The study team also needed to find ways to identify individuals in villages where many people had the same formal names and where it might not be appropriate to call elders by their first names. For this, we used photographs instead of names of consenting adults. We also needed to change the currency from money to rice to avoid envy that could jeopardize our long-term relationship with the communities (58). Many of these are the same kinds of challenges that researchers regularly have to resolve when adapting protocols to new contexts outside of standard laboratory settings (59, 60).

For the purposes of this paper, we focus on the challenges of assessing a key independent variable, "social closeness," and a dependent variable, "amount foregone to give something to another person." The process of piloting and adapting these aspects of the protocol revealed key assumptions about participants that did not hold in this rural region of Bangladesh. We identified these potential misunderstandings by piloting the task with community members using cognitive interviews. Specifically, at each stage of the protocol, we would ask participants if they understood what they were being asked to do, and if so, to re-state their view of what was expected (26). We were fortunate to have cultivated a sufficiently good relationship with participants that they were willing to tell us when they did not understand something and a sufficiently good relationship with the study team that they would tell us when they thought something was amiss. In addition to this proactive approach, we responded to behavioral cues during piloting, such as long hesitation, laughing out loud, and repeated questioning about "what am I supposed to do again?" as an impetus to ask more about problems with protocols. Through collaborative brainstorming with participants and the local study team, we learned how to appropriately ask about these concepts and triangulate our assessments to have some confidence that we were assessing the concepts of interest.

The process of adaptation was part of a project approved by the Institutional Review Boards at both Arizona State University and the LAMB Project for Health and Development in Bangladesh, which included adults who provided verbal consent to participate in the project.

Assessing Closeness. Before we began asking Bangladeshi villagers about relationship qualities ("who they felt close to"), it was necessary to find out how they talked about social relationships and what linguistic tools they used to characterize differences in

the quality of their relationships. In open-ended interviews, respondents often used an adjective, *ghonishto*, to characterize the quality of their relationships with relatives and friends. Derived from the root *ghon*, meaning variously thick, solid, or viscous, *ghonishto* describes relationships that are intimate, close, or familiar. In interviews, respondents mentioned that *ghonishto* friends and relatives feel comfortable around each other; enjoy visiting each other; help each other with chores, loans, and advice; and trust each other enough to talk freely, or "open their minds," to each other about important, sensitive, and secretive matters. The factors described as making people feel less close included envy; regular failures to help when needed; conflicts over lending and borrowing; fights between one's children; and disputes over farmland, livestock, and situations where another's cattle ate one's crops (61). Another potential source of conflict was worry that an envious friend had used supernatural tools (amulets and spells) to cause injury or illness. In addition to *ghonishto*, respondents sometimes used a spatial metaphor of nearness (*kache*) and an ordinal measure of firstness (*prio*), but not with the same frequency; thus, we used *ghonishto* as the crucial term in the protocol.

In the social and behavioral sciences, there are two common methods of assessing social closeness. The first involves imagining an abstract list of 100 people and then selecting individuals at specific locations along that list (e.g., 1, 2, 5, 10, ...) (46). The second uses a seven-point Likert item, with each item depicting two circles with varying degrees of overlap. This task involves imagining oneself and another as abstract circles on a sheet of paper and the overlap between those circles as a measure of nearness or closeness to that person (47). In early piloting, both methods caused serious confusion and hesitation among participants.

For the Likert item, we went through five modifications while attempting to make the task more concrete and to encourage fluid responding. This involved changing the overlapping circles to overlapping stick figures; changing the orientation of the seven items from a left-to-right orientation to a top-to-bottom orientation to encourage thinking about quantities in terms of up and down; reducing the number of items from seven to five; and having respondents place photographs of other villagers in five baskets in a line from left to right, each corresponding to a different level of closeness (Fig. 1). Each of these solutions created its own problems and confusions. Even when respondents said they understood the tasks, they often took a long time to make decisions about each partner.

A breakthrough came when we decided to arrange the baskets in a line in front of the respondent, with each successive basket a bit farther away. Arranging the baskets in terms of distance from ego apparently fixed the problems experienced with earlier versions. With this new version, respondents usually expressed little difficulty, responded fluently, and were able to classify the entire adult population of the village (~100 in each village) in 5–10 min.

Based on this insight, we used a similar distance-from-ego approach to rank social partners from 1 to 100. However, we encountered another snag, as participants hesitated when identifying person 5, person 10, or person 20 in an imaginary list. To deal with this, we gave participants a pile of all photographs of all individuals in the village and asked them to choose the 20 they felt closest to. We then asked them to place those 20 photographs in the order of how *ghonishto* their relationship was, with the photograph closest to them representing person 1 and the photograph farthest from them representing person 20. The researcher then picked the appropriate photographs from that ordering and used those as the targets for questions about sacrifice. Organized this way, the task posed few difficulties. In most cases, respondents identified rankings quickly and with few follow-up questions.
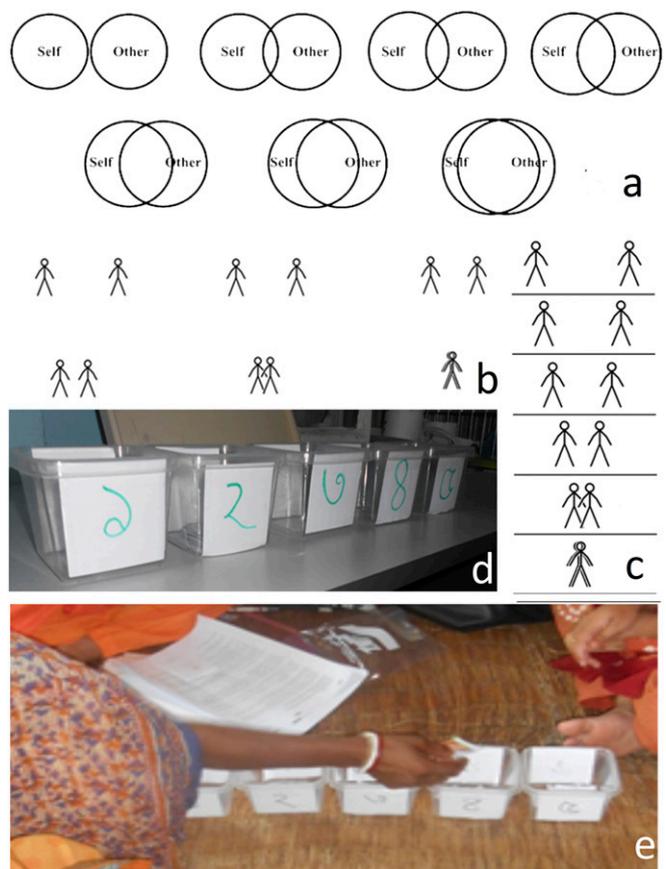
Fig. 1. Five ways of asking participants about social closeness: original Aron inclusion of other in self scale (*A*), modified version with stick figures (*B*), modified version vertical (*C*), numbered buckets for placing partner photographs (*D*), and final version with buckets arranged moving away from a participant (*E*; seated at the left of the photograph and reaching to the right of the photograph).

With these tools in hand, we could efficiently elicit judgments of social closeness (*ghonishto*) that captured key aspects of local usage. Most respondents (94%) placed their spouses, their closest genetic relatives ($R = 0.5$; 90%), and their closest friends (68%) in the closest bin, even though respondents, on average, used that bin for only 20% of general community members. Placement in the top 20 showed a similar pattern, with 95% of spouses, 91% of closest genetic relatives, and 86% of closest friends chosen for the top 20, and 95%, 72%, and 64% of the same relationships chosen for the top 10. When ranking and binning were compared directly, persons 1, 2, 5, 10, 15, and 20 on a respondent's ranked list were placed in the closest of the five bins 100%, 88%, 82%, 70%, 52%, and 32% of the time, respectively. In contrast, when asked about people outside the community who they did not know, respondents nearly always placed them in the farthest bin. The measures of closeness also showed strong independent associations with helping in the previous year (61).

This process revealed several tacit assumptions underlying the standard protocols about how participants perceive and respond to stimuli that did not hold up well in this setting. Standard protocols assumed that moving from left to right (or vice versa) on a Likert-type item was a natural way to represent magnitude or distance. They assumed that it was natural to represent individuals with abstract geometric shapes and to represent the relationship between two people as the relative positions of those abstract geometric figures. They also assumed that individuals

could imagine an ordinal list of 20 or more people ranked by some property and name people at specific locations on that list. The adaptation process also revealed aspects of the protocol that we might expect to cause confusion but, ultimately, did not. For example, in this case, mixing metaphors of spatial distance in the task with thickness in the linguistic description of the task (*ghonishto*) did not appear to pose serious problems for respondents. This illustrates that it can be difficult to make a priori predictions about what details of a protocol will or will not work in a particular setting, and why extensive piloting is a necessary first step when developing protocols for new contexts.

Assessing Generosity Toward Others. In a similar process, we arrived at a procedure for asking participants about their preferred choice across a sequence of decisions to sacrifice some amount of rice (e.g., 0.5 kg, 1 kg, 2 kg, 3 kg, 4 kg, 5 kg) to give another person 5 kg of rice. Once a respondent has made all decisions toward a given partner, a standard procedure in the literature is to identify the point in the sequence of choices at which a person crosses over from not sacrificing to sacrificing. This is variously called a crossover, indifference, or equivalence point (46, 48). If a participant crosses more than once (e.g., says he or she would sacrifice 2 kg and 0.5 kg, but not 1 kg), this poses problems for the researcher because there is no single point. In some cases, the participant or choice set is labeled as inconsistent and discarded from analysis (46). In others, the researcher neglects any repeat crossovers by focusing on the first crossover point that meets some condition (e.g., that has two subsequent choices in the same direction) (62). In whatever way the researcher chooses to identify the crossover point for an individual, it requires discarding participants, choice sets, or certain choices because they do not fit a model of a single crossover point.

In the social discounting literature, researchers frequently discard these participants or choice sets (51). This decision relies on a model of responding whereby respondents treat each decision as independent of the others and where they apply the same utility function to each independent decision. These assumptions are not particularly problematic in the vast majority of discounting studies among college students. In most cases, fewer than 10% of participants are discarded, with the underlying assumption that those participants did not understand the task or were not paying attention (51).

However, there is some evidence that when we leave college samples and present different kinds of discounting decisions to children or truck drivers in the Midwest, we find much higher levels of inconsistent responding (63, 64). When we applied this task in rural Bangladesh, we also found high rates of inconsistent responding (80% of all respondents with at least one crossover point) (51). The few studies that have found comparably high levels of inconsistent responding have attributed this to participant deficiencies (65). However, when our tacit models of responding force us to exclude the majority of participants, we need to ask whether it is a problem with the participants, a problem with the task, or a problem with our tacit assumptions about how participants should respond.

Although we cannot be sure which is the culprit in the Bangladesh case, we have indications that there are at least some problems with the underlying theory of responding. First, during the task, participants often talked out loud when making decisions. Among these statements, participants would consider their decisions in previous choices ("Well, I didn't give up 1 kg in the last decision, so I'll give up 2 kg this time" or "I already gave up 5 kg and 2 kg rice, so I won't give up 3 kg this time"). Moreover, in postchoice interviews about their decisions, respondents frequently used the phrase "give some, take some," which describes the kind of back-and-forth behavior that would lead to multiple

crossover points. This suggests that at least some respondents were not treating these as independent decisions but rather as aggregate contributions whereby the previous decision might have bearing on the subsequent decision. In doing so, it is completely reasonable for participants to move back and forth between giving and taking in a way that looks inconsistent from a model assuming a single crossover point.

Based on this insight, we developed alternative estimates of amount foregone that did not require discarding inconsistent responders. Using this approach, we found no evidence that inconsistent or "consistent" responders differed in their level of sacrifice as a function of their partner's social distance to a partner (51). This suggests that inconsistency in these settings is not a good marker of failure to understand the protocol. Notably, if we had discarded "inconsistent responders," we would have thrown out 80% of rural Bangladeshi respondents who gave anything toward any social partner.

Exporting the Adapted Protocol. When we applied this Bangladesh-adapted protocol (and several variants) over four field seasons in rural Bangladesh, we repeatedly arrived at a puzzling result: a completely flat relationship between perceived closeness and generosity. This was puzzling because it contradicted findings from over 50 published studies on social discounting in the United States and among college students in other countries. A number of checks suggested that this puzzling finding could not be solely explained by measurement issues or misunderstanding (51, 61). Another possibility was that we had adapted the protocol to such an extent that it would no longer demonstrate social discounting in any population, even US college students. To assess this possibility, we exported and translated the adapted protocol as directly as possible for use with US college students (51). We also exported the protocol to another low-income, rural, and semi-literate setting (rural Indonesia) to examine how well the adapted protocol could translate to other settings and to determine if the finding of no social discounting was unique to rural Bangladesh. Notably, the Bangladesh-adapted protocol produced the same results among US college students as the traditional US-adapted protocol (e.g., strong social discounting), but it also produced no social discounting among the rural Indonesia population. This suggests that the finding of no social discounting in rural Bangladesh is (*i*) not due simply to the process of adapting the protocol to the rural Bangladeshi context and (*ii*) not unique to Bangladesh. Interestingly, these findings were largely consistent with the postdecision justifications provided by the respondents. US respondents nearly unanimously justified their choices based on personal feelings about the relationship (e.g., love, closeness, liking), but less than half of Indonesian and almost none of Bangladeshi respondents made relationship-based arguments. In contrast, another factor, the donor's perceived need relative to the recipient's need, appeared to play a recurring role in decisions across all three samples. Specifically, relative need was associated with participants' decisions to sacrifice in all three samples, and the vast majority of participants in all three samples mentioned relative need in their postdecision justifications (51).

Discussion

The vast majority of tools that social and behavioral scientists use to learn from their fellow humans have been honed on a rarified set of humans possessing unique skills, motivations, and expected modes of social interaction. Through this process of methodological honing and refinement, researchers have produced tools that efficiently generate reliable data on this same set of participants. In turn, the success of these tools now depends on the unique skills, motivations, and expected modes of social interaction of the humans upon which these protocols were honed.

While rare case studies illustrate this point, we currently know very little about how the requisites for working efficiently with common protocols, like the social discounting protocol, are distributed around the globe. Literacy is probably the best-studied requisite (65, 66), but as our case study illustrates, we need to know about how people treat abstract geometric objects and imaginary situations of varying types, how they work with numbers, and how they perceive and assign meaning to different spatial orientations and arrangements. We also know next to nothing about the kinds of practiced competencies and learned motivations that people possess in the diverse settings not represented by US college students (13).

That said, the case studies reviewed here suggest that even the most basic tools of the social and behavioral sciences rely on tacit and largely untested expectations about how humans should perceive, think, interact, and respond when confronted with those tools. To illustrate the multiple ways that tacit researcher assumptions fail when exporting a refined protocol, we walked through the informal process used to identify problems and to overcome them. Notably, we identified key but untenable assumptions embedded in our protocols about how humans should be able to respond to stimuli: that moving from left to right (or vice versa) on a Likert-type item was a natural way to represent magnitude or distance, that it was natural to represent people with abstract geometric shapes and to represent the relationship between two individuals as the relative positions of those abstract geometric figures, and that individuals were willing to imagine an ordinal list of 20 or more people ranked by some property and then name individuals at specific locations on that list. Nonetheless, our study was not designed to formally examine the piloting process. Since we did not collect systematic data on how badly these assumptions were violated, they remain hypotheses to be examined in more detail in the future. However, they suggest that more scientific and transparent study of the protocol-adaptation process across contexts will refine our understanding of the tacit expectations embedded in existing protocols and improve our ability to craft protocols that participants respond to willingly, fluidly, and meaningfully.

Although the process we followed was informal, it relied on many tools for inquiry that we suspect will be generally useful for exploring how humans from diverse backgrounds interact with imported tools, procedures, and situations. First, in-depth engagement with the population, through observation, conversations, and interviews, is important for identifying how people use key terms of interest in everyday life (e.g., *ghonishto* as one analog to closeness), and how certain techniques and modes of asking will not work (e.g., using names to identify targets in rural Bangladesh), while others will (e.g., using photographs to identify targets) (14, 19, 67–69). This is especially important for identifying locally sensitive topics that might require special kinds of inquiry (70). Second, cultivating trusting relationships with participants and local researchers can encourage them to point out misunderstandings and to argue about the feasibility and meaning of different aspects of situations and protocols. Conversely, cultivating a habit of listening to local researchers and participants and discussing the benefits and costs of different suggestions ensures that such feedback is used when appropriate. This willingness to criticize, listen, and argue is especially important in contexts where people have a tendency to defer to the authority of the researcher or to avoid conflict. Third, cognitive interviewing, or systematically asking respondents about how they are perceiving and responding to situations, questions, and stimuli, provides an important check on all aspects of a protocol, and can help point to solutions when a roadblock is encountered (26). Fourth, by reexporting the Bangladesh-adapted protocol to US college students, we were able to examine whether the modifications changed how any population might respond to the

protocol. Finally, multiple triangulation of constructs, through what people say, through different elicitation formats, and with diverse validation checks, is crucial in cross-cultural settings, where researchers need to be especially cautious about their subjective assessment of the face validity of different measures (71, 72).

These are all steps that individual teams can take to identify and remedy failures of protocol in their specific field sites. However, stopping there would miss an opportunity to learn from each other's failures to identify commonly violated assumptions and move toward discovering general principles that underlie cross-culturally successful protocols. Such a pursuit would benefit from a shared platform for documenting and sharing how and why specific elicitation formats, stimuli, and social interactions succeed or fail. The US Centers for Disease Control and Prevention's Q-Bank provides one model, focused on a subset of protocols: survey questions. It permits survey methodologists to share their findings in a systematic, easily searchable platform, so that data analysts can understand the underlying construct and questionnaire designers can weigh evidence for alternative versions of questions. A comparable platform that can accommodate the broader set of tools in the social and behavioral sciences would be an important step toward ensuring that individual learning about protocols in diverse settings is documented and shared.

The ubiquity of failures at all levels of the research process may lead to skepticism about the merits of cross-cultural surveys that do not engage in such intensive piloting. However, given the currently limited knowledge about human cognitive and behavioral diversity, the social and behavioral sciences can still benefit from broad studies of how people worldwide respond to similar formats, stimuli, and situations. By identifying overt commonalities and differences, these studies provide raw material for further hypothesis generation and investigation about the root causes of variation. That said, such studies are not sufficient: In-depth investigation of how participants interact with these researcher-created technologies serves as an important complement to ensure that hypotheses and interpretations are consistent with local realities.

How fragile protocols will be in new contexts and populations is an empirical question, but documented cases suggest that fragility may be the rule rather than the exception. In part, this will stem from the tacit models or theories of how humans normally perceive, respond, and interact that underlie existing protocols. These may not be the models that we are interested in testing, but getting these models wrong can have serious consequences for efforts to test the theories that we actually care about (18, 19, 43). In contrast, getting these models right will be an important step toward developing better ways of adapting protocols across different human groups. It is difficult to imagine getting such models right without careful on-the-ground observation and long-term conversations with participants about what is meaningful and acceptable. This process is particularly important, since cumulative cultural learning over generations constructs lifeways, practices, knowledge, and ways of perceiving the world that are unique and difficult to predict a priori. With more systematic work designed to understand how people and protocols interact in the broadest range of human settings, we might learn that what appear to be mere methodological concerns can reveal crucial theoretical insights about human cognitive and behavioral diversity.

1. Likert R (1932) A technique for the measurement of attitudes. *Arch Psychol* 22:5–55.
2. Hills P, Argyle M (2002) The Oxford happiness questionnaire: A compact scale for the measurement of psychological well-being. *Pers Individ Dif* 33:1073–1082.
3. Flaskerud JH (2012) Cultural bias and Likert-type scales revisited. *Issues Ment Health Nurs* 33:130–132.
4. Johnson KB, Diego-Rosell P (2015) Assessing the cognitive validity of the Women's Empowerment in Agriculture Index instrument in the Haiti Multi-Sectoral Baseline Survey. *Surv Pract* 8:1–11.
5. Barrett HC, et al. (2016) Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proc Natl Acad Sci USA* 113:4688–4693.
6. Fessler DM, et al. (2015) Moral parochialism and contextual contingency across seven societies. *Proc Biol Sci* 282:20150907.
7. Cheung GW, Rensvold RB (2000) Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *J Cross Cult Psychol* 31:187–212.
8. Bernal H, Wooley S, Schensul JJ (1997) The challenge of using Likert-type scales with low-literate ethnic populations. *Nurs Res* 46:179–181.
9. Lee JW, Jones PS, Mineyama Y, Zhang XE (2002) Cultural differences in responses to a Likert scale. *Res Nurs Health* 25:295–306.
10. Kishor S, Johnson K (2004) *Profiling Domestic Violence: A Multi-Country Study* (ORC Macro, Calverton, MD).
11. Fincham FD, Cui M, Braithwaite S, Pasley K (2008) Attitudes toward intimate partner violence in dating relationships. *Psychol Assess* 20:260–269.
12. Fiedler K (2011) Voodoo correlations are everywhere—Not only in neuroscience. *Perspect Psychol Sci* 6:163–171.
13. Medin D, Bennis W, Chandler M (2010) Culture and the home-field disadvantage. *Perspect Psychol Sci* 5:708–713.
14. Zuilkowski SS, McCoy DC, Serpell R, Matafwali B, Fink G (2016) Dimensionality and the development of cognitive assessments for children in Sub-Saharan Africa. *J Cross Cult Psychol* 47:341–354.
15. Serpell R, Deregowski JB (1980) The skill of pictorial perception: An interpretation of cross-cultural evidence. *Int J Psychol* 15:145–180.
16. Medin D, Ojalehto B, Marin A, Bang M (2017) Systems of (non-) diversity. *Nat Hum Behav* 1:0088.
17. Ciborowski TJ (1979) Cross-cultural aspects of cognitive functioning: Culture and knowledge. *Perspectives in Cross-Cultural Psychology*, eds Marsella AJ, Tharp RG, Ciborowski TJ (Academic, New York), pp 101–116.
18. Cole M, Scribner S (1974) *Culture & Thought: A Psychological Introduction* (Wiley, New York).
19. Briggs CL (1986) *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research* (Cambridge Univ Press, New York).
20. Wober M (1969) Distinguishing centri-cultural from cross-cultural tests and research. *Percept Mot Skills* 28:488.
21. Greenfield PM (1997) You can't take it with you: Why ability assessments don't cross cultures. *Am Psychol* 52:1115–1124.
22. Mistry J, Rogoff B (1994) Remembering in cultural context. *Psychology and Culture*, eds Lonner W, Malpass R (Allyn and Bacon, Needham Heights, MA), pp 139–144.
23. Brislin RW (1970) Back-translation for cross-cultural research. *J Cross Cult Psychol* 1:185–216.
24. Mullen MR (1995) Diagnosing measurement equivalence in cross-national research. *J Int Bus Stud* 26:573–596.
25. Sechrest L, Fay TL, Zaidi SH (1972) Problems of translation in cross-cultural research. *J Cross Cult Psychol* 3:41–56.
26. Willis GB (2004) *Cognitive Interviewing: A Tool for Improving Questionnaire Design* (Sage Publications, Beverly Hills, CA).
27. Solano-Flores G (2006) Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teach Coll Rec* 108:2354–2379.
28. Delhey J, Newton K, Welzel C (2011) How general is trust in "most people"? Solving the radius of trust problem. *Am Sociol Rev* 76:786–807.
29. DeKlerk HM, Dada S, Alant E (2014) Children's identification of graphic symbols representing four basic emotions: Comparison of Afrikaans-speaking and Sepedi-speaking children. *J Commun Disord* 52:1–15.
30. Deregowski JB (1989) Real space and represented space: Cross-cultural perspectives. *Behav Brain Sci* 12:51–74.
31. Henrich J, Heine SJ, Norenzayan A (2010) Beyond WEIRD: Towards a broad-based behavioral science. *Behav Brain Sci* 33:111–135.
32. Chan AH, Ng AW (2010) Effects of sign characteristics and training methods on safety sign training effectiveness. *Ergonomics* 53:1325–1346.
33. Ou Y-K, Liu Y-C (2012) Effects of sign design features and training on comprehension of traffic signs in Taiwanese and Vietnamese user groups. *Int J Ind Ergon* 42:1–7.
34. Jahoda G (1978) Cross-cultural study of factors influencing orientation errors in the reproduction of Kohs-type figures. *Br J Psychol* 69:45–57.
35. Willats J (1997) *Art and Representation: New Principles in the Analysis of Pictures* (Princeton Univ Press, Princeton).
36. March B (1927) A note on perspective in Chinese painting. *China J* 7:69–72.
37. Otto H, Keller H (2014) *Different Faces of Attachment: Cultural Variations on a Universal Human Need* (Cambridge Univ Press, Cambridge, UK).
38. Mesman J, et al. (2018) Universality without uniformity: A culturally inclusive approach to sensitive responsiveness in infant caregiving. *Child Dev* 89:837–850.
39. Keller H, et al. (January 23, 2018) The myth of universal sensitive responsiveness: Comment on Mesman et al. (2017). *Child Dev*, 10.1111/cdev.13031.
40. Mesman J (January 23, 2018) Sense and sensitivity: A response to the commentary by Keller et al. (2018). *Child Dev*, 10.1111/cdev.13030.
41. Baumard N, Sperber D (2010) Weird people, yes, but also weird experiments. *Behav Brain Sci* 33:84–85.
42. Shweder RA (2010) Donald Campbell's doubt: Cultural difference or failure of communication? *Behav Brain Sci* 33:109–110.
43. Blalock HM (1982) *Conceptualization and Measurement in the Social Sciences* (Sage Publications, Beverly Hills, CA).
44. Johnson-Hanks J (2005) When the future decides: Uncertainty and intentional action in contemporary Cameroon. *Curr Anthropol* 46:363–385.
45. Irwin MH, Schafer GN, Feiden CP (1974) Emic and unfamiliar category sorting of Mano farmers and US undergraduates. *J Cross Cult Psychol* 5:407–423.
46. Jones B, Rachlin H (2006) Social discounting. *Psychol Sci* 17:283–286.
47. Aron A, Aron EN, Smollan D (1992) Inclusion of other in the self scale and the structure of interpersonal closeness. *J Pers Soc Psychol* 63:596–612.
48. Jones BA, Rachlin H (2009) Delay, probability, and social discounting in a public goods game. *J Exp Anal Behav* 91:61–73.
49. Goeree JK, McConnell MA, Mitchell T, Tromp T, Yariv L (2010) The 1/d law of giving. *Am Econ J Microecon* 2:183–203.
50. Strombach T, et al. (2015) Social discounting involves modulation of neural value signals by temporoparietal junction. *Proc Natl Acad Sci USA* 112:1619–1624.
51. Tiokhin L, Hackman J, Hruschka D (January 2018) Why replication is not enough: Insights from a cross-cultural study of social discounting, 10.17605/OSF.IO/F5S84.
52. Strombach T, et al. (2014) Charity begins at home: Cultural differences in social discounting and generosity. *J Behav Decis Making* 27:235–245.
53. Hackman J, Danvers A, Hruschka DJ (2015) Closeness is enough for friends, but not mates or kin: Mate and kinship premiums in India and US. *Evol Hum Behav* 36:137–145.
54. Pornpattananangkul N, Chowdhury A, Feng L, Yu R (April 5, 2017) Social discounting in the elderly: Senior citizens are good Samaritans to strangers. *J Gerontol B Psychol Sci Soc Sci*, 10.1093/geronb/gbx040.
55. Miller JG, Bersoff DM (1998) The role of liking in perceptions of the moral responsibility to help: A cultural perspective. *J Exp Soc Psychol* 34:443–469.
56. Pasick RJ, et al. (2009) Behavioral theory in a diverse society: Like a compass on Mars. *Health Educ Behav* 36(5 Suppl):11S–35S.
57. Vekaria KM, Brethel-Haurwitz KM, Cardinale EM, Stoycos SA, Marsh AA (2017) Social discounting and distance perceptions in costly altruism. *Nature Human Behaviour* 1:1–7.
58. Hruschka D, et al. (2014) Impartial institutions, pathogen stress and the expanding social network. *Hum Nat* 25:567–579.
59. Apicella CL, Marlowe FW, Fowler JH, Christakis NA (2012) Social networks and co-operation in hunter-gatherers. *Nature* 481:497–501.
60. Henrich J, et al. (2001) In search of homo economicus: Behavioral experiments in 15 small-scale societies. *Am Econ Rev* 91:73–78.
61. Hackman J, Munira S, Jasmin K, Hruschka D (2017) Revisiting psychological mechanisms in the anthropology of altruism. *Hum Nat* 28:76–91.
62. Green L, Myerson J, Lichtman D, Rosen S, Fry A (1996) Temporal discounting in choice between delayed rewards: The role of age and income. *Psychol Aging* 11:79–84.
63. Sharp C, et al. (2012) Social discounting and externalizing behavior problems in boys. *J Behav Decis Making* 25:239–247.
64. Burks SV, Carpenter JP, Goette L, Rustichini A (2009) Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proc Natl Acad Sci USA* 106:7745–7750.
65. Scribner S, Cole M (1981) *The Psychology of Literacy* (Harvard Univ Press, Cambridge, MA).
66. Olson DR (1994) Literacy and the making of the Western mind. *Functional Literacy: Theoretical Issues and Educational Implications*, ed Verhoeven L (John Benjamins, Amsterdam), Vol 1, pp 135–150.
67. Berland JC (1982) *No Five Fingers Are Alike: Cognitive Amplifiers in Social Context* (Harvard Univ Press, Cambridge, MA).
68. Serpell R (2011) Social responsibility as a dimension of intelligence, and as an educational goal: Insights from programmatic research in an African society. *Child Dev Perspect* 5:126–133.
69. Serpell R (2017) How the study of cognitive growth can benefit from a cultural lens. *Perspect Psychol Sci* 12:889–899.
70. Tourangeau R, Yan T (2007) Sensitive questions in surveys. *Psychol Bull* 133:859–883.
71. Morse JM (1991) Approaches to qualitative-quantitative methodological triangulation. *Nurs Res* 40:120–123.
72. Pasick RJ, Stewart SL, Bird JA, D'Onofrio CN (2001) Quality of data in multiethnic health surveys. *Public Health Rep* 116(Suppl 1):223–243.