Evolutionary Human Sciences (2025), **7**, e6, page 1 of 20
doi:10.1017/ehs.2024.48

METHODS PAPER

The cover of randomness: validating implicit methods for the study of sensitive topics

Charles Efferson and Sonja Vogt

Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland
Corresponding author: Charles Efferson; Email: charles.efferson@unil.ch

Abstract
We review the methods we developed to study female genital cutting in Sudan and sex-selective abortion in Armenia. These methods were untested at the time of our original research, and here we compare the distinct but overlapping approaches we used to validate our methods for each of the two countries. Additionally, we repeat a number of analyses, including those related to validation, with previously unpublished data from Sudan. All results replicate previous findings. Replicating previous results is encouraging, but we nonetheless argue that validation for Armenia is more convincing than for Sudan. Specifically, even if female genital cutting and the preferential abortion of females are equally sensitive as research topics, son bias is inherently easier to study than cutting because biological sex determination is a random process with no natural analogue in the case of cutting. This randomness provides a kind of cover for research participants who are son-biased but want to create the impression that they are not. This cover, in turn, allows the researcher to resolve any trade-off between methods that produce explicit granular data and methods that produce untraceable, highly aggregated data in favour of methods producing the explicit and granular.

Keywords: Female genital cutting; sex-selective abortion; implicit association test; harmful traditions

Social media summary: Validating implicit attitudes on sensitive topics works best when a random device obscures the choices people make.

1. Introduction

In 2011, we began what would turn out to be eight years of fieldwork on female genital cutting in Sudan (Efferson & Vogt, 2018; Efferson et al., 2015; Vogt et al., 2016, 2017) and sex-selective abortion in Armenia (Schief et al., 2021, 2024). For both projects, we worked closely with a diverse array of governmental organizations, non-governmental organizations, and inter-governmental organizations. We also collaborated extensively with local study coordinators, field assistants, writers, artists, actors, and film-makers. Most importantly for present purposes, we began both projects with nearly immobilizing worries about the practical challenges such sensitive research topics create in terms of collecting reliable data. To work around these challenges, we developed and eventually validated new methods tailor-made for each topic. Our primary task here is to review these methods and examine why we suspect we were in a better position to validate our approach in Armenia than we were in Sudan.

When we first went to Sudan, the country already had a long history of programming efforts intended to reduce or eliminate female genital cutting (Boddy, 2007; Gruenbaum, 2001). This history

had presumably created some poorly understood scope for biases related to social desirability or backlash (De Cao & Lutz, 2018; Gibson et al., 2018; Merry, 2006; Shell-Duncan, 2016; Thomas, 2000). Any mix of these two different types of bias poses a serious problem for empirical researchers, us included (Cloward, 2016; Efferson et al., 2020, 2023; Lawson et al., 2024; Shell-Duncan & Hernlund, 2000).

Armenia, in contrast, did not have a long history of programming to reduce or eliminate the preferential abortion of female foetuses. In fact, Armenians only started relying on sex-selective abortions in the early 1990s (Schief et al., 2021, 2024). Nonetheless, the pioneering research of Duthé et al. (2012), Guilmoto and Duthé (2013), and Guilmoto (2013) showed that prenatal sex selection in Armenia had been producing increasingly distorted sex ratios at birth since the collapse of the Soviet Union. When we first went to Armenia in 2013, the recent publications of Duthé, Guilmoto, and colleagues ensured that son bias and sex-selective abortion were newly prominent and in our experience sometimes acrimonious topics of public discourse.

To meet the challenges of working on female genital cutting and sex-selective abortion, we developed some entirely new methods, and we also developed bespoke implementations of existing methods. As a result, nearly every method we used was in some way untested. This created a validation problem. We were convinced that traditional methods like the questionnaire items used for the Demographic and Health Survey (Shell-Duncan, 2016), for example, were rife with potential for bias, and so we largely shunned such approaches. Shunning traditional methods, however, did not mean we had the luxury of studying and validating our own methods in minute detail before beginning our primary research. The inevitable constraints when running large projects in the field ensured that we did not.

Our approaches to this validation problem were in some ways similar in Sudan and Armenia and in some ways different. In both cases, our most important validation exercises centred on correlating results from implicit association tests with the behaviours of interest, namely cutting in Sudan and family planning based on a preference for sons in Armenia. In spite of this commonality, measuring all attitudes and behaviours related to cutting was challenging, whereas measuring some attitudes and behaviours related to son bias was not. This discrepancy lies at the root of why we think validation in Armenia was more convincing than in Sudan.

That said, we do not necessarily wish we had designed the methods we used in Sudan to be more like the methods we used in Armenia. Instead, in both countries we faced important trade-offs between what we will call 'precision' and 'discretion'. By 'precision', we mean any methods that involve asking individual decision-makers explicit questions about themselves on the sensitive topic at hand. By 'discretion', we mean any collection of methods that deliberately avoids such questions. If biases like social desirability bias were not a concern, any researcher would presumably prefer precision and the explicit disaggregated data it produces. If biased responses are a concern, however, the researcher may try to reduce bias by shifting towards discretion and simply accepting the costs this brings. As we argue below, because the sex of one's offspring is to some extent random, we leaned towards precision in Armenia. As researchers we cannot definitively infer son bias by observing a family, for example, with three sons and no daughters because such an outcome can occur simply by chance. For the same reason, son-biased parents with three sons and no daughters need not hide the sex of their children to please the researcher, the government, or some locally active development agency. The randomness of sex determination provides them with cover for their son bias. Female genital cutting has no analogous source of randomness, and so in Sudan we leaned towards discretion.

To examine the discrepancy between validation in Sudan versus Armenia, we first summarize the bespoke implicit association tests we developed for each of the two countries. We then summarize the novel method we developed for estimating cutting rates by community in Sudan. We go on to repeat our basic validation exercise from Sudan (Efferson et al., 2015) with previously unpublished data from 120 communities. Along the way, we interpret specific patterns in the data in light of a prominent cultural evolutionary model of female genital cutting, and we arrive at the same conclusions we arrived at previously (Efferson et al., 2015). Although arriving at the same conclusions is encouraging,

we nonetheless suspect our approach to validation in Sudan is fundamentally limited compared to our approach in Armenia. We discuss why to isolate some of the key questions researchers working on culturally sensitive topics might consider when hinging their empirical strategy on a collection of new and largely untested methods.

2. Implicit association tests

Implicit association tests are computerized psychometric tests based on relative response times under different categorization schemes (Greenwald et al., 1998; Nosek et al., 2007). We can think of an implicit association test, loosely speaking, as an attitudinal measure designed to be less susceptible to social desirability bias than traditional explicit measures like questionnaires. To illustrate how such a test might work, imagine we are interested in how people feel about flowers versus insects (Baron & Banaji, 2006). With insects, we do not mean, for example, Monarch butterflies; we mean bugs. Further imagine that many people are in fact disgusted by insects, but for whatever reason they believe they should hide their disgust from researchers. Thus, if you simply ask people, they tend to say they are just fine with both insects and flowers. To work around this problem, an implicit association test does not simply ask people; it takes an entirely different approach.

One way to develop such a test for this example would be to assemble a collection of flower images and a matched collection of insect images. Further assemble a collection of audio recordings of positive words (e.g. beautiful, charming) and a matched collection of audio recordings of negative words (e.g. ugly, disgusting). The images are target stimuli in the sense that we somehow want to know how people evaluate the content the images represent, namely insects and flowers. The words are the valenced stimuli. We do not question the meaning of these stimuli. Rather, we assume the positive words are positive, and the negative words are negative. The question is, how do people associate the positive and negative words, which have a valence we assume, with the images of flowers and insects, which have a valence we want to infer via association?

To answer this question, the implicit association test randomly presents different stimuli and measures how fast a given participant categorizes each stimulus. Ignoring a number of technical details (Nosek et al., 2007), the test compares response times under two crucial categorization schemes.

- Under the **insect-negative, flower-positive** scheme, a participant might receive the following instructions, 'If you see an image of an insect or if you hear a negative word, press the D key on your keyboard as fast as possible. If you see an image of flowers or if you hear a positive word, press the K key on your keyboard as fast as possible.'
- Under the **flower-negative, insect-positive** scheme, the participant would receive different instructions. 'If you see an image of flowers or if you hear a negative word, press the D key on your keyboard as fast as possible. If you see an image of an insect or if you hear a positive word, press the K key on your keyboard as fast as possible.'

Whatever the precise wording, each participant categorizes many stimuli under each of the two schemes. The key comparison is thus a within-subject comparison of response times under two opposing rules for categorizing stimuli. If a given subject has negative associations with insects relative to flowers, the participant should be fastest on average under the insect-negative, flower-positive scheme. If the subject has relatively negative associations with flowers relative to insects, response times should exhibit the opposite pattern.

A common scoring algorithm converts all response times into a single D_i score for individual i (Greenwald et al., 2003; Nosek et al., 2007). The D_i score is a normalized measure of relative response times. It can take any value in the interval $[-2, 2]$, and it is designed, among other objectives, to minimize the potential for participants to manipulate their scores. The researcher can calculate D_i values in two different ways. On the one hand, the researcher can calculate so that $D_i > 0$ would

Figure 1. One set of stimuli used for the implicit association test in Sudan. The pictures are of a specific girl, distinguished by the absence of pigtails, in a dress made from firka cloth. Firka cloth is associated with the cutting ceremony and is thus a strong mnemonic device reminding participants that this girl is cut. All nuisance variables were counterbalanced across participants (Vogt et al., 2016). For example, roughly half of participants completed a version of the implicit association test in which the girl shown here, namely the girl without pigtails, was not cut and wearing a dress made from the saleema cloth (see Figure 2).

Figure 2. Another set of stimuli used for the implicit association test in Sudan. The pictures are of a specific girl, distinguished by the presence of pigtails, in a dress made from cloth with the saleema graphic design. The saleema campaign is a national campaign promoting the abandonment of cutting, and thus this graphic design is a strong mnemonic device reminding participants that this girl is not cut. All nuisance variables were counterbalanced across participants (Vogt et al., 2016). For example, roughly half of participants completed a version of the implicit association test in which the girl shown here, namely the girl with pigtails, was cut and wearing a dress made from the firka cloth (see Figure 1).

indicate that i is more positive about insects than flowers, whereas $D_i < 0$ would indicate the opposite. On the other hand, the researcher can calculate so that $D_i > 0$ would indicate that i is more positive about flowers than insects, whereas $D_i < 0$ would indicate the opposite. The two approaches are equivalent; one just needs to know which approach has been used. In either case, $D_i = 0$ means that response times are the same on average under the two categorization schemes, which is the analogue of indifference from the perspective of the implicit association test. Indifference, as a result, is not an arbitrary point on the scale. Rather, indifference implies a specific score that separates, for everyone in the same way, anti-insect and pro-flower attitudes from anti-flower and pro-insect attitudes.

In Sudan, we developed one implicit association test based on drawings suggesting a girl is cut and drawings suggesting a girl is uncut (Figures 1 and 2). We collaborated closely with a local artist to generate target stimuli that fulfilled a number of key criteria. The online supplements for Efferson et al. (2015) and Vogt et al. (2016) provide extensive details, but here we would simply like to highlight one point. To represent a cut girl and an uncut girl in a compelling and memorable way for our participants, we relied on the fabrics used for their dresses. The cut girl was shown in a dress made from firka cloth, which is closely associated with the cutting ceremony in the region where we worked in Sudan (Figure 1). The uncut girl, in contrast, was shown in a dress made from a fabric with the saleema graphic design (Figure 2). This graphic design is closely associated with a large-scale campaign promoting the abandonment of female genital cutting in Sudan.

Figure 3. One set of stimuli used for the implicit association tests in Armenia. The pictures are of a specific set of parents who have daughters but no sons. All nuisance variables were counterbalanced across participants (Schief et al., 2021). For example, roughly half of the participants completed a version of an implicit association test in which the parents shown in Figure 4 had daughters but no sons.

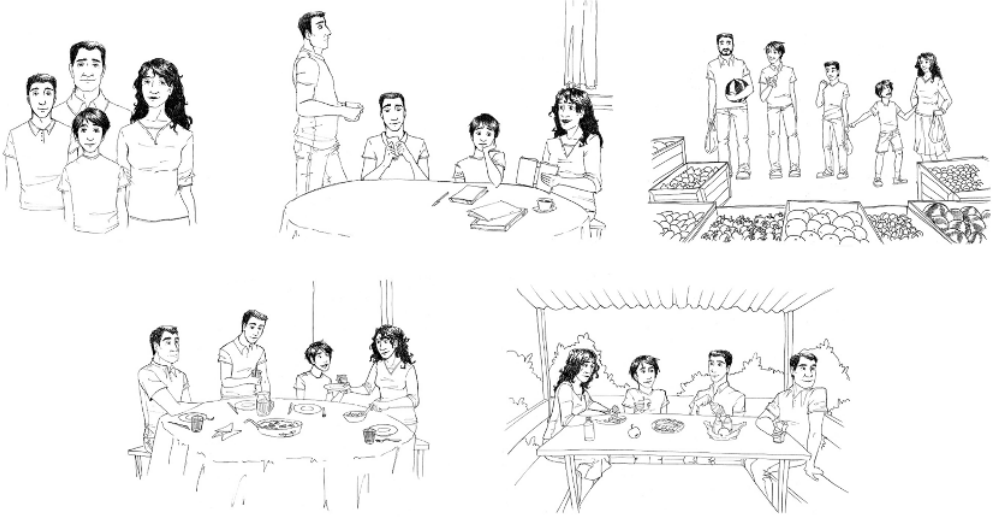

Figure 4. Another set of stimuli used for the implicit association tests in Armenia. The pictures are of a specific set of parents who have sons but no daughters. All nuisance variables were counterbalanced across participants (Schief et al., 2021). For example, roughly half of the participants completed a version of an implicit association test in which the parents shown in Figure 3 had sons but no daughters.

In Armenia, we developed two extremely similar implicit association tests based on drawings of a family with only daughters and drawings of a family with only sons (Figures 3 and 4). We again collaborated closely with a local artist to produce the images. Details are available in the online supplement to Schief et al. (2021). In practice, the two implicit association tests in Armenia produced nearly identical results, and we pooled them for all analyses in Schief et al. (2021).

We designed the implicit association tests in both Sudan and Armenia with the same principles in mind. In particular, the target stimuli were images of people in everyday scenes (Figures 1–4), and these images were perfectly matched in terms of the variable of interest and perfectly counterbalanced in terms of nuisance variables. To illustrate, in Sudan the variable of interest was cut versus uncut. Nuisance variables consisted of which specific girl was cut, and the order in which the participant experienced the two key categorization schemes. The variable of interest was matched in the sense that, if one of the target stimuli was an image of the cut girl feeding chickens (Figure 1), another target stimulus was an extremely similar image of the uncut girl feeding chickens (Figure 2). With this kind of matching, we isolated the effect of cut versus uncut.

However, for a given participant the cut girl and uncut girl needed to be distinct individuals. We thus worked with images of a girl without pigtails and images of a girl with pigtails, but this distinction was of no interest to us. Analogously, a given participant could only experience the two key categorization schemes in one of two orders, and any order effects were also of no interest to us. Because these two variables were of no interest, we counterbalanced them across participants. Approximately 25% of participants worked with a version of the test in which the girl without pigtails was cut and the uncut-negative, cut-positive scheme came first. Approximately 25% worked with a version in which the girl without pigtails was cut, and the cut-negative, uncut-positive scheme came first. For the remaining 50% of participants, we counterbalanced the orderings in the same way, but in this case the girl with pigtails was cut, and the girl without pigtails was not. Our approach in Armenia was precisely analogous (Figures 3 and 4). The variable of interest was a family with only daughters versus a family with only sons. The nuisance variables were which specific set of parents had only daughters and the order of the two key categorization schemes.

A potential problem with implicit association tests is interpretation. Some results have put the test forward as an attitudinal measure, specifically a measure of implicit associations, that is less susceptible to social desirability bias than traditional explicit methods like questionnaires (Greenwald et al., 1998; Kim, 2003). For this reason, implicit association tests have been widely used to study, for example, racial bias (Morehouse & Banaji, 2024). The intuition is that everyone may know the correct answers when a university researcher comes knocking with questions about race. Give an ostensibly enlightened participant an implicit association test, however, and the participant will betray the racist within.

Nonetheless, the fact remains that an implicit association test places participants in an extremely artificial setting and measures the milliseconds required to categorize stimuli. The link between these milliseconds, covert racist attitudes, and the racist behaviours that actually affect people's daily lives is not always obvious. Although our implicit associations tests were not about race, analogous forms of ambiguity were especially important in our case because we had little if any interest in the psychometric subtleties of our tests. Our interest was – and remains – only about measured attitudes in relation to the behaviours that actually matter. For this reason, in both Sudan and Armenia we estimated correlations between D_i scores produced by implicit association tests and parental decision-making related to cutting and son bias.

3. Behaviour and validation in Sudan

The traditional approach to collecting data on cutting (Shell-Duncan, 2016) is simply to ask parents about cutting their daughters. Rightly or wrongly, we were convinced that this approach is unreliable, and we spent several months looking for new alternatives in Sudan. We considered, discussed, and rejected many proposals. We were on the verge of giving up when a woman with whom we had been working closely revealed that girls have henna applied to their feet the day they are cut, and that in our study region being recently cut was the only reason a young girl would have henna on her feet. Interestingly, the woman was not hiding this information previously. She just never thought to mention it, perhaps because she assumed we would have asked if we were interested. We never asked,

Table 1. Comparisons between different definitions of cutting

	NC (2)	C (2)		NC (3)	C (3)		NC (3)	C (3)
NC (1)	3460	952	NC (1)	2686	1635	NC (2)	2139	1247
C (1)	0	1207	C (1)	0	1207	C (2)	544	1594

Each comparison shows the counts of girls coded as not cut (NC) or cut (C) under pairwise combinations of Definition 1 (1), Definition 2 (2) and Definition 3 (3). Because the definitions vary in terms of the raw data they use and how they use these data, the total counts are not the same across the three comparisons. The exact computations are available in the code provided as supplementary materials. Because cut under Definition 1 implies cut under Definition 2, the two definitions are not independent, which is why one cell has a zero. Similarly, cut under Definition 1 also implies cut under Definition 3, and so these two definitions are also not independent. For this reason, when comparing Definition 1 to either Definition 2 or Definition 3, we do not test for statistical independence. When comparing Definitions 2 and 3 to each other, we can safely reject the hypothesis that they are independent (Fisher's exact test, odds ratio of 5.024 with a 99% confidence interval of $[4.289, 5.897]$, two-sided $p < 2.2 \times 10^{-16}$).

of course, because we were clueless about what exactly needed asking. In any case, once henna was under discussion, it quickly became the basis for a new approach to measuring cutting behaviour.

At the time, the state of Gezira sponsored health exams for children entering primary school. We added two steps to these exams. First, a team of photographers photographed the feet of girls while measuring the girls' shoe sizes as one of the activities associated with the health exam. Second, the physicians administering the actual health checks asked each girl if she had been 'purified', which is a literal translation of the term used for cutting in this part of Sudan. The supplement to Efferson et al. (2015) provides additional details.

We received two kinds of anonymized data. For each girl, we received a photograph of her foot with a sticker that included a community identifier and a unique number for the girl herself. We also received the answer each girl provided when asked by the physician if she had been purified. We did not know who the girls were or who their parents were. The identifiers only allowed us to identify in which geographically defined community a given girl lived, which is what we mean by 'community', and to match the photograph of her foot to her response to the question about being purified. In Switzerland, we then hired two coders, blind to all details of the study, to view the photographs and say whether henna was present. We hired a third coder, also blind to all details, to digitize the data from the medical exams.

We coded these data in three different ways (Table 1) to estimate cutting rates by community (Efferson et al., 2015).

- **Definition 1.** If both coders stated that henna was present on the foot, we coded the girl as cut. If at least one coder stated that henna was not present, we coded her as uncut.
- **Definition 2.** If at least one coder stated that henna was present, we coded the girl as cut. If both coders stated that henna was not present, we coded her as uncut.
- **Definition 3.** If the girl was coded as cut under Definition 1 or if the girl stated that she had been purified, we coded the girl as cut. If the girl was coded as uncut under Definition 1 and she stated that she had not been purified, we coded her as uncut.

We used the distribution of cutting rates by community to examine a prominent model of why parents cut their daughters and, by extension, how practitioners might design programmes to convince parents to stop.

Specifically, the model posits that families face incentives to coordinate their choices, and the coordination game families are playing generates three steady states (Bellemare et al., 2015; Camilotti, 2016; Efferson et al., 2015, 2020, 2023; Mackie, 1996; Novak, 2020; Platteau et al., 2018; Shell-Duncan et al., 2011). One locally stable steady state has all families cutting. Another locally stable steady state has all families not cutting. Finally, an unstable steady state, often referred to as a 'tipping point', involves a mix, with some families cutting and other families not cutting. Because the tipping point is unstable, we do not expect to observe it. Instead, it represents the boundary between two opposing cultural evolutionary regimes. If the proportion cutting is above this boundary, cutting increases until

all families cut. If the proportion cutting is below the boundary, cutting decreases until no families cut.

Under this model, if the practitioner observes a cutting society, she assumes the society is trapped in the cutting equilibrium. Families cut because this choice, given the social landscape, is the best choice they can make for their daughters. However, if a sufficiently large group of families abandon the practice together, this collective deviation can push the society from one side of the unstable steady state to the other. Coordination incentives switch from favouring the decision to cut to favouring the decision not to cut, which induces all remaining families who cut to abandon the practice on their own. The model predicts this dramatic turn of events precisely because everyone not cutting and everyone cutting are both locally stable. If a population gets close enough to either equilibrium, whatever the reason, coordination incentives push the population the rest of the way.

A crucial question, however, concerns whether the model actually captures how real incentives and associated cultural evolutionary dynamics work. One straightforward check is to look for evidence of two stable equilibria at each of the two extremes. In a metapopulation in equilibrium with variation in attitudes and practices, the model implies that the variation should mainly be between groups, not within groups. Attitudes and practices should be relatively homogeneous at local scales even though they are heterogeneous globally (Young, 2015). Many groups with a mix of cutting and non-cutting families living together would be difficult to reconcile with the notion that coordination incentives push any group, depending on its current state, towards either the cutting equilibrium or the non-cutting equilibrium. More practically, such a mix would be difficult to reconcile with the notion that a practitioner can intervene to tip a cutting group to the non-cutting equilibrium.

We checked for local homogeneity and global heterogeneity, especially local homogeneity, in an initial sample of 45 communities in Sudan using the henna method described above (Efferson et al., 2015). We found that, in direct contrast to the predictions of the coordination game model, most communities seemed to have a mix of cutting and non-cutting families. Communities were, in short, locally heterogeneous. This finding called into question any simple version of the coordination game model of cutting (cf. Efferson et al., 2020; Platteau et al., 2018). By extension, it called into question the potential for a practitioner to rely on a circumscribed intervention to tip a community from the cutting equilibrium to the non-cutting equilibrium.

As an important caveat, however, we estimated cutting rates as the proportion of girls cut in geographically defined communities. Geographically defined communities might have been the wrong social unit. If families faced incentives to coordinate with each other in some social unit unrelated to community boundaries, the predicted pattern of local homogeneity and global heterogeneity might have held in spite of our results. Under this explanation, we simply estimated the proportion cutting in inappropriate social groups. Local was not a geographic concept. Some other notion of social proximity was at work.

This possibility represents a challenge to our interpretation of local heterogeneity. The best way to meet such a challenge would be to couple an analysis that measures cutting at the individual level with social network data that identify the social groups relevant for coordination (Myers et al., 2024). We did not have the latter type of data. We thus cannot rule out the possibility that the basic model of coordination and tipping were appropriate, but geographically defined communities were the wrong way to think about groups. We can clarify, however, what exactly this latter possibility would imply.

If we are to reconcile the standard prediction of local homogeneity and global heterogeneity with our results by saying geographic communities were the wrong unit, the implication is that nearly all geographic communities in our sample were subdivided into two or more subgroups. Moreover, in each geographic community, at least one of the subgroups was in the cutting equilibrium, and at least one was in the non-cutting equilibrium. Geographic communities were not just the wrong way to group people; they were about as far as possible from the correct way to group people. Although not impossible, this explanation seems a touch unlikely. In addition, anecdotal evidence, the overwhelming predominance of endogamy, and modelling results suggest that the geographically

defined community was, at least approximately, the correct social unit for our study site (Efferson et al., 2015).

The fact remains, however, that we could not and cannot rule out the possibility that we estimated cutting rates for inappropriate social groupings. For this reason, we used our implicit association test to examine the question in a different way. Imagine that all the families were grouped into social units, and each social unit was coordinating on one of the two pure-strategy equilibria of the game. Thus, whatever the groups were, each group was unambiguously either a cutting or a non-cutting group. Further assume that the people in each group had internalized the values associated with their group's respective equilibrium. People in cutting groups were pro-cutting. People in non-cutting groups were anti-cutting. If our implicit association test captured these attitudes, we expected a bimodal distribution of D_i scores. Specifically, because $D_i = 0$ is effectively like indifference under the test, we should have seen one mode on one side of $D_i = 0$ for the pro-cutting participants in cutting groups and another mode on the other side of $D_i = 0$ for the anti-cutting participants in non-cutting groups. Crucially, this reasoning did not require us to know what a group was. It simply required us to assume that individuals had internalized the value systems associated with their groups, however these groups were defined and regardless of whether a given group was pro-cutting or anti-cutting.

We found that the distribution of D_i scores was unequivocally unimodal, which means we found no evidence for two identifiable value systems (Efferson et al., 2015). All in all, when relying on photographs of girls' feet and associated medical exams, we found no indication that the 45 communities in our sample could be readily categorized as either cutting communities or non-cutting communities. When relying on implicit association tests among randomly selected adults in the same 45 communities, we found no indication of two distinct value systems, one pro-cutting and the other anti-cutting (Efferson et al., 2015). Both of these results were somehow inconsistent with the idea that coordination incentives both sustain cutting and create scope for social tipping as a dramatic path to abandonment of the practice.

That said, our results were based on new and unproven methods. Our henna feet method was entirely novel. Implicit association tests had an established if controversial history (Nosek et al., 2007), but our specific test was of course entirely new. Apart from some pre-testing to improve the instructions and protocols used when administering the test, we were not able to validate our implicit association test in detail before we began our primary research on the decision-making mechanisms underlying cutting. Instead, we relied on a kind of expost validation exercise. We correlated mean D_i scores, averaged over participants in a community, with estimated cutting rates by community. We found a robust and highly significant correlation between mean D_i scores and cutting rates under Definition 3, but only under Definition 3 (Efferson et al., 2015).

Figures 5–7 repeat our analyses with new data different from those used in Efferson et al. (2015). Specifically, the data related to henna feet and medical exams have not been published before. The implicit association data are from the baseline data collection for the randomized controlled trial in Vogt et al. (2016). The henna feet data collection did not take place in exactly the same communities as the randomized controlled trial, but considerable overlap exists. Altogether, we have henna feet data and implicit association data from 120 communities in the state of Gezira, Sudan. This is the sample of communities we are using for Figures 5–7.

Our new results are essentially identical to our previous results (Efferson et al., 2015; Vogt et al., 2016). First, regardless of which definition we use to code a girl as cut, we find that most communities have a non-degenerate mix of cut and uncut girls (Figure 5). Second, our distribution of D_i scores for the same set of 120 communities is clearly unimodal (Figure 6), and in this sense we find no evidence for two identifiable value systems. Third, implicit association scores are on average slightly positive. Under our coding, this means implicit attitudes were slightly more favourable towards the uncut girl than towards the cut girl. Finally, across the 120 communities in the sample, the cutting rates and mean implicit association scores are correlated in the expected direction, but only using Definition 3. All of these results replicate earlier findings (Efferson et al., 2015; Vogt et al., 2016). Because we again

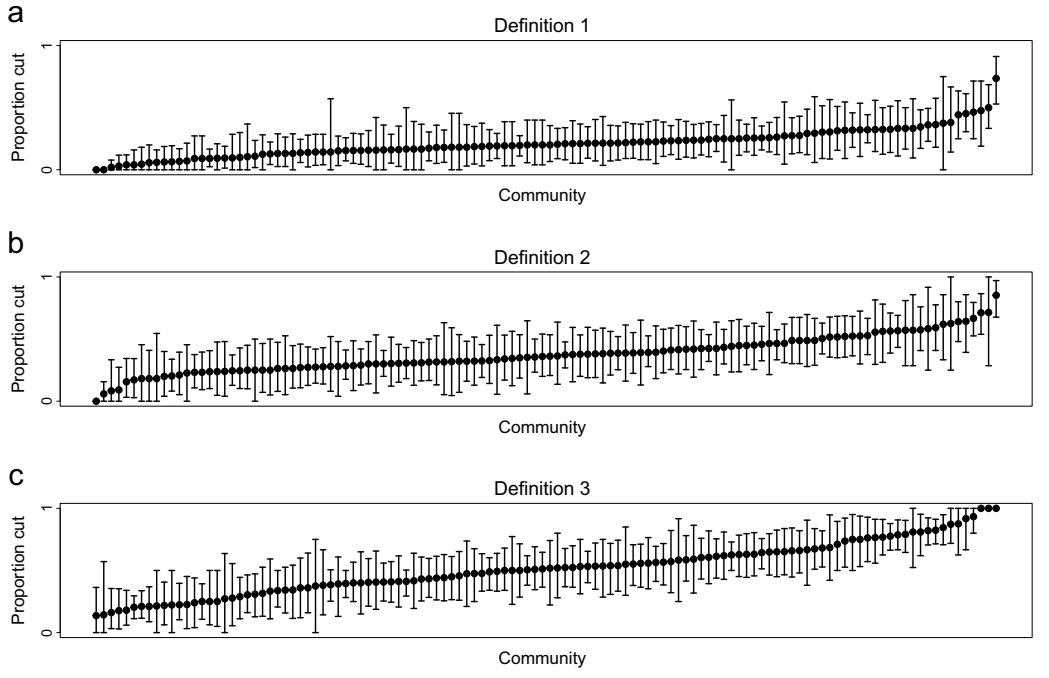

Figure 5. Estimated proportion cutting by community for 120 communities in Sudan. To estimate the cutting rates, we used each of the three definitions explained in the main text (**a**–**c**). Communities are ordered based on estimated proportions cutting from lowest to highest. Error bars denote 99% confidence intervals based on non-parametric bootstrapping.

work with geographically defined communities, the uncertainty about whether this is the right social grouping remains a concern.

4. Explicit attitudes, behaviour, and validation in Armenia

In addition to our implicit association tests in Armenia, we also collected two types of explicit data, one about fertility choices and the other about preferences for daughters versus sons (Schief et al., 2021). In terms of explicit data on fertility choices, we elicited the full reproductive histories of the mothers in our sample. In particular, we collected detailed data about children born, children who died after birth, and abortions. Of the 2858 children reported born, 2780 of them were reported alive at the time of data collection. To reduce possible bias, we implemented several protocols to maximize participant anonymity in a way that was salient and transparent to our participants. In spite of our efforts, however, we still found clear evidence that mothers under-reported sex-selective abortions. Specifically, self-reported abortions were common in our sample, which suggests that abortions per se were not an especially sensitive topic for our participants (Schief et al., 2021). However, self-reported abortions conditional on the woman expecting a girl were well below the number required to account for the missing girls in our sample (Schief et al., 2021; Figure 2), and this result suggests that sex-selective abortions were a sensitive topic. For this reason, we assumed that data about children born were accurate, and we used these data to examine son bias in ways that did not depend on mothers explicitly revealing the selective abortion of females.

Here we rely heavily on the assumption that mothers provided accurate data on children born. This assumption is the basis for our claim in the Introduction that some ways of measuring son bias in Armenia were not especially challenging. We would, however, like to be explicit that we are making an assumption. We do not have any independent evidence to justify it. We simply assume that most

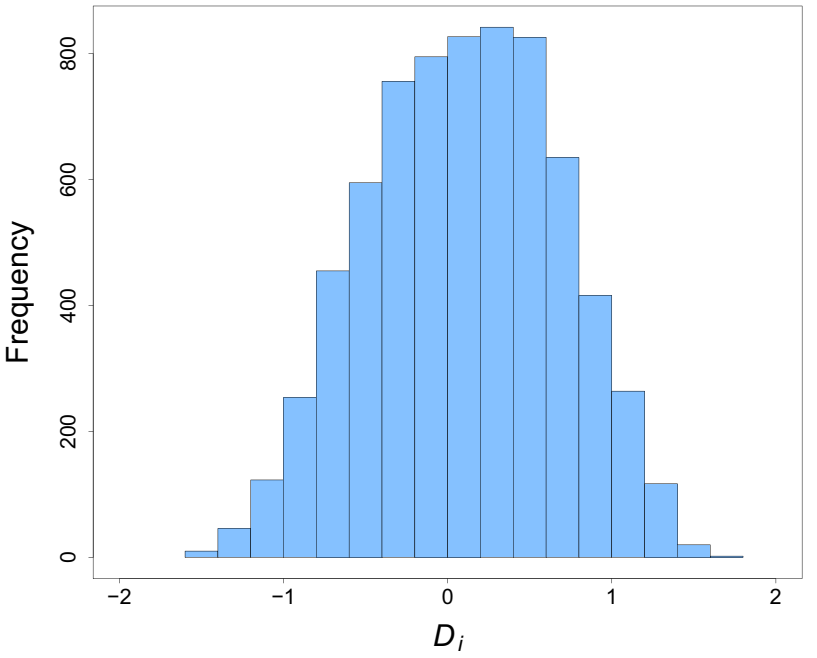

Figure 6. The distribution of D_i scores for 6983 participants in 120 communities in Sudan. Using Hartigans' Dip Test, we cannot reject the null hypothesis that the distribution is unimodal ($p = 0.9843$). The mean D_i score is 0.099, which is significantly positive (Wilcoxon signed rank test, two-sided $p < 2.2 \times 10^{-16}$).

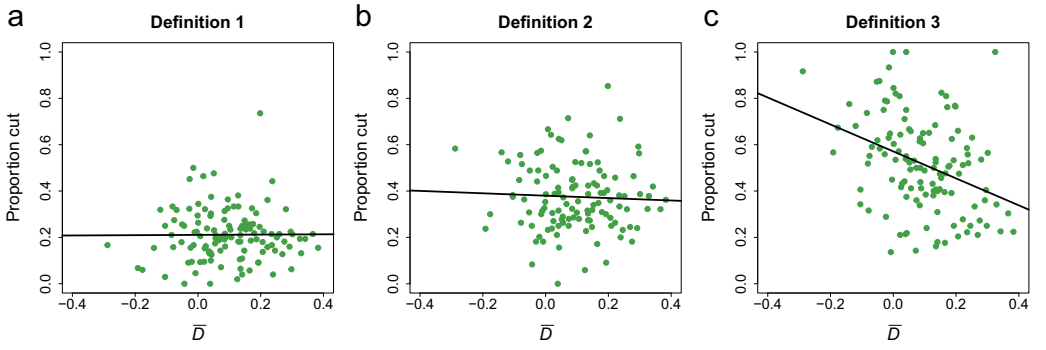

Figure 7. Correlations between average scores on the implicit association test and estimated cutting rates for 120 communities in Sudan. For each of the three definitions of cutting (**a**–**c**) explained in the main text, the plots show the relation between average implicit association scores by community and the estimated proportions of families cutting by community. Under Definition 1 (**a**), the Pearson correlation between the two aggregate-level variables is 0.007, and the 99% confidence interval based on a two-dimensional non-parametric bootstrap (Efferson et al., 2015) is $[-0.156, 0.161]$. Under Definition 2 (**b**), the correlation is −0.045, and the 99% confidence interval is $[-0.188, 0.119]$. Under Definition 3 (**c**), the correlation is −0.356, and the 99% confidence interval is $[-0.415, -0.153]$.

mothers did not mis-report children born, especially if nearly all children born were likely alive at the time of data collection, and especially if the mother recorded these data under conditions of maximum anonymity (Schief et al., 2021). If the assumption is wrong, then our claims about validation in Armenia being more convincing than in Sudan may also be wrong. As explained below, we relied on

data about children born to check for correlations between implicit association test scores, explicit preference measures, and observed fertility outcomes.

In terms of explicit data on preferences for daughters versus sons, we developed the methods of Jayachandran (2017) to produce a metric of son bias based on explicit questioning (Schief et al., 2021). Like Jayachandran (2017), however, our explicit questions were notably different from traditional approaches. The traditional approach, as in the Demographic and Health Survey, for example (Arnold, 1997), first asks a participant to go back to when she did not have any children. The facilitator then asks the participant how many children she would have in her entire life if she could choose exactly the number of children she wants. Finally, conditional on the response to the first question, the facilitator asks the participant how many of these children the participant would like to be boys, how many girls, and for how many the sex would not matter.

Apart from social desirability bias, this kind of approach involves considerable scope for at least two additional biases. First, the participant may answer the questions in a way that rationalizes the family she has, which means the participant anchors her responses on the number and gender composition of her actual children. We call this 'rationalization' bias. Second, the participant is forced to answer the gender composition question conditional on her response to the question about the ideal family size, and this requirement means that some participants cannot express certain types of preferences they may actually have. We call this 'conditioning' bias. To illustrate, imagine a participant who is son-biased, but only weakly so, and this participant answers the first question with two children. This participant can only answer the second question in a way that reflects no gender bias or extreme son bias. Expressing weak son bias is not possible. Conditioning bias could be an especially serious problem if comparing preferences for sons versus daughters between groups of people who also have different notions about ideal family sizes.

In an effort to attenuate or eliminate these biases, we adopted the following strategy (Schief et al., 2021). To minimize rationalization bias, we did not ask participants to imagine themselves before having children. Instead, we asked each participant to imagine the future family of her youngest child, a family that did not exist at the time of data collection but was presumably important to the participant. To minimize conditioning bias, we conditioned on family size for the participant. Specifically, we asked each participant to imagine a future in which her youngest child had one child. We also posed analogous scenarios imagining a future family with two children, three children, and four children. For each of these four possibilities, we asked the participant to specify the ideal composition of the hypothetical family. Because son bias can manifest itself not only in terms of the proportion of boys versus girls, but also in terms of the sequence of births by gender, we asked the participant for the ideal sequence of boys and girls conditional on the family size under consideration.

To avoid average order effects, we randomized the ordering of the four family sizes across participants, and we also separated the four questions from each other with intervening questions. After a participant had provided the ideal sequence of boys and girls for each of the four family sizes, we presented four ideal families to the participant and asked her to choose which would be best for the future family of her youngest child. With these data, we calculated an explicit index of son bias for each participant. Intuitively, the index was designed to capture two basic ways in which a son bias might manifest itself (Schief et al., 2021). Namely, a son-biased participant might think that having sons before daughters is best because having sons early quickly removes the worry that one might not have a son at all. This form of son bias is about the ordering of daughters versus sons. A son-biased participant might also think that most or all of one's children should be sons. This form of son bias is about the number of daughters versus sons. Our index could take values from zero to nine, with large numbers reflecting strong son bias (Schief et al., 2021).

To validate our measures, we estimated both the internal consistency of our preference measures and the external consistency (Schief et al., 2021). In terms of internal consistency, we correlated our implicit association test scores with our explicit index values. The two preference measures were

statistically significantly correlated in the right direction. So long as our implicit association test minimized social desirability bias, as the test and scoring algorithm are designed to do, the correlation we found further suggests that our explicit measure of son bias was not completely contaminated by social desirability bias.

In terms of external consistency, we correlated both our implicit and explicit preference measures with observed fertility outcomes, specifically the proportion of sons versus daughters a participant had and the probability the participant's youngest child was a boy (Schief et al., 2021). Although we know that the mothers in our sample under-reported sex-selective abortions, son-biased participants preferentially aborting female foetuses will of course tend to have a high proportion of sons. This is where our assumption about the accuracy of data on children born becomes important. Even with under-reporting of sex-selective abortions, accurate data on children born allow the researcher to check how observed fertility outcomes relate to both implicit and explicit preference measures. We found statistically significant correlations between the proportion of sons and both our implicit and explicit measures of son bias. Interestingly, we were able to show that causation was almost certainly running in both directions. Having daughters made participants less son-biased, and son bias made participants have a relatively high number of sons.

Even without sex-selective abortion and distorted sex ratios, son bias can manifest itself via differential stopping rules. An example of a son-biased stopping rule would be the following. If you have the desired number of children, and if at least one of them is a son, stop having children. If you have the desired number of children, and if none of them are sons, continue having children even though you would otherwise stop. Son-biased stopping rules do not affect the probability that any given child is a son, and by extension they do not affect the sex ratio at birth in the population. They do, however, change the probabilities of observing different family compositions. Families in which the youngest child is a son appear more often than expected. This is true because son-biased families that start off with multiple daughters are exactly those families forced to rely on a son-biased stopping rule. Such families tend to have more children than they want, and they stop once they have a son.

Before Armenians began relying on sex-selective abortion, ethnic Armenians living in Russia in Soviet times relied on son-biased stopping rules (Schief et al., 2024). Because son bias could have also manifested itself in this way when we were working in contemporary Armenia, we also correlated our implicit and explicit preference measures with a dummy variable coding whether the youngest child was a son. Similar to our analysis based on the number of daughters versus sons, we again found statistically significant correlations in the right direction (Schief et al., 2021). Our assumption that data about children born were accurate was, of course, also relevant for the validation exercise based on stopping rules.

5. Discussion of two paths to validation

As explained in the Introduction, our paths to validation shared broad similarities in Sudan and Armenia. In both countries, we analysed the relationship between implicit attitudinal measures, which we designed to limit bias, and behavioural data. In both countries, we found correlations in the correct direction. These results are encouraging, and they give us confidence that our efforts to develop new methods and reliable protocols were worth the trouble. Nonetheless, as explained in the Introduction, we ourselves find validation in Armenia more convincing than in Sudan, largely because we favoured precision in Armenia but discretion in Sudan.

Indeed, we were extremely, perhaps overly cautious in Sudan. We were so preoccupied with social desirability bias (Gibson et al., 2018), we chose not to collect explicit data on preferences related to cutting one's own daughters. Whether our scepticism was justified or not, refusing these kinds of data had two important consequences. First, we could not examine the relation between implicit and explicit preference measures because we did not have explicit preference measures. We fully appreciated this

limitation when designing our study, and we accepted it because we had little or no faith in explicit measures related to female genital cutting. Second, and for the same reason, we could not examine the relation between explicit preference measures and behaviour. We also fully understood and accepted this limitation.

One could argue that we should have asked parents, one way or another, about cutting their daughters and about the prospects of cut versus uncut future wives for their sons. If we had done so, we would at least have the data now, just in case they might prove useful. This argument, of course, ignores the costs. We collected our attitudinal data from parents in the context of a large randomized controlled trial with multiple data collections over the course of several weeks (Vogt et al., 2016). Minimizing attrition is a major concern in such studies, and our study was no exception. As part of our larger strategy to limit attrition, we did not want to take up the time of our participants unnecessarily. Equally important, because cutting is a sensitive topic, we did not want to be unnecessarily obtrusive about the inner workings of each participant's family. Our worry was that, if we had been too obtrusive, participants would simply not have returned after the first data collection. For this reason, and in direct contrast to traditional methods, we did not ask parents anything about whether they cut, how they felt about cutting their daughters, and whether they hoped for their sons to have cut or uncut wives. Perhaps collecting such data would not have increased attrition in our field experiment, but without knowing we would probably evaluate the risks in the same way again.

Our excess of caution, however, created another limitation that we did not fully appreciate at first. Because of our preoccupation with social desirability, we took several steps to maximize both participant anonymity and the extent to which participant anonymity was salient and transparent to the participants themselves (Vogt et al., 2016). We imagined that, if participants were anonymous, and if they knew they were anonymous, biased responses would be less of a problem than otherwise. Because of our commitment to anonymity, however, we now have almost no information about who our participants were. For the girls in our henna feet study, we know the communities they lived in, and we know when they began primary school, but nothing more. For the parents in our field experiment, we have slightly more information, but not much. In particular, we have no data that would allow us to link, at the family level, the parents in our field experiment to the girls in our henna feet study. Although the two samples are independent of each other in any given community, at least some of the adults in our field experiment were probably parents of girls in our henna feet study. We nevertheless have no way to make this connection now, and so we can only correlate mean implicit association scores by community with cutting rates by community. This is a direct consequence of our insistence on extreme anonymity coupled with the fact that we have two different samples in each community, a sample of girls and a sample of adults. This is the cost of favouring discretion in Sudan.

The upshot is that the geographically defined community is now the only information that links the two data sets. We thus created a situation in which we are unlikely to find a relationship between attitudes and behaviour even if such a relationship exists. Our sample size, for example, is effectively small. The data used for Figure 7 consist of observations for 5642 girls and 6983 adults. To calculate the confidence intervals for the correlations in the figure, we used a bootstrapping algorithm (Vogt et al., 2016, Supplementary Information) that makes use of all these observations by resampling girls and adults within each community. Ultimately, however, we have only correlated 120 mean implicit association scores with 120 cutting rates. In addition, a community-level analysis necessarily depends on variation among communities, which of course must be interpreted relative to variation within communities. Figures 5 and 7 reveal that between-community variation surely existed when we collected our data. All else equal, however, due to the surprising degree of within-community variation related to cutting that we and others have found (Bellemare et al., 2015; Efferson et al., 2015, 2020; Hayford, 2005; Howard & Gibson, 2017; Myers et al., 2024; Novak, 2020; Platteau et al., 2018), identifying community-level patterns with confidence is likely to be difficult in many places.

We did not fully appreciate these limitations when we designed our study because we did not initially plan analyses of the sort shown in Figure 7. Our fieldwork in Sudan took place from 2011 to 2014, just before the replication crisis in the behavioural sciences and the corresponding shift to pre-registration. We only realized that we could and should correlate mean implicit association scores and cutting rates after we had committed ourselves methodologically, especially in terms of the henna feet method. Our methodological commitments grew out of other concerns, and we only appreciated later how our focus on discretion had forced us into validation as a community-level exercise in a setting where each community tended to be highly heterogeneous.

A positive interpretation would be that, because our own methods undercut our ability to find a correlation between attitudes and behaviour, the one correlation we found (Figure 7c) is especially convincing. Given that we have now found the same correlation in two different data sets, we have confidence in the correlation itself. The question, however, is what exactly is correlated with what? We used a new method for estimating cutting rates, and we used a new method for measuring attitudes related to cutting. The correlation we found (Figure 7c) indicates some degree of mutual consistency between these two new methods. Mutual consistency, however, does not preclude the possibility of similar types of bias, although the scoring algorithm we used for the implicit association test was designed to reduce the potential for participants to manipulate results (Greenwald et al., 2003). In the final analysis, we are left with novel methods and the methodological uncertainty that goes along with this novelty.

This brings us to the negative interpretation. We are limited to fewer analyses for validation in Sudan than in Armenia, and the analyses we have conducted show less consistency for Sudan than for Armenia. This is why we find validation in Armenia relatively convincing. We have three definitions of cutting in Sudan, and so we have three different ways of correlating attitudes and behaviour. Of these three, we can say with confidence that only one correlation is not zero. In Armenia, we have both explicit and implicit attitudinal measures, and we have two key indicators of decision-making, namely the number of boys versus girls in the family and whether the youngest child is a boy. By comparing the two attitudinal measures with each other, and by comparing each attitudinal measure with each behavioural measure, we have five possible correlations. Moreover, because we can link the attitudinal data to the behavioural data at the family level, all analyses can treat the family as the unit of analysis. All correlations are highly statistically significant (Schief et al., 2021), and so our validation exercises tell a consistent story from every perspective.

This kind of repeated mutual consistency creates an extra degree of confidence. Although we are certainly confident in the methods we used and data we collected in Sudan, we are happy to say that we are more confident for Armenia. One might imagine that our recommendation is thus to make sure you collect explicit attitudinal data and ensure that your methods allow you to work at the most disaggregated level possible. Unfortunately, we cannot support these recommendations. In both countries, we were trading precision against discretion, and we resolved this trade-off in different ways. Sudan has a legacy of colonial powers and international organisations judging Sudanese people because of female genital cutting (Gruenbaum, 2001), and this legacy meant that minimizing the risk of data contaminated by social desirability bias was an overriding priority.

In contrast, we began our work in Armenia shortly after publication of the first papers showing the importance of sex-selective abortion (Duthé et al., 2012; Guilmoto & Duthé, 2013). Although judgement was certainly present in the offices of the international organizations in Yerevan, we had little reason to assume that knowledge of this judgement had pervaded the general population. More importantly, although we correctly suspected that direct questions about sex-selective abortions would produce data tainted by social desirability (Schief et al., 2021), we knew we had good alternatives for eliciting information on attitudes and behaviours. In particular, we suspected and continue to assume that asking people about the ideal hypothetical family for the youngest child in the family under conditions of maximum anonymity would produce relatively reliable data. Similarly, we suspected and continue to assume that asking people about children born to the family would produce

relatively reliable data. For these reasons, we had straightforward options to measure explicit attitudes and decision making, and we resolved the trade-off between discretion and precision with a relatively strong focus on precision. Finally, in Armenia we collected all data at a single point in time with adults in a given household. We thus had an automatic link between individual attitudes and household decision-making, and we could perform all analyses at the family level as a result.

As a result, even if female genital cutting and sex-selective abortion are equally sensitive as research topics, our task in Armenia was easier than our task in Sudan. The reason is because the sex of a child is a realization of a random process, and this randomness gives the research participant a ready form of plausible deniability. Plausible deniability, in turn, allows the participant to manage the impression she creates, both with respect to herself and with respect to the researcher. This is especially clear when asking about children born to the participant.

To illustrate, imagine you are son-biased, and we are researchers studying son bias. We ask you about the children you have, in order, and you accurately answer a daughter then two sons. Presumably, you answer accurately because you have little incentive to mis-report. First, you probably love your children, and thus you hesitate to pretend that one or more of them does not exist. Second, one can have a daughter then two sons without son bias or sex-selective abortion. Because you know this, you also know that we cannot definitively infer your attitudes or past choices from your answer. You can answer honestly and hide behind the obscurity randomness creates. This obscurity, however, only exists at the individual level. If families with twice as many sons as daughters are over-represented in the sample, we know that families are engaging in sex-selective abortion even if we cannot identify which families are doing so. More subtle still, imagine you actually had two daughters and then a son. Your incentive to mis-report is even weaker here because you actually have twice as many daughters as sons. Again, however, if families in which the youngest child is a son are over-represented, we know that families are engaging in son-biased stopping even if we cannot identify the specific families in question.

In this sense, measuring son bias by asking about children born is analogous to measuring lying with the widely used die-rolling task (Fischbacher & Föllmi-Heusi, 2013). The die-rolling task creates incentives to lie, and it uses a random device that allows individual participants to lie without being detected. This tips the balance in favour of the incentives to lie, which increases the probability that participants who would otherwise tell the truth actually lie. Researchers lose the ability to detect individual lies, but they gain confidence in their ability to detect aggregate lies (Fischbacher & Föllmi-Heusi, 2013; Hruschka et al., 2014). For son-biased participants worried about providing socially desirable responses, incentives also favour lying. The randomness of childbirth, however, tips the balance in favour of telling the truth, and participants who would otherwise lie thus tell the truth. At least that is our thinking when we assume that data on children born in Armenia are accurate.

A key difference between childbirth and female genital cutting is that a useful random device is built into the former but not the latter. In the case of cutting, a researcher who wants to use randomness to increase the reliability of data must build the randomness into the elicitation methods themselves. We did not do so, opting instead to rely on the implicit association test. Other researchers, however, have developed and effectively used elicitation methods with a random component to improve the quality of data on female genital cutting (De Cao & Lutz, 2018; Gibson et al., 2018). This is an excellent development. Ultimately, we do not find validation in Armenia more convincing than in Sudan because we have explicit attitudinal data from Armenia and the possibility of family-level analysis. Rather, we find it more convincing because the decision-making domain in Armenia comes with a ready random device that allows participants to reveal their son bias without actually revealing their son bias.

Most importantly, how can researchers and practitioners use these ideas to inform their methods? We speculate that, in terms of collecting reliable data, a topic like son bias represents an ideal scenario. Because the link between decision-making and outcomes has a strong random component, the cover randomness provides should be obvious and salient to most participants. Researchers and

practitioners should focus on exploiting this cover. For example, a study that simply asks participants to evaluate sex-selective abortion on a five-point scale would represent a wasted opportunity because the study would ignore the randomness intrinsic to sex determination.

Most sensitive topics, however, do not come with a built-in random device. Researchers and practitioners can expect difficult trade-offs as a result. The implicit association test does not incorporate a random device. It instead relies on indirect measures that limit the scope for participants to misrepresent themselves (Nosek et al., 2007). The cost is that these measures are highly artificial. Response times in milliseconds are entirely divorced from how we normally express opinions, and they are entirely divorced from everyday decision-making. Thus, if the task is to use data for programming, validation is essential. An interested practitioner probably requires some understanding of how implicit measures relate to real decision-making. If so, researchers and practitioners should design validation ex ante to account for the demands of their specific programme. They should decide how they will link implicit attitudes and behaviour, whether validation will use granular or aggregate data, and how in general they will trade precision against discretion.

In particular, researchers and practitioners should consider explicit methods that incorporate randomness into the data elicitation itself. To illustrate, list experiments (e.g. De Cao & Lutz, 2018; Gibson et al., 2018) and coin flipping (e.g. Fischbacher & Föllmi-Heusi, 2013; Hruschka et al., 2014) both create situations in which individual participants cannot reveal anything definitive about themselves. For this reason, researchers and practitioners must work at the aggregate level to some extent, but at least they can link implicit and explicit data at the individual level, as we did in Armenia but could not do in Sudan. A key concern with random elicitation methods is to implement them in a way that makes the cover they provide transparent to participants.

Researchers and practitioners should also be clear about how damaging bias may or may not be. Applied cultural evolution represents an activity in which bias can be especially damaging. Applied cultural evolution refers to any situation in which a social planner wants to manage the cultural evolutionary process to produce widespread, socially beneficial changes in behaviour (Efferson et al., 2023, 2024). Social tipping based on coordination incentives and conformity is a classic example that has generated considerable enthusiasm in many domains related to social justice, health, and environmental problems (Cloward, 2016; Constantino et al., 2022; Efferson et al., 2023; Mackie, 1996; Nyborg et al., 2016; Platteau et al., 2018; Travers et al., 2021). For a practitioner who wants to initiate social tipping, reliable information about the attitudes in place before intervention is essential.

Specifically, the shape of the distribution of attitudes directly influences whether social tipping is even possible (Efferson et al., 2023; Ehret et al., 2022), and we can easily imagine biases that would lead to the wrong conclusion. To illustrate, consider a practitioner who rapidly assesses the distribution of attitudes in a population before intervention (Janas et al., 2024) in order to see if tipping is feasible. The population includes groups with two distinct sets of attitudes. One group of people is relatively favourable towards change, and another group is relatively opposed. Bimodal attitudes of this sort tend to undermine social tipping (Efferson et al., 2020, Ehret et al., 2022). Further suppose that those who oppose change, and only those who oppose change, are susceptible to social desirability bias. When the practitioner asks, they claim they support change. A heterogeneous bias of this sort can turn an actual distribution of attitudes that is bimodal into a measured distribution that is unimodal. The practitioner could incorrectly infer that social tipping is feasible (Efferson et al., 2023), and could spend a lot of time and money chasing a self-reinforcing process of cultural change that never comes.

Moreover, even if social tipping is actually possible, using an intervention to target people amenable to change is often the worst strategy from the practitioner's perspective (Efferson et al., 2020, 2023, 2024; Schimmelpfennig et al., 2021). The practitioner may need to identify these people to avoid wasting limited resources on those ready for change anyway. Again imagine that only those resistant to change are prone to social desirability bias. Although the practitioner has correctly concluded that tipping is possible, whether tipping actually occurs can depend on whom the practitioner targets with the intervention (Efferson et al., 2023). Because of the bias, however, the practitioner

cannot distinguish between those amenable to change and those resistant. Targeting the wrong segment of the population could again mean a lot of time and money chasing a self-reinforcing process of cultural change that never comes.

In these hypothetical examples, bias is potentially devastating precisely because it is heterogeneous. Some people provide biased responses, whereas others do not. In general, ordinary forms of heterogeneity make applied cultural evolution much more challenging than it would otherwise be (Andreoni et al., 2021; Efferson et al., 2020, 2023, 2024; Ehret et al., 2022; Platteau et al., 2018). This is true when people differ in terms of their preferences and in terms of whom they know. We suspect it is also true if people differ in terms of providing biased data to researchers and practitioners. For this reason, if the long-run objective is to spark some process of cultural change, researchers and practitioners should resolve trade-offs squarely in favour of relatively disaggregated data untainted by bias. Such data are necessary to manage cultural change in heterogeneous populations. Cultural engineering can be ethically questionable in general. Collectively, as a minimum requirement, we should at least be prepared to invest in the data we need to know what we are doing.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/ehs.2024.48.

Acknowledgements. We thank Amy Elhadi, Nadia Ahmed Mohmmed Zaid, Hilal El Fadil Ahmed, Nada Sayed Ahmed, Waleed Omer Babikir Alalfi, and Omer Abdazeez Abdarazig Abdalla for their insights and assistance when conducting the original fieldwork in Sudan. We thank the Women's Resource Center of Armenia for their insights and assistance when conducting the original fieldwork in Armenia. We would also like to thank Sarah Myers and two anonymous referees for comments on an earlier version of this paper.

Author contributions. C.E. and S.V. contributed equally to the conceptualization, design, and implementation of the projects in Sudan and Armenia. C.E. analysed the previously unpublished henna feet data and wrote the paper with input from S.V.

Financial support. C.E. and S.V. are funded by the Swiss National Science Foundation (Grant Nrs. 100018_185417 and 100018_215540).

Competing interests. C.E. is a member of the editorial board of *Evolutionary Human Sciences*. S.V. declares none.

Research transparency and reproducibility interest. Data and code are included as supplementary materials.

References

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, *118*(16), e2014893118.

Arnold, F. (1997). Gender preferences for children. DHS comparative studies no. 23. Calverton, Maryland, USA: Macro International. http://dhsprogram.com/pubs/pdf/CS23/CS23.pdf.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 to 10 and adulthood. *Psychological Science*, *17*(1), 53–58.

Bellemare, M. F., Novak, L., & Steinmetz, T. L. (2015). All in the family: Explaining the persistence of female genital cutting in West Africa. *Journal of Development Economics*, *116*, 252–265.

Boddy, J. (2007). *Civilizing women: British Crusades in Colonial Sudan*. Princeton University Press.

Camilotti, G. (2016). Interventions to stop female genital cutting and the evolution of the custom: Evidence on age at cutting in Senegal. *Journal of African Economies*, *25*(1), 133–158.

Cloward, K. (2016). *When norms collide: Local responses to activism against female genital mutilation and early marriage*. Oxford University Press.

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S., & Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest*, *23*(2), 50–97.

De Cao, E., & Lutz, C. (2018). Sensitive survey questions: Measuring attitudes regarding female genital cutting through a list experiment. *Oxford Bulletin of Economics and Statistics*, *80*(5), 871–892.

Duthé, G., Meslé, F., Vallin, J., Badurashvili, I., & Kuyumjyan, K. (2012). High sex ratios at birth in the Caucasus: Modern technology to satisfy old desires. *Population and Development Review*, *38*(3), 487–501.

Efferson, C., Ehret, S., von Flüe, L., & Vogt, S. (2024). When norm change hurts. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230039.

Efferson, C., & Vogt, S. (2018). Behavioural homogenization with spillovers in a normative domain. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1879), 20180492.
Efferson, C., Vogt, S., Elhadi, A., Ahmed, H. E. F., & Fehr, E. (2015). Female genital cutting is not a social coordination norm. *Science*, *349*(6255), 1446–1447.
Efferson, C., Vogt, S., & Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour*, *4*(1), 55–68.
Efferson, C., Vogt, S., & von Flüe, L. (2023). Activating cultural evolution for good when people differ from each other. In J. Kendal, R. Kendal, & J. Tehrani (Eds.), *The Oxford handbook of cultural evolution* (online edn, Oxford Academic). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198869252.013.69.
Ehret, S., Constantino, S., Weber, E., Efferson, C., & Vogt, S. (2022). Group identities can undermine social tipping after intervention. *Nature Human Behaviour*, *6*(12), 1669–1679.
Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise – An experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.
Gibson, M. A., Gurmu, E., Cobo, B., Rueda, M. M., & Scott, I. M. (2018). Indirect questioning method reveals hidden support for female genital cutting in south central ethiopia. *PLoS One*, *13*(5), e0193985.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.
Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216.
Gruenbaum, E. (2001). *The female circumcision controversy: An anthropological perspective*. University of Pennsylvania Press.
Guilmoto, C. Z. (2013). Sex imbalanced at birth in Armenia: Demographic evidence and analysis. Technical report, CEPED Paris.
Guilmoto, C. Z., & Duthé, G. (2013). Masculinization of births in eastern europe. *Population & Societies*, *11*, 1–4.
Hayford, S. R. (2005). Conformity and change: Community effects on female genital cutting in Kenya. *Journal of Health and Social Behavior*, *46*(2), 121–140.
Howard, J. A., & Gibson, M. A. (2017). Frequency-dependent female genital cutting behaviour confers evolutionary fitness benefits. *Nature Ecology and Evolution*, *1*(3), 0049.
Hruschka, D., Efferson, C., Jiang, T., Falletta-Cowden, A., Sigurdsson, S., McNamara, R., Sands, M., Munira, S., Slingerland, E., & Henrich, J. (2014). Impartial institutions, pathogen stress and the expanding social network. *Human Nature*, *25*(4), 567–579.
Janas, M., Nikiforakis, N., & Siegenthaler, S. (2024). Eliciting thresholds for interdependent behavior. Technical report, National Bureau of Economic Research.
Jayachandran, S. (2017). Fertility decline and missing women. *American Economic Journal: Applied Economics*, *9*(1), 118–139.
Kim, D.-Y. (2003). Voluntary controllability of the implicit association test (IAT). *Social Psychology Quarterly*, *66*(1), 83–96.
Lawson, D. W., Chen, Z., Kilgallen, J. A., Brand, C. O., Ishungisa, A. M., Schaffnit, S. B., Kumogola, Y., & Urassa, M. (2024). Misperception of peer beliefs reinforces inequitable gender norms among tanzanian men. *Evolutionary Human Sciences*, *6*, e17.
Mackie, G. (1996). Ending footbinding and infibulation: A convention account. *American Sociological Review*, *61*(6), 999–1017.
Merry, S. E. (2006). *Human rights and gender violence: Translating international law into local justice*. University of Chicago Press.
Morehouse, K. N., & Banaji, M. R. (2024). The science of implicit race bias: Evidence from the implicit association test. *Daedalus*, *153*(1), 21–50.
Myers, S., Gurmu, E., Alvergne, A., Redhead, D., Howard, J. A., & Gibson, M. A. (2024). Social clustering of preference for female genital mutilation/cutting (FGMC) in South-Central Ethiopia. https://osf.io/preprints/osf/szq6x.
Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). Psychology Press.
Novak, L. (2020). Persistent norms and tipping points: The case of female genital cutting. *Journal of Economic Behavior & Organization*, *177*, 433–474.
Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W. N., Arrow, K. J., Barrett, S., Carpenter, S., Chapin, III, F. S., Crépin, A.-S., Daily, G., Ehrlich, P., Folke, C., Jager, W., Kautsky, N., Levin, S. A., Madsen, O. J., Polasky, S., & Scheffer, M. (2016). Social norms as solutions. *Science*, *354*(6308), 42–43.
Platteau, J.-P., Camilotti, G., & Auriol, E. (2018). Eradicating women-hurting customs. In S. Anderson, L. Beaman & J. Platteau (Eds.), *Towards gender equity in development* (pp. 319–356). Oxford University Press.
Schief, M., Vogt, S., Churilova, E., & Efferson, C. (2024). Isolating a culture of son preference among Armenian, Georgian and Azeri parents in Soviet-era Russia. *Evolutionary Human Sciences*, *6*, e19.
Schief, M., Vogt, S., & Efferson, C. (2021). Investigating the structure of son bias in Armenia with novel measures of individual preferences. *Demography*, *58*(5), 1737–1764.
Schimmelpfennig, R., Vogt, S., Ehret, S., & Efferson, C. (2021). Promotion of behavioural change for health in a heterogeneous population. *Bulletin of the World Health Organization*, *99*(11), 819.

Shell-Duncan, B. (2016). Considerations on the use and interpretation of survey data on FGM/C. Addendum to the SOTA. November 2016. *Evidence to End FGM/C: Research to Help Women Thrive*. New York: Population Council.

Shell-Duncan, B., & Hernlund, Y. (2000). Female "circumcision" in Africa: Dimensions of the practice and debates. In B. Shell-Duncan & Y. Hernlund (Eds.), *Female "circumcision" in Africa: Culture, controversy, and change* (pp. 1–40). Lynne Rienner.

Shell-Duncan, B., Wander, K., Hernlund, Y., & Moreau, A. (2011). Dynamics of change in the practice of female genital cutting in Senegambia. *Social Science & Medicine*, *73*(8), 1275–1283.

Thomas, L. (2000). "Ngaitana (I will circumcise myself)": Lessons from colonial campaigns to ban excision in Meru, Kenya. In B. Shell-Duncan & Y. Hernlund (Eds.), *Female "circumcision" in Africa: Culture, controversy, and change* (pp. 129–150). Lynne Rienner.

Travers, H., Walsh, J., Vogt, S., Clements, T., & Milner-Gulland, E. (2021). Delivering behavioural change at scale: What conservation can learn from other fields. *Biological Conservation*, *257*, 109092.

Vogt, S., Efferson, C., & Fehr, E. (2017). The risk of female genital cutting in Europe: Comparing immigrant attitudes toward uncut girls with attitudes in a practicing country. *SSM-Population Health*, *3*, 283–293.

Vogt, S., Zaid, N. A. M., Ahmed, H. E. F., Fehr, E., & Efferson, C. (2016). Changing cultural attitudes towards female genital cutting. *Nature*, *538*(7626), 506–509.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, *7*(1), 359–387.

Cite this article: Efferson C., & Vogt S. (2025). The cover of randomness: validating implicit methods for the study of sensitive topics. *Evolutionary Human Sciences* 7, e6, 1–20. https://doi.org/10.1017/ehs.2024.48