How should we translate survey questionnaires?

An analysis of Kenyan DHS data

Alexander A. Weinreb

Mariano Sana

To be presented at the UAPS Meetings, Arusha, Tanzania, Dec. 10-14, 2007

Alexander A. Weinreb, Department of Sociology and Anthropology, Hebrew University, Jerusalem 91905, Israel. awein@mscc.huji.ac.il; tel. +972 2 588 3038; fax. +972 2 588 3549

Mariano Sana, Department of Sociology and Louisiana Population Data Center, Louisiana State University. 126 Stubbs Hall, Baton Rouge LA 70803, United States. msana@lsu.edu; tel. +1 225 578 1115; fax: +1 225 578 5102

How should we translate survey questionnaires?

An analysis of Kenyan DHS data

ABSTRACT

The collection of survey data in developing and, increasingly, developed countries often requires the translation of the survey instrument. This article addresses the implications for data and analysis of two of the most common approaches to translation. The first, the officially sanctioned—though not empirically verified—method, involves the pre-fieldwork production of a standardized translation of the template questionnaire into all or most languages in which interviews are expected to be conducted. The second, rarely acknowledged in the literature but quite common in the field, is more *ad hoc*: the interviewer spontaneously translates from the language of the questionnaire to the language of the interview. Using the 1998 Kenya DHS, in which 23% of interviews were conducted using spontaneous translation, we explore the effects of these two translation modes on four indicators of measurement error and on estimated multivariate relations. In general we find that moderate effects of spontaneous translation on univariate statistics—including higher-order variance structures—become magnified in multivariate analysis. This suggests that, although under certain circumstances it may be justified to allow well-trained interviewers to spontaneously translate their questionnaires, standardized translation should continue to be the norm.

Introduction

Demographers, economists and, increasingly, sociologists often use data collected in settings which are linguistically different from their own. Where they themselves, or some other group of outsiders, play lead roles in project design and data collection, the instrument used to collect those data must be translated. Typically, the translation is from an original template questionnaire constructed in a "global" language like English or French into a "local" language (or multiple local languages). The problem is, notwithstanding its central role in the research enterprise—like sampling and instrument design, or interviewer selection and training, questionnaire translation shapes the data we use—very little research has been conducted on how researchers should translate a questionnaire. Nor is there much guidance available for those who have to translate. There is no mention of translation, for example, in the otherwise exhaustive monographs by Converse and Presser (1986), Groves (1989), Creswell (2002), Sapsford and Jupp (2006), and Babbie (2006). Rather, the extant literature in which researchers are encouraged to translate their questionnaires—all of which, somewhat reasonably, appears to be focused on data collection in non-Western societies—is based less on empirical studies and evaluations than on more informal types of observation (e.g., Mitchell 1965; Ware 1977), or on descriptions of the long and careful series of translations/back-translations that some researchers have gone through before finalizing their instrument (e.g., Axinn, Fricke, and Thornton 1991; Chen, Liu and Ennis 1997). Together, these have come to represent a type of best-translation practice prescribed by methodologists (e.g., Bernard 2000: 246-7; Overton and van Dierman 2003: 39).

This general lack of research on translation issues is worrisome for a number of reasons. First, it suggests that standard research practice during an important stage of data collection is not based on a proven experimental record, nor even on a weaker non-experimental empirical record. Rather, it is based on a simple methodological norm: we *ought* to formally translate a questionnaire, and the process of translation ought to include a series of translations/back-translations, because that is what researchers have done in the past. This absence of prior research is less than ideal on an epistemological level.

Second, even casual observation of the translation process in action shows that current translation norms can be quite difficult to enact with confidence since, for any given question on the "template" questionnaire, multiple translation outcomes are often legitimate. In other words, both translation process and outcomes are imbued with greater uncertainty than that which is ideally associated with good science. In the first author's own fieldwork in sub-Saharan Africa, for example, he has often seen considerable discord between translators—where each is translating into their first language—over how to ask a given question or communicate a given idea. While some of the disagreements between translators have reflected differences between "high" language and more colloquial formulations, making them relatively easy to resolve, others have reflected real differences in opinion among the translators about how to best render the global language into the local. Or they have reflected differences in dialect within the local language that were only finally resolved by interviewers hired at the actual field sites—that is, not by the college-educated translators. Finally and most worrying, they have reflected the fact that not all terms or ideas exist in all languages. Thus the difficulty in translating concepts such as "likely" or "likelihood," important in contemporary demographic studies of HIV-related behavior into a number of African languages in Malawi (e.g., Helleringer and Kohler 2005; though see Delavande and Kohler [2007] for an alternative method of collecting subjective probabilities in such settings which relies much less on direct translation of the underlying concept).

The third reason that the lack of research on translation issues is worrisome is specific to certain settings. In particular, current translation ideals are much harder to implement in settings where the questionnaire needs to be translated into multiple languages. This is often the case in multiethnic states in sub-Saharan Africa, many of which have been the focus of considerable demographic interest and research. It is also increasingly the case in urban areas in developed countries, in many of which there are large pockets of language-minority immigrants. In all multilingual contexts, in short, the complicated translation process discussed above must be replicated over multiple languages, with each language-specific translation-team charged with maintaining the meaning intended in the original template questionnaire.

Research Questions and Plan

In an effort to initiate more formal research focus on translation issues—and provide some baseline results—this article addresses a simple question. Are the gains associated with formally translating a questionnaire, as opposed to letting proficient interviewers spontaneously translate from a template questionnaire, worth the hassle of translating? There are conceptual grounds for thinking they may not be. In particular, over the last two decades, the key principle that underlies the practice of formal translation—the idea of “stimulus equivalence,” that we should maximize the standardization of data collection procedures across respondents (Fowler and Mangione 1990)—has been increasingly challenged in the data collection literature. Stimulus equivalence approaches do not, its critics assert, automatically generate the most valid data. Rather, non-standardized, “conversational,” or “personalized” interviewing practices often perform better (Dijkstra 1987; Suchman and Jordan 1990; Schober and Conrad 1997; Schaeffer and Presser 2003). These differences can be mapped onto different types of translation procedures since stimulus equivalence requires formal and standardized translation into all interview languages. A conversational approach does not, relying instead on spontaneous translation on the part of the interviewers.

More specifically, our aim in this article is to explore the impact, if any, that deviating from the stimulus equivalence principle, as it applies to the translation of survey questions, has on data and analysis. In official practice, as sanctioned by the principle of stimulus equivalence, the interviewer uses a standardized translation of the template questionnaire into the language of the interview, yielding *language correspondence* between the printed questionnaire and the interview. In an alternative scenario, more compatible with the conversational interviewing approach, the interviewer is not equipped with a questionnaire written in the same language as that of the interview interaction. Consequently, the interview is based on the interviewer’s *spontaneous translation* of a survey instrument that is written in another language into the language of the interview.

Our analysis uses the 1998 wave of the Kenyan Demographic and Health Survey (KDHS). This data source may initially seem surprising given that, like other Demographic and Health Survey (DHS) questionnaires, the KDHS instrument was translated and back-translated into all major national languages in Kenya. However, it is less surprising when one moves beyond formal DHS translation protocols and delves into actual KDHS data. These show that in 23 percent of interviews there was no correspondence between the language of the interview and the language of the questionnaire. Rather, interviewers appear to have resorted to spontaneous translation. We explore the circumstances that led to this contingency below. For now, suffice to say that, with proper controls for respondent selectivity, it allows us to evaluate the relative effects of the two different translation modes.

We address two discrete analytic questions. To begin, we focus on whether the different translation modes are associated with different levels of measurement (i.e. non-sampling) error. Here we use four indicators. The first three are specific to variables: the proportion of overall response variance which can be attributed to interviewers, systematic differences in response values, and the level of complex variation across interviewers and districts. The fourth—the length of the interview—is a more general indicator of data quality. The second question assesses whether observed effects on these indicators of measurement error matter. That is, to what extent do they affect actual analyses of relationships among measured variables in substantive ways?

Our analysis has considerable practical implications for the collection of demographic data in multilingual settings. In particular, research projects could save considerable time and money if researchers could be sure that, with only a minimal and acceptable increase in non-sampling error, they could either translate a questionnaire into a single *lingua franca* rather than into every major language in a given society, or pay less attention to strictly matching respondents and questionnaires on given languages (in addition to matching respondents and interviewers, which must be the case). On the other hand, if spontaneous translation during an interview does increase non-sampling error beyond a level judged to be acceptable, survey researchers would have to consider investing more resources—time and money—into making sure there is language correspondence, whether this involves translating survey instruments into more languages or simply making more translated questionnaires available. This last point, as shown below, appears to have been the key cause of the KDHS' deviation from the standard translation protocol.

The paper is divided into five main sections. In the first we introduce our data and describe the frequency of spontaneous translation. In the second, we review some selectivity issues related to the types of respondents who were interviewed with a spontaneously translated questionnaire as opposed to those in which there was complete language correspondence. In the third, we compare the relative effects of the different translation procedures on measurement error. In the fourth, we explore their relative impact on analysis. In the fifth we close with some relatively conservative conclusions.

Data

Barring North Eastern Province, a large and politically unstable area that contains less than 5 percent of Kenya's population, the 1998 KDHS data are nationally representative survey data collected from 7,881 women.

Data collection followed standard DHS procedures. Two factors are relevant to this analysis. The first is that questionnaires were made available in ten languages in addition to English and Kiswahili: Kalenjin, Kamba, Kikuyu, Kisii, Luhya, Luo, Meru, Embu, Mijikenda, and Masai. All ten are homonymous with large Kenyan ethnic groups and with the exception of the Masai, each is the dominant ethnic group in at least one district sampled by the 1998 KDHS. The second is that interviewers—all women—were assigned to language-specific teams that were then allotted responsibility for specific ethno-linguistic regions, eleven in total. In almost all cases, these ethno-linguistic regions are coterminous with administrative districts—at the time of the data collection there were 43 districts in Kenya—though there are a few cases where a second interviewer team was brought into areas of districts associated with language minorities (usually a group associated with a neighboring district). In either case, all members of a given interviewer team appeared to have covered the same sample clusters, and no sample cluster was distributed to more than one interviewer team. So there is some interpenetration between interviewers and respondents within sample clusters, though it was not systematically randomized. Steps taken to maximize the equivalence of interviewers' target population and workload within teams prior to analysis are described below.

Table 1 presents basic data on the distribution of interviewers and their teams by districts and languages, as well as the percent of interviews per team without language correspondence between questionnaire and interview. It shows that the 7,702 interviews used in this analysis—the reasons for reduction from the full 7,881 will be presented below—were collected by 64 interviewers working in twelve teams. In all but one of those teams (team 12) there were between five and seven interviewers. Each was assigned to between two-to-four districts in which a single language was dominant (with the exception of teams 10 and 12). Overall, the interviewers used in this analysis worked in all the listed districts and conducted an average of 120 interviews each. Finally, although 23 percent of all interviews were conducted using spontaneous translation of the questionnaire, there is considerable variation between districts, ranging from a low of 10 percent for team 10 to a high of 50 percent for team 9.

Table 1 about here

Who was interviewed using spontaneous translation?

Since our analysis does not use experimental data, it is important to understand the differences between the 77 percent of women interviewed in the same language as that of the questionnaire, and the remaining 23 percent interviewed in a different language.

The levels of spontaneous translation found in the KDHS are a direct product of fieldwork-related decisions. As we now describe, these appear to have been reasonable decisions. However, to the extent that DHS' aim was to maximize the number of interviews conducted with complete language correspondence, the frequency of spontaneous translation does signal a problem with underlying DHS field methods.

We begin by describing the overall contextual characteristics in which field decisions are made. We then discuss field decisions, the relationship between spontaneous translation and timing of fieldwork, and analytic concerns stemming from these timing issues.

Contextual characteristics

It is very difficult to avoid some level of deviation from full linguistic correspondence in a truly random population-based sample in Africa. The reason lies at the nexus of two contextual characteristics that are common across many African settings. First is the combination of linguistic heterogeneity in most African states—though there are notable exceptions (e.g., Botswana, Rwanda and Burundi)—with the fact that since the colonial era, given administrative areas have been associated with particular ethnic groups. The second characteristic is the relatively high mobility of many Africans. In other words, notwithstanding the association between given areas and given ethnic (and usually linguistic) groups, there has long been substantial movement of individuals for trade, temporary labor, or marriage (Cordell, Gregory, and Piche 1996). Moreover, this movement tends to be concentrated in people's prime working and reproductive ages, the period in which individuals become the focus of data collection for social and behavioral research.

These two characteristics give rise to geographic areas with a significant minority of individuals who have crossed administrative boundaries and who speak the dominant local language with only some skill. From an office in Maryland or Nairobi, researchers might imagine them having sufficient command of the dominant local language(s) to allow them to work on a large farm or estate, or trade in the local market, but not yet sufficient to be able to comfortably understand questions with detailed health-related vocabulary, or to accurately express opinions about such topics when approached by an interviewer. Following on from this, researchers might imagine how a team of interviewers outfitted with questionnaires in only one local language would run afoul of some of these individuals. There would inevitably be a mismatch between the available questionnaires and the potential respondents' own language skills.

Field decisions

The distribution of questionnaires by language in the KDHS appears to reflect sensitivity to this problem, since the KDHS also makes questionnaires available in the national *lingua franca,* in Kenya's case, Kiswahili. This can be seen in columns 2, 3, and 4 of Table 2, which shows the distribution of questionnaires by language in the 1998 KDHS. A majority of the questionnaires given to ten of the twelve field-teams were in the dominant local language (field team 9, which had 50 percent spontaneous translation [see Table 1] was the exception). But they were also given a substantial number of questionnaires in Kiswahili, the back-up language.

Table 2 about here

This seems like a reasonable solution to the language diversity issue. Two problems remain, however. The first is that until the field team actually gets to the field, no-one knows the distribution of linguistic ability among sampled individuals. So printing decisions—in themselves time-consuming and nerve-wracking in settings where paper, toner, electricity, and other supplies are unreliable and expensive—are necessarily based on estimates of how many people the survey project can expect to interview in a given language in each area. The second problem, specific to the KDHS, lies with the expectation that Kiswahili can be a reliable back-up language. This belief fits with common stereotypes among Kenyan urban and educated elites—among whom are the local survey practitioners who provide foreign specialists with counsel on matters such as local linguistic proficiency in different parts of the country—that all folks in rural areas in Kenya speak Kiswahili (Susan Watkins, personal communication). This belief does not appear to be supported by actual linguistic proficiency among rural Kenyans, certainly not when it comes to the type of detailed or sensitive questions asked by the KDHS. Even on the level of self-reported linguistic competence, for example, data from the Kenya Diffusion and Ideational Change Project, collected in rural Nyanza Province, show that only 55 percent of women from the all-Luo and Suba sample claimed to be able to speak Kiswahili (own calculation).[1]

In our view, the high levels of spontaneous translation observed in the KDHS are first and foremost a product of the KDHS overestimate of how much Kiswahili can serve as a backup language. This seems clear when, using Table 2, we compare the distribution of completed questionnaires by language with completed interviews by language. In relation to field teams 1 – 3 and 5 – 9 there is a substantial mismatch between the distribution of questionnaires and interview languages. Take the Luo team (7), for example. Although 61.5 and 37.9 percent of their questionnaires were, respectively, in Luo and Kiswahili, 94.3 percent of their actual interviews were conducted in Luo. In other words, like their colleagues in teams 1 – 3, 5, 6, 8, and 9, the Luo team was not given enough questionnaires in the dominant local language (in this case Luo). Instead, they were given too many Kiswahili questionnaires, forcing the team's interviewers to spontaneously translate from Kiswahili questionnaires into the local language.

Spontaneous translation and the timing of fieldwork

More support for this suggestion can be found when we look at the relationship between spontaneous translation of questionnaires and timing of fieldwork. Simple tabulations show that women interviewed using spontaneously translated questionnaires were much more likely to be interviewed in the later stages of fieldwork than in the earlier stages. This can be seen on two dimensions in Table 3. The first column shows how the percentage of spontaneously translated interviews increases from around 10 in month 1, to 11, 14, 29, 46, and 81 percent in months 2-to-6 respectively.

[1] See http://kenya.pop.upenn.edu for more on the KDICP (including access to data).

The second column of Table 3 shows that although this may have begun as an individual-level phenomenon—that is, specific to particular individuals in a given setting—it began to vary by sampling cluster in the later months of fieldwork. Thus, in the first two months of fieldwork we see that in only 2 of the 219 sample clusters were more than 75 percent of the women interviewed using a spontaneously translated questionnaire. In contrast, in months 3 and 4 this occurred in 6 and 18 of 131 clusters, respectively, and in months 5 and 6, it occurred in 22 and 17 sampling clusters respectively. The latter represented 20 and 55 percent of all clusters visited in those months. We interpret this as a sign that field teams had run out of their local language questionnaires, or were desperately trying to conserve them. Either way, as discussed below, our analysis controls for month of fieldwork since it should be correlated with both interviewers' skills and the geographic area in which fieldwork was conducted.

Table 3 about here

Impact of translation on measurement error

We assess the impact of the translation modes on three indicators of measurement error: the contribution of interviewers to total survey error, systematic differences in response values, and complex variation in interviewer- (or cluster-) and district-level estimates. We also consider the relationship between translation mode and the overall length of the interview, which beyond the first period of fieldwork during which interviewers are becoming increasingly familiar with the research instrument, we interpret as a general indicator of the quality of the interaction.

Since each of these requires a somewhat different model specification, we introduce the models and discuss the results in distinct subsections, each of which is devoted to one of the indicators. We begin, however, by noting one global change that we made to the dataset, and by describing the group of variables on which we chose to focus analytic attention.

We changed the original data set by dropping all observations corresponding to sixteen relatively low productivity interviewers, on the assumption that they were not full-time interviewers or were relieved of their positions after producing a small number of unsatisfactory questionnaires. Together, these sixteen individuals completed 179 interviews, less than one-tenth of the average for the remaining 62 interviewers. Discarding these interviews reduced the original sample size from 7,881 to the 7,702 interviews reported earlier in Table 1. In addition, we dropped 222 interviews collected by the two interviewers in team 12 (ID numbers 111 and 112), since they worked in districts with quite distinct ethnic-linguistic profiles (Nairobi, Kajiado and Narok) and there are too few of them to meaningfully distinguish in-group patterns. This has implications for the analytic models used below. Discarding the 222 observations from team 12 further reduced our sample size to 7,480 interviews.

We selected 24 variables representing four key categories of survey questions: respondent's household and background characteristics, fertility and contraceptive use, more general fertility-related knowledge and attitudes, and some AIDS-related information. The specific variables within these categories can be seen in the Appendix. They include many core variables used in demographic analysis over the last four decades. In relation to each of them, and across the two translation modes, we evaluated the contribution of interviewers to total survey error, systematic differences in response values, and complex variation in interviewer- (or cluster-) and district-level estimates.

Does interviewer-related error vary across the two translation modes?

We evaluate interviewer-related error using a random intercept model, that is, by parsing the variance on a given variable between respondents from the variance between interviewers and the variance between districts, as in:

$$Y_{ijk} = \gamma_{ooo} + V_{ook} + U_{ojk} + R_{ijk} \qquad [1]$$

where γ_{ooo} is the population grand mean, V_{ook} is the specific effect of district k, U_{0jk} is the specific effect of interviewer j within district k, and R_{ijk} is the residual effect for respondent i within interviewer j within district k.[2]

Because of the non-experimental structure of our data some basic controls are needed to implement model [1]. Specifically, given the relationship between timing of fieldwork and type of translation procedure used in any given interview, we add dummy indicators for month of fieldwork, as well as a dummy indicator for urban residence, so that the

[2] Throughout this section we draw heavily on Snijders and Bosker (1999). All multilevel analyses described in this paper were implemented in Stata 9 using the "xtmixed" command (do-files available from the authors upon request). Two details about model specification are worth noting here. First, all models were estimated using restricted maximum likelihood (REML) over maximum likelihood (ML) since the latter is more sensitive to loss of degrees of freedom when dealing with a small number of groups (see Snijders and Bosker 2003[1999]: 56). As level-3 specification is set to the district level and there are 33 districts, this is of concern, especially in final models that have a relatively large number of regression parameters. Second, no assumptions were made about the structure of the covariance matrix. Rather, all variances and covariances were distinctly estimated.

model we estimate in order to answer this first question is:

$$Y_{ijk} = \gamma_{000} + \gamma_{100}\, x_{ijk} + V_{00k} + U_{0jk} + R_{ijk} \qquad [2]$$

where x indexes a set of dummy variables for month of fieldwork and for urban residence.

Two sets of model [2] were estimated for each variable of interest, the first restricted to interviews with language correspondence (superscript LC), the second to interviews with spontaneous translation (superscript ST). Since

$$\text{Total Variance (TV)}^{\text{ST}} = \text{var}(R_{ijk})^{\text{ST}} + \text{var}(U_{ojk})^{\text{ST}} + \text{var}(V_{ook})^{\text{ST}} \qquad [3a]$$

and

$$\text{Total Variance (TV)}^{\text{LC}} = \text{var}(R_{ijk})^{\text{LC}} + \text{var}(U_{ojk})^{\text{LC}} + \text{var}(V_{ook})^{\text{LC}} \qquad [3b]$$

we can see whether the relative contribution of interviewers to TV is greater when interviews are based on spontaneous translation than when they are done based on language correspondence by simply looking at the difference $\text{var}(U_{ojk})^{\text{ST}}/\text{TV}^{\text{ST}} - \text{var}(U_{ojk})^{\text{LC}}/\text{TV}^{\text{LC}}$. Since successful standardization reduces the interviewers' contribution to variance (Maynard and Schaeffer 2002:5), our expectation is that this difference will be positive.

Figure 1 about here

Results across 24 variables are graphed in Figure 1 and are consistent with expectations. Specifically, an average of 3.9 percent of the total variance was estimated to be due to interviewers where there was language correspondence between the language of the questionnaire and the interview. Where there was spontaneous translation of the questionnaires, this increased to 5.5 percent of the variance. More specifically, increases in the proportion of variance attributable to interviewers were recorded in 20 of the 24 variables. Of the four exceptions, the only variable where interviewer-related variance declined by more than 1 percent of the total is in relation to whether the respondent reported hearing a family planning-related message on the radio in the last few months. Overall, therefore, interviewer-related variance is greater where interviewers spontaneously translated, but it nevertheless represents a small portion of the total variance.

Are there systematic differences in response values across the two translation modes? In order to answer this question we add a term to the fixed part of model [2] that indexes "spontaneous translation," yielding a full model

$$Y_{ijk} = \gamma_{000} + \gamma_{100}\, x_{1ijk} + \gamma_{200}\, x_{2ijk} + V_{00k} + U_{0jk} + R_{ijk} \qquad [4]$$

where the new term x_{2ijk} indexes the dichotomous variable for spontaneous translation. As above, this model was estimated in relation to all 24 selected variables, under a number of different specifications. In particular, we alternatively used interviewer and sample cluster as the level-2 group-identifier. We also tried each of these with extra controls for respondent's age and urban residence. As results concerning the estimated coefficient x_{2ijk} did not vary substantively across these different specifications, we only present results from a three-level model with level-2 identification as the sample cluster (this had a better model fit than where level-2 was identified as the interviewer). Results from other specifications are available from the authors upon request.

Table 4 about here

Results are presented in Table 4. They suggest that there is very little effect of differential translation procedures on response values. Specifically, of the 24 selected variables, there is a statistically significant difference in mean response value in only 2 cases (at the 5% level): the question about the number of children 5 and under in the household, and whether the respondent had heard a FP-related message on radio in the last few months. Relaxing the significance threshold to the 10% level yields only one more significant difference: the overall number of household members.

Which of the two translation modes generates more higher-order variation?

Model [4] ignores possible heterogeneity in translation effects across level 2 (interviewers or clusters) or level 3 (districts). That is, although Table 4 shows that differences in translation procedures may generate no obvious bias at the population level, there may nonetheless be significant differences across interviewers or across communities. Respectively, these could arise where, for example, each interviewer spontaneously translates a given question in a slightly different manner, or interactional styles associated with particular communities or districts differ in ways which affect peoples' response to fully standardized *versus* spontaneously translated questions. We explore whether either of these is occurring by allowing the variance associated with spontaneous translation to vary in the random parts of the model. That is, we parse the level-2 and level-3 variance in model [4] into three terms each (leaving the variance at level 1 as a "fixed" parameter). Thus, U_{ojk} represents baseline level-2 variance, U_{2jk} represents additional variance in the intercept associated with spontaneously translated questionnaires, and the covariance term (U_{ojk},U_{2jk}) indexes differential slope of variance across level 2. Similarly, V_{ook}, represents baseline level-3 variance, V_{2ok} represents additional variance in the intercept associated with spontaneously translated questionnaires, and the covariance term (V_{00k}, V_{20k}), indexes differential slope of variance across districts. In short, the model that we estimate is:

$$
\begin{aligned}
Y_{ijk} = {} & \gamma_{000} + \gamma_{100}\, x_{1ijk} + \gamma_{200}\, x_{2ijk} \\
& + (V_{00k} + V_{20k} + \mathrm{cov}(V_{00k}, V_{20k})) \\
& + (U_{0jk} + U_{2jk} + \mathrm{cov}(U_{0jk}, U_{2jk})) + R_{ijk} \qquad [5]
\end{aligned}
$$

The addition of these new random parameters allows us to explore how much heterogeneity there is in the effect of questionnaire translation across sample clusters/interviewers and districts. That is, how much do the *cluster-/interviewer-* or *district-specific* intercepts and slopes vary from each other in relation to types of questionnaire translation. As above, for the presentation of results, we chose clusters over interviewers to index level 2.[3]

In order to assess whether the changes to the model embodied in equation [5] explain part of the variability in survey responses we compared the fit of model [5] and model [4] across the 24 dependent variables. Figure 2 shows the results. The vertical axis

[3] Two discrete series of models based on equation [5] were estimated. In the first, level-2 was defined as the sample cluster. In the second, it was defined as the interviewer. We use results from the cluster since they provide better model fit than those from the interviewer. In addition, we also ran two other complete series of models. In the first we allowed variance associated with spontaneous translation to vary only at level-2 (restricting level-3 variance to one term). In the second, we allowed variance associated with spontaneous translation to vary only at level-3 (restricting level-2 variance to one term). Results of all these models are available upon request. We have omitted them for brevity and because model 5 yields the most comprehensive information.

corresponds to minus twice the difference between the Log Likelihoods of Model 5 and Model 4. The two horizontal lines indicate the thresholds for significance at the .01 and .05 level with four degrees of freedom.

Figure 2 about here

As the figure shows, in relation to 12 of the 24 variables, we find a statistically significant difference between models in which we allow for complex variation at the cluster or district level (model [5]), and those in which we do not (model [4]). These include some core variables in demography such as household size, educational attainment, number of daughters who have died, intention to use family planning in the next 12 months, awareness that a healthy-looking person can have AIDS, and assessment of the own risk of catching AIDS.

In relation to the other 12 variables, adding complex variation does not improve model fit. Here too we find some core demographic variables, such as current marital status, number of children ever born, number of living children, and currently using a method of family planning.

Translation mode and length of interview

Length of interview is an indicator of the interviewer's familiarity with the research instrument, as well as of the quality of the interview interaction. Net of the former, the latter works in two ways. First, a shorter interview may signal less engagement and motivation on the part of the respondent. Alternatively, it may also be a product of a greater tendency to abbreviate responses to particular questions, or a tendency to shape one's responses in order to skip anticipated sections. In addition, there is also an interviewer-related component. On sensitive questions, in particular, interviewers can sometimes rush through the wording, with deleterious effects on the quality of the respondent's answer (for detailed examples, see Thompson, Nawab Ali, and Casterline [1982]; for a more general review, see Fowler and Mangione [1990]).

We evaluate whether the length of the KDHS interviews varied across the two translation modes using models that are structurally equivalent to equation [4] above. Our dependent variable in the models was the length of interview in minutes (set to its grand centered mean) for interviews completed in one visit. Because DHS does not make length of interview available for the 971 (12.6 percent) women whose interviews were conducted over more than one visit (out of the 7,702 women from all 12 teams listed in Table 1), these data were missing for an additional 71 of these women, and other control variables (in particular language data on another 204 women, our analysis is limited to 6,456 women. Explanatory variables included dummy identifiers for time of fieldwork, ethnicity, region, and urban residence.

Results are presented in Table 5. Controls have a largely expected impact. In particular, the length of the interview falls significantly from month 1 of fieldwork to month 3 (there is no difference from months 4 to 6). There are some regional and ethnic differences which likely reflect behavioral differences of relevance to the questionnaires. Interviews in Central Province, for example, were an average of 6 minutes shorter than those conducted in Nyanza and Rift Valley provinces (the reference category) , and 14 and 11 minutes shorter than interviews conducted in Coastal and Western provinces, respectively. This is consistent with the fact that Central province is the lowest fertility (rural) area in Kenya, allowing interviewers to skip lengthy sections on each child's health.[4]

Table 5 about here

In terms of the translation modes, the effects are minimal. In a baseline OLS regression (not shown in Table 5), spontaneously translated questionnaires are about two minutes shorter, and this is statistically significant. In the fuller three-level model shown in Table 5 they are less than half a minute shorter, and this difference is not statistically significant. On the other hand, the fuller model also shows that, net of controls in the fixed part of the model, the random intercept of length of the interview varies at both the cluster and district levels, though further modeling of this along the lines of equation [5] show that there is no significant difference in random *slope* effects across either clusters or districts (results not shown but available from the authors).

Summary

Overall, there appear to be some differences between the data collected in interviews in which there was full correspondence between the language of the questionnaire and the language of the interview and those in which the questionnaire was spontaneously translated by the interviewer. On the one hand, some of the results suggest little reason

[4] It is also possible that these ethnic and regional differences reflect linguistic differences in which accurate translations of the global template questionnaire could be communicated at different speeds in different languages. This seems much less important than the behavioral answer, however, since there are significant differences in length of interview between ethnic groups whose languages are closely affiliated (e.g., Kikuyu and Meru.)

for concern over spontaneously translated questionnaires. To begin, while interviewers' contribution to total survey error appears to be somewhat greater in the spontaneous translation mode, it remains at a low level. In addition, only in 2 out of 24 variables is there evidence of systematic difference in mean response values between interviews with language correspondence and spontaneous translation. Furthermore, there is no general difference in the length of the interview across the translation modes, nor are there signs of any complex variation in the relationship between length of interview and translation mode. On the other hand, there are much wider effects of translation mode on complex variation across clusters, interviewers and districts. This suggests quite heterogeneous effects of translation across these aggregates.

Impact of differential translation modes on analysis

In this final section we evaluate the extent to which translation modes affect analysis, focusing on two substantive questions long considered important in fertility-related research: what influences ideal family size (that is, ideal number of children)? And what are the characteristics of current contraceptive users? Note that our aim is not to address these questions from any new conceptual position. Rather, it is to specify models with a standard set of explanatory variables, and then explore whether analytic inference from such models varies when one takes into account the effects of the two translation modes that we compared.

The models are similar to those specified in equations [2] and [4], i.e., three-level models in which discrete measures of variance are estimated at the district, cluster and individual level. In relation to each of the dependent variables, two models are estimated. The first, which we refer to as the baseline model, specifies a number of standard explanatory variables (in addition to controls for timing of fieldwork, urban residence, region of residence, and ethnicity). Explanatory variables that the two models share are age, not currently married, number of children ever born, years of schooling, and a measure of wealth. The analysis of ideal family size adds an identifier for women who have ever used contraception. And the analysis of current use of contraception adds variables for currently pregnant, number of sons who have died, number of daughters who have died, and number of children under 5 in the respondent's household.

The second model, which we call the full model, adds a main effect for spontaneous translation of the questionnaire, as well as interaction terms between spontaneous translation and *each* of the explanatory variables used in the baseline model.

Table 6 about here

Results for both baseline and full models are presented in Table 6. We do not dwell on the baseline models since all effects are as expected from the fertility literature. Rather, our interest is in the translation-related variables added in each of the full models, and in the overall model fit.

With respect to the specific translation-related coefficients, the full model shows that a number of the interaction terms are statistically significant. For example, the negative relationship between ideal number of children and years of schooling is more pronounced for women interviewed with language correspondence than with spontaneous translation. In contrast, the negative relationship between ideal number of children and ever use of contraception is markedly more pronounced for women who were interviewed with spontaneous translation. Similarly, the indicator of wealth has a much stronger positive effect on current use of contraception where there is language correspondence (.04 per unit of wealth) than where the interview was spontaneously translated (.040 - .028 = .012 per unit of wealth). And the inverse is true—though with only borderline significance—for the effect of number of daughters who have died on contraceptive use (-.067 and -.067 + .043 = -.024 per deceased daughter, respectively). Finally, the positive effect of marriage on contraceptive use is more pronounced among those interviewed under spontaneous translation.

With respect to overall model fit, the bottom rows of Table 6 present Wald statistics and tests for difference between the baseline and full models, where degrees of freedom are the number of additional variables used in the full model over the baseline (respectively 7 and 10 across the two dependent variables). Notwithstanding the fact that a number of the interaction terms are not significant in themselves, the overall improvement in model fit is statistically significant at the .001 level in the analysis of ideal number of children, and .01 level in the analysis of contraceptive use.[5]

We interpret these effects as the product of differences in translation modes for two reasons. The first is that these are net of an array of controls (for urban residence, region of residence, timing of fieldwork, and ethnicity) which should have eliminated most of the selectivity issues discussed earlier. The second is that, as noted in Table 4, the different translation modes have no significant effect on response values on variables like years of schooling, currently married, children ever born, number of daughters who

[5] The .01 significance can easily be improved to .001 by dropping some of the non-performing interaction terms—results available from the authors—but for comparability we use the full model.

have died, ever used contraception, and wealth, where we see differential analytic effects by translation mode. In other words, we have no evidence that spontaneous translation had an effect on the measurement of each of these variables, but it does appear to have had an effect on the measurement of the relationship between each of them and one of the two dependent variables just analyzed.

Conclusion

Results do not point unambiguously in one direction or the other. On one hand we have shown that data collected in the spontaneous translation mode are only marginally different from those collected using the standardized translation mode. Specifically, we recorded differences in mean response value on only two of the 24 variables at the standard 5 percent significance threshold. In addition, there are borderline differences in the length of the interview. Finally, across all 24 variables interviewer-related error is 5.5 percent of the total variance in data generated with spontaneously translated questionnaires, and 3.9 percent of the total variance in data generated with full correspondence between the languages of the questionnaire and the interview. While this is not a negligible increase in relative terms, and may yield substantial "design effects" where the average number of interviews per interviewer is high (Fowler and Mangione 1990), these are relatively low levels of interviewer-related error in absolute terms. That is, they are also only a small part of total variance.

On the other hand, we have also shown that across the two modes there are statistically significant differences in higher-order variance on 12 variables, and that there is also some impact of translation mode on analysis. These are important results. The first suggests that there are substantial differences in translation mode effects across groups. In other words, in some groups—here represented by clusters, interviewers, or districts—the difference in translation mode has no effect on data. In other groups, it does affect data. Similarly, the analytic results suggest that the relative profiles of current contraceptive users and people with different ideal family sizes varies somewhat across the two translation modes in ways that may be important for both conceptual and policy-related reasons. For example, other things being equal, the wealth effect on contraceptive use among those interviewed with spontaneous translation was only 30% of what it was in interviews with language correspondence, and the positive effect of being married on contraceptive use was 29% greater in data from spontaneously translated questionnaires. Similarly, spontaneous translation increased the well-established effects of ever having used contraception (negative) and children ever born (positive) on ideal family size. However, a negative coefficient on the interaction term attenuated the known negative relationship between years of education and ideal family size.

Overall, then, the results suggest that marginal effects of translation mode on measurement error are somewhat magnified as the level of analysis becomes complex. In terms of practical advice, this can be crudely summarized in the following way: spontaneous translation seems to be relatively inconsequential for univariate statistics, but somewhat riskier when it comes to multivariate analysis. This has clear implications for the choice of data collection methods. Simply, if the ultimate goal of a particular data collection endeavor is to obtain general univariate statistics for a given population, and especially if this is to be done at low cost and by focusing efforts on maximizing sample size rather than on collecting a large amount of data per respondent, then it looks like spontaneous translation can legitimately be used. This is because, given careful interviewer training (of the type conducted by DHS), spontaneous translation can efficiently simplify data collection with no significant loss in data quality. This is a potentially important result given that policymakers in developing countries often need simple basic statistics to determine policy priorities.

Alternatively, if the goal is to collect data for more complex multivariate analysis, in particular where that analysis extends across districts and ethno-linguistic boundaries, results presented here suggest that spontaneous translation should be avoided. Rather, project leaders should insist on standardized translation of questionnaires. This is not, it should be noted, a ringing endorsement of the status quo as currently *practiced* in the field, since as seen here, a substantial proportion of data do not meet this standard, even where it is the declared policy of the data collection organization. But it is an endorsement of the current *ideals* that underlie standardized data collection.

In short, the results suggest that, as applied to DHS field practice or to that of other large scale data collection in multilingual settings (e.g., the World Bank's Living Standards Measurement Study), there are gains to formal translation. Project managers therefore need to provide more local-language questionnaires in each target area and, to the extent that they want to reduce measurement error—for example, of the type represented by interviewer-related error—lean toward a more standardized approach to interviewing over a more flexible, conversational approach. Similarly, results also suggest that, when engaging in analysis of DHS and equivalent data, researchers need to be conscious of how field decisions—however reasonable in their context—are often made under the radar of official practice. These decisions leave their imprint on data, shaping means and variances in ways that can impact estimated relationships among parameters.

Finally, and more generally yet, our analysis has also shown that DHS data, while not explicitly designed to explore methodological issues, can still shed light on them. This is important. After a long time-out, data collection issues in developing countries have recently begun to move back to the center of research focus: thus the growth in experiments on mode effects (Plummer et al 2004; Mensch, Hewett and Erulkar 2003) and also evaluations of the darker underbelly of developing country research—the widely practiced methods which violate established methodological norms, sometimes with surprising consequences (e.g., Sana and Weinreb 2005; Weinreb 2006). Overall, however, there remains a clear dearth of methodological research in developing country settings, all the more worrying given that the absence of other types of data collection infrastructures make surveys an all-the-more vital source of information for analysts and policy makers alike.

Appendix

The 24 DHS variables used in the analysis, with the reference numbers used in the figures, are listed below.

Respondent's household and background characteristics

1 Number of household members (C)

2 Number of children 5 and under in the household (C)

3 Number of women in the household eligible for the interview (C)

4 Drinking water is from a river or stream (D)

5 Education in single years (C)

6 Ownership of durables (C)

7 Currently married (D)

Fertility and contraceptive use

8 Age of respondent at first birth (C)

9 Total children ever born (C)

10 Number of living children (C)

11 Number of daughters who have died (C)

12 Ever terminated a pregnancy (D)

13 Has ever used any method of family planning (FP) (D)

14 Living children at first use of FP (C)

15 Currently using a method of FP (D)

16 Intends to use FP in the next 12 months (D)

General fertility-related knowledge and attitudes

17 Reports menstruation in last six weeks (D)

18 Considers radio messages about FP acceptable (D)

19 Does not know a source for FP method (D)

20 Heard FP-related message on radio in the last months (D)

AIDS-related information

21 Has heard of AIDS from friends or family (D)

22 Thinks a healthy person can have AIDS (D)

23 Ranks own chances of catching AIDS (C)

24 Reports only one sex partner as behavioral response to AIDS (D)

(C) Continuous variable

(D) Dichotomous variable

References

Axinn, William G., Thomas E. Fricke, and Arland Thornton. 1991. “The Microdemographic Community-Study Approach: Improving Survey Data by Integrating the Ethnographic Method.” *Sociological Methods and Research* 20:187–217.

Babbie, Earl. 2006. *The Practice of Social Research (11th edition)*. Wadsworth Publishing.

Bernard, H. Russell. 2000. *Social Research Methods: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage Publications.

Chen, Ang, Z. Liu, and C. D. Ennis. 1997. "Universality and uniqueness of teacher educational value orientations: A cross-cultural comparison between USA and China." *Journal of Research and Development in Education* 30(3):135-143.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage Publications, Sage Series of Quantitative Applications in the Social Sciences, No. 63.

Cordell, Dennis D., Joel W. Gregory, and Victor Piché. 1996. *Hoe and Wage: A Social History of a Circular Migration System in West Africa*. Boulder, CO: Westview Press.

Creswell, John W. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (2nd edition)*. Thousand Oaks, CA: Sage Publications.

Delavande, Adeline, and Hans-Peter Kohler. 2007. “Subjective Expectations in the Context of HIV/AIDS in Malawi.” University of Pennsylvania, Population Aging Research Center, PARC Working Paper 07-06.

Dijkstra, Wil. 1987. “Interviewing Style and Respondent Behavior: An Experimental Study of the Survey-Interview.” *Sociological Methods and Research* 16(2):309-334.

Fowler, Floyd J., and Thomas W. Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park, CA: Sage Publications.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Helleringer, Stéphane, and Hans-Peter Kohler. 2005. Social Networks, Risk Perceptions and

Changing Attitudes Towards HIV/AIDS: New Evidence from a Longitudinal Study Using Fixed-Effect Estimation. *Population Studies*, 59(3): 265-282.

Maynard, Douglas W., and Nora C. Schaeffer. 2002. "Standardization and Its Discontents." Pp. 3-45 in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, and H. van der Zouwen. New York: Wiley.

Mensch, Barbara S., Paul C. Hewett, and Annabel S. Erulkar. 2003. "The Reporting of Sensitive Behavior Among Adolescents: A Methodological Experiment in Kenya." *Demography* 40: 247–68.

Mitchell, Robert E. 1965. "Survey Materials Collected in the Developing Countries: Sampling, Measurement and Interviewing Obstacles in International Comparisons." *International Social Science Journal* 17:665-85.

Overton, John, and Peter van Dierman. 2003. "Using quantitative techniques," pp. 37-56 in Regina Scheyvens and Donovan Storey (eds.) *Development Fieldwork: A Practical Guide*. Thousand Oaks, CA: Sage Publications.

Plummer, Mary L., Daniel Wight, David A. Ross, Rebecca Balira, Alessandra Anemona, Jim Todd, Zachayo Salamba, Angela I. N. Obasi, Heiner Grosskurth, John Changalunga and Richard J. Hayes. 2004. "Asking Semi-Literate Adolescents about Sexual Behaviour: The Validity of Assisted Self-Completion Questionnaire (ASCQ) Data in Rural Tanzania." *Tropical Medicine and International Health* 9:737–754.

Sana, Mariano, and Alexander A. Weinreb. 2005. "An Experiment on Expert Field Guess." Paper presented at the 37th World Congress of the International Institute of Sociology. Stockholm, 5-9 July.

Sapsford, Roger, and Victor Jupp. 2006. *Data Collection and Analysis (2nd edition)*. Thousand Oaks, CA: Sage Publications.

Schaeffer, Nora C. and Stanley Presser. 2003. "The Science of Asking Questions." *Annual*

Review of Sociology 29:65–88.

Schober, Michael and Frederick G. Conrad. 1997. “Does Conversational Interviewing Reduce Survey Measurement Error?” *Public Opinion Quarterly* 61:576–602.

Snijders, Tom A. B.; Roel J. Bosker. 2003[1999]. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications.

Suchman, Lucy and Brigitte Jordan. 1990. “Interactional Troubles in Face-to-Face Survey Interviews.” *Journal of American Statistical Association* 85:232–253.

Thompson, L.V., M. Nawab Ali, and J. B. Casterline. 1982. *Collecting Demographic Data in Bangladesh: Evidence From Tape-Recorded Interviews*. London: International Statistical Institute & World Fertility Survey, Scientific Report No.41.

Ware, Helen. 1977. *Language Problems in Demographic Field Work in Africa: The Case of the Cameroon Fertility Survey*. London: International Statistical Institute & World Fertility Survey, Scientific Report No.2.

Weinreb, Alexander A. 2006. “The Limitations of Stranger-Interviewers in Rural Kenya.” *American Sociological Review* 71(6):1014-1039.

Table 1

Interviewers, interviewer assignments, completed interviews, and other pertinent frequencies, by interviewer team

Team Number	Total number of Intervie-wers	District assignments	Largest ethnic group (% of DHS respondents)	Total num. of interviews	Percent of interviews *without* language correspondence	Average number of interviews per interviewer
1	6	Kericho/Baringo/Trans-Nzoia/W.Pokot	Kalenjin (78.1)	705	18.2	117.5
2	5	Nandi/Elgeyo-Marakwet/Uasin Gishu	Kalenjin (77.7)	847	24.3	169.4
3	7	Kitui/Machakos	Kamba (95.1)	673	19.9	96.1
4	6	Nyandarua/Laikipia/Nakuru	Kikuyu (71.4)	433	16.4	72.2
5	5	Kiambu/Kirinyaga/Muranga/Nyeri	Kikuyu (93.5)	648	28.7	129.6
6	6	Bungoma/Busia/Kakamega	Luhya (86.1)	896	13.6	149.3
7	6	Kisumu/Siaya/South Nyanza	Luo (91.4)	812	39.3	135.3
8	5	Meru/Embu	Meru/Embu (93.7)	478	24.7	95.6
9	5	Kilifi/Kwale	Mijikenda/Kiswahili (91.1)	467	50.1	93.4
10	6	Mombasa/Taita Taveta	Taita/Taveta (35.9)	732	10.3	122.0
11	5	Nairobi/Kisii/South Nyanza/Nyamira	Kisii (69.5)	789	14.1	157.8
12	2	Nairobi/Kajiado/Narok	Kikuyu (27.8)	222	29.7	111.0
Total	64			7,702	23.0	120.3

Table 2
Distribution of questionnaires and interviews by language, by interviewer teams

Interviewer	% Questionnaires by language				% Interviews by language		
Team	Main	Other 1	Other 2	%	Main	Other	%
1	Kalenjin (57.9)	Kiswahili (42.1)		100.0	Kalenjin (65.6)	Kiswahili (33.5)	99.1
2	Kalenjin (64.0)	Kiswahili (35.7)		99.7	Kalenjin (72.9)	Kiswahili (26.7)	99.6
3	Kamba (79.3)	Kiswahili (20.5)		99.8	Kamba (94.0)	Kiswahili (4.3)	98.3
4	Kikuyu (81.9)	Kiswahili (17.9)		99.8	Kikuyu (70.6)	Kiswahili (26.8)	97.4
5	Kikuyu (69.9)	Kiswahili (30.1)		100.0	Kikuyu (96.0)	Kiswahili (3.4)	99.4
6	Luhya (80.7)	Kiswahili (19.1)		99.8	Luhya (83.9)	Kiswahili (15.4)	99.3
7	Luo (61.5)	Kiswahili (37.9)		99.6	Luo (94.3)	Kiswahili (4.6)	98.9
8	Meru/Embu (67.4)	Kiswahili (32.6)		100.0	Meru/Embu (88.7)	Kiswahili (11.3)	100.0
9	Mijikenda (47.3)	Kiswahili (39.0)	Masai (12.6)	98.9	Mijikenda (69.6)	Kiswahili (28.9)	98.5
10	Kiswahili (95.1)	Masai (4.9)		100.0	Kiswahili (94.7)	English (4.2)	98.9
11	Kisii (62.0)	Kiswahili (28.5)	Masai (5.2)	95.7	Kisii (63.8)	Kiswahili (34.2)	98.0
12	Kiswahili (49.1)	Masai (44.1)		93.2	Kiswahili (73.4)	Masai (21.2)	94.6

Table 3
Fieldwork Progression and Spontaneous Translation

		Clusters		
Fieldwork Month	Proportion of Interviews with ST	Number Visited	Number with over 75% of interviews with ST	Proportion with over 75% of interviews with ST
1	0.10	0	64	0.00
2	0.11	2	155	0.01
3	0.14	6	131	0.05
4	0.29	18	131	0.14
5	0.46	22	112	0.20
6	0.81	17	31	0.55

ST: Spontaneous Translation.

Table 4
Results from 24 Multilevel Regressions for the Spontaneous Translation Variable

Dependent variable		Coefficient	Standard error	Sig. level
1	Number of household members	0.152	0.092	#
2	Number of children 5 and under in the household	0.078	0.035	*
3	Number of eligible women in the household	0.033	0.030	ns
4	Drinking water is from a river or stream	0.007	0.013	ns
5	Education in single years	-0.071	0.115	ns
6	Wealth	-0.010	0.022	ns
7	Currently married	0.005	0.014	ns
8	Age of respondent at first birth	0.120	0.124	ns
9	Total children ever born	-0.034	0.057	ns
10	Number of living children	-0.024	0.051	ns
11	Number daughters who have died	-0.016	0.015	ns
12	Ever terminated a pregnancy	0.007	0.008	ns
13	Has ever used any method of family planning (FP)	0.002	0.015	ns
14	Living children at first use of FP	-0.059	0.083	ns
15	Currently using a method of FP	-0.002	0.014	ns
16	Intends to use FP in the next 12 months	0.011	0.018	ns
17	Reports menstruation in last six weeks	0.003	0.015	ns
18	Considers radio messages about FP acceptable	-0.013	0.010	ns
19	Does not know a source for FP method	0.022	0.014	ns
20	Heard FP-related message on radio in the last months	-0.043	0.016	**
21	Has heard of AIDS from friends or family	0.009	0.016	ns
22	Thinks a healthy person can have AIDS	-0.003	0.013	ns
23	Ranks own chances of catching AIDS	-0.027	0.029	ns
24	Reports only 1 sex partner as behavioral response	0.003	0.016	ns

** p <. 01

* .01 < p <. 05

# .05 < p < .10

Table 5
Results from Multilevel Regression of Length of Interview on Selected Controls

Controls	Coefficient	Standard error	Sig. level
Interview spontaneously translated (ST)	-0.44	0.59	ns
Month of fieldwork			
1	17.00	1.13	***
2	6.69	0.97	***
3	2.21	0.89	*
4-6	*reference*		
Region of residence			
Central	-5.57	2.58	*
Coast	8.21	2.63	**
Eastern	-3.82	2.66	ns
Western	5.30	3.01	#
Other regions (1)	*reference*		
Urban resident	-1.09	1.00	ns
Ethnicity			
Kalenjin	-3.30	1.21	**
Kikuyu	-3.10	1.13	**
Kisii	-5.97	1.58	***
Luhya	-1.43	1.10	ns
Other (2)	*reference*		
Variance estimates			
District	18.44	6.37	
Cluster	8.81	2.00	
Individual	264.20	4.84	
Constant	-2.22	1.44	
N	6,456		
Wald chi(2)	341.49		

*** p <. 001
** .001 < p <. 01
* .01 < p <. 05
# .05 < p < .10
(1) Other regions include Rift Valley, Nyanza, and Northern.
(2) Other ethnic groups include Embu, Kamba, Luo, Maasai, Meru, Mijikenda, Taita/Taveta, and other.

Table 6
Results from Multilevel Regressions of Ideal Number of Children and Current Use of Contraception on Selected Controls, Baseline Model and Full Model Including Spontaneous Translation Variable and All Interactions

	Ideal Number of Children				Currently Using Contraception			
Explanatory variables	Base model		Full model		Base model		Full model	
Fixed estimates								
Interview spontaneously translated (ST)			-0.018				-0.047	
			(0.146)				(0.040)	
Number of children <5 in household (HH)					-0.013	**	-0.009	
					(0.005)		(0.006)	
ST x children in HH							-0.015	
							(0.011)	
Age	0.017	***	0.016	***	0.002	**	0.002	*
	(0.003)		(0.004)		(0.001)		(0.001)	
ST x age			-0.001				0.000	
			(0.007)				(0.002)	
Years of schooling	-0.072	***	-0.078	***	0.023	***	0.022	***
	(0.006)		(0.006)		(0.001)		(0.002)	
ST x years schooling			0.029	*			0.005	
			(0.013)				(0.003)	
Wealth (1)	-0.101	***	-0.083	**	0.035	***	0.040	***
	(0.026)		(0.028)		(0.007)		(0.008)	
ST x wealth			-0.077				-0.028	*
			(0.011)				(0.014)	
Currently pregnant					-0.322	***	-0.333	***
					(0.018)		(0.021)	
ST x pregnant							0.046	
							(0.041)	
Currently married	0.201	***	0.208	***	0.198	***	0.211	***
	(0.041)		(0.047)		(0.011)		(0.013)	
ST x married			-0.039				0.062	*
			(0.096)				(0.027)	
Total children ever born (CEB)	0.100	***	0.091	***	0.028	***	0.027	***
	(0.011)		(0.012)		(0.003)		(0.004)	

ST x CEB			0.043	#			0.002	
			(0.024)				(0.008)	
Number of sons who have died					-0.041	***	-0.031	*
					(0.010)		(0.012)	
ST x sons died							-0.033	
							(0.023)	
Number of daughters who have died					-0.054	***	-0.067	***
					(0.011)		(0.013)	
ST x daughters who have died							0.043	#
							(0.024)	
Ever used contraception	-0.325	***	-0.269	***				
	(0.039)		(0.043)					
ST x ever used contraception			-0.248	***				
			(0.091)					
Variance estimates								
District	0.056		0.053		0.001		0.001	
	(0.022)		(0.021)		(0.001)		(0.001)	
Cluster	0.019		0.021		.0015		0.002	
	(0.011)		(0.011)		(0.001)		(0.001)	
Individual	2.100		2.100		0.159		0.158	
	(0.036)		(0.036)		(0.003)		(0.003)	
N	7466		7466		7470		7470	
Wald chi(2)	1568.04		1596.43		1690.39		1714.69	
Difference in Wald chi(2)			28.39				24.30	
Significance of difference in Wald chi(2)			<.001 [7 d.f.]				<.01 [10 d.f.]	

These estimates are net of urban/rural residence, region, ethnic identity, and timing of fieldwork.

(1) Measure of wealth is a 0-3 additive scale based on whether a household has electricity, piped water, and the respondent reports watching TV weekly.

Significance levels: *** p < .001 ; ** .001<p<. 01; * .01<p<. 05; # .05<p<.10. Standard errors in parenthesis.

FIGURE 1. Proportion of Variance Explained by Interviewers under Spontaneous Translation and under Language Correspondence

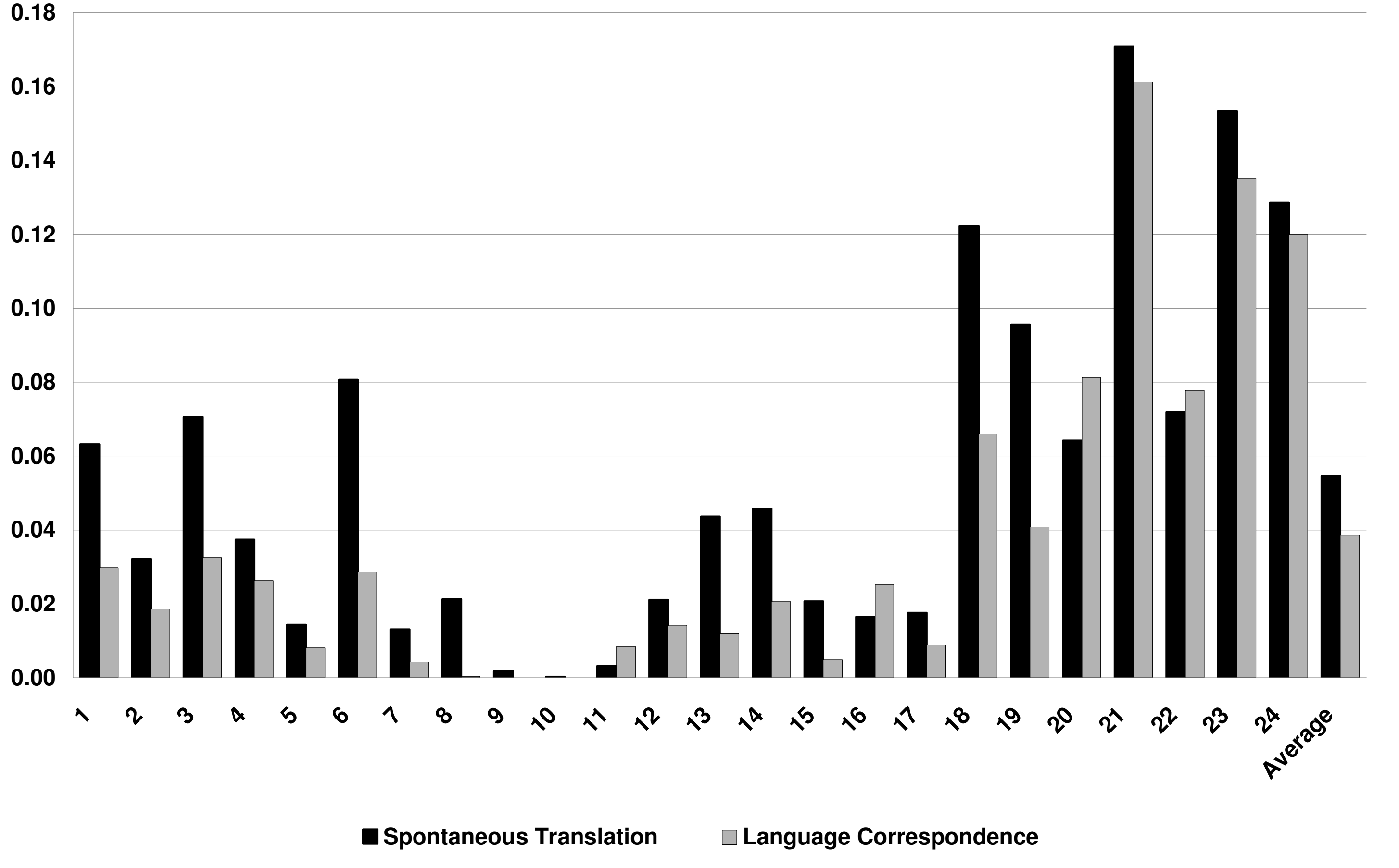
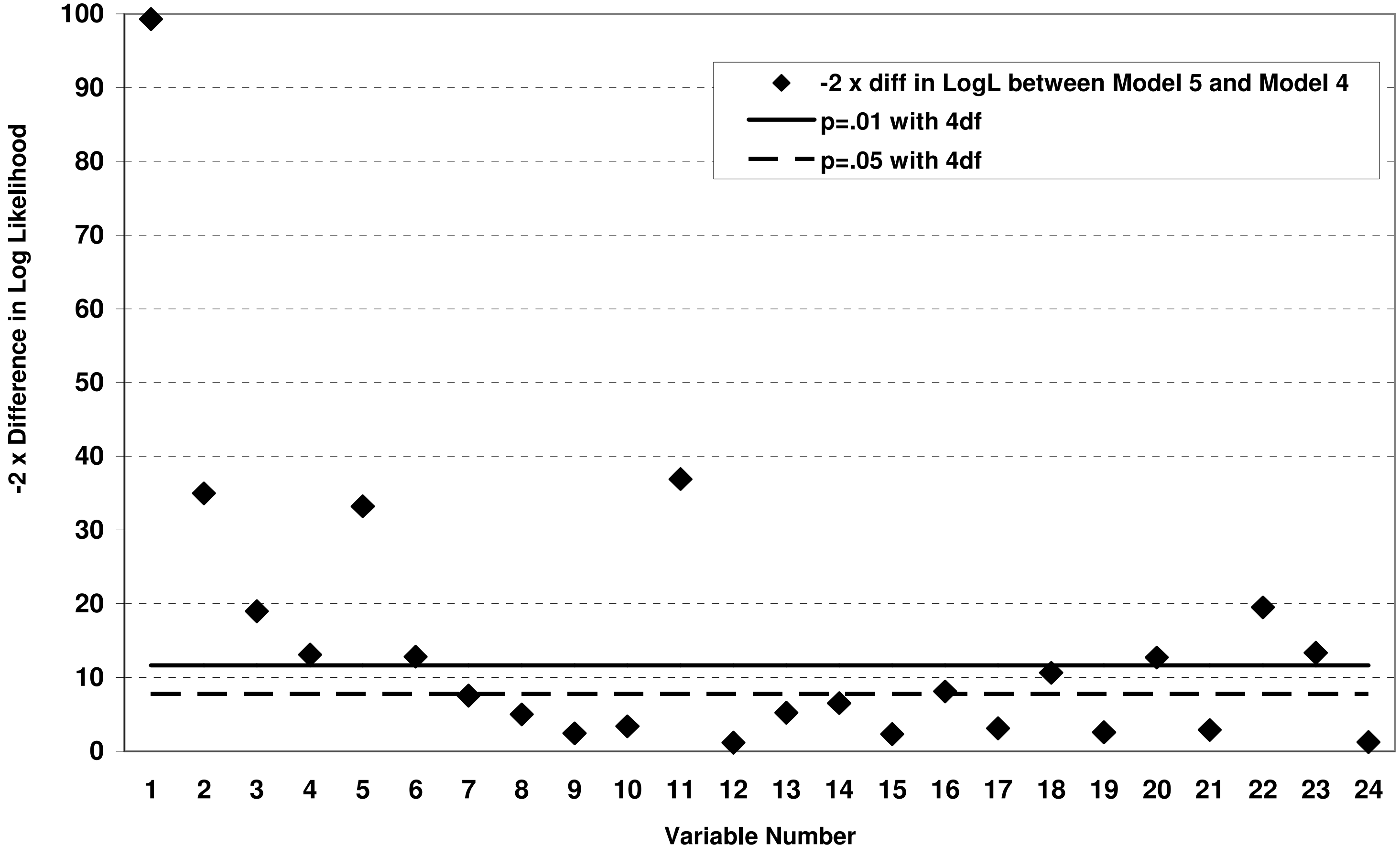


FIGURE 2. Comparison of Fit of Model 5 (complex variation at the cluster or district level) with Model 4 (no complex variation)


◆ -2 x diff in LogL between Model 5 and Model 4
p=.01 with 4df
p=.05 with 4df
-2 x Difference in Log Likelihood
Variable Number