STANDARDIZED VERSUS SPONTANEOUS TRANSLATION OF SURVEY QUESTIONS: AN ANALYSIS OF KENYAN DHS DATA

Alexander A. Weinreb
Mariano Sana

DRAFT, WORK IN PROGRESS

22 September 2006

Do not cite without the authors' permission

Submitted for presentation at PAA 2007

ABSTRACT

Demographic surveys implemented in multilingual settings routinely rely on a template questionnaire and its standardized translation to the main languages spoken by the respondents. This approach is consistent with the recommended pursuit of stimulus equivalence, and it intends to ensure that all respondents answer exactly the same questions. In some settings, however, the number of ethnic groups, and their corresponding languages, is too large to make this practice feasible. We analyze data from the Kenya 1998 DHS, in which about a quarter of interviews were conducted in a language other than that of the questionnaire held by the interviewer, so that the interviewers had to rely on their own spontaneous translation of the questions rather than on the project's standardized translation. We explore the effects of spontaneous translation on the proportion of variance that is across interviewers (the intraclass correlation coefficient rho) and on systematic differences in responses (bias). Preliminary results suggest that, although the former tends to increase with spontaneous translation, there are no clear bias effects. In general, results are heterogeneous and vary on a variable-by-variable basis.

Alexander A. Weinreb, Department of Sociology and Anthropology, Hebrew University, Jerusalem 91905. Email: awein@mscc.huji.ac.il. Phone: 972 3 529 1495

Mariano Sana, Department of Sociology and Louisiana Population Data Center, Louisiana State University. 126 Stubbs Hall, Baton Rouge LA 70803. Email: msana@lsu.edu. Phone: 225 578 1115

STANDARDIZED VERSUS SPONTANEOUS TRANSLATION OF SURVEY QUESTIONS: AN ANALYSIS OF KENYAN DHS DATA

Introduction

Over the last two decades there has been increasing recognition that old "stimulus equivalence" approaches, the key aim of which is to maximize the standardization of data collection procedures across respondents (Fowler and Mangione 1991), do not always generate the most valid data. Rather, non-standardized, "conversational," or "personalized" interviewing practices often perform better (Suchman and Jordan 1990; Schober and Conrad 1997; Maynard and Schaeffer 2002; Schaeffer and Presser 2003).

Although aspects of these new practices have been incorporated into data collection procedures in certain stand-alone survey projects in Less Developed Countries (LDCs) (e.g., the Malawi Diffusion and Ideation Change Project, MDICP—Watkins et al 2003, or the Mexican Migration Project, MMP—Durand and Massey 2004), larger gold-standard projects like the multinational and multiwave Demographic and Health Surveys (DHS) have yet to embrace them. Such, at least, is the official position. Unofficially, that is, during actual data collection, the situation is a little more complicated. The combination of high linguistic diversity and resource constraints has forced survey organizations in many LDC populations, especially in ethnically and linguistically diverse areas of sub-Saharan Africa, to quietly embrace a somewhat less standardized position than the one reported in official publications. Specifically, although questionnaires are carefully translated and back-translated into all major national languages, data-files show that in a substantial proportion of interviews, there is no correspondence between the language of the interview and the language of the questionnaire. Rather, interviewers spontaneously translate questions to respondents using an alternative shared language, or move back and forth between languages. Either way, both spontaneous translation and language-hopping are antithetical to older notions of stimulus equivalence, since they raise questions about the comparability of the data across interviews conducted with and without complete language correspondence. This in turn raises the possibility that a portion of the observed response variance on any given variable is a product of different interview-"mode" effects. In the standard mode, a project's template questionnaire is carefully translated into the language of the interview, yielding *full correspondence* between the language of the printed questionnaire and the language

of the interview. In the alternative mode, interviewers spontaneously translate from the language of their questionnaire into the language of the interview interaction. Here there is an *absence of correspondence* between the language of the printed questionnaire and the language of the interview. In both modes, it should be noted, questions are translated—the interaction could not proceed otherwise. The distinction lies in the type of translation.

Using Demographic and Health Survey (DHS) data from Kenya, our key aim in this article is to evaluate the effects of this difference in interview-mode on two indicators of measurement (i.e. non-sampling) error: the correlated component of response variance and systematic differences in response values. Our analysis has important practical implications for data collection in multilingual settings. Specifically, survey 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, interviewers and questionnaires on given languages. On the other hand, if spontaneous translation during an interview did increase non-sampling error beyond a level judged to be acceptable, as existing methodological norms claim, survey researchers should consider investing more resources—time and money—into translating survey instruments into more languages, and in training more language teams.

The Specific Problem

To administer a questionnaire to a nationally representative sample in multilingual countries while applying standard stimulus equivalence approaches, a research project should carefully translate its questionnaire(s) into all languages represented in the sample, and also match respondents who speak and prefer a given language to interviewers with the right language skills and the correct language questionnaire. Budgetary and other constraints make it nigh impossible to achieve this level of standardization. In actual field practice, therefore, a compromise position is typically—albeit quietly—adopted. First, questionnaires are translated into the languages of all major ethnic-linguistic groups rather than into all languages. And second, language-specific interviewer teams are assigned responsibility for given ethno-linguistic regions. Thus, where *A* is the dominant language in an area, all interviewers assigned to that area are *A*-speakers, and are outfitted with *A*-language questionnaires.

Three unstated assumptions underlie this compromise position: first, all residents of the area in which *A* is dominant are proficient speakers of *A*; second, the dialect of *A* spoken by the interviewers, many of whom are urban and college-educated, is easily understood by *A*-speakers in sampled rural areas; and third, as long as respondents are proficient in the language of the interview, the choice of language has no influence on their willingness to discuss, and the terms in which they will discuss, personal issues, behaviors, and attitudes.

There is little doubt that the last of these assumptions is flawed. There is considerable evidence that people feel more comfortable talking about personal matters in their first language. Moreover, language proficiency often carries distinct political messages. This is especially the case in various multiethnic African societies where political mobilization and competition are based on ethno-linguistic identity (Nash 1989; Horowitz 2001). As seen below, however, the first two assumptions are also problematic, at least when considering the Kenyan data used below, where the main language of the interview did not match the language of the questionnaire in 23.8 percent of cases.

Notwithstanding its impact on survey data there has been little formal evaluation of these types of language effects on survey research in developing countries. Rather, the explicit recommendation that researchers translate the questionnaire (e.g., Mitchell 1965; Ware 1977; Iyengar 1993) appears to draw on two sources. The first is the principle of stimulus equivalence. Based on that principle it seems reasonable to assume that in order to ensure that a given question will mean the same thing within and across diverse groups of respondents, we need to provide all respondents with the same standardized translation (and interviewing techniques, and so on).

The second source is the authority of precedent. Because large scale survey projects with an international imprimatur have translated questionnaires in this way since the early days of developing country survey research in the 1960s, it has become imbued with a level of "procedural legitimacy" (Suchman 1995:579). This in turn means that spontaneous translation is a type of deviation, with all the attendant risks associated with other types of deviant behavior: damage to professional reputations, to the authority of data and scientific claims, and so on.

Claiming scientific validity on the basis of either of these two sources, in the absence of direct empirical tests, is clearly problematic. Moreover, suggestive evidence that questionnaires need not be translated can be seen in the only study (of which we are aware) that evaluates

questionnaire-translation effects on survey data in a sub-Saharan African setting. Using data from a rural sample in Kenya, Bignami-van Assche, Reniers & Weinreb (2003) show that intraclass correlation coefficients (*rho*) associated with interviewers do not significantly differ "where a questionnaire was formally translated and where it was rendered into local language by the interviewers themselves." (p.60). While, for reasons outlined below, this finding must be taken with care, it is nevertheless an intriguing result which begs more focused analytic attention. [1]

Hypotheses

We define two alternative hypotheses, each drawing on an established literature, and each associated with different results

The "conventional hypothesis:" The mainstream methodological literature asserts that standardizing a question's wording across a given population increases data quality since it ensures that all respondents are provided with an equivalent stimulus. Among other things, this protects interviewers from feeling and projecting embarrassment, something to which respondents tend to react, to the detriment of reliable and accurate reporting. To the extent that this can be applied to the Kenyan setting, the conventional hypothesis suggests that we can find higher levels of non-sampling error in interviews which were spontaneously translated than in those which were translated as part of the project's standardization process.

The "conversational hypothesis:" Based on the literature on conversational interviewing, this hypothesis posits that we should find no systematic difference in levels of non-sampling error between these two groups of interviews. Alternatively, if differences are to be found,

[1] This case of questionnaire translation is part of a more general phenomenon with regards to data collection in developing countries. Not only is there a growing awareness that too little methodological research has been conducted in developing countries, with implications for the scale of measurement error in developing country data. There is also a growing awareness that field teams in developing countries, irrespective of what is written in official fieldwork protocols, develop non-standardized ad hoc solutions to a wide array of emergent problems while in the field, and that some of these solutions are at odds with accepted methodological principles even if they appear, both conceptually and empirically, to better suited to local interactional contexts. See Weinreb (2006) and Sana and Weinreb (2005) for some specific examples.

spontaneous translators will do better than their formal, standardized counterparts.

In order to evaluate the relative merits of these competing hypotheses, we explore two indicators of data quality across the two interview modes: the intraclass correlation coefficient *rho*, a standard measure of variation in role-restricted interviewer effects (Sudman and Bradburn 1974); and systematic differences in response values, an indicator of response bias.

Data

We use 1998 Kenya DHS data. Barring North Eastern Province, a large and politically unstable area that contains less than 5 percent of Kenya's population, this is a nationally representative survey with data collected from 7,881 women.

Data collection followed standard DHS procedures. Of relevance to this analysis, the research instruments were translated into the major languages in Kenya, and interviewers—all women—were assigned to language-specific teams that were then allotted responsibility for specific ethno-linguistic regions. 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. In either case, all members of a given interviewer team covered the same sample clusters and areas, and no sample cluster was distributed to more than one interviewer team. So there is some level of randomized interpenetration between interviewers and respondents within sample clusters, even if the randomization is not systematic. 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 6,614 interviews used in this analysis—the reasons for reduction from the full 7,881 are discussed below—were collected by 59 interviewers working in ten teams of between five and eight individuals, yielding an average of 114 interviews per interviewer. Each team was assigned to between two-to-four districts in which a single language was dominant. All interviewers in a given team, with the exception of some who were dropped for the purposes of this analysis (discussed below), worked in all the listed districts.

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 number of interviews	Percent of Interviews without language corres-pondence	Average number of interviews per Intervie-wer
1	6	Kericho/Baringo/Trans-Nzoia/W.Pokot	Kalenjin (78.1)	632	16.3	105.3
2	5	Nandi/Elgeyo-Marakwet/Uasin Gishu	Kalenjin (77.7)	847	24.3	169.4
3	8	Kitui/Machakos	Kamba (95.1)	697	19.9	87.1
4	6	Nyandarua/Laikipia/Nakuru	Kikuyu (71.4)	381	18.1	63.5
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	7	Mombasa/Taita Taveta	Taita/Taveta (35.9)	756	10.2	108.0
Total	59			6,614	23.8	113.7

KDHS questionnaires from the 1998 wave of data collection 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.

Table 1 also shows that in roughly 24 percent of the interviews—ranging from 10 percent in interviewer group 10 to 50 percent in interviewer group 9—there was no correspondence between the language of the questionnaire and that of the interview. The reasons for this can be inferred from data presented in Table 2, which compares the distribution of questionnaires by language with interviews by language. It suggests that one of the main reasons for the linguistic fluctuation appears to have been the inadequate supply of questionnaires in the local language. This was presumably based on an assumption that everyone in Kenya can speak the *lingua franca*, Kiswahili, generating a "compromise position" in which data collection managers erred on the side of printing too many Kiswahili questionnaires rather than too many questionnaires in languages associated with the specific ethnic groups.[2] Thus, the proportion of Kiswahili

[2] This fits with common stereotypes among Kenyan urban elites—among whom are the local survey practitioners who would provide foreign specialists with counsel on local matters such as linguistic proficiency in different parts of the country—that all folks in rural areas in Kenya speak Kiswahili (Watkins, personal communication). This is patently untrue. Data from the Kenya Diffusion and Ideational Change Project, for example, collected in rural Nyanza, show that only 55 percent of women

questionnaires given to all field teams exceeds the proportion of interviews conducted in Kiswahili.

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	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	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	Meru/Embu (88.7)	Kiswahili (11.3)	100
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	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

The data were manipulated in a number of ways prior to analysis, the net result of which was to reduce total sample size from 7,881 to 6,614 women.

First, in order to allow for reliable comparisons within teams and research areas, all interviews conducted by interviewer team 11 were dropped (n=789). This included areas in Kisii, Nyamira, South Nyanza and Nairobi districts, in all of which a different local language predominates, making interpretation of differences difficult. For similar reasons, two interviewers who worked in three districts but with some difference in proportion of interviews belonging to each place were also dropped from the data.

Second, because estimates of *rho* are sensitive to interviewer quality, we did not include data collected by thirteen relatively low productivity interviewers, on the assumption that they were not full-time interviewers or had been relieved of their positions after producing a small number of unsatisfactory questionnaires. In total, these thirteen individuals completed 114 interviews, roughly one-tenth of the average for the remaining 59 interviewers.

It is important to note the key weakness of these DHS data but equally their advantage over all other data collected in similar settings. The weakness is simply that they are from a non-experimental study, rather than a study that systematically randomized interviewer assignments. We deal with this by estimating discrete models for each of the ten interviewer teams identified

claimed to be able to speak Kiswahili (own calculation – see http://kenya.pop.upenn.edu for more on the KDICP and for access to data).

above. Given the standard field procedure used in the Kenya DHS—as mentioned earlier, all members of a given interviewer team covered the same sample clusters and areas—we can reasonably assume that systematic differences in overall response variance across all interviewers from a single team are a product of interviewer-related measurement error rather than a manifestation of simple or random response variance. In addition, the DHS structure also yields a particular advantage over the KDICP data used to explore this issue in another sub-Saharan setting. Specifically, Bignami-van Assche et al (2003) compared the intraclass correlation coefficient *rho* across waves of a longitudinal survey, with different teams of interviewers, slightly different training procedures and length of training, and so on. Here, in contrast, we compare levels of measurement error generated by the same individual interviewers across two different modes of data collection: a questionnaire in the language of the interview, and a questionnaire in a different language. In addition, here the comparison is within the same data collection period rather than across two different survey waves.

Methods

We selected 22 variables representing three key categories of survey questions: Household and background characteristics; fertility and contraceptive use; and more general fertility-related knowledge and attitudes. The specific variables within these categories can be seen in Appendix A. They include many core variables used in demographic analysis over the last four decades. In relation to each of these 22 variables, we then evaluated two possible interview-mode effects associated with the correspondence, or lack thereof, between language of interview and language of questionnaire. First we focused on role-restricted interviewer effects, then on systematic differences in response values. Figure 1 illustrates the simple structure of the data.

Figure 1
Structure of the data

Team N (N=1,2...11), Variable Y (Y=1,2... 22)						
Interviewer 1		Interviewer 2		...	Interviewer j	
Language correspondence	Spontaneous translation	Language correspondence	Spontaneous translation		Language correspondence	Spontaneous translation
Respondent 1	Respondent 1	Respondent 1	Respondent 1		Respondent 1	Respondent 1
Respondent 2	Respondent 2	Respondent 2	Respondent 2		Respondent 2	Respondent 2
...	...	...	...		...	...
Respondent *i*	Respondent *i*	Respondent *i*	Respondent *i*		Respondent *i*	Respondent *i*

i. Role-restricted interviewer effects

A standard measure of role-restricted interviewer effect, that is, the variance on a given variable which arises from interviewer's professional behavior, is the correlated component of response variance, also frequently referred to as the intraclass correlation coefficient or *rho*. This is typically estimated in a one-way random effects analysis of variance (ANOVA), which can be expressed as (Snijders & Bosker 2003[1999]:17):

$$Y_{ij} = \mu + U_j + R_{ij} \qquad (1)$$

where each observation concerning variable Y for individual *i* within group *j* results from the sum of the population grand mean μ, a group-specific effect U_j and a residual effect for individual *i* within group *j*. The group effects Uj have a mean of zero and variance τ^2 (the between-group variance), while the residuals have mean zero and variance σ^2 (the within-group variance). The total variance of Y_{ij} is equal to the sum of these two variances, and the intraclass correlation coefficient for a given variable Y is defined as:

$$rho(Y) = \tau^2 / (\tau^2 + \sigma^2) \qquad (2)$$

or the between-group variance as a proportion of the total variance.

We generated a pair of *rhos* per variable for each team of interviewers. Where interviews were conducted in the same language as the questionnaire, we refer to rho^S. Where they were conducted in a language other than that of the questionnaire, we refer to rho^D. In each case, a team-specific *rho* estimated the proportion of the total variance in variable *Y* that was due to correlations between all interviewers working in the team of reference under given language conditions: standardized translation (rho^S) or spontaneous translation (rho^D).

We estimated two sets of rho^S and rho^D. The first set consisted of team- and variable-specific estimates using all data from the 59 interviewers net of the data modifications already mentioned. That is, discrete ANOVAs were specified for each of the ten interviewer-teams on each variable of interest, and under the two interviewing conditions: with (rho^S) and without (rho^D) language correspondence between the questionnaire and the interview. The second set was equivalent, but limited the sample population to members of the dominant ethnic group in

areas interviewed by teams 1-3 and 5-7. In each of these areas, between 78 - 95% of the sampled population self-identified as a member of a single dominant ethnic group (see Table 1). Restricting the analysis to these six areas allows us to evaluate the extent to which our initial results were affected by the study's non-experimental design. Specifically, on the assumption that each self-identifying member of the dominant local ethnic group—into whose language DHS translated the questionnaire—would speak the local language as a mother-tongue, the only reason for the lack of correspondence between the language of the interview and the language of the questionnaire would be the lack of a local language questionnaire. Information laid out in tables 1 and 2 suggests that this was the case. In other words, this second set of estimated *rhos* allows us to discount the effects of unobserved ethnic heterogeneity—which has implications for actual behavior about which the questions ask—on the first set of estimates.

Throughout this analysis, the "conventional hypothesis" predicts that *rho* will be lower where the language of the questionnaire and the language of the interview are the same (rho^S), than where the languages are different (rho^D). This results from assuming that interviewers translate uniformly across their own interviews, so that within-group variance is not affected by spontaneous translation, but they translate differently from each other, therefore increasing between-group variance. The conversational hypothesis, on the other hand, predicts that there will not be a difference between team- and variable-specific rho^S and rho^D.

ii. Systematic differences in response values

In each setting (or team of interviewers), and for each of the 22 variables, we used a simple regression framework to identify systematic differences in response values associated with the change in interview-mode. For any given variable *Y*, a systematic difference is indicated by a statistically significant estimate for a dummy variable indexing *correspondence* between the language of the questionnaire and the language of the interview. The specific choice of model depends on the type of variable: We used probit models where the dependent variable was dichotomous and OLS regression where it was not. Age and years of education were included as additional controls. We adjusted standard errors on all models to control for DHS cluster sampling.

Notice that the regression coefficients are not informative for our purposes. Given the difference in models and in units, the coefficients are not comparable across models. Instead, we

count the number of statistically significant regression coefficients for the dummy variable across models (to get an idea of to what extent language correspondence matters) and we focus on the sign of those coefficients (in order to evaluate whether the direction of change, for a given variable, varies across teams).

Results

i. Role-restricted interviewer effects

For each combination of interviewer team and variable, *rhos* were estimated twice: first, where there was language correspondence, yielding rho^S; and second, where the interviewer spontaneously translated from a questionnaire written in a different language, yielding rho^D. Over the ten teams, this generated a total of 440 models of the type presented as equation (1) with their corresponding *rhos*. As discussed above we then generated a parallel set of rho^S and rho^D using the restricted sample of respondents—that is, members of the dominant ethnic group where 78 - 95% of the district's sampled population self-identified as a member of a single dominant ethnic group— limiting us to interviewer teams 1-3 and 5-7. This second set comprised an additional 264 models. Given the number of models, we focus on aggregates and averages rather than specific models.

The most general result is depicted in Table 3, which presents the mean *rho* across all 22 variables, under the two conditions—with and without language correspondence—and with the full and restricted samples. Where there is correspondence between the language of the interview and the language of the questionnaire, mean *rho* across all 22 variables was .049 across the full sample and .037 across the more restricted sample. In the absence of correspondence between the language of the interview and the language of the questionnaire—that is, when these same interviewers spontaneously translated questions—mean *rho* was .083 and .067 respectively. At this general level, therefore, the proportion of total variance that derives from variation among

Table 3
Average rho across 22 variables, by language correspondence and type of sample

Language correspondence	Full sample	Restricted sample
Yes (*rhoS*)	0.049	0.037
No (*rhoD*)	0.083	0.067

interviewers is considerably greater where an interviewer spontaneously translates from an existing questionnaire than where she is reading off the standard project translation of the questionnaire. This is consistent with the "conventional hypothesis."

More detail is shown in Figure 1, which presents the mean *rho* across all 22 variables by interviewer team. Specifically, discrete mean *rhos* for the full sample are presented for all ten teams: the top bar in each cluster represents mean rho^S; the second bar from the top represents mean rho^D. The seven clusters where analysis was repeated on a restricted sample (numbered 1-3, and 5-7) have an additional two bars representing, respectively, rho^S and rho^D. Overall, Figure 1 shows that mean *rho* is higher in the absence of language correspondence (rho^D) in nine out of ten interviewer teams—we discuss the exception, interviewer group 9, below. It is also higher in all six of the teams used in the restricted analysis. Here too, therefore, overall results are consistent with the conventional hypothesis, though there are clearly unexplained differences among teams. For example, on the full sample, excess *rho* associated with the lack of language correspondence is .053 for one of the Kikuyu language teams (team 4, where 0.093 – 0.040 = 0.053), while it is only .006 for its linguistic counterpart (team 5). Similarly, it is .044 for Kiswahili team 10 and -.020 for Kiswahili team 9.

Figure 1. Mean *rho* across 22 variables, by interviewer team-dominant language, language correspondence, and type of sample

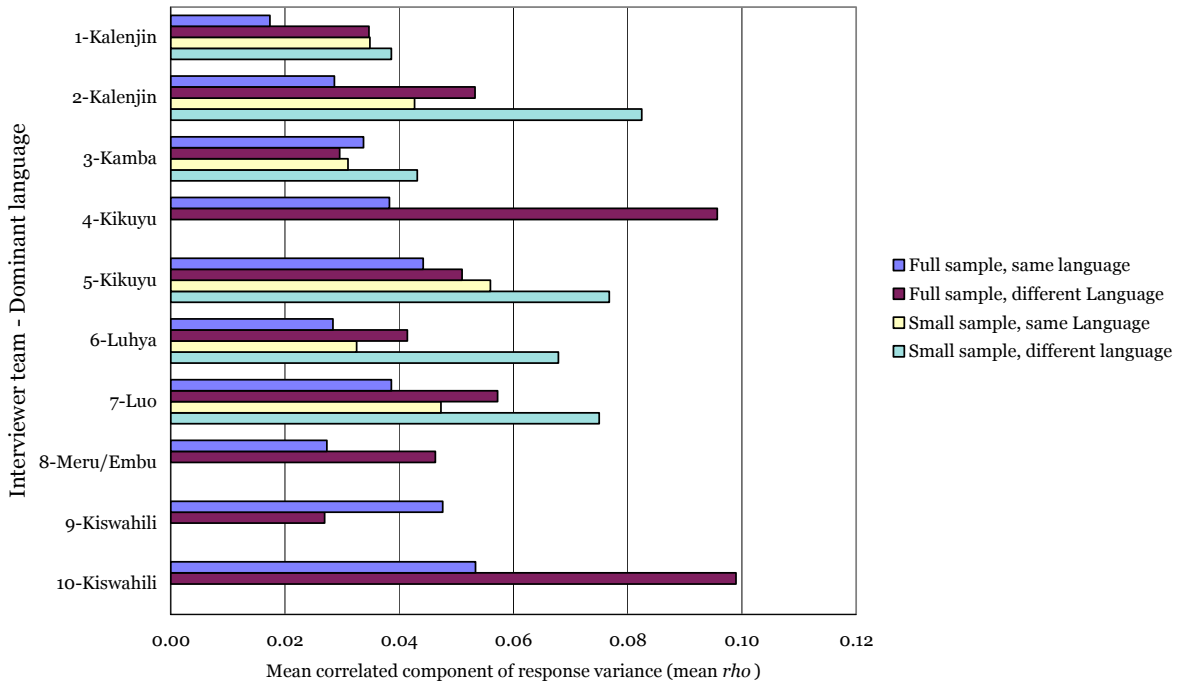
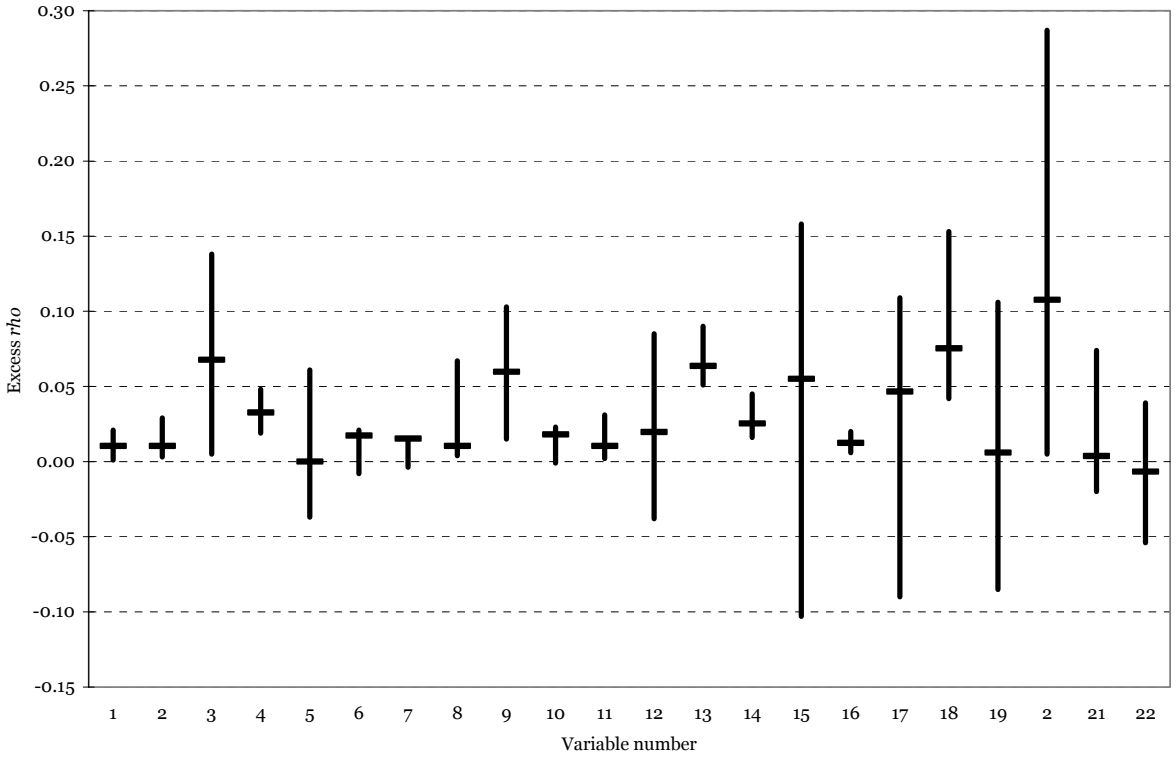


Figure 2 presents, by variable and across interviewers, the full range of changes in *rho* associated with the substitution of spontaneously translated questionnaires for the official DHS instruments. On 12 variables *rho* was higher on spontaneously translated questionnaires across all interviewer groups (i.e. the full range of values on the vertical axis are positive), and on another four variables, the lowest excess rho is only slightly below the zero line. This is consistent with the general trend towards larger rhos when there is no language correspondence between the interviews and the questionnaire. Yet, some variables show a very wide range of values for excess rhos, with a substantial part of this range falling below zero. In other words, for these variables, spontaneous questionnaire translation seems to have mixed effects, often reducing, rather than increasing, non-sampling error. More interesting, these variables do not appear to be randomly distributed across the 22 questions, but they are clustered on the right side of the figure, where we have placed the variables that deal with family planning or fertility-related questions. The four with the lowest low excess rhos -- which we define as those where $rho^D - rho^S \leq -.05$ –concern respondents' preferred future method of family planning, their

Figure 2. Mean, highest and lowest excess *rho* across interviewers, by variable



knowledge of the ovulatory cycle, and the acceptability of TV messages about family planning. The presence of low *rhos* on spontaneously translated questions of this type—generally thought to be more sensitive questions than those about household composition and characteristics explored by the variables plotted on the left side of the figure—is somewhat contrary to the conventional hypothesis.

In general, however, although there is some variation across questions and at least one exceptional team (discussed below), interviewers generated higher levels of *rho* when they spontaneously translated questionnaires than when they used questionnaires which matched the language in which they conducted the interview.

ii. Systematic differences in response values

Differences in the value of responses across the two interviewing modes were evaluated separately for each interviewer team in relation to each of the selected 22 variables. As explained in the methods section, the result of interest is given by the significance level of the estimated coefficient of a dummy variable indexing language correspondence in the OLS and Probit models. A significant and positive (or negative) estimate for the dummy variable will result in a higher (or lower) predicted value for the variable in question. Summary results for the 220 models in this analysis are presented in Table 4 and Figure 3.

Table 4 shows the number of interviewer teams (or, equivalently, the number of regressions) in which we estimated statistically significant differences in predicted values between the two interviewing modes, for each of the 22 variables. First and foremost, we see that significant differences were not very common. Overall, significant differences were found in 28 of the 220 regressions (12.7%). Table 4 also hints at some heterogeneity across variables. While language correspondence (or lack there of) between interview and questionnaire had no significant effect on seven variables, on three of the 22 variables—all describing household or background characteristics—the effect of interview-questionnaire language correspondence was recorded to be both significantly positive *and* significantly negative, depending on the interviewer team.

This heterogeneity emerges more clearly in Figure 3, which presents, for each of the 22 variables, the estimated coefficient for the dummy variable indexing language correspondence in each of the ten regressions (interviewer teams). The values on the vertical axis are of no interest, as the estimated coefficients are interpreted differently in OLS and Probit regressions and the

Table 4
Number of interviewer teams in which language correspondence between the interview and the questionnaire significantly affected the predicted outcome, by variable and direction of change

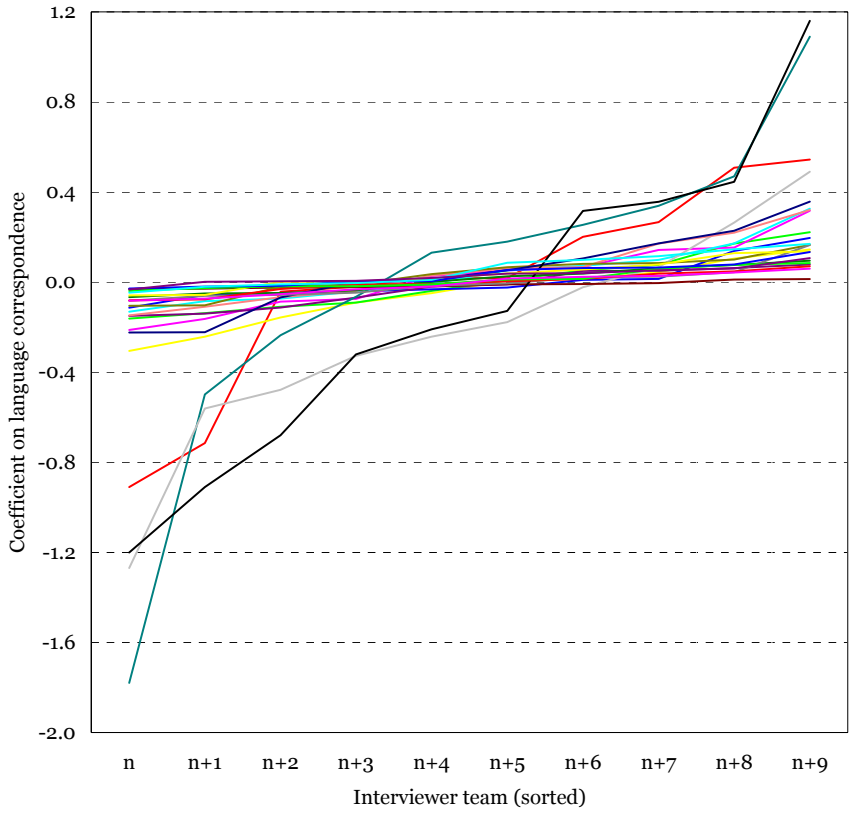
	Number of interviewer teams were effect was:			
Variable	Significant and positive	Significant and negative	Significant in either direction	Non-significant
Household/Background characteristics				
1. Number of household (HH) members (1)	0	0	0	10
2. Number of children 5 and under in the HH (1)	0	0	0	10
3. Number of eligible women in the HH (1)	1	0	1	9
4. Drinking water is piped into home (2)	2	1	3	7
5. Drinking water is from a river or stream (2)	2	2	4	6
6. Time to get to water source (1)	2	0	2	8
7. HH owns a radio (2)	0	2	2	8
8. Education in single years (1)	1	1	2	8
9. Age of respondent at first birth (1)	0	1	1	9
Fertility & contraceptive use				
10. Total children ever born (1)	0	0	0	10
11. Ever terminated a pregnancy (2)	0	0	0	10
12. Has ever used any method of family planning (FP) (2)	1	0	1	9
13. Living children at first use of FP (1)	0	0	0	10
14. Intend to use FP in the next 12 months (2)	1	0	1	9
15. Would prefer to use method (coded 1-13) (1)	1	0	1	9
Fertility-related knowledge & attitudes				
16. Reports menstruation in last six weeks (2)	0	0	0	10
17. Claims *not* to know ovulatory cycle (2)	3	0	3	7
18. Considers radio messages about FP acceptable (2)	1	0	1	9
19. Considers TV messages about FP acceptable (2)	1	0	1	9
20. Does not know a source for FP method (2)	0	1	1	9
21. Heard FP-related message on radio in the last months (2)	4	0	4	6
22. Heard FP-related message on TV in the last months (2)	0	0	0	10
Total	20	8	28	192

Notes:
(1) Estimated coefficient from OLS regression. (2) Estimated coefficient from probit model.
The reference group in both OLS and probit models are survey responses where the language of questionnaire and language of interview are different. Data are from Kenya DHS 1998. Statistical significance was established at the conventional alpha=0.05 level. Control variables on all models with the exception of (8) and (9) are age and schooling, with standard errors adjusted to control for DHS cluster sampling.

variables are measured in a variety of units. Instead, our interest focuses on the range of effects recorded. To see this clearly we have sorted the regression coefficients for the dummy variable from smallest to largest. In other words, each line represents one of the 22 variables and each point on the line represents the coefficient for the language correspondence dummy for a given interviewer group. Confirming the hints of heterogeneity presented in Table 4, Figure 3 shows that, as all lines cross the zero divide, on none of the 22 variables is there evidence of systematic bias across all 10 interviewer groups. Rather, on all variables, respondents interviewed by some interviewer groups give lower value responses where there is language correspondence than

where the interview is conducted in a different language from the questionnaire, while respondents interviewed by other interviewer groups give higher response values.

Figure 3. Estimated coefficients for the dummy variable indexing language correspondence, across 22 dependent variables and 10 interviewer teams sorted by smallest to largest coefficient



Discussion

Overall results tend to support what we have referred to as the "conventional hypothesis." That is, consistent with the mainstream methodological literature outside sub-Saharan Africa, interviews which were spontaneously translated tended to reduce the reliability of measurement, as indicated by the increase in the correlated component of response error associated with interviewers. On the whole, this reduction in reliability was not associated with any concomitant shift in systematic differences in response value, meaning that there are no grounds for thinking that more flexible, spontaneous translations increased the accuracy of measurement in any systematic way. On the flipside, the same absence of effects means that spontaneously translated interviews do not appear to have systematically reduced the accuracy of measurement. In terms

of the two competing hypotheses described above, therefore, our results appear closer to the "conventional" than to the "conversational" hypothesis.

Results also point to two types of heterogeneity, however. First, certain interviewer teams or, since we have no way of distinguishing between these aggregate identifiers, perhaps the areas or the cultural or linguistic groups among which they worked, appeared more or less prone to differences in *rho* or bias than others. For example, the difference in interviewing mode generated much greater differential *rho* in the Meru & Embu area (interviewer team 8) or Kalenjin areas (interviewer team 2) than in the neighboring Kamba area (interview team 3). Similarly, there was one occasion on which the difference in interview mode generated a statistically significant difference in response values in one of the Kalenjin team of interviewers (team 1), in comparison to six occasions in the other Kalenjin team. Assuming that all teams underwent standardized training, we suspect that these differences are a product of differential response behavior rather than interviewer behavior. For example, the subtext associated with a given question or its method of asking would vary from one place to another. Either way, it is clear that these meanings are the same within a given ethnic group—for example, there are considerable differences between the two Kalenjin and Kikuyu interviewer teams. Similarly, there seems to be no obvious link between these two types of error and the areas' or ethnic groups' national political associations, historically considered important determinants of identity and behavior in Kenya (e.g., Bates . . . Haugerud . . . ; Weinreb 2001). In particular, the Luo areas—political outsiders from the late late 1960s until the late 1990s—seem no more nor less prone to these effects than other groups which have been more favored (e.g., the Kikuyu and Kalenjin).

A second type of heterogeneity was related to variables themselves. There were at least a few variables on which there appeared to be systematic bias across the two interview modes. Four of the ten interviewer teams, for example, received higher reports that the respondent "had heard a family planning message on the radio in the last few months" when there was language correspondence between the interview and questionnaire. Three had higher estimated time to a water source. Three had lower reported ownership of a radio. In each of these cases, no team received reports that were statistically significant in the opposite direction, suggesting that a social desirability bias on these particular variables tends to act in one direction only, though not across all teams, as can be inferred from Figure 3. On the other hand, to the extent that this bias

exists, it does so at the district rather than national level. We return to this point below.

Conclusion

The implications of our analysis are clear. Notwithstanding the increasing acceptance of conversational styles of interviewing in general, in this Kenyan setting it still appears better for research projects which, like the Demographic and Health Surveys, draw on a nationally representative sample, to formally translate questionnaires than to rely on interviewers to translate them spontaneously. Given the multiethnic and multilingual nature of most other countries in sub-Saharan Africa, we feel comfortable generalizing from this one national setting to others, at least until evidence to the contrary becomes available.

Taking this one step further, our analysis suggests that DHS-type surveys should reduce the number of spontaneously translated interviews through a more careful matching of respondents of given language characteristics and translated questionnaires. In the 1998 KDHS used here, 23.8 percent of all questionnaires—and 39 percent in Luo areas (Table 1)—were spontaneously translated. Given that DHS fieldwork procedures, like those of equivalent international surveys, result in interviewers typically conducting a high number of interviews—more than 100 in these data—we suspect that the increase in *rho* associated with the spontaneous translation of questionnaires generates an unacceptable increase in non-sampling error. The reason is that, as is well known, the contribution of *rho* to overall variance increases exponentially as the mean number of interviews increases (see Fowler and Mangione 1991).The inevitable result of this is frustration of analysts' attempts to identify relationships between measured variables (a "type-1" error).

On the other hand, while this may all be true for national samples, the heterogenous patterns among teams / areas suggests that there may be areas within a given country in which distinct interactional patterns favor spontaneous translations over formal, written ones. Else, there may be particular types of questions in particular types of areas in which this is the case. This is of little concern for nationally focused studies. But the increasing frequency with which local studies are being conducted in demography and related disciplines—in particular, longitudinal studies—means that researchers should not automatically assume that a formally translated questionnaire will generate lower *rho* on all variables of interest in their single culture area. Rather, it seems wise to explore the interactional patterns through which *rho* is generated during

fieldtesting. That seems to be the best way to assess whether their site of choice fits the normal pattern, in which they would carefully translate the template questionnaire, or whether it is one of the exceptional cases.

Finally, and more generally yet, our analysis also shows that DHS data, while not designed to explore methodological issues, can still shed light on them. Given the dearth of such research in developing countries where, ironically, the absence of other types of data collection infrastructures make surveys an all-the-more vital source of information for analysts and policy makers, this is important to acknowledge. There is yet much to learn about data collection and measurement error in such settings.

REFERENCES

Bignami-van Assche, Simona, Georges Reniers, and Alexander A. Weinreb. 2003. "An Assessment of the KDICP and MDICP Data Quality: Interviewer Effects, Question Reliability and Sample Attrition." *Demographic Research* SC1:31–75.

Durand, Jorge and Douglas S. Massey. 2004. "Appendix: The Mexican Migration Project", pp.321-336 in Durand, Jorge and Douglas S. Massey (editors), *Crossing the Border: Research from the Mexican Migration Project*. New York: Russell Sage Foundation.

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

Horowitz, Donald L. 2001. *Ethnic Groups in Conflict*. Berkeley: University of California Press.

Iyengar, Shanto. 1983/1993. "Assessing Linguistic Equivalence in Multilingual Surveys." Pp. 173-182 in *Social Research in Developing Countries: Surveys and Censuses in the Third World*, edited by Martin Bulmer and Donald P. Warwick. London: UCL Press.

Maynard, Douglas W. and Nora Cate Schaeffer. 2002. "Standardization and Its Discontents." Pp. 3–46 in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by Douglas W. Maynard, Hanneke Houtkoop-Steenstra, Nora Cate Schaeffer, and Johannes van der Zouwen. New York: John Wiley and Sons.

Mitchell, Robert E. 1965. "Survey Materials Collected in Developing Countries: Sampling, Measurement, and Interviewing Obstacles to Intranational and International Comparisons." *International Social Science Journal* 17: 665-685.

Nash, Manning. 1989. *The Cauldron of Ethnicity in the Modern World*. Chicago: University of Chicago Press.

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.

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.

Suchman, Lucy and Brigitte Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of American Statistical Association* 85:232–53.

Sudman, Seymour and Norman M. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis*. Chicago, IL: Aldine Publishing Company.

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

Watkins, Susan C., Hans-Peter Kohler, Jere R. Behrman, and Eliya Zulu. 2003. "Introduction to the MDICP." *Demographic Research* SC1:1– . . .

Weinreb, Alexander A. 2006 (forthcoming). "The Limitations of Stranger-Interviewers in Rural Kenya." *American Sociological Review* 70(6).