Using vignettes to improve cross-population comparability of health surveys: concepts, design, and evaluation techniques

Joshua A Salomon

Ajay Tandon

Christopher JL Murray

Global Programme on Evidence for Health Policy Discussion Paper No. 41

World Health Organization
November 2001

I. INTRODUCTION

One of the key challenges in the analysis and interpretation of health survey data is the comparability of answers to questions that use ordered categorical response scales. Even for instruments with established reliability and validity, the problem of cross-population comparability remains as a consequence of differences in the ways that individuals understand and use the available responses for a given question. We may conceptualise these differences as resulting from individual variation in the mapping from an unobserved continuous latent scale (for example, level of mobility) into a set of discrete categorical responses. In this framework, an individual's observed characterization of a particular level on the latent variable will depend on that individual's cutpoints, which are threshold values on the latent scale that mark the transition from one categorical response to the next. There are numerous empirical examples that suggest that response category cutpoint shifts hinder the meaningful interpretation of health survey results [1-3].

Strategies for enhancing the cross-population comparability of health surveys require the augmentation of both existing instruments for data collection and existing statistical models for data analysis. In this paper, we introduce the concept of vignettes as a new component of survey instruments that allows adjustment for response category cutpoint differences in ordinal self-reported data in order to improve the comparability of these data.

Standard statistical models for ordinal data, such as the ordered probit model, cannot allow for variation in response category cutpoints. Tandon et al. [4] have described adaptations of these standard models to incorporate systematic cutpoint shifts as functions of some defined set of covariates. Without the introduction of exogenous information, however, these models could not allow cutpoints to vary in relation to the same variables as those used in modeling mean values on the latent variable of interest. In other words, these models applied to self-report survey data alone do not allow us to recognize that individuals in Denmark may have both different levels of health status and different expectations for health status relative to individuals in Morocco.

This paper describes the use of vignettes as a source of additional information that may be used in conjunction with the hierarchical ordered probit (HOPIT) model [4] in order to adjust self-reported responses into cross-population comparable measures. We present the concept of vignettes generally and give examples of vignettes from the WHO Multi-Country Household Survey Study [5], then explore a range of practical issues on the design, application and formal evaluation of vignettes. In this paper, we will refer specifically to applications in measuring health and assessing the responsiveness of the health system, but the general approach described here would apply to a wide range of analytical problems that rely on self-reported ordinal data.

II. DEFINITION

A vignette is a description of a concrete level on a given domain that respondents are asked to evaluate with relation to the main self-report question on that domain using the same categorical response scale for that question. Vignettes fix the level of ability so that variation in categorical responses is attributable to variation in response category cutpoints. The introduction of exogenous information in the form of vignette ratings allows identification of

the effects of different covariates on both the level of the underlying latent variable as well as on the cutpoints.

We define two key requirements for the use of vignettes as:

(a) *response equivalence*, which states that individuals use the response categories for a particular question in the same way when they evaluate hypothetical scenarios as they do when they provide self-reported assessments of their own health or their own experiences of health system responsiveness;

(b) *vignette equivalence*, which states that the domain levels represented in each vignette are understood in the same way by all respondents, irrespective of their age, sex, income, education, country of residence or other sociodemographic variables.

III. EXAMPLES FROM THE WHO HEALTH AND RESPONSIVENESS SURVEY INSTRUMENTS

Following are examples of vignettes in one domain of health and one domain of responsiveness in the WHO Multi-Country Study. The instrument includes a range of six to eight vignettes in the different domains of health and responsiveness.

A. Mobility vignettes

The survey instrument includes six vignettes for the domain of mobility:

Vignette 1: [Paul] is an active athlete who runs long distance races of 20 kilometers twice a week and engages in soccer with no problems.

Vignette 2: [Mary] has no problems with moving around or using her hands, arms and legs. She jogs 4 kilometers twice a week without any problems.

Vignette 3: [Rob] is able to walk distances of up to 200 meters without any problems but feels breathless after walking one kilometer or climbing up more than one flight of stairs. He has no problems with day-to-day physical activities, such as carrying food from the market.

Vignette 4: [Margaret] feels chest pain and gets breathless after walking distances of up to 200 meters, but is able to do so without assistance. Bending and lifting objects such as groceries produces pain.

Vignette 5: [Louis] is able to move his arms and legs, but requires assistance in standing up from a chair or walking around the house. Any bending is painful and lifting is impossible.

Vignette 6: [David] is paralyzed from the neck down. He is confined to bed and must be fed and bathed by somebody else.

For each vignette, respondents are asked the main question on mobility in the survey: "How much difficulty did [*name*] have in moving around?" The response categories are the same as those used for the self-reports: (1) extreme difficulty / unable to move around, (2) severe difficulty, (3) moderate difficulty, (4) mild difficulty, and (5) no difficulty.

B. Dignity vignettes

The survey includes seven vignettes on dignity:

Vignette 1: [Conrad] is suffering from AIDS. When he enters the health care unit the doctor shakes his hand. He asks him to sit down and inquires what his problems are. The nurses are concerned about Conrad. They give him advice about improving his health.

Vignette 2: [Anya] took her three-month old infant for her vaccination. The nurse asked her why she had not been to the clinic before, and was sympathetic to hear that Anya had a problem finding transport. She advised her about the importance of regularly monitoring the growth of her baby.

Vignette 3: [Julia] visits the health care centre for treatment at a time when the centre is very crowded. The patients are all impatient to get their treatment and are reluctant to queue and wait for their turn. The nurses are very patient most of the time about asking patients to wait their turn, but occasionally they get angry and shout at her for breaking the queue.

Vignette 4: [Patricia] goes to a health care unit close to her home regularly. The nurses there are very busy, but they always speak pleasantly to her. The receptionist however is often in a bad mood, and when she is in a bad mood she shouts at Patricia, and at other patients. All appointments to meet doctors and nurses have to be made through this receptionist so the patients put up with her rudeness.

Vignette 5: [Kim] took her six month old infant to the health centre for her regular check-up. The nurse was very annoyed when she found that Kim had forgotten to bring the baby's growth chart with her. She scolded her loudly in the hearing of all the other mothers who had come to the clinic, and kept grumbling about inconsiderate forgetful mothers who caused extra work as she weighed the baby.

Vignette 6: [Said] has AIDS. When he goes to his health centre he feels that all the doctors and nurses are unfriendly towards him. They do not talk to him freely. Often they deliberately ignore him. He often has to beg them to answer his questions.

Vignette 7: [Florence] goes to the hospital as she has a pain in her stomach. The nurse shouts at her for not bringing her health card. Two other nurses who are standing by make rude comments about Florence's family and those from her village. Though Florence is in pain, and moaning she is not asked to sit down while her personal details are entered in the register.

For each vignette, respondents are asked the main question on dignity: "How would you rate [*name*]'s experience of getting treated with dignity?" with the same response categories as in the main self-report question on dignity.

IV. VIGNETTE RATINGS: EMPIRICAL RESULTS

When survey respondents provide ratings for a series of different vignettes on a particular domain, we may visualize the responses in terms of the distribution of categorical ratings for each vignette across different groups of respondents. Figure 1 presents an example showing the distribution of responses for the mobility vignettes in China and Morocco. In this figure,

each stacked bar shows the categorical responses for one vignette, and the series of vignettes is ordered from higher mobility levels to lower ones.

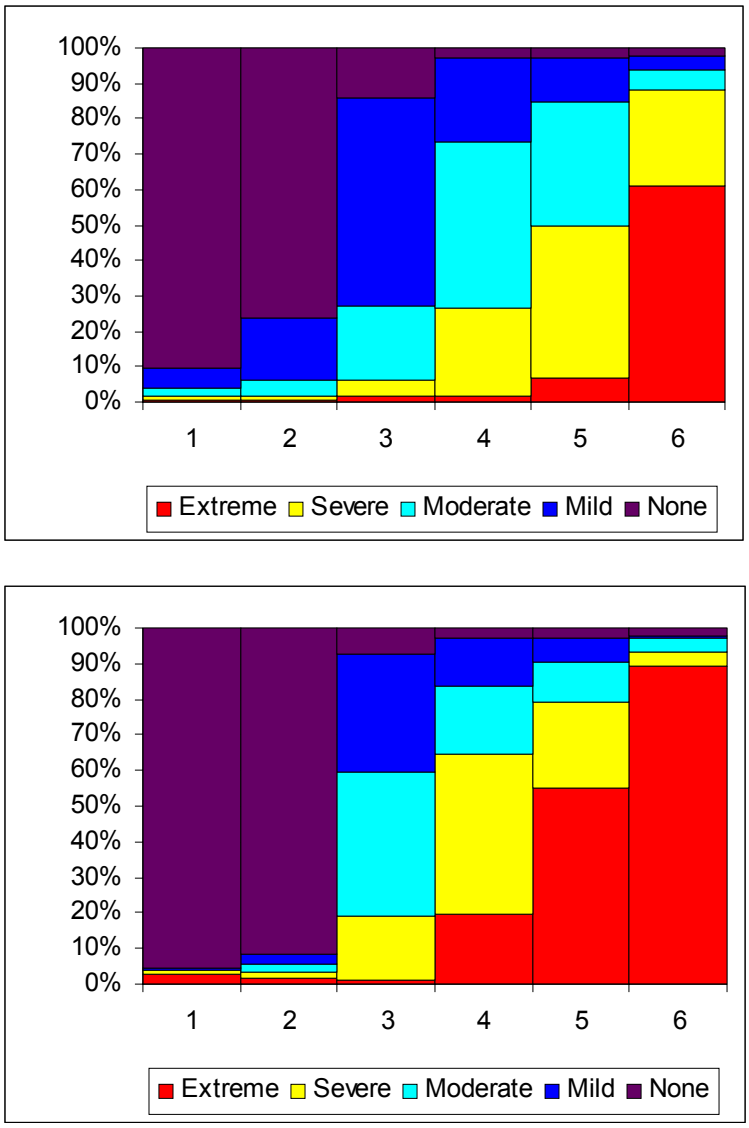


Figure 1. Ratings for mobility vignettes in China (top panel) and Morocco (bottom panel).

These figures allow some general insights into differences in the uses of categorical response categories that are formalized in the statistical models described elsewhere [4]. This simple example offers a comparison of the distributions in two countries, but it is important to note that the models will also allow analyses of differences within countries, for example, across age, sex, income, education, or other covariates of interest.

From the distributions of responses, it is evident that individuals in China on average are less likely than individuals in Morocco to use either the best mobility category (no difficulty) or the worst category (extreme difficulty / unable to move around). The use of the category "mild

difficulty" is also more prevalent in China then Morocco. Note that it is not simply the case that individuals in China tend to rate vignettes as either "better" or "worse" than in Morocco, but rather that there are more shaded differences in the use of the same categorical scale in the two countries.

V. PRACTICAL CONSIDERATIONS

The two key requirements for the use of vignettes – response equivalence and vignette equivalence – along with statistical considerations in estimating the analytical models, lead to a series of practical concerns.

A. Number, range and spacing of vignettes

A minimum number of vignettes are required in order to provide enough information to estimate differences in all categorical cutpoints. The vignettes should cover a range of different levels on the domain of interest in order to ensure that, across ratings of the complete vignette set, each response category contains an adequate number of observations from each subgroup defined by the set of explanatory variables. Given a fixed number of vignettes, the information content of the vignette ratings will be optimized if the vignettes are spaced at sufficiently large intervals along the range of the latent variable. In other words, if one vignette represents a particular level on the latent scale, a second vignette at this same level will provide little information on the cutpoint locations of different individuals; spacing the vignettes at different levels of the latent scale, on the other hand, will produce higher marginal information value for each vignette and therefore maximize the amount of inference that may be gained from a fixed number of questions.

B. Ensuring equivalence of vignettes

The requirement of vignette equivalence demands careful attention to both the design and translation of vignettes.

In the design stage, it is important to ensure to the extent possible that the concepts described in each vignette will allow equivalence to be established across different populations. It may be useful for vignettes to include concrete terms rather than vague phrases that are subject to different interpretations. For example, in the mobility vignettes, descriptions of distances are concrete, such as "20 kilometers" rather than imprecise, such as "long distances." A competing concern, however, is that individuals with different degrees of numeracy or different frames of reference may not regard these defined quantities in the same way. The distance of 20 kilometers may be understood differently by a long distance runner as compared to a truck driver or a math professor or a subsistence farmer.

An alternative is to refer to specific examples rather than numerical quantities, as in the mobility vignettes describing "carrying food from the market" or "lifting groceries," as opposed to defining weights in terms of kilograms or pounds. While these examples may be more relevant to individuals with lower levels of numeracy, they remain subject to variation in interpretation, *i.e.*, they will evoke different quantities in individuals depending on how much food or how many groceries they imagine. The tradeoff between concreteness of numerical specifications versus relevance and comprehensibility of non-numerical examples

suggests that it may be ideal to include both types of information across the range of vignettes, or even within a specific vignette. While it may be impossible to ensure complete equivalence in the interpretation of a particular vignette across all respondents, the use of a set of vignettes will minimize the impact of error introduced by any single vignette.

Once the set of vignettes has been designed, it is crucial to adopt a rigorous protocol for translating vignettes in order to ensure that minimal variation is introduced in the concrete domain level represented by each vignette through inexact translations. It is useful to consider translation issues during the design of vignettes rather than treating the two sets of concerns separately. It may be possible, for example, to anticipate that a particular concept or word will be difficult to translate into different languages and therefore to choose a different vignette specification at the design stage in order to obviate this problem.

C. Framing and ordering effects

The framing and ordering of the vignette questions may be used to increase the likelihood that individuals use the categorical responses in the same way for vignettes as for self-reports.

By presenting the set of vignettes in random order, as has been implemented in the WHO surveys, we may reduce the tendency for respondents to resort to arbitrary sorting of the vignettes into ordered categories without considering the meaning of the categorical labels. In this way, respondents are more likely to consider their categorical responses to each vignette in the same way they do their self-reports.

A key issue in framing vignettes is whether the age and sex of the individual in the vignette should be specified explicitly, and if so, what should be the reference age and sex. There are at least 3 different possibilities for framing questions in terms of a specific age and sex:

- No reference to age and sex
- Refer to somebody of "your age and sex" in each vignette
- Refer to some specific age and sex for each vignette, fixed across respondents

To the extent that we will use the vignettes to adjust for norms that may depend on age and sex, it may be useful to have the vignettes matched to an individual's own characteristics (as in case 2). Matching vignettes to an individual's own characteristics may improve response equivalence by increasing the similarity of the question to the self-reported question of interest. On the other hand, it is critical to ensure that, as much as possible, the domain levels described in each vignette are fixed across respondents. There is a danger that the introduction of variation in terms of age and sex will implicitly introduce variation in the latent variable level evoked by a particular vignette, thus compromising vignette equivalence. Thus, there are important tradeoffs between standardization of the domain levels across respondents and establishing scale equivalence for self-reports and ratings of vignettes for a given respondent.

VI. FORMAL EVALUATION OF VIGNETTES

A series of formal assessments of vignettes may be used to address questions of reliability and validity. We present a brief overview of the evaluation techniques in this paper. Practical appli-

cation of these techniques to empirical data on health and responsiveness from the WHO Multi-Country Study is currently in progress.

A. Test-retest reliability

One key measure of reliability is the extent to which individual responses are stable in repeated measures. The table below shows average kappa statistics for the health vignettes by domain across 9 countries. Kappa statistics are a measure of agreement between two different observations that accounts for the level of agreement that would be expected from chance alone and also includes weights allowing for partial credit. A value of 1 indicates perfect agreement, while a value of 0 would be the level of agreement expected by chance.

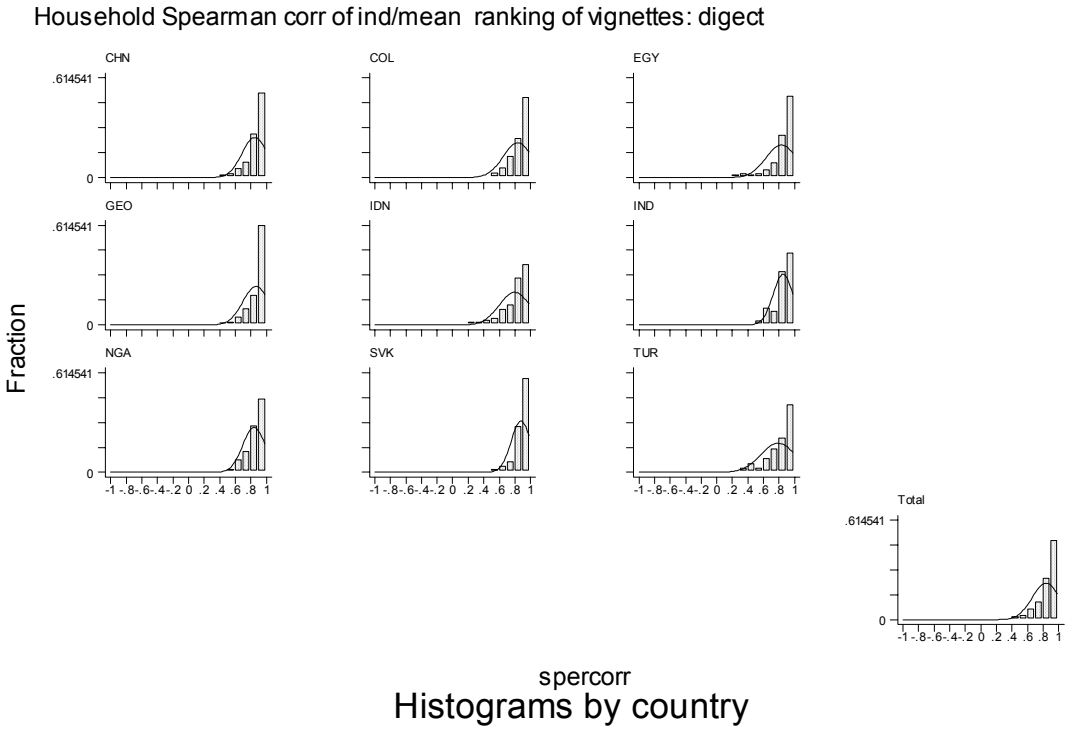
Domain	Mean	Std. Dev
Pain	0.605	0.032
Self-care	0.592	0.042
Affect	0.606	0.021
Mobility	0.569	0.054
Cognition	0.645	0.033
Usual activities	0.623	0.035

On all domains, average kappa statistics are reasonably high across countries, and there is little variation across particular vignettes within a given domain.

B. Rank order correlations

One way to assess the performance of a set of vignettes is to examine the correlations between individual rankings of the vignettes with the overall average rankings. This provides a weak measure of the requirement that vignettes evoke the same concepts and convey the same fixed domain levels across respondents.

The figure below presents the distribution of correlation coefficients by country for the dignity vignettes. The correlations are quite high overall in all countries, although there is some variation in the distributions of these measures in different countries.



We may also examine how the levels of correlation may vary depending on particular sociodemographic characterisitics of the respondents. In so doing, we can analyze the extent to which individual characteristics such as age or education produce differences in interpretation of vignettes on a domain.

C. Evaluations within the analytical models

Within the context of the statistical models described in Tandon et al. [4], there are additional tests that may be used for formal evaluation of vignettes. For example, variation in the domain level evoked by a particular vignette may be examined formally within the HOPIT model by allowing for the coefficient on a given vignette to vary across countries while holding the others fixed across countries. The variance in the estimated vignette level across countries may be compared for different vignettes as a measure of vignette-specific equivalence.

After the statistical models have been estimated, a further evaluation technique relies on visual inspection of the range and spacing of vignettes along the latent variable scale and in reference to the distribution of cutpoints in the survey populations. It is useful to have a range of vignettes with levels that are spaced along the full range of the latent variable and particularly in areas that have proximity to the cutpoint distributions in the population. It is also important to ensure that there are vignettes at both the high end and low end of the range of the latent variable in the population, in order to allow sufficient information with which to

estimate the cutpoints that define the extreme categorical responses. As described above, vignettes that are closely spaced result in an efficiency loss in the use of the survey instrument.

VII. STRATEGIES FOR DESIGNING AND CHOOSING VIGNETTES

The recommended strategies for selecting a range of different vignettes is to design a large number (around 40 to 50) of vignettes for each domain and then to test out the various properties of subsets of these.

Tests of vignettes would include the ones described above, as well as qualitative assessments of comprehensibility and cognitive interviewing to develop a better understanding of respondents' interpretations of the vignettes and the vignette rating exercise. Goals of the testing are:

- To ensure that vignettes have equivalent meanings cross-culturally.
- To ensure that rankings of vignettes are highly correlated across respondents.
- To ensure that the choice of vignettes includes a sufficient number and covers a sufficient range on the latent variable in order to provide the statistical power needed to draw inferences about response category cutpoint shifts.

VIII. CONCLUSIONS

The use of vignettes is part of an integrated strategy of instrument design and analytical methods for enhancing cross-population comparability of health surveys. Vignettes may be applied to many different analytical problems where ordered categorical self-reported responses are observed.

Vignettes provide a means of examining systematic differences in categorical cutpoints between populations or within populations across different sociodemographic groups. The vignette approach depends on the two key requirements of response equivalence (i.e., that individuals use response scales for a particular question in the same way for themselves as for hypothetical individuals described in vignettes) and vignette equivalence (i.e., that vignettes describe domain levels that are fixed across respondents, so that variation in their ratings gives information on cutpoint shifts). These two assumptions may sometimes conflict with each other in considering various design issues, so a careful strategy is required in order to optimize the tradeoffs between these two requirements.

IX. REFERENCES

1. Sadana R, Mathers CD, Lopez AD, Murray CJL, Iburg KM. Comparative analysis of more than 50 household surveys of health status. In: Murray CJL, Salomon JA, Mathers CD, Lopez AD, Editors. *Summary measures of population health: concepts, ethics, measurement and applications* . Geneva: World Health Organization, 2002.

2. Murray CJL, Chen LC. Understanding morbidity change. *Population and Development Review* 1992; 18(3):481-503.

3. Mathers CD, Douglas RM. Measuring progress in population health and well-being. In: Eckersley R, Editors. *Measuring progress: Is life getting better*. Collingwood, Victoria, Australia: CSIRO Publishing, 1998.p. 125-55.

4. Tandon A, Murray CJL, Salomon JA, King G. Statistical methods to enhance cross-population comparability (Global Programme on Evidence for Health Policy Discussion Paper No. 42). Geneva: World Health Organization, 2001.

5. Üstün TB, Chatterji S, Villanueva M et al. WHO Multi-country Household Survey Study on Health and Responsiveness, 2000-2001 (Global Programme on Evidence for Health Policy Discussion Paper No. 37). Geneva: World Health Organization, 2001.