A comparison of simple score and latent class approaches: Application to HIV knowledge data in Chinese and multi-country contexts

Olga Maslovskaya[a], Peter W.F. Smith[b,c], Sabu S. Padmadas[b]

[a] ESRC National Centre for Research Methods (NCRM), University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
[b] Department of Social Statistics and Demography, Social Sciences, Faculty of Social, Human and Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
[c] ESRC Administrative Data Research Centre for England (ADRC-E), Social Sciences, University of Southampton, UK

Abstract

Knowledge about different health-related attitudes, beliefs, and risks is of significant interest to scholars in different Social Science disciplines. Usually knowledge is collected in a form of multiple variables and then constructed as a composite indicator. The question any researcher working with knowledge-related variables faces is: what is the best way to measure and summarise different dimensions of health-related knowledge? The main goal of this paper is to evaluate and compare simple score and latent class approaches to measuring and summarising health-related knowledge using population data on HIV knowledge collected in five selected countries (China, India, Kenya, Malawi, and Ukraine). The advantages and shortcomings of both approaches (simple score and latent class approaches) to measuring and summarising health-related knowledge are evaluated and discussed.

(Word count: 123 words)

Key words: latent class analysis, composite indicators, simple score measures, HIV knowledge, China.

Data Statement

This study uses the following datasets: the China National Family Planning and Reproductive Health Surveys 1997 and 2001, the China UNFPA Reproductive Health and Family Planning Survey 2005, Kenya DHS 2003, Malawi DHS 2004, India DHS 2006, and Ukraine DHS 2007. The access to the first three datasets was obtained through the China Population and Development Research Centre (CPDRC) in Beijing, China. The access to the DHS datasets was enabled through Measure DHS (Demographic and Health Surveys) project (http://www.measuredhs.com/).

Acknowledgements

Dr Olga Maslovskaya was funded by the Economic and Social Research Council (ESRC 1+3 scholarship: PTA-031-2006-00185), which facilitated her contribution. The authors thank the editorial team and the anonymous referees for providing useful suggestions and comments. They also thank Professor Nyovani J. Madise and Professor Monica Magadi for their valuable advice. They are grateful to Professor Li Bohua from the China Population and Development Research Centre (CPDRC) in Beijing, China, for access to the China National Population and Reproductive Health Survey 1997 and the China National Family Planning and Reproductive Health Survey 2001.

Introduction

Different Social Science disciplines study various types of knowledge. To summarise knowledge is a complex task which is often fraught with methodological challenges. For example, HIV knowledge as other types of knowledge cannot be observed or measured directly but it can be measured with the help of the various components of HIV knowledge which are measurable directly in surveys such as knowledge of specific routes of HIV transmission. It is usually collected in a form of multiple variables and then is summarised in a form of a composite indicator/score/measure. There are different approaches to calculating composite measures: simple score approach, latent trait approach, latent class approach, principal component analysis, factor analysis and others. All researchers studying knowledge of different issues or impact of the knowledge on some other phenomena will be faced with the decisions which measure is the best for their specific purpose. The choice of summary measure would depend on the type of available indicator variables (continuous or categorical) and the desired type of the summary variables (continuous or categorical). Simple score approach, latent trait analysis (LTA) and latent class analysis (LCA) are widely used in different health-related disciplines and areas when observed variables are categorical. LTA is frequently referred to as (explanatory) factor analysis in the literature. By LTA we mean here the analysis when unobserved latent variable is continuous and observed variables are categorical. As this paper focuses on categorical summary measures, it, therefore, compares simple score approach and LCA to summarising different components of composite measures of knowledge. Population data of HIV knowledge collected in selected countries (China, India, Kenya, Malawi, and Ukraine) are used for the analysis. In this paper we provide advantages and disadvantages of the two approaches to summarising HIV knowledge. HIV knowledge provides a perfect case study for comparison of different composite indicators as it is always derived on the basis of multiple questions collected as categorical variables. However, the results of the current analysis are transferable and can be applied in other substantive areas.

The majority of studies in the area of HIV knowledge use a simple score approach (Barden-O'Fallon *et al*. 2004; Feldman *et al.* 2011; Lou *et al*. 2006; Stephenson 2009). They add up correct answers to HIV knowledge questions in order to calculate a score of HIV knowledge for each respondent. There are slight deviations from this approach in some research. Studies which use a simple score approach to measuring HIV knowledge use either a true/false scale (Ambati *et al*. 1997; Barden-O'Fallon *et al*. 2004; Benotsch *et al*. 2004; Hu *et al*. 2006; Lau *et al*. 2002; Li *et al*. 2004; Loue *et al*. 2003; Pinkerton *et al*. 2003; Stephenson 2009; Zimet 1998) or a Likert-type scale when questions about HIV knowledge are presented in a form of attitudinal questions (Fisher *et al*. 1996; Fisher *et al*. 2002). In some studies both scales are used together depending on a specific question (Brown and Bocarnea 1998; Koch and Singer 1998; Koopman and Reid 1998). In some of the studies reviewed, respondents who answered "no" to a filter question - if they had ever heard of HIV/AIDS - scored 0 on HIV knowledge as their HIV knowledge is assumed to be inadequate (Lagarde *et al*. 1997).

LCA is widely used in summarising knowledge about health-related attitudes, beliefs and risks in different health-related areas of research: HIV-related research (Beadnell *et al*. 2003; Dias 2001; Kattumuri 2003), sexually transmitted infections and risky behaviours (Stuart and Hinde 2009), smoking and tobacco dependence (Chen *et al*. 2004; Storr *et al*. 2004), drug use (Monga *et al*. 2007), depression (Sullivan *et al*. 1998), eating disorders (Bulik *et al*. 2000), and others. The main aim of the studies which used LCA was to determine the number of more homogenous segments in a heterogeneous population and to split population into these homogenous groups or classes.

To date, no studies compared two different approaches to measuring knowledge. Therefore, the current research will fill this gap as it aims to assess and compare simple score approach and LCA approach to measuring HIV knowledge. The results can then be extended to other substantive areas where different types of knowledge are of main interest.

To compare simple score and LCA approaches, we consider two study contexts: the Chinese context over time and so-called five country context. Within the two study contexts we employ seven different datasets from countries representing different socioeconomic and geographic criteria. The use of different contexts ensures the robustness of the results of the comparison of the two approaches.

Each HIV knowledge measure can be assessed from the point of view of its so-called qualitative and quantitative meanings (Agresti 2013). The quantitative or ordinal meaning of any measure is the number of components of the measure known by a respondent. In other words, a quantitative meaning of a measure can help to identify the place of a respondent on a continuous spectrum of HIV knowledge between no knowledge and prefect knowledge. The qualitative or nominal meaning of any score measure is a specificity of knowledge of each respondent in each category, i.e. which specific elements of HIV knowledge are known by a respondent in a specific group. In other words, a qualitative meaning of a measure can help identify specific groups of respondents which exist in a population on the basis of their HIV knowledge. Ideally a perfect measure of HIV knowledge should have both precise qualitative and quantitative meanings. However, this might not be achievable and, therefore, more suitable for a specific research context measure should be selected. In some specific research contexts only one of the two requirements or even a part of one requirement might be sufficient.

Assessment and comparison of the two different measures of HIV knowledge can help researchers from a wide range of disciplines to decide more easily which measure is the most suitable in their study context. Specifically, it can help researchers in health-related disciplines to decide which summary measure is more suitable when the effective implementation of public health interventions is of interest or when evaluation of progress of educational and information campaigns is conducted.

Data

The analysis presented is a part of a larger project which studied evolution of HIV awareness and knowledge in China. One of the aims of this project was to assess whether China has succeeded in improving women's HIV knowledge over time, and if China is a relative success story in improving women's HIV knowledge when compared with other countries in the world with generalised and non-generalised HIV epidemics. This aim determined the choice of the datasets used for the analysis.

In order to illustrate the two different approaches to measuring HIV knowledge, two study contexts were employed: the first one studied HIV knowledge in China over time and the second compared HIV knowledge in China to the levels of knowledge in four other countries (India, Kenya, Malawi, and Ukraine).

Three datasets were used for the first study context: the China National Family Planning and Reproductive Health Surveys 1997 and 2001, and the China UNFPA Reproductive Health and Family Planning Survey 2005. The first two surveys are representative of the whole of China, whereas the third survey is representative of 30 counties that were purposefully selected from the 30 provinces in China (Jiang, 2000; Li *et al.*, 2004; Li *et al.*, 2008; Pan *et al.*, 2003). Further details about surveys can be found in Table A.1 of the Appendix as well as in Jiang (2000), Pan *et al*. (2003), Li *et al*. (2008), and Maslovskaya *et al*. (2014). The following five datasets were used for the second study context: the China UNFPA Reproductive Health and Family Planning Surveys 2005, Kenya DHS 2003, Malawi DHS 2004, India DHS 2006 and Ukraine DHS 2007. Further details about surveys can be found in Table A.1 as well as in CBS *et al*. (2004), NSO and ORC Marco (2005), IIPS and Marco International (2007), and UCSR *et al*. (2008). Demographic and Health Surveys (DHS) are nationally-representative household surveys that provide data for a wide range of indicators. For this study context, the four different surveys were selected in order to be able to compare HIV knowledge in China with HIV knowledge in other parts of the world. Countries with generalised HIV epidemics (Kenya and Malawi) as well as those with non-

generalised epidemics and similar to China characteristics, e.g. population size (India), educational system (Ukraine), were selected. The choice of years of surveys was determined by data availability and by the attempt to use the most recent available data which are chronologically closer to the most recent Chinese survey from 2005.

HIV knowledge cannot be measured directly and has to be derived on the basis of different components. Simple score and LCA measures were derived on the basis of the availability of questions found in different surveys (see Table A.2 in Appendix). All questions in the first study context are HIV knowledge questions. However, in the second study context there is a combination of HIV knowledge and HIV prevention questions. The choice of questions for the analysis was determined by the availability of questions across different surveys and by the components of the Millennium Development Goals (MDG) indicator which combines the ability to correctly identify ways of preventing the sexual transmission of HIV and to reject major misconceptions about HIV transmission, and is used in the cross-country contexts (UNAIDS 2008; UNAIDS 2010). The first four questions in both study contexts (see Table A.2) are included to measure the knowledge of correct routes of HIV transmission, whereas three last questions are included to measure the knowledge of incorrect routes of HIV transmission. The wordings of questions differ slightly from survey to survey. The exact wordings of questions can be found in Table A.3 and Table A.4 of the Appendix. All these variables have three response categories: “yes”, “no” and “do not know”. All answers were recoded into two categories “correct” and “incorrect” which were coded as 1 and 0 respectively to ease calculation of scores. “Do not know” answer options were coded as “incorrect” as people who answered “do not know” were assumed to have no correct knowledge about a specific question. The correct response for the items 1-7 and 12 in Table A.2 is “yes”, the correct response for the items 8-11 in Table A.2 is “no”.

The datasets used for the analysis contain only respondents who were asked HIV knowledge related questions. Only women who answered all seven HIV knowledge component questions (see Table A.2) were retained in the datasets used for the analysis. Table A.5 in Appendix presents

stages of preparation of datasets for the analyses and shows percentages of women who are excluded from the final datasets owing to missing values for HIV knowledge items.

As mentioned above, observations which contain missing values for HIV knowledge items are removed. Missing values are assumed to be missing at random. Table A.5 shows that proportions of removed observations with missing values for HIV knowledge items are zero in the Chinese context and are low in all other contexts and, therefore, observations with missing values are not examined further.

The absence of comparable sets of questions across two study contexts does not represent a limitation for this analysis as the study contexts are used just for the illustration of the usage of different measures of HIV knowledge and not for the purpose of the comparison of results across the two study contexts.

Methodology

HIV knowledge data are collected frequently as a part of larger surveys. HIV knowledge has various components and cannot be collected as a single variable. As mentioned earlier, there are two main approaches to measuring HIV knowledge which exist in literature: a so-called simple score and a LCA approaches.

By simple score approach we mean adding up correct answers to multiple questions related to a single concept in order to calculate a final score for the knowledge about the overall concept for each respondent. The two main assumptions are that all components are equally weighted and that the highest score represents the perfect knowledge about the phenomena or in other words, the higher the score, the better the HIV knowledge.

HIV knowledge is a latent variable as it cannot be measured directly but only through the indicator variables such as knowledge of separate routes of HIV transmission. Latent variables help to reduce the number of dimensions and to summarise data into one (or more) summary measure(s) without significant loss of information (Bartholomew *et al*. 2008).

In order to conduct LCA and extract categorical latent variables the M-Plus statistical software version 5.1 (Muthén and Muthén 2009) was employed.

LCA is a model-based technique which allows inferences about a population to be made and helps summarise data in the form of one latent variable with a number of classes without significant loss of information (Bartholomew *et al*. 2008; Beadnell *et al*. 2003; Hagenaars and McCutcheon 2002; McCutcheon 1987). LCA can help split a heterogeneous sample into classes which are more homogeneous with respect to the latent variable, and within these classes people will share similar characteristics (Clogg and Goodman 1984). The main aim of LCA is to determine the smallest number of latent classes that is sufficient to explain relationships between the observed variables (Magidson and Vermunt 2004) and to provide classification of individuals (Muthén and Muthén 2009). LCA can help identifying groups of individuals which should be targeted by interventions as they require specific information about HIV. Details about latent class analysis, model fit criteria used for selection of the best model as well as allocation of observations to classes can be found in Durrant et al. (2017), Muthén (1998-2004) and Muthén and Muthén (2009).

It is important to acknowledge that latent class values have the following limitation: they ignore potential uncertainty in class predictions. For the work presented here, this limitation does not represent a problem as the focus of the work is on the comparison of the two approaches and not on the specific results obtained from the LCA. In order to address this limitation, Structural Equation Modelling (SEM) framework can be used by researchers to correct for this uncertainty. Alternatively, methodology outlined in WebNote 15 for MPlus version 7 can be employed (Asparouhov and Muthén 2013).

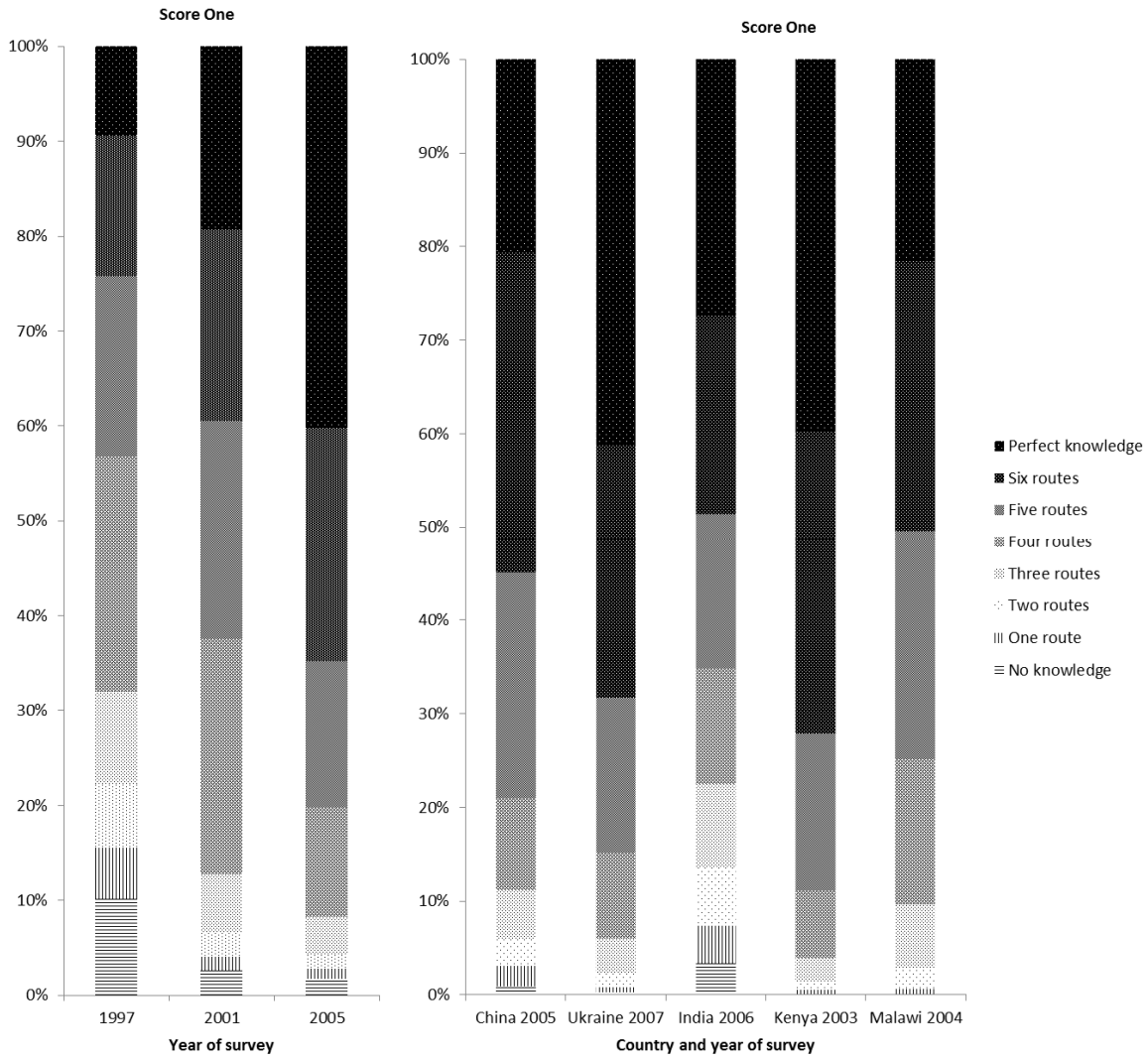
Results

Simple score measures

Three simple score measures were obtained for each study context. Score One (S1) is a simple score measure for combined knowledge about HIV. Score Two (S2) is a simple score measure for knowledge of correct routes of HIV transmission in the Chinese context and a measure for knowledge about ways to prevent HIV in the five country context. Score Three (S3) is a simple measure for knowledge of incorrect routes of HIV transmission in China and a measure for knowledge about misconceptions about HIV transmission in the five countries. Table A.2 contains the information about each item which is included into calculation of a certain score.

Figure 1 shows distribution of overall HIV knowledge (S1) in two study contexts. Figure 1 left panel shows distribution of S1 in China over time. It suggests that with time the proportion of people having perfect knowledge increases and the proportion of people with no knowledge decreases. Figure 1 right panel shows distribution of S1 by countries in five country context. It suggests that the proportion of people with perfect HIV knowledge is the highest in Ukraine and Kenya when compared to other countries and the lowest in China and Malawi. The proportion of people with no knowledge is the highest in India when compared to other four countries.

Figure 1: Distribution of Score One in both study contexts



Note: Score One measures overall HIV knowledge in Chinese context (left panel) and in five country context (right panel). Different questions are taken into account for score calculation in China in two contexts.
Sources: China National Family Planning and Reproductive Health Surveys 1997 and 2001, China UNFPA Reproductive Health and Family Planning Survey 2005, Kenya DHS 2003, Malawi DHS 2004, India DHS 2006, Ukraine DHS 2007.

Simple scores can provide information about the location of a specific group of respondents on HIV knowledge spectrum ranging between no knowledge and perfect knowledge as well as useful insights into understanding of evolution of HIV knowledge. However, this approach to summarising HIV knowledge has limitations. The clear meaning exists only for the first and the last categories. It is not clear what type of knowledge those in the middle categories have. The first score option always suggests that a respondent does not have any knowledge and the last score option suggests that a respondent has knowledge of all components of HIV knowledge. All other score options suggest the number of known components but they do not indicate specific

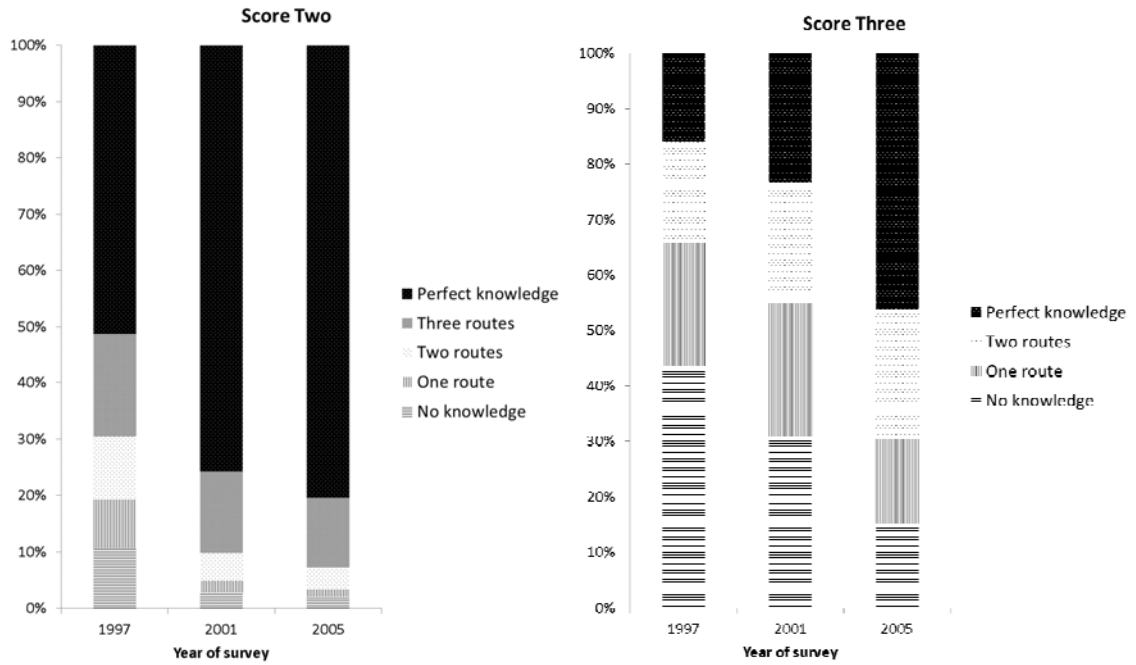
components known by respondents. For different individuals having the same number of correct answers can represent different patterns of HIV knowledge, e.g. respondents who have knowledge about one route, either mother-to-child transmission (MTCT) or handshaking will end up in the same group in S1. Women with knowledge of three routes might have different combinations of correct answers.

The main limitation of S1 is its inability to discriminate between different patterns of knowledge for the majority of groups of respondents apart from quantitative patterns. Therefore, they are best used when respondents should be located on a continuum of HIV knowledge.

In order to overcome this limitation, different types of HIV knowledge can be studied separately by creating separate simple scores for different types of knowledge. Figures 2 and 3 present knowledge of correct routes of HIV transmission (left panels) (S2) and knowledge of incorrect routes and misconceptions about HIV transmission (right panels) (S3) in China and in five countries respectively.

Figure 2 suggests that the proportion of people with no knowledge reduces with time in China. However, it is higher for knowledge about incorrect routes of HIV transmission at every point in time when compared to the knowledge about correct routes. The proportion of people with perfect knowledge increases with time and it is higher at every point in time for knowledge of correct routes in comparison to the knowledge about incorrect routes.

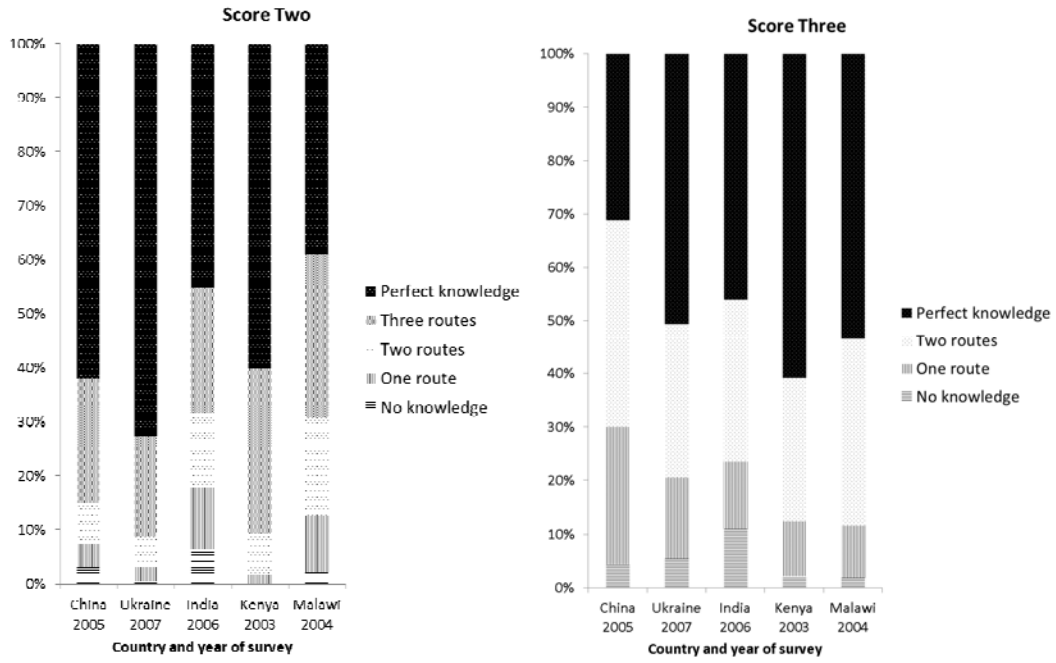
Figure 2: Distributions of Score Two (left panel) and Score Three (right panel) in China over time



Note: Score Two measures knowledge about correct routes of HIV transmission and **Score Three** measures knowledge about incorrect routes of HIV transmission in the Chinese context.
Sources: China National Family Planning and Reproductive Health Surveys 1997 and 2001, China UNFPA Reproductive Health and Family Planning Survey 2005.

Figure 3 left panel shows that the proportion of people with perfect knowledge in five country context is the highest for knowledge about correct routes when compared to other knowledge scores. The highest knowledge of correct routes is observed in Ukraine, China and Kenya and the highest level of knowledge about incorrect routes is observed in Kenya and Ukraine with China lagging behind when compared to other four countries. India has the highest proportions of people with no knowledge about correct and incorrect when compared to other countries. Malawi has the lowest proportion of women with the perfect knowledge about the correct routes when compared to other four countries.

Figure 3: Distributions of Score Two (left panel) and Score Three (right panel) in the five country context



Note: Score Two measures knowledge about ways to prevent HIV and **Score Three** measures knowledge about misconceptions about HIV transmission in the five country context.
Sources: China UNFPA Reproductive Health and Family Planning Surveys 2005, Kenya DHS 2003, Malawi DHS 2004, India DHS 2006, Ukraine DHS 2007.

These separate measures of HIV knowledge (S2 and S3) help overcome some limitations of overall HIV knowledge measure (S1). They allow having more precise information about specific types of knowledge in the population. However, some limitations still remain. There are still problems with clear definitions of middle categories. This approach also has its own limitations. There is a need to predefine dimensions and the approach does not allow data driven definitions of the dimensions. When HIV knowledge is of interest, two dimensions are quite clear. However, in some other contexts this might not be the case. When using this approach to measuring HIV knowledge it is only possible to study different dimensions of knowledge separately. It is not possible to identify individuals with high scores on one dimension but low scores on the other. Therefore, in order not to miss important information, first overall knowledge should be studied and then different types of knowledge separately.

It is a well-known fact in survey research that there will always be people answering "yes" to all questions (Bentler *et al*. 1971; Billiet and Davidov 2008; Cloud and Vaughan 1970; Hamilton 1968; Watson 1992). This effect is called acquiescence or agreeing-response bias (Bentler *et al*. 1971; Billiet and Davidov 2008; Cloud and Vaughan 1970; Hamilton 1968; Watson 1992). In the Chinese context the correct answers to S2 are "yes", whereas for S3 are "no". In the five country context the correct answers to S2 and one question from S3 (healthy-looking person question) are "yes", whereas for the remaining questions for S3 the correct answers are "no". Unfortunately, S1 in both contexts cannot help identifying individuals who had a tendency to display acquiescence as S1 does not provide precise qualitative meaning for each individual as for different individuals having the same number of correct answers can represent different patterns. This is another limitation of a simple score approach.

The main limitation of all three measures is the inability of the simple score measures to take weights into account for each contributing component which might be important in some research contexts. There might be knowledge of one or two routes of HIV transmission which is sufficient for successful prevention programs and might have larger weights than other routes. It is difficult to use this approach for specific interventions as it would be impossible to identify the specific information needs of specific groups in a population.

The main advantage of a simple score approach is that the measures are easily calculated. On many occasions the information which is provided by these measures is sufficient for the researchers. These measures provide general picture of the levels of HIV knowledge and can be used successfully for the assessment of the evolution of HIV knowledge as well as for studying positions of different groups on the continuum between no knowledge and perfect knowledge. However, a more detailed measure of HIV knowledge is required in order to be able to further validate the survey responses and to take into account the complicated nature of HIV knowledge.

All the limitations mentioned above suggest that in order to understand the distribution of HIV knowledge across population better, it is important to create a more complex measure of HIV

knowledge. In order to overcome some of these limitations of the simple score approach, latent class approach to obtain a summary of HIV knowledge will be used.

Latent measures

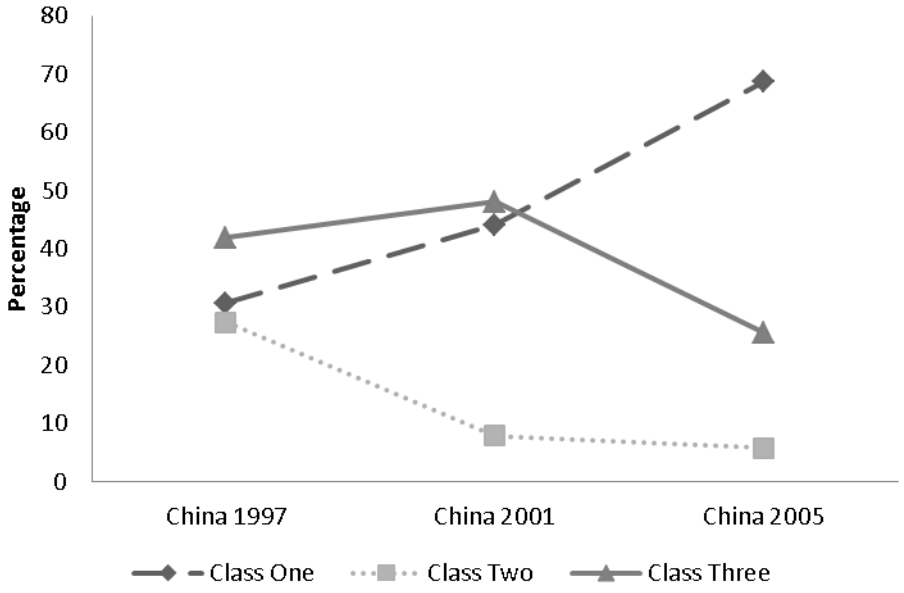
Two Latent Class variables were derived using MPlus software package to identify latent classes of HIV knowledge first in China and then in the five country context.

If the three criteria for determination of the number of classes (discussed in Durrant et al. 2017) are taken into consideration (AIC, BIC and classification quality) then the model which fits the Chinese data best is the three-class model (AIC and BIC are lower in comparison to the two-class solution and classification quality is the highest for three-class solution than for any other solutions studied) (Table A.6 in Appendix). This three-class solution will be interpreted here.

This solution extracts the following classes on the basis of conditional response probabilities: Class One (high overall knowledge), Class Two (low overall knowledge), Class Three (high level of knowledge of correct routes and low level of knowledge of incorrect routes (some women in this group answered "yes" to all questions)).

Figure 4 shows prevalence of different classes at different points in time in the Chinese context. It suggests that the proportion of people in the first class increases with time and proportion of people in the second class reduces with time. The proportion of people in the third class first increases (misconceptions arise when the epidemic picks up) and then decreases.

Figure 4: Prevalence of different classes at different points in time in China



Note: **Class One** is high overall knowledge; **Class Two** is low overall knowledge; **Class Three** is high level of knowledge of correct routes and low level of knowledge of incorrect routes of HIV transmission.
Sources: China National Family Planning and Reproductive Health Surveys 1997 and 2001, China UNFPA Reproductive Health and Family Planning Survey 2005.

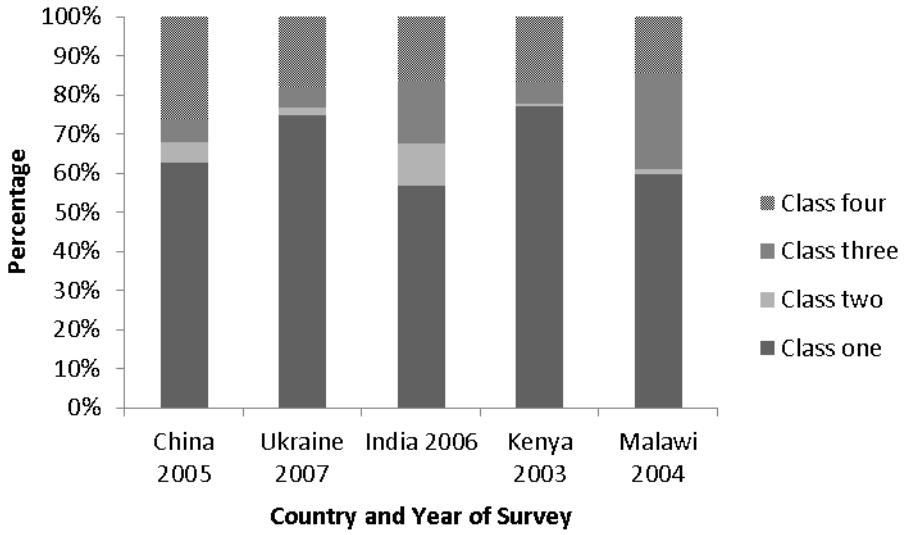
In the five country context the substantial drop in AIC and BIC statistics is observed between two- and three-class solutions and then between three- and four-class solutions (see Table A.6 in Appendix for details). Between four- and five-class solutions the drop in AIC and BIC statistics is not as sizable as between previous solutions (Table A.6 in Appendix). Classification quality is the lowest for the four-class solution and it is also slightly lower than for the five-class solution. However, as AIC and BIC do not change much between the four- and five-class solutions and, as the final model should be more parsimonious and interpretable, the four-class solution is considered to be the best solution and will be interpreted here and used here.

In the four-class solution, the following classes are extracted on the basis of conditional response probabilities: Class One (high overall knowledge), Class Two (low overall knowledge), Class Three (high level of knowledge about misconceptions and MTCT and low level of knowledge of the remaining ways to prevent HIV), Class Four (high level of knowledge about ways to prevent

HIV and about the healthy-looking person question and low level of knowledge about remaining misconceptions (some of them answered "yes" to all questions)).

Figure 5 shows prevalence of different classes in five different countries. The highest proportion of people belonging to high overall knowledge class is observed in Kenya and Ukraine whereas the highest proportion of people in low overall knowledge class is observed in India. Highest proportion of people in the fourth class is observed in China and in the third class in Malawi. As Malawi is a country with a generalised HIV epidemic, these findings about the third class being very prevalent are crucial for the formulation of necessary interventions in the country.

Figure 5: Prevalence of different classes in Five Countries



Note: **Class one** is high overall knowledge; **Class two** is low overall knowledge; **Class three** is high level of knowledge about misconceptions and MTCT and low level of knowledge about remaining ways to prevent HIV; **Class four**: high level of knowledge about ways to prevent HIV and about health-looking person question and low level of knowledge about remaining misconceptions.
Sources: China UNFPA Reproductive Health and Family Planning Surveys 2005, Kenya DHS 2003, Malawi DHS 2004, India DHS 2006, Ukraine DHS 2007.

In the Chinese context some women in Class Three are so-called "yes-women" who agree with every question and this class can be a product of acquiescence discussed earlier. According to Table A.7, women who belong to this class are more likely to be from rural areas, with either no or

only primary education and they are more likely to be 30-49 years old and married, remarried, widowed or divorced.

The latent class measures provide information about the main classes or segments of HIV knowledge observed in a population. These measures do not have a clear quantitative meaning as it is not clear how many correct answers respondents in each class obtained (this number might differ from respondent to respondent (tables with frequencies are not presented here)). However, these measures have a very clear qualitative meaning for each class as it provides information about the specific type of HIV knowledge for members of each class. Furthermore, some of the classes are impossible to identify using simple score measures. For example, simple score measures are not able to provide information about Class Three in the Chinese context or about Class Three and Class Four in the five country context. Latent class measures can be useful in partitioning populations into segment on the basis of their specific knowledge in different research situations and this information can be helpful when designing necessary interventions.

The main advantages of the latent class analysis to measuring HIV knowledge are that the approach takes into account all data included in the analysis and that it takes into account the dimensions of HIV knowledge and that it can identify the dimensions of it. The dimensions are obtained from the data rather than they are predefined by a researcher (as in a simple score approach). Therefore, different classes might be obtained in different research contexts. These different sets of classes reflect differences in available questions across different datasets as well as contextual qualitative differences in specific research contexts. This approach to measuring HIV knowledge can help partitioning populations into the segments on the basis of their specific HIV knowledge and the information about classes might be useful in some research situations and possibly interventions. Despite all the advantages of the latent class approach to measuring HIV knowledge, there are number of limitations too. Sometimes it can prove difficult to define classes obtained. From the results obtained it is not clear how many correct answers respondents in each

class obtained as these numbers might differ from respondent to respondent. It is not possible to clearly place a respondent on the HIV knowledge continuum.

Discussion

As HIV knowledge cannot be measured directly, researchers always face the choice between different measures they can use in their analysis. Assessment of different measures of HIV knowledge available can help researchers to decide more easily which measure is most suitable in their study context. Different aspects of measure suitability can be taken into account when a decision about a usage of a certain measure is taken.

This analysis is useful for the researchers who use composite variables as both response and explanatory in different areas of public health, demography and other areas of Social Sciences.

In order to understand the evolution of HIV knowledge in China two approaches were used to measure HIV knowledge: a simple score variable approach and latent class variable approach. The simple score variable approach enables us to obtain detailed information about two groups: no knowledge and perfect knowledge. For other groups, information about specific routes is not available. For example, a person with the knowledge of just one route of HIV transmission can have knowledge about one correct route or one incorrect route and the variable would not provide information about which route is known by the respondent. This group of respondents (with one correct answer) combines seven different possibilities of knowledge. It would be difficult to use the results of this analysis for specific interventions as it would be impossible to identify the specific information needs of specific groups in a population. The simple score variable approach does not allow for identification of those who, for example, belong to Latent Class Three in China or Latent Class Three and Latent Class Four in the five country context. However, simple score measures can provide useful information about the location of a specific group of respondents on HIV knowledge spectrum between no knowledge and perfect knowledge. For many research projects and general knowledge interventions this information can be of the main interest. It can be concluded that this

approach to measuring HIV knowledge provides useful insights into understanding of the evolution of HIV knowledge in China and other countries used for the analysis.

Latent class analysis was used to assess the overall knowledge, but it was possible to identify specific problems with knowledge about ways to prevent HIV in Malawi when compared with other countries and to obtain an important conclusion about the need for sustainable interventions even in contexts where HIV epidemics are generalised. It was impossible to obtain this result when a simple score of combined knowledge (S1) was analysed. Separation of two types of knowledge into two simple score variables (knowledge of correct routes (S2) and incorrect routes (S3) of HIV transmission) was required to identify this potential problem. It can be concluded that it would be appropriate to study overall knowledge and then to study specific types of HIV knowledge separately when a simple score approach is employed in order not to miss important findings.

The latent class variable approach provides information about predominant classes which exist in the sample and can help to identify specific knowledge patterns existing in populations and, therefore, to address specific information needs within each group. Variables obtained through latent class analysis are more suitable when the overall level of HIV knowledge needs to be assessed and when the purpose of analysis is to help designing interventions or to create more homogenous groups on the basis of their knowledge about HIV in the heterogeneous population. Once the needs of every specific group are identified, these groups can be targeted for specially designed educational interventions and campaigns. However, this approach to measuring HIV knowledge does not provide information about the number of known components which might be of interest in some research situations. This approach also would not allow us to precisely place groups of women on the HIV knowledge continuum on the basis of their HIV knowledge. When this specific information is of interest for designs of interventions or assessment of progress of educational and information campaigns, this particular approach to measuring HIV knowledge might not be appropriate.

The latent class variable will often have a smaller number of categories than the simple score response variable (in this study latent class variables have 3-4 categories whereas comparable simple score variable had 8 categories) and, therefore, the model will usually have less parameters to estimate. However, it is easier to obtain simple score measures. Simple score measures can be modelled using ordinal, partial proportional odds, and multinomial logistic regression models as well as linear regression models if the scores are normally distributed, whereas multinomial logistic regression models are required to model latent class variables.

Another limitation of the latent class approach is that classes are dataset specific and not easily generalisable which makes the comparison of classes across datasets difficult or even impossible whereas simple scores are easily generalisable across different datasets.

It can be concluded that both approaches provide complimentary information as they split populations in different ways and both are useful for analysing the evolution of HIV knowledge. In some research situation a specific approach can be more beneficial. For example, when health promotion interventions are of interest, latent class analysis approach can be more useful as it helps to split populations for targeted interventions and campaigns. However, for example, for evaluation purposes simple score measures could be more suitable.

The results of the analyses suggest that no measure can provide both a clear qualitative and clear quantitative meanings. However, in some specific research contexts only one of the two requirements or even a part of one requirement is sufficient.

The findings suggest that simple score measures produce ordinal variables and have clear quantitative or ordinal meaning of the categories, whereas latent class measures produce nominal variables and have clear qualitative or nominal meaning of the categories. Both measures split populations into different groups. Different types of measures can be useful in different types of research situations. Therefore, it is important for researchers to have clear objectives and then to decide which measure is sufficient and will provide them with the information needed. Table A.8 in Appendix presents the summary of comparisons of the two approaches to measuring HIV

knowledge. This table can be useful for researchers to make decision about the approach to measuring and summarising of knowledge components or similar variables. It can help them decide which approach would be more appropriate in a specific research situation.

References

Agresti, A. (ed.) (2013), *Categorical data analysis* (3rd ed.). New York: Wiley.

Ambati, B. K., Ambati, J., and Rao, A. M. (1997), "Dynamics of knowledge and attitudes about AIDS among the educated in southern India," *AIDS Care*, 9, 319-330.

Asparouhov, T. and Muthén, B.O. 2013. *Auxiliary variables in mixture modeling: 3-step approaches using Mplus*. Mplus Web Notes: No. 15. (Available from http://www.statmodel.com/download/AppendicesOct28.pdf).

Barden-O'Fallon, J. L., deGraft-Johnson, J., Bisika, T., Sulzbach, S., Benson, A., and Tsui, A. O. (2004), "Factors associated with HIV/AIDS knowledge and risk perception in rural Malawi," *AIDS Behavavior*, 8, 131-140.

Bartholomew, D. J., Steele, F., Moustaki, I., and Galbraith, J. (eds.) (2008), *Analysis of multivariate social science data* (2nd ed.), Boca Raton: Chapman & Hall.

Beadnell, B., Baker, S., Knox, K., Stielstra, S., Morrison, D. M., DeGooyer, E., Wickizer, L., Doyle A., and Oxford, M. (2003), "The influence of psychosocial difficulties on women s attrition in an HIV/STD prevention program," *AIDS Care*, 15, 807-820.

Benotsch, E. G., Pinkerton, S. D., Dyatlov, R. V., DiFranceisco, W., Smirnova, T. S., Swain, G. R., Dudko, V. Y., and Kozlov, A. P. (2004), "A comparison of HIV/AIDS knowledge and attitudes of STD clinic clients in St. Petersburg, Russia and Milwaukee, WI, USA," *Journal of Community Health*, 29, 451-465.

Bentler, P. M., Jackson, D. N., and Messick, S. (1971), "Identification of content and style: A two-dimensional interpretation of acquiescence," *Psycholological Bulletin*, 76, 186-204.

Billiet, J. B., and Davidov, E. (2008), “Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design,” *Sociological Methods and Research,* 36, 542-562.

Brown, W. J., and Bocarnea, M. C. (1998), “Assessing AIDS-related concern, beliefs, and communication behavior,” in *Handbook of sexuality-related measures* (eds. C. M. Davis, W. L. Yarber, R. Bauserman, G. Schreer, and S. L. Davis), 310-312, London: SAGE publications.

Bulik, C. M., Sullivan, P. F., and Kendler, K. S. (2000), “An empirical study of the classification of eating disorders,” *American Journal of Psychiatry,* 157, 886-895.

Central Bureau of Statistics (CBS) [Kenya], Ministry of Health (MOH) [Kenya], and ORC Macro. (2004), “Kenya Demographic and Health Survey 2003,” Calverton, Maryland: CBS, MOH, and ORC Macro.

Chen, X., Li, X., Stanton, B., Mao, R., Sun, Z., Zhang, H., Qu, M., Wang, J. and Thomas, R.(2004), “Patterns of cigarette smoking among students from 19 colleges and universities in Jiangsu Province, China: A latent class analysis,” *Drug and Alcohol Dependence*, 76, 153-163.

Clogg, C. C., and Goodman, L. A. (1984), “Latent structure analysis of a set of multidimensional contingency tables,” *Journal of the American Statistical Association*, 79, 762-771.

Cloud, J., and Vaughan, G. M. (1970), “Using balanced scales to control acquiescence,” *Sociometry,* 33, 193-212.

Dias, J. G. (2001), *Components of knowledge on AIDS in Brazil: Identifying information needs using a segmented approach*. Population Research Centre Working Paper. University of Groningen: Population Research Centre.

Durrant, G., Maslovskaya, O. and Smith, P.W.F. (2017), “Using prior wave information and paradata: Can they help to predict response outcomes and call sequence length in a longitudinal study?” *Journal of Official Statistics,* 33(3): 1-33.

Feldman, B. S., Kark, J. D., Zarka, S., Ankol, O., Letyagina, V., and Shtarkshall, R. A. (2011), "Behavioral surveillance of knowledge about HIV/AIDS transmission and perceived need for additional knowledge in a national sample of young Israeli men and women between 1993 and 2005," *AIDS Behavior*, 15, 193-203.

Fisher, J. D., Fisher, W. A., Bryan, A. D., and Misovich, S. J. (2002), "Information-motivation-behavioral skills model-based HIV risk behavior change intervention for inner-city high school youth," *Health Psychology*, 21, 177-186.

Fisher, J. D., Fisher, W. A., Misovich, S. J., Kimble, D. L., and Malloy, T. E. (1996), "Changing AIDS risk behavior: Effects of an intervention emphasizing AIDS risk reduction information, motivation, and behavioral skills in a college student population," *Health Psychology*, 15, 114-123.

Hagenaars, J. A., and McCutcheon, A. L. (ed.) (2002), *Applied latent class analysis* (1st ed.), Cambridge: Cambridge University Press.

Hamilton, D. L. (1968), "Personality attributes associated with extreme response style," *Psychological Bulletin*, 69, 192-203.

Hu, Z., Liu, H., Li, X., Stanton, B., and Chen, X. (2006), "HIV-related sexual behaviour among migrants and non-migrants in a rural area of China: Role of rural-to-urban migration," *Public Health*, 120, 339-345.

International Institute for Population Sciences (IIPS) and Macro International. (2007), "National Family Health Survey (NFHS-3), 2005–06: India: Volume II," Mumbai: IIPS.

Jiang, Z. (2000), *Data collection of 1997 National Population and Reproductive Health Survey*. China Population Publishing House. (in Chinese)

Kattumuri, R. (2003), "One-and-a-half decades of HIV/AIDS in Tamil Nadu: How much do patients know now?" *International Journal of STD and AIDS*, 14, 552-559.

Koch, P. B., and Singer, M. D. (1998), "HIV/AIDS knowledge and attitudes scales for teachers," in *Handbook of sexuality-related measures* (eds. C. M. Davis, W. L. Yarber, R. Bauserman, G. Schreer, and S. L. Davis), 317-320, London: SAGE publications.

Koopman, C., and Reid, H. (1998), "Assessment of knowledge and beliefs about HIV/AIDS among adolescents," in *Handbook of sexuality-related measures* (eds. C. M. Davis, W. L. Yarber, R. Bauserman, G. Schreer, and S. L. Davis), 321-324, London: SAGE publications.

Lagarde, E., Pison, G., and Enel, C. (1997), "Improvement in AIDS knowledge, perceptions and risk behaviours over a short period in a rural community of Senegal," *International Journal of STD and AIDS*, 8, 681.

Lau, J. T. F., Tsui, H. Y., Siah, P. C., and Zhang, K. L. (2002), "A study on female sex workers in southern China (Shenzhen): HIV-related knowledge, condom use and STD history," *AIDS Care*, 14, 219-233.

Li, X., Lin, C., Gao, Z., Stanton, B., Fang, X., Yin, Q., and Wu, Y. (2004), "HIV/AIDS knowledge and implications for health promotion programs among Chinese college students: Geographic, gender and age differences," *Health Promotion International*, 19, 345-356.

Li, B., Qi, J., Gao, W., Wu, J., Wu, Q., Brown, J. J., and Padmadas, S. S. (2008), *Baseline survey technical report: Focus on young people, gender and HIV/AIDS, UNFPA/CHINA Sixth Country Programme (2006-2010).* China Population & Development Research Centre, National Centre for Women and Children Health, Chinese Centre for Disease Control and Prevention and Southampton Statistical Sciences Research Institute, UK.

Lou, C.-h. Q., Zhao, Q., Gao, E. S., and Shah, I. H. (2006), "Can the Internet be used effectively to provide sex education to young people in China?" *Journal of Adolescent Health*, 39, 720-728.

Loue, S., Cooper, M., and Fiedler, J. (2003), "HIV knowledge among a sample of Puerto Rican and Mexican men and women," *Journal of Immigrant Health*, 5, 59-65.

Magidson, J., and Vermunt, J. K. (2004), "Latent class models," in *The SAGE handbook of quantitative methodology for social sciences* (ed. D. Kaplan), 175-198, Thousand Oaks: SAGE publications.

Maslovskaya, O., Brown, J. J., Smith, P. W. F., and Padmadas, S. S. (2014), "HIV awareness in China among women of reproductive age (1997–2005): a decomposition analysis," *Journal of Biosocial Science,* 46, 178-198.

McCutcheon, A. L. (ed.) (1987), *Latent class analysis* (1st ed.), Newbury Park: SAGE publications.

Monga, N., Rehm, J., Fischer, B., Brissette, S., Bruneau, J., El-Guebaly, N., Noël, L., Tyndall, M., Wild, C., Leri, F., Fallu, J.-S., and Bahl, S. (2007), "Using latent class analysis (LCA) to analyze patterns of drug use in a population of illegal opioid users," *Drug and alcohol dependence,* 88, 1-8.

Muthén, B. O. (1998-2004), *Mplus Technical Appendices*, Los Angeles, CA: Muthén & Muthén. (Available from http://www.statmodel.com/download/techappen.pdf).

Muthén, L. K., and Muthén, B. O. (eds.) (2009), *Mplus user's guide* (5th ed.), Los Angeles, CA: Muthén and Muthén.

National Statistical Office (NSO) [Malawi], and ORC Macro. (2005), "Malawi Demographic and Health Survey 2004," Calverton, Maryland: NSO and ORC Macro.

Pan, G., Zhang, H., and Wang, Q. (2003), *Collection of data of 2001 National Family Planning and Reproductive Health Survey. Volume 1*. China Population Publishing House. (in Chinese)

Pinkerton, S. D., Dyatlov, R. V., DiFranceisco, W., Benotsch, E. G., Smirnova, T. S., Dudko, V. Y., Belyanin, D. V., and Kozlov, A. (2003), "HIV/AIDS knowledge and attitudes of STD clinic attendees in St. Petersburg, Russia," *AIDS Behavior*, **7**, 221-228.

Stephenson, R. (2009), "A community perspective on young people's knowledge of HIV/AIDS in three African countries," *AIDS Care,* 21, 378-383.

Storr, C. L., Zhou, H., Liang, K. Y., and Anthony, J. C. (2004), "Empirically derived latent classes of tobacco dependence syndromes observed in recent-onset tobacco smokers: epidemiological

evidence from a national probability sample survey," *Nicotine and Tobacco Research*, 6, 533-545.

Stuart, B., and Hinde, A. (2009), "Identifying individuals engaging in risky sexual behaviour for chlamydia infection in the UK: A latent class approach," *Journal of Biosocial Science*, 42, 27-42.

Sullivan, P. F., Kessler, R. C., and Kendler, K. S. (1998), "Latent class analysis of lifetime depressive symptoms in the national comorbidity survey," *American Journal of Psychiatry*, 155, 1398-1406.

Ukrainian Center for Social Reforms (UCSR), State Statistical Committee (SSC) [Ukraine], Ministry of Health (MOH) [Ukraine], and Macro International Inc. (2008), "Ukraine Demographic and Health Survey 2007," Calverton, Maryland, USA: UCSR and Macro International.

UNAIDS (2008), *Report on the Global AIDS epidemic – Executive summary*. (Available from http://data.unaids.org/pub/GlobalReport/2008/JC1511_GR08_ExecutiveSummary_en.pdf).

UNAIDS (2010), *Global report: UNAIDS report on the global AIDS epidemic - 2010.* (Available from http://www.unaids.org/documents/20101123_GlobalReport_em.pdf).

Vermunt, J.K. (2010), "Latent class modeling with covariates: Two improved three-step approaches," *Political Analysis*, 18, 450-469.

Watson, D. (1992), "Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness," *Sociological Methods and Research,* 21, 52-88.

Zimet, G. D. (1998), "Adolescent AIDS knowledge score," in *Handbook of sexuality-related measures* (eds. C. M. Davis, W. L. Yarber, R. Bauserman, G. Schreer and S. L. Davis), 365-366, London: SAGE publications.

Appendices

Table A.1: Details about surveys used for the analysis

Survey	Data collection period	Data collection organisation	Sample size	Response rate
National Population and Reproductive Health Survey 1997	September-November 1997	National Population and Family Planning Commission	15,213	98.7%
National Family Planning and Reproductive Health Survey 2001	July-September 2001	National Population and Family Planning Commission	39,586	98.3%
China UNFPA Reproductive Health and Family Planning Survey 2005	November 2005	China Population and Development Research Centre [CPDRC] and National Centre for Women and Children's Health [NCWCH]	7,356	95.0%
Kenya DHS 2003	April-September 2003	Central Bureau of Statistics (CBS)	8,195	94.0%
Malawi DHS 2004	October 2004-January 2005	National Statistics Office (NSO)	11,698	96.0%
India DHS 2006	November 2005-August 2006	Ministry of Health and Family Welfare	124,385	94.5%
Ukraine DHS 2007	July-November 2007	Ukrainian Centre for Social Reforms and State Statistical Committee of Ukraine	6,841	92.0%

Sources: Jiang (2000); Pan *et al*. (2003); Li *et al*. (2008); CBS *et al*. (2004); NSO and ORC Marco (2005); IIPS and Marco International (2007); UCSR *et al*. (2008).

Table A.2: Availability of questions in different study contexts

Item number	Survey question	Chinese context	Five country context
	Correct routes of HIV transmission		
1	Can HIV be transmitted through sexual intercourse?	Yes (S1, S2)	
2	Can HIV be transmitted through sharing needles?	Yes (S1, S2)	
3	Can HIV be transmitted through blood transfusion?	Yes (S1, S2)	
4	Can HIV be transmitted through MTCT?	Yes	Yes (S1, S2)
5	Can HIV be prevented by the use of condom?	(S1, S2)	Yes (S1, S2)
6	Can HIV be prevented by abstinence?		Yes (S1, S2)
7	Can HIV be prevented by having one sexual partner?		Yes (S1, S2)
	Misconceptions about HIV transmission		
8	Can HIV be transmitted through handshaking?	Yes (S1, S3)	
9	Can HIV be transmitted through sharing utensils or eating with HIV+ person?	Yes (S1, S3)	Yes (S1, S3)
10	Can HIV be transmitted through kissing?	Yes (S1, S3)	
11	Can HIV be transmitted through mosquito bite?		Yes (S1, S3)
12	Can a healthy-looking person be HIV+?		Yes (S1, S3)

Note: S1 – Score One, S2 – Score Two, S3 – Score Three.

Table A.3: Wording of questions for variables included in the calculation of different HIV knowledge measures in China.

	China 1997	China 2001	China 2005
Blood transfusion	Can HIV be transmitted through blood transfusion?	Can HIV be transmitted through blood transfusion?	Can HIV be transmitted through blood transfusion?
Sharing needles	Can HIV be transmitted through injections?	Can HIV be transmitted through contaminated needles?	Can HIV be transmitted through contaminated needles?
MTCT	Can HIV be transmitted from the infected mother to her child?	Can HIV be transmitted from the infected mother to her child?	Can HIV be transmitted from the infected mother to her child?
Sexual transmission	Can having multiple sex partners transmit HIV?	Can having multiple sex partners transmit HIV?	Can HIV be transmitted through sexual intercourse?
Handshaking	Can HIV be transmitted through a handshake?	Can HIV be transmitted through a handshake?	Can HIV be transmitted through a handshake?
Sharing food or utensils	Can HIV be transmitted through sharing food and utensils with an HIV infected person?	Can HIV be transmitted through sharing food and utensils with an HIV infected person?	Can HIV be transmitted through sharing food and utensils with an HIV infected person?
Kissing	Can HIV be transmitted through kissing?	Can HIV be transmitted through kissing?	Can HIV be transmitted through kissing?

Sources: the China National Population and Reproductive Health Survey 1997, the China National Family Planning and Reproductive Health Survey 2001, and the UNFPA Reproductive Health and Family Planning Surveys 2005 questionnaires.

Table A.4: Wording of questions for variables included in the calculation of different HIV knowledge measures in the five countries.

	China 2005	India 2006	Kenya 2003	Malawi 2004
One partner	Would any of the following help reduce the risk of getting HIV/AIDS? Limit sex to single partner	In your opinion, can people reduce their chances of getting HIV/AIDS by having just one uninfected sex partner who has no other sex partners?	Can people reduce their chances of getting the AIDS virus by having just one sex partner who has no other partners?	Can people re chances of ge AIDS virus b just one sex p has no other p
Condom	Would any of the following help reduce the risk of getting HIV/AIDS? Condom	In your opinion, can people reduce their chances of getting HIV/AIDS by using a condom every time they have sex?	Can people reduce their chances of getting the AIDS virus by using a condom every time they have sex?	Can people re chances of ge AIDS virus b condom every have sex?
Abstinence	Would any of the following help reduce the risk of getting HIV/AIDS? Abstain from sex	In your opinion, can people reduce their chance of getting HIV/AIDS by abstaining from sexual intercourse?	Can people reduce their chance of getting the AIDS virus by not having sex at all?	Can people re chance of get AIDS virus b sex at all?
MTCT	Which of the following ways is possible for HIV/AIDS transmission? From mother to baby	Can HIV be transmitted from a mother to her baby?	Can the virus that causes AIDS be transmitted from a mother to a child?	Can the virus AIDS be trans from a mothe
Healthy-looking person	Can you tell if a healthy-looking person has the HIV virus?	Is it possible for a healthy-looking person to have HIV/AIDS?	Is it possible for a healthy-looking person to have the AIDS virus?	Is it possible healthy-looki have the AID
Mosquito bites	Which of the following ways is possible for HIV/AIDS transmission? Mosquito bite	In your opinion, can people get HIV/AIDS from mosquito bites?	Can people get the AIDS virus from mosquito or other insect bites?	Can people ge virus from mo bites?
Sharing food	Which of the following ways is possible for HIV/AIDS transmission? Sharing utensils with HIV positive person	In your opinion, can people get HIV/AIDS by sharing food with a person who has AIDS?	Can people get the AIDS virus by sharing utensils with a person who has AIDS?	Can people ge virus by shari with a person AIDS?

Sources: the UNFPA Reproductive Health and Family Planning Surveys 2005, India DHS 2006, Kenya DHS 2003, Malawi DHS 2004 and Ukraine DHS 2007 questionnaires.

Table A.5: Percentage of women who were excluded from the analysis due to missing values.

	All women	All women who are aware of HIV	All women who were asked HIV knowledge questions	Women who answered all components of HIV knowledge questions	Percentages of women with missing HIV knowledge items removed (%)
Chinese context					
China 1997	15,213	9,653	9,653	9,653	0.0
China 2001	39,586	28,781	26,138	26,138	0.0
China 2005	7,356	6,837	6,837	6,837	0.0
Five country context					
China 2005	7,356	6,837	6,837	6,837	0.0
Kenya 2003	8,195	8,052	6,912	6,877	0.5
Malawi 2004	11,698	11,548	11,024	11,002	0.2
India 2006	124,385	88,395	88,395	88,223	0.2
Ukraine 2007	6,841	6,729	6,729	6,638	1.3

Table A.6: Model fit statistics in the Chinese context (LCA) and Five Country context (LCA).

Study context	Number of classes	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)	Classification quality (%)
China	2	268534	268663	77.1
	3	249637	249836	80.9
	4	246547	246816	80.5
	5	245498	245836	77.6
5 countries	2	875453	875599	77.5
	3	864876	865099	70.9
	4	856255	856556	68.7
	5	854396	854774	71.2

Table A.7: Characteristics of women in China (1997-2005) by different classes (%).

Covariate	Class One (A: High overall knowledge)	Class Two (B: High overall knowledge)	Class Three (C: High level of knowledge about correct routes and low level of knowledge about incorrect routes)
	45.9	**11.0**	**43.1**
Year of survey			
China 1997	32.8	25.0	42.2
China 2001	44.7	7.3	48.0
China 2005	69.0	5.2	25.8
Residence			
Urban	60.2	4.7	35.1
Rural	38.8	14.1	47.1
Ethnicity			
Han	45.8	10.9	43.3
Minority	46.7	12.0	41.3
Age group			
15-19	48.5	11.0	40.5
20-29	49.1	10.0	40.9
30-39	45.4	10.3	44.3
40-49	41.4	13.4	45.2
Education			
No education	19.1	28.0	52.8
Primary	31.3	17.8	50.9
Secondary	51.3	7.4	41.3
Higher	77.1	1.0	21.9
Marital status			
Widowed or divorced	46.2	9.3	44.5
Married or remarried	44.6	11.3	44.1
Never married	52.0	9.6	38.4

Table A.8: Summary of two approaches to measuring HIV knowledge.

	Simple Score Measure	Latent Class Measure
Number of categories	8 or (5 for correct and 4 for incorrect).	2-5 (depends on the decision).
Meaning of categories	0 – no knowledge, 7 – perfect knowledge, other categories represent number of routes correctly answered without specification of which routes are known, just provides numerical value of correctly answered routes.	Each category has a specific clear meaning.
Quantitative meaning	For every group.	No groups.
Qualitative meaning	Only for 2 extreme groups.	Provide most probable patterns of answers in each group.
Ways of modelling	Ordinal, partial proportional odds, multinomial or linear regression if scores are normally distributed	Multinomial.
Difficulty of obtaining variable	Easy.	Difficult (specialised software is required).
Research context	Location of a specific group of people on a knowledge spectrum between no knowledge and perfect knowledge. Can be used for evaluations and general knowledge interventions.	Identification of predominant classes which exist in a population and their specific knowledge patterns. Can be used for designing specific interventions.
Comparability across datasets	Easy	Difficult (as data driven approach)
Limitations	No qualitative details about categories which are not extreme. Does not take into account weights for each contributing route.	No quantitative details about number of routes known.
Summary of measure	Good quantitative details and qualitative details for two extreme categories.	No quantitative details, good quantitative details about specific patterns of responses.