or socially undesirable behavior. This is also the case for hypothetical questions about stigma. While it is not surprising that there are no studies measuring enacted stigma at the general population level, the same review found surprisingly few that measure the actual experience of stigma among PLHA (Fife and Wright 2000; Berger et al. 2001; Asia Pacific Network of People Living With HIV/AIDS 2004; Swendeman et al. 2004). This study seeks to overcome this gap by investigating the occurrence of enacted stigma among three population groups: general community members, PLHA, and health care providers.

A recent study of the causes, forms, and consequences of HIV stigma in Africa untangled the complexities of stigma and identified discrete domains (Nyblade et al. 2003). Most studies of stigma measure only one or a few domains of stigma and not all of them. In addition, the more comprehensive studies reviewed by Nyblade were usually conducted in small samples, or with very narrow groups of respondents (e.g., undergraduate students in the United States), while studies with larger, more representative samples only asked a few, often ambiguous, questions related to stigma (Nyblade 2004). Two aspects of HIV stigma stand out as lacking measurement at the population level: enacted stigma and compound stigma (HIV stigma that is layered on top of pre-existing stigmas, frequently toward homosexuals, commercial sex workers, injecting drug users, women, and youth). This study undertakes a far more comprehensive investigation of stigma by including indicators in numerous domains among a broad sample of the general population and two specific populations (PLHA and health care providers).

2. METHODS

As described in the previous section, HIV-related stigma is a complex construct with multiple dimensions. Therefore, a set of items or questions (as opposed to a single one) is tested to try to capture the complexity of each key dimension. Based on the existing literature and data, we measured items in four key domains: fear of casual transmission and avoidance of casual contact with PLHA; values and attitudes, including shame, blame, and judgment; the experience of stigma and discrimination (enacted stigma); and disclosure of HIV status. The first two domains are latent, or not directly observable, while the last two are manifest or observable.

Scales were developed and tested to measure the two latent domains, while an index and single-item indicators were tested for the manifest domains. Developing scales or indices is important when a single item or question may not capture the complexity of the phenomena. A scale composed of several items offers greater validity and precision when measuring an underlying, unobservable, or latent construct. Where we cannot measure the construct directly (e.g., stigma due to attitudes and values), we assess the relationships between a set of items that we believe reflect the latent or unobservable variable, such as responses to a series of attitudinal or value statements that we expect reflect HIV-related stigma (Spector 1992; DeVellis 2003; Netemeyer et al. 2003).

The complexity of stigma also indicates the need to develop indicators to measure stigma with specific groups. While some indicators may work across multiple sub-groups of the population, others will be critical to only one or a few groups, or will need to be measured in different ways for different groups. For example, although enacted stigma is an important indicator across all groups within a population, it will be measured differently among PLHA as opposed to the general population. Additionally, some indicators may be more important for women as opposed to men if, for example, one gender experiences different forms of stigma than the other. Similarly, there are added dimensions among health care providers (e.g., work-related exposure) as compared to the general population that need to be measured, along with indicators such as fear of casual transmission of HIV. The scope of this project allowed us to focus on three groups: community/general population, people living with HIV and AIDS, and health care providers. Across all groups, a “good” indicator is one that is:

- **Valid:** an accurate measure of a behavior, practice, or task
- **Reliable:** consistently measurable, in the same way, by different observers
- **Precise:** operationally defined in clear terms
- **Independent:** non-directional and unidimensional, depicting a specific, definite value at one point in time
- **Measurable:** quantifiable, using available tools and methods
- **Timely:** provides a measurement at time intervals relevant and appropriate in terms of program goals and activities
- **Programmatically important:** linked to a public health impact or to achieving the objectives that are needed for impact

The focus of this project is to test and evaluate the indicators proposed by the S&DIWG and the Blue Book for each of four key domains, with a focus on evaluating reliability and validity of indicators that are programmatically important, timely, independent, and measurable.

Reliability is a statistical measure of the reproducibility of a survey instrument. As no measure is perfectly reliable, there is always some possibility of measurement error. When assessing the quality of a data set, one usually begins with an examination of the reliability characteristics of the measurement instrument, using three different techniques: test–re-test, alternate form, and internal consistency.

Test–re-test reliability examines the correlation in responses to the same questions, asked of the same respondent, by the same interviewer, at different points in time. The scope of this project does not allow for assessment of standard test–re-test reliability, as it did not allow for interviewers to return to the field a second time to ask the same questions. A selected few questions were asked twice within the same questionnaire/ interview, and the responses to these are compared for reliability. Certain limitations of this comparison should be noted: (1) the time elapsed between repeat questions is relatively brief (i.e., respondents are likely to remember what they answered before and question why the question is being repeated) and, (2) to ensure some “distance” between questions, repeat questions were posed at the end of the interview, when respondents often suffer from fatigue.

Inter-rater reliability examines the consistency of responses to a single question that is assessed twice with the same respondent but by different interviewers. The scope of this project did not allow for inter-rater reliability measurement.

Internal reliability examines how highly inter-correlated items within a scale are to each other. The more highly correlated, the higher the reliability of the scale. We assess internal reliability for the two latent domains (fear/refusal of contact and attitudes/values). Internal reliability was assessed using *Cronbach's Alpha*.[2]

Reliability examines to which degree items are measuring the same construct (e.g., stigma caused by attitudes and values), as opposed to *validity*, which focuses on whether the underlying variable (HIV stigma due to attitudes and values) is the true cause of the co-variation between the items being assessed (i.e., whether an item or scale is measuring what it is supposed to measure, such as HIV stigma related to attitudes and values). It is possible to have a reliable scale (all items measuring the same construct–items highly correlated) that is not necessarily valid (e.g., scale is measuring a different construct from the one intended). Validity is typically inferred from all or one of the following: content, criterion, or construct validity.

Content validity relates to the extent to which an item, or specific set of items, truly reflect the content of a particular domain. In particular, content validity focuses on the manner in which items are chosen or the scale is constructed. Content validity is typically assured by choosing items that are supported by existing data and by having experts review items. The choice of items for this project was based on existing data, in particular the collective work and expertise of the members of the S&DIWG and the questionnaires developed by a small group of experts.

Criterion-related validity, sometimes referred to as *predictive validity*, is an assessment of an item or scale (typically by correlation coefficient) against an existing criterion or gold standard. Given the nascent nature of measurement of HIV-related stigma, there is no gold standard against which to compare our data and indicators. We hope our work contributes to the development of such a gold standard.

Construct validity examines the extent to which a given measure behaves in the manner expected, given theory, hypotheses, and experience vis-à-vis other variables. In particular, it is the relationship between the item or scale being evaluated and other known/established variables in the expected direction and magnitude. For example, we might expect that a scale of fear of casual transmission of HIV will be related to knowledge about how HIV is or is not transmitted. We might expect that individuals with incomplete or incorrect knowledge of HIV will be more fearful of casual transmission of HIV than individuals with complete, correct knowledge of HIV transmission. For each of the four domains, where possible, we will detail and then examine the construct validity of the individual items or scales we are testing by examining the expected relationships of the indicator against other variables.

[2] Coefficient of reliability (consistency) measuring how well a set of items measures a single unidimensional latent construct