Assessing health-related stigma and social participation

Stevelink, S.A.M.

2011

document version
Publisher's PDF, also known as Version of record

CHAPTER 7

General discussion

"To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science" (Albert Einstein).

I. General discussion

The demand for instruments to measure stigma and (social) participation increased during the last 20 years. As a result, many measures have been developed and this gave rise to two important practical needs. The first is the need to test the psychometric properties of these measures and compare these to international benchmarks. The second need is to be able to test whether measures are culturally valid for use in the particular study context.

Two important developments in the field of health measurement provided practical tools for addressing these two needs. The first was the development of quality criteria for good psychometric properties of health measurement instruments; the second the realisation of a cultural equivalence framework (1;2). In the present thesis we investigated how these tools could be operationalized in the field of participation and stigma measures in low and middle-income country settings.

In this chapter, the research questions formulated in the 'General introduction' will be answered briefly, after which implications of these findings are discussed.

The section 'Health-related stigma and participation' describes the findings of the stigma comparison study in India, thereby answering the first research question: What are the differences and commonalities in the level of stigma and participation restrictions among persons living with HIV/AIDS and leprosy-affected persons in Southern India? (Chapter 2) This is followed by an extended description of the relationship between health-related stigma and social participation.

The third section, 'Psychometric properties' answers the second research question: What are the psychometric properties of the P-scale in a study population consisting of people with various disabilities living in Eastern Nepal? (Chapter 3) The results found during the further validation study of the P-scale in Nepal are discussed. In this study the quality criteria for 'good' psychometric properties of Terwee et al. were operationalized and tested for applicability under low-income countries circumstances (2). Furthermore, a ranking of the psychometric properties in order of importance is suggested.

The third research question concerns the conceptualization of internalized stigma and the adequate measurement of this concept (Chapter 4). The findings of this question are discussed in the fourth section, 'The assessment and conceptualization of stigma', after which the different types of stigma are conceptualized in the context of the ICF.

The section 'Participation Scale Short' answers the following research questions: 1) Can the P-scale be shortened, without negatively affecting the psychometric properties of the scale?; 2) What factor structure best fits the P-scale data from various study populations? (Chapter 5) After a discussion of the results found, practical implications for the use of the PSS are provided.

Section six, 'Cultural equivalence framework', answers the question to what extent cultural equivalence testing is addressed in cross-cultural validation studies of participation measures (Chapter 6). It also discusses the integrated framework for testing both cultural equivalence and measurement properties, presented in Chapter 6, and suggests a priority ranking of the different types of the different types of equivalence.

After the take-home messages and a reflection on the methodology, this chapter ends with an outline of future research priorities.

II. Health-related stigma and participation

This section starts with a short summary of the results found during our study in India, thereby answering the research question concerning the differences and commonalities in the level of stigma and participation restrictions among persons living with HIV/AIDS and persons affected by leprosy (Chapter 2). This is followed by a general discussion of the relationship between stigma and participation. A paragraph describing disability in relation to activity limitations, participation restrictions and impairments concludes this section.

The findings from our study conducted in India suggested that both participant groups, people living with HIV/AIDS and leprosy-affected people, reported substantial levels of internalized and perceived stigma, and participation restrictions (Chapter 2). Feelings of shame, disappointment in oneself, inferiority and lack of respect were often identified. Most restrictions were reported in the domain of work-related participation, whereas fewer restrictions were experienced in the social and mobility domains of participation.

The relationship between stigma, participation and disabling health conditions

As described in the 'General introduction', stigma and participation restrictions are known to be present across health conditions (3). In India I took a closer look at stigma and participation related to leprosy and HIV/AIDS. However, the results and its implications may be generalized to other health conditions also, as the effects of stigma found are often remarkably similar (3-5). In addition, similar items are often used across measures to examine these effects (3;6;7).

We showed an association between stigma and participation and suggested that the types of stigma investigated, internalized and perceived, may result in participation restrictions. This can be illustrated by the example given in Chapter 1, where the situation of Anand is described. Due to the beliefs and attitudes of his community members, Anand starts believing the stereotypes about leprosy-affected people and excludes himself from community and social life. Self-exclusion and other participation problems in major life domains, such as loss of social support, and difficulties in employment, have also been reported in the area of HIV/AIDS and epilepsy related stigma (8-10).

However, the question arises; can restrictions in participation also result in, or increase the level of stigma? According to findings from a study among leprosy-affected people in Nepal, affected people experienced lower educational and working opportunities, and restrictions in community life, resulting in feelings of shame and isolation (11).

There is also an interaction between stigma and participation. Stigma may result in or increase the level of participation restrictions, or the other way around, as illustrated in the following example found in the literature of an HIV-positive Antillean woman. “Anytime you want a relationship with someone, you have to think, I am sick. I have HIV and how am I going to tell this person? After a while, you reach a point that you don’t want to do it. ...I am alone and I don’t have a partner, and I don’t want a partner, because you’ll tell that person and maybe he will push you away because you have HIV” (Stutterheim et al. in press).

An association between stigma and participation was confirmed quantitatively in several other studies (11-13). Findings from a study conducted in Nepal among people with leprosy-related disabilities, showed a significant association between perceived stigma, quality of life and participation restriction (11). In a study conducted by Rensen et al. strong correlations were found between instruments measuring stigma and participation in a leprosy-affected population (12). Our studies in Southern India and Nepal confirmed these findings (see also Chapter 2 and 3). We were unable to find any other literature than what was cited above, describing associations between stigma and participation in other disabling conditions.

These examples showed that stigma may result in participation restrictions and that restrictions in participation may also lead to (an increase in) stigma. Instead of a unidirectional relationship between perceived and/or internalized stigma and participation restriction, a reciprocal relationship may be a better way to describe this relation. Since these findings came from cross-sectional studies,

no causal inference can be made. Further research needs to confirm this reciprocal relationship between stigma and participation (restrictions).

It is important to note that having a disabling condition, such as leprosy, physical disability, mental illness, will not always result in stigma or participation restrictions (14). Several studies showed that persons affected by disabling conditions, such as a physical disability or facial differences (cleft lip and palate), had no lower self-esteem compared to the general public (15;16). Furthermore, the personal response towards stigma or restrictions in participation may differ across situations and in people with different personalities.

For a detailed conceptualization of internalized stigma please refer to Chapter 4 and to section 4 in this chapter, 'The assessment and conceptualization of stigma'.

Activity limitations, participation restrictions, impairment and disability

A disabling condition may result in stigma and affect body functions and structures, activities and social participation; however, when do we classify someone as 'having a disability'? Often the physical impairments come first. Personal factors in combination with environmental factors are determinants of disability, whereas activity limitations and participation restrictions are difficulties in functioning that define disability (17). It is the interaction between these components that cause a person to experience disability.

Impairment is closely related to activity and social participation; however, the role of impairment deserves separate consideration. Impairment is often seen as being part of the definition of 'disability' (17). In other words, if someone has an impairment, they also have a disability. However, if the experience of the person is the deciding factor with regard to the presence or absence of disability, then a definition of disability that includes impairment cannot hold. There are many situations where impairment does not result in a disability, because it is somehow compensated. For example, a person who has visual impairment due to near-sightedness of both eyes may, under certain circumstances, have considerable activity limitations and even participations restrictions and thus experience a disability. However, wearing the right glasses, this 'impairment' has no adverse effects on his performance, because the impairment is fully compensated. In such a situation, the person would not classify himself as 'having a disability'. Therefore, a person having an impairment cannot always be classified as 'having a disability'. This is subject to the experience of the person concerned and depends on the interaction with activity and participation.

III. Psychometric properties

This section answers the research question concerning the psychometric properties of the P-scale among people with various disabilities (Chapter 3). During this study in Nepal a quality criteria framework was used (2). This framework is described briefly, after which a summary is provided of the results found on the psychometric properties of the P-scale. This section closes with a priority ranking of which psychometric properties to address, when time and resources are limited.

The quality criteria framework used during the study in Nepal was proposed by Terwee et al (2). In this framework, nine quality criteria for what constitute 'good' psychometric properties were described: content, construct and criterion validity, internal consistency, agreement, reliability, responsiveness, interpretability, and floor- and ceiling effects (2). An elaboration is given on which statistical methods to use to assess a particular property. In addition, criteria were described to rate the outcome as 'positive', 'negative', 'indeterminate', or 'no information' (2). A detailed description of this framework can be found in Chapter 3 and 4 (Appendix 1) of this thesis.

We operationalized this framework during the validation study of the P-scale in Nepal. The results confirmed the validity and the reliability of the P-scale among people with various disabilities. Construct validity, internal consistency, test-retest reliability, and floor- and ceiling effects were good according to the quality criteria proposed by Terwee et al. (Chapter 3). Furthermore, no problems were identified during the operationalization of the quality criteria framework in this low-income setting.

Priority ranking of the psychometric properties

From our own experience, we are well aware of the tension between the need to adequately (cross-) validate an instrument and the reasons why investigators often only test a few psychometric properties. These reasons may include financial constraints, lack of time, lack of knowledge or expertise for testing these properties, or a lack of awareness of the importance of testing measurement properties. We decided therefore to recommend a priority ranking of the most important properties to address during an instrument validation study.

The measurement aim of the instrument is important to make a well-argued decision regarding which psychometric properties to test. However, regardless of the measurement aim, it is always important to address content validity. A construct can only be assessed thoroughly when the items of the concerned instrument represent the construct adequately and when the items are relevant to

the target population. If this is not the case, it is not useful to assess any other psychometric properties (2).

We can distinguish three different measurement aims: evaluative, discriminative and predictive (2;18). Each is briefly discussed below.

An *evaluative instrument* may aim to measure the effect over time of an intervention concerning the construct of interest in a group of respondents (2;18). This is important when effects of stigma reduction interventions have to be evaluated with a particular instrument. For that reason the instrument should be sensitive to changes over time. This can be addressed by investigating the responsiveness of an instrument, defined as "the ability of an instrument to detect change over time in the construct to be measured" (19). Therefore, to use an instrument for evaluative purposes, the smallest change the instrument is able to detect should exceed the minimal change in score that is considered as important. (2;18). 'Agreement', paradoxically also called 'measurement error', should be assessed, since this property will show whether the instrument is able to distinguish between the measurement error of the instrument and the real change in score (2). These statistics are derived from the absolute measurement error, which has to be smaller than the changes one wishes to be able to measure. Furthermore, floor- and ceiling effects have to be investigated. The responsiveness of a particular instrument may be hampered when floor- or ceiling effects are present (2). For example, if there are many respondents that already scored the highest possible score (e.g. high levels of stigma) and stigma levels are worsening over time, the instrument will not be able to detect this change. Or the other way around, respondents that already had the best possible score may improve over time, but this would remain undetected.

Discriminative instruments intend to distinguish between respondents or groups of respondents regarding the construct of interest (2;18). A possible application may be the quantification of participation restrictions across communities. For discriminative purposes the assessment of reliability is very important because the differences identified between respondents should be reproducible and remain stable over time (2;18). Therefore the within-person changes should be small. Otherwise it is not possible to distinguish differences *between* respondents from those *within* the respondents. Floor- and ceiling effects are also important to address for a discriminative instrument. If too many respondents score the highest or lowest possible score, one will not be able to distinguish adequately between respondents (2).

Predictive instruments are of particular interest in measurement areas where a golden standard exists. However, in our area of health measurement (stigma and participation), this is not the case. A situation in which we do use this perspective is the comparison between the short version of an instrument with the full version. In this case, the latter serves as the golden standard (20). A predictive instrument is typically used for screening purposes (18). For example, we may use a short instrument to select people who may be in need of rehabilitation assistance; the gold standard would be a detailed needs assessment. The most important property to address will be criterion validity, showing whether the instrument correlates with the golden standard.

Based on the measurement aim of the instrument, it is therefore possible to make a balanced decision about which psychometric properties should have priority in the investigation. However, if time and other resources allow, the other psychometric properties should also be evaluated. Internal consistency, item-total correlations and missing values can be assessed without any additional testing, and should therefore be assessed anyway. If existing instruments are selected for use in a survey or other context, it is important to base the selection on available evidence of psychometric properties considered essential for the particular aim.

IV. The assessment and conceptualization of stigma

This section the following research addresses the question: How is internalized stigma conceptualized and to what extent is it being measured adequately? (Chapter 4) To answer this question, a systematic review was conducted of internalized stigma measures. During this review, the psychometric properties of these measures were assessed using the quality criteria proposed by Terwee et al. (2). After the summary of the results found in the systematic review, the rationale for using these particular quality criteria is addressed briefly; after which a detailed conceptualization is provided, positioning individual and public stigma in the ICF framework.

For details on the ICF framework, see the 'General introduction'. The quality criteria used were briefly described in the section 'Psychometric properties' of this chapter. For an extended overview of these criteria please refer to Chapter 4, Appendix 1.

Psychometric properties of internalized stigma measures

The results of the systematic review suggested that the testing of psychometric properties of internalized stigma measures leaves much to be desired. Content and construct validity, as well as internal consistency were most often addressed during these studies. Agreement and responsiveness were not investigated in most cases. Of the 21 instruments included, only two scored three positive

ratings, out of a maximum of nine. Therefore we recommended further validation of these instruments (Chapter 4).

During this review, we used the quality criteria proposed by Terwee et al. to evaluate the internalized stigma instruments (See section 'Psychometric properties' of this Chapter) (2). Similar, but also some additional criteria (e.g. administration time, respondent burden) have been proposed for evaluative purposes by the Scientific Advisory Committee of the Medical Outcomes Trust, McDowell & Jenkinson, and Andresen (21-23). However, while these issues are important, the utility of these criteria is limited for rating the methodological quality of studies, because no exact statements are made about what constitutes 'good' psychometric properties.

In 2010, a consortium of experts published the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist (19;20;24). The checklist was established during an international Delphi study comprising most major research groups working on health measurement instruments and contains criteria and a checklist for testing that can be used to assess the methodological quality of psychometric properties (19;20;24). The majority of the criteria correspond with those proposed by Terwee et al. (2). However, some properties are named differently (e.g. agreement vs. measurement error) or comprise sub-domains. An example for the latter includes the presence of floor- and ceiling effects (2;24). In the quality criteria framework this is a separate property, whereas the COSMIN consortium classifies floor- and ceiling effects as part of 'interpretability' (2;24).

An important advantage of the framework proposed by Terwee et al. is the presence of a scoring system (2). With this scoring system it is possible to classify outcomes on the particular properties as positive, indeterminate, negative or no information. However, no scoring system was available for the COSMIN checklist and therefore we decided to use the quality criteria as proposed by Terwee et al. (2). Only very recently a scoring system was published for the COSMIN checklist (25).

Stigma and the ICF

The conceptualization of stigma in relation to the ICF warrants further discussion. The ICF comprises three components namely the health condition, functioning, and contextual factors. Contextual factors comprise 'Personal factors' and 'Environmental factors' (Chapter 1). Stigma has been described as an environmental factor in the ICF framework. However, stigma is a multi-faceted construct and based on our results, this is insufficiently conceptualized in the ICF. We propose that

certain types of stigma, especially from the perspective of the stigmatized, are more appropriately located in the component 'Personal factors'.

Different types of stigma can be identified from the perspective of those who stigmatize (stigmatizers) and those who are stigmatized (5). Public stigma occurs once "large social groups endorse stereotypes about and acting against a stigmatized group" (26). In our case, this relates to people affected by various disabling conditions. The behaviour and (emotional) reactions of the stigmatizers reflect their attitudes, values and beliefs about the affected person. These external factors that are part of the environment often have a strong negative impact on the affected person.

At the individual level, several types of stigma occur, such as internalized stigma, perceived stigma and experienced stigma. Internalized stigma, also called 'self-stigma', occurs when people internalize the public stigma, consider the beliefs and stereotypes as relevant to themselves and experience loss in self-esteem and self-efficacy (14;26). Perceived stigma refers to the person's awareness of negative perceptions held by others about disabling conditions (27). Experienced stigma refers to the actual experience of discrimination (3;28). I propose that these types of stigma fit better within the ICF component 'Personal factors', than in the external 'Environmental factors'.

These types of stigma can be described as psychological phenomena or mechanisms. The effect of actual experiences of stigma and the awareness of stereotypes depend on a person's internal coping style and individual characteristics. Perceived and experienced stigma will not always result in diminished self-esteem and self-efficacy, as also suggested by Corrigan et al. (14;27). However, once a person applies such stigma to himself and experiences such feelings, internalization of the stigma has occurred (27).

In conclusion, internalized, perceived and experienced stigma should be included in the component 'Personal factors' of the ICF, whereas public stigma fits better within the component 'Environmental factors'.

V. Participation Scale Short (PSS)

This section answers two research questions: 1) Can the P-scale be shortened, without negatively affecting the psychometric properties of the scale?; 2) What factor structure best fits the P-scale data from various study populations? (Chapter 5) First the development of the PSS is discussed, after which practical implications are provided for its use.

In Chapter 5 of this thesis, the development of the 13-item PSS using a large multi-country database is described. During this study, also several psychometric properties of the PSS were tested, showing good results. Factor analysis suggested the best fit for a two-factor model, 'work-related participation' and 'general participation' (CFI = 0.983, TLI = 0.979, RMSEA = 0.061). These fit indices were even higher than in the full P-scale. The Cronbach's alphas of the PSS were good. For the whole scale the alpha was 0.91; the subscales 'work-related participation' and 'general participation' showed alphas of 0.83 and 0.90, respectively. These were only slightly lower compared to the full P-scale (α = 0.92, 0.83 and 0.92, respectively). A very high correlation was found between the two factors (*r* = 0.75), suggesting that both are part of one overall factor, 'participation'. As expected, both versions of the scale were very highly correlated (*r* = 0.99).

The main advantages of the PSS compared to the full version of the P-scale are a reduced administration time and respondent burden. Especially the latter is important. Often a toolkit of instruments is used during a survey and even a slight reduction in number of items will reduce the burden of the respondent. However, a potential disadvantage of the shortening may be reduced internal consistency and the occurrence of floor- and ceiling effects (29). Therefore it is important that the quality of the data be ascertained using appropriate analyses.

Utility of the Participation Scale Short

The psychometric properties of the P-scale in this study, as well as other studies, showed good results (11-13;30). The P-sale can be used for several purposes, for example to provide a profile of participation restriction or as a screening tool. An example of the former is shown in Chapter 2, where a visual comparison of the pattern of restrictions between respondent groups or between conditions was made. However, the P-scale can also be used as a screening tool to investigate which respondents may be in need of an intervention. For a screening purpose like this, the shorter the instrument, the better. Thus the PSS will have preference over the full P-scale. Another situation, in which the PSS may be preferred, would be during a large survey using a battery of instruments and aiming to cover all the different components of the ICF. In these cases, the PSS will be a good choice.

However, it is important to note that the P-scale itself is not a needs assessment tool. After the P-scale results indicate a possible need, an in-depth exploration is necessary to identify the exact needs of the respondent. This may involve interviews with the affected person, family members and other stakeholders. In addition, the intervention(s) that might benefit the respondent could be discussed.

In Chapter 5 the suggestion was offered that besides using the P-scale as a general measure of participation (restrictions), the 'work-related participation' subscale, may be useful as an indicator instrument for specific work-related problems. Work-related problems are very common and are often the highest scoring items in the scale. This indicates that a 'work indicator' may be particularly relevant. In this study, the subscale showed good internal consistency ($\alpha = 0.83$). The subscale is very short and can be useful to check whether persons are in need of specific work-related interventions or to monitor the effect of such interventions.

Furthermore, the other subscale, 'general participation', may be of use in certain study populations in which employment is not relevant yet, or not any more. Examples are children and elderly people. Also this subscale showed good internal consistency in short and full form ($\alpha = 0.90$ and 0.92, respectively).

To be able to use separate subscales and a shorter version of the P-scale has potential advantages. However, their utility will be need to be confirmed by empirical studies before application can be recommended more widely.

VI. Cultural equivalence framework

This section starts with a summary of the findings from our systematic review that described the process of cultural equivalence testing in participation instruments. It addresses the following research question: To what extent is cultural equivalence testing addressed in cross-cultural validation studies of participation measures? (Chapter 6) This is followed by a description of the cultural equivalence framework used and a priority ranking of the different types of equivalence. The difference between cultural equivalence and cultural validity is discussed, followed by a paragraph on the involvement of the target population. This section ends with a paragraph describing the usefulness of mixed methods.

In Chapter 6 we proposed an extended cultural equivalence framework to facilitate the process of cross-cultural equivalence testing. The framework comprises conceptual, semantic, item, operational and measurement equivalence, and integrates theoretical perspectives with the practical knowledge obtained during the research for this thesis. Based on the findings presented in the particular study, the categories of equivalence can be rated 'extensively', 'partially', or 'none/minimally'.

This framework serves several purposes. An important purpose may be the assessment of cultural equivalence testing in cross-validation studies, as done in our review. In that review, we concluded

that cultural equivalence testing of participation instruments was not performed adequately. Eight cross-cultural validation studies were included. Of these eight studies only three received at least two 'extensive' ratings for the different categories of equivalence. Conceptual, item and semantic equivalence were most often rated 'extensively'. The majority of the ratings given were 'none/minimal' and 'partial'. Therefore we concluded that cultural equivalence testing was performed insufficiently.

Cultural equivalence testing

The framework can be applied also for designing a (cross-)validation study for a particular instrument that is to be used in a different cultural setting.

I will illustrate the need for cross-cultural adaptation with a story (van Brakel, personal communication). A scale developed in North America was to be used in an African country. It contained an item asking the respondent whether their current problem was "like the tip of an iceberg". This was readily understood to mean "a small sign of a much bigger problem" by most respondents in North America. The item was carefully translated in the local language. Not surprisingly, however, the item missed the point completely, since many people had never heard of icebergs or of their characteristics. An African co-worker then suggested a more culturally appropriate translation; instead of "like the tip of an iceberg" the translation read "like the eyes of a hippopotamus". This conveyed the meaning of the item perfectly and was readily understood by the respondents.

The metaphor of the hippopotamus may be used to illustrate another aspect of the process of cross-cultural validation of an instrument. For many investigators, it is clear that an instrument has to be translated into the appropriate language – the eyes of the hippo are obvious. However, there is more than meets the eye. Beneath the water surface the size and the shape of the hippo illustrate that many other aspects need attention. Examples of these are the different categories of equivalence: conceptual, item, semantic, operational and measurement equivalence.

Prioritizing the different types of equivalence

Taking these findings into account and considering this demand, the question arises which categories of equivalence are most important; which categories should have preference if time and other resources are scarce? According to Herdman, Fox-Rushby and Badia, the different categories of equivalence have to be investigated in the following order: conceptual, semantic, item, operational,

and measurement equivalence (1). However, each component comprises several subcategories, which may not all be equally important (see Chapter 6).

Based on our results, all three subcategories of conceptual equivalence have to be addressed; 1) 'local populations' conceptualization', 2) 'appropriateness of the measure' and 3) 'theoretical arguments'. This is because concepts, such as participation and disability, may differ across cultures (1). Expressions of the concept under investigation may therefore vary. For that reason I stress the importance of evaluating the local populations' conceptualization of the construct, before the assessment of the suitability of the instrument in the light of the cultural meaning of the concept. Sufficient information on conceptual equivalence at an early stage of the validation process will prevent problems in operationalization and measurement at a later stage. This will help to justify the use of the instrument in a different culture than it was initially developed for (31). Especially input from local people is important, because they are the experts on the local culture. They can provide input for the local population's conceptualization as well as the appropriateness of the measure. In addition, local researchers can be asked to judge the appropriateness of the measures. Furthermore, discussion between the researchers may contribute to the theoretical foundation of the instrument and may highlight the practical implications of the input from the local people.

The findings regarding the conceptual equivalence of an instrument indicate the suitability of the measure for adaptation, after which translation and/or adaptation can take place. The 'translation procedure' is a subcategory of semantic equivalence. Semantic equivalence consists of five subcategories; 1) 'contact with initial developer', 2) 'use of translation protocol', 3) 'translation procedure', 4) 'initial meaning of key words and phrases' and 5) 'description of translation problems or difficulties experienced'. Adequate translation is important, otherwise interviewers will experience problems with the administration of the scale, respondents will get confused, and/or the original meaning behind items will not be reflected. Therefore, the subcategory 'translation procedure', and 'initial meaning of key words and phrases' of the semantic equivalence component should be addressed. The other categories of semantic equivalence, number 1, 2 and 5, were considered less important, because they only contribute indirectly to the appropriate translation of an instrument.

The reasons why adequate translation is important apply also to the category item equivalence. Therefore, all three sub-categories of item equivalence should be examined. These are; 1) 'assessment relevancy and acceptability of the items', 2) 'discussion of findings on relevancy and acceptability of the items', and 3) 'well-argued adaptations made of items'. Examples for investigating item equivalence include in-depth interviews or focus-group discussions with the target

population, discussion among researchers and research assistants and quantitative methods such as Rasch analysis.

The fourth category of equivalence, operational equivalence, comprises 3 subcategories; 1) 'missing data', 2) 'administration format' and 3) 'pre-testing'. The subcategory 'administration format' of operational equivalence is essential to address, and especially the following three aspects: 'suitability of the questionnaire format', 'item format' and 'response options'. It is important that respondents understand the response options used, as well as the framing of the items. Otherwise the answers will not reflect the real situation and results will be useless. In addition, using a self-report instrument in a population with low literacy rates is not possible and therefore the choice for a particular administration method needs careful consideration. The other two aspects of the category 'administration format' were identified as less important ('educational levels' and 'instructions interviewers'). In addition to 'administration format', pilot-testing is essential to tailor the instrument to the needs of the population under study. The sub-category 'missing data', was identified as less important.

Measurement equivalence is important to assess, because psychometrics show whether a measure is able to assess a given construct accurately and reliably. For a detailed description of the most important psychometric properties to address, please refer to the paragraph 'Psychometric properties' in this chapter.

If circumstances so dictate, we suggest that testing could be limited to the categories indicated as 'priority categories'.

In addition, I would like to highlight three important issues related to the cultural equivalence framework, namely the difference between 'culturally valid' and 'culturally equivalent', the involvement of the target population and the use of 'mixed methods'.

Cultural validity vs. cultural equivalence

I am well-aware that the conceptualization of social participation and health-related stigma may differ across cultures. This has direct consequences for the validity of instruments to assess these constructs. We used the term 'cultural equivalence testing' to indicate the process of testing to what extent a given adaptation has resulted in an instrument that is equivalent to the original. An instrument is culturally equivalent, once the complete instrument can be transferred to the new culture, without major adaptations to its original form and content (besides translation and contextualisation). Cultural equivalence is required for direct comparison of instrument scores

between languages, cultures or countries. This is a stricter requirement, however, than establishing cultural validity. Sometimes, results from a pilot-study may indicate that rigorously adaptations are needed, before the instrument can be used in the particular study setting. This was done, for example, in the study by Misajon et al. (Chapter 6). They omitted a particular subscale and collapsed the item response categories from six to four (32). With help of these adaptations, the instrument was made culturally valid for use in Thailand. An instrument can be said to be 'culturally valid' if it has, good construct and content validity, adequate psychometric properties and has been shown to be culturally acceptable in the local context. If major adaptations have been necessary the instrument may be culturally valid for use in the new context, but it cannot be said to be culturally equivalent to the original. The concept of cultural validity applies to an instrument that has been newly development as well as to instruments adapted for use in a new cultural context.

Involvement of the target population

During the contextualisation process, it is important to check whether the main construct on which the instrument was based, is culturally valid and relevant in the particular target setting. Therefore knowledge is necessary about behaviour patterns, community structures and social rules. The local population's conceptualization of the construct is natural and logical for people living in that culture, however, for external researchers the understanding may be quite complex (31). Local people are the best source of information on local beliefs, values and understanding of language and their involvement is essential. Ideally, instrument development or adaptation should be undertaken by local researchers in close collaboration with the local people themselves. This can be done using focus groups discussions, participant observation and in-depth interviews (1). Local people may indicate the relevance of particular domains, difficulties in understanding items, acceptability of items, frequently used words and how they are understood. Local people should be involved during the item selection phase of an instrument. Sometimes, some of this information can also be found by reviewing the relevant local literature, local instruments (if available), or through discussion with local researchers (1). If local concepts have not been assessed adequately, it is difficult to check whether responses truly reflect the respondent's situation and give a complete picture of the construct under investigation. Therefore it is important that the target population is involved during a cross-validation or instrument development study.

The use of mixed methods

The combination of qualitative and quantitative methods strengthens this framework. Qualitative methods provide richer data and a contextual basis for the interpretation of the data. This is important since cultural equivalence testing aims to facilitate context-specific research. Qualitative

methods are essential to investigate the cultural basis for conceptual equivalence, content and construct validity, and validity related to the understanding and relevance of the items of a scale. The detailed findings from semi-structured interviews or focus group discussions can be used as input for an instrument development or adaptation study; provide leads for the development of interventions or campaigns, and ascertain that instruments used or interventions designed are culturally relevant and acceptable.

The limited generalizability of the qualitative findings can be compensated by using quantitative methods (33). Quantitative data from a representative sample of people can capture the heterogeneity and variability of the population under study. Quantitative methods use statistical procedures to investigate the validity and reliability of a particular instrument. The application of, for example, the quality criteria of Terwee et al. such as internal consistency, construct validity, reliability, agreement and responsiveness, and also the application of Rasch analysis, will minimize the subjective interpretation of findings and allow for the assessment of longitudinal validity (2). In addition, quantitative methods allow validation of an instrument in large and diverse samples, which helps in establishing the cross-condition validity of the instrument. The application of both methods increases the interpretability and allows triangulation of the findings.

The sequence of the application of both methods depends on the purpose of the study (31). For example, during an instrument development study, qualitative methods may help to identify the core domains and develop underlying theoretical concepts and hypotheses that are most relevant for the construct under investigation. Thus the combination of quantitative and qualitative methods is important to develop a meaningful and relevant instrument.

VII. Take-home messages

Several field studies and systematic reviews were conducted to investigate whether the quality criteria for psychometric properties and the cultural equivalence testing framework were applicable in the field of participation and stigma measures, and in low- and middle-income country settings. Based on our results, several take-home messages were formulated, which are listed in Box 1 on the next page.

Box 1: Take-home messages

1. Persons living with HIV/AIDS and leprosy-affected persons reported substantial levels of stigma and participation restrictions.

2. There may be a reciprocal relationship between stigma and social participation.

3. The Participation Scale showed sufficient psychometric properties across study populations.

4. The subscales 'work-related participation' and 'general participation' of the Participation scale may be useful as separate instruments.

5. The Participation Scale Short is a promising instrument.

6. The different types of individual stigma can be incorporated in the component 'personal factors' of the International Classification of Functioning, Disability and Health.

7. Psychometric property testing in internalized stigma measures is inadequate.

8. Psychometric properties can be prioritized based on the measurement aim of the instrument.

9. Cultural equivalence and cultural validity are important concepts to address and needs more attention.

10. The cultural equivalence of participation instruments should be tested more fully, before they are used in other countries or cultural settings.

VIII. Reflection on methodology

This thesis has some important limitations that I would like to highlight before providing future research priorities. One of these is the focus on the quantitative aspects of measuring stigma and participation. During our field studies we used only quantitative methods, such as the P-scale, ISMI and EMIC, to assess stigma and participation. The application of qualitative methods would have given us in-depth insights in the experiences of the people we interviewed. Especially for the conceptualization of both constructs in the local context, this would have been an advantage. In addition, it would have allowed us to triangulate our data, which may have strengthened our findings. Therefore detailed qualitative studies of this nature are a research priority, as outlined in section nine. However, taking into account the quality criteria proposed by Terwee et al. we are convinced that our results from the field studies are valid and reliable, thereby providing an accurate impression of the level of stigma and participation restrictions in the populations under study (2). Our findings also confirm the utility of the P-scale in different target settings.

Another methodological limitation is the cross-sectional nature of the field studies performed. As a result we were unable to study measurement properties related to change over time, such as responsiveness. This property is very important to assess the effectiveness of for example stigma reduction interventions. Therefore, I recommend an assessment of the longitudinal validity of the concerned measures included as another priority for future research. Furthermore, due to this research design, conclusions about the causality of particular associations between stigma and participation could not be drawn.

Besides these limitations, this thesis also has an important methodological strength. By applying the quality criteria of Terwee et al. during the validation of the P-scale in Nepal, developing the PSS and investigating the psychometric properties of internalized stigma measures in a systematic review, we made use of state-of-the-art statistical methods (2).

IX. Future research priorities

In the previous chapters, I outlined several future research priorities. In this section I address the ones that are in my opinion very important.

In the first place, a detailed assessment of the scaling properties of the P-scale should be done with the help of Rasch analysis. Such an analysis would identify the extent to which items are statistically equivalent across the different translation versions and in different study populations.

Second, an extensive conceptualization of participation and stigma with help of qualitative methods will provide useful insights in the local conceptualization, meaning, and interpretation of both concepts in different study populations and across culture. This will also contribute to the adequate interpretation of the quantitative results found. In addition, with the help of in-depth interviews or focus group discussions, the relevancy and acceptability of the measures can be investigated in detail.

Third, the shortened version of the P-scale, the PSS, looks promising. However, this instrument should be validated in different study populations. Field testing needs to confirm our expectations about the validity of the PSS, the reduced respondent burden and decreased administration time.

Fourth, the P-scale is especially developed for use in low and middle-income countries. Other well-known participation instruments may also be suitable for use in this particular context. Therefore it is important to conduct instrument comparison studies to investigate the relative utility of instruments that measure participation in low and middle-income countries.

Fifth, we suggested that the testing of cultural equivalence is very important. However, the complete cultural equivalence framework proposed is very extensive. Therefore the development of a (shortened and) practical field protocol for the assessment of cultural validity would be helpful. We already suggested a priority ranking for the most important (sub)-categories of equivalence to be addressed. This framework needs to be tested in practice during a (cross-) validation study to improve its utility. Such a study can be used also to further shorten the protocol.

Finally, two generic stigma instruments should be developed with broad inter-disciplinary consensus, one for the stigmatizers' perspective and one for the perspective of the stigmatized. Nowadays, health condition-specific instruments are often used. This is of interest to obtain specific knowledge that can be used to design health condition-specific interventions. An example is items that relate to

sexual behaviour in a HIV/AIDS stigma scale. However, in many situations, such as a stigma assessment in a public health context, generic instruments will be preferred. Generic instruments also allow comparison of data across conditions, which may be important in determining the effectiveness of interventions targeting people with a variety of stigmatized health conditions.

Reference List

(1) Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. Quality of Life Research 1998 May;7(4):323-35.

(2) Terwee CB, Bot SDM, De Boer MR, van der Windt DAWM, Knol DL, Dekker J et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 2007 January;60(1):34-42.

(3) Van Brakel WH. Measuring health-related stigma--a literature review. Psychology Health and Medicine 2006 August;11(3):307-34.

(4) Weiss MG, Ramakrishna J, Somma D. Health-related stigma: rethinking concepts and interventions. Psychology Health and Medicine 2006 August;11(3):277-87.

(5) Weiss MG. Stigma and the social burden of neglected tropical diseases. PLoS Neglected Tropical Diseases 2008;2(5):e237.

(6) Brohan E, Slade M, Clement S, Thornicroft G. Experiences of mental illness stigma, prejudice and discrimination: a review of measures. BMC Health Services Research 2010 March 25;10.

(7) Noonan VK, Kopec JA, Noreau L, Singer J, Chan A, Masse LC et al. Comparing the content of participation instruments using the International Classification of Functioning, Disability and Health. Health and Quality of Life Outcomes 2009 November 13;7.

(8) Malcolm A, Aggleton P, Bronfman M, Galvão J, Mane P, Verrall J. HIV-related stigmatization and discrimination: its forms and contexts. Critical Public Health 1998;8(4):347-70.

(9) Varas-Diaz N, Serrano-Garcia I, Toro-Alfonso J. AIDS-related stigma and social interaction: Puerto Ricans living with HIV/AIDS. Quality of Health Research 2005 February;15(2):169-87.

(10) Allotey P, Reidpath D. Epilepsy, culture, identity and well-being: a study of the social, cultural and environmental context of epilepsy in Cameroon. Journal of Health Psychology 2007 May;12(3):431-43.

(11) Brouwers C, Van Brakel WH, Cornielje H, Pokhrel P, Dhakal KP, Banstola N. Quality of life, perceived stigma, activity and participation of people with leprosy-related disabilities in south-east Nepal. Disability, CBR and inclusive development 2011;22(1):16-34.

(12) Rensen C, Bandyopadhyay S, Gopal PK, Van Brakel WH. Measuring leprosy-related stigma - a pilot study to validate a toolkit of instruments. Disability and Rehabilitation 2010 August 7;33(9):711-9.

(13) Stevelink SAM, Van Brakel WH, Augustine V. Stigma and social participation in Southern India: Differences and commonalities among persons affected by leprosy and persons living with HIV/AIDS. Psychology Health and Medicine 2011 February 28;1-13.

(14) Corrigan PW, Watson AC. The paradox of self-stigma and mental illness. Clinical Psychology-Science and Practice 2002;9(1):35-53.

(15) Arnold P, Chapman M. Self-Esteem, Aspirations and Expectations of Adolescents with Physical-Disability. Developmental Medicine and Child Neurology 1992 February;34(2):97-102.

(16) Clifford E, Clifford M. Social and psychological problems associated with clefts: Motivations for cleft palate treatment. International Dental Journal 1986;36(115):119.

(17) World Health Organization. World report on disability. Geneva: WHO; 2011.

(18) Kirshner B, Guyatt G. A Methodological Framework for Assessing Health Indexes. Journal of Chronic Diseases 1985;38(1):27-36.

(19) Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology 2010 July;63(7):737-45.

(20) Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Medical Research Methodology 2010;10:22.

(21) Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E et al. Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research 2002 May;11(3):193-205.

(22) Andresen EM. Criteria for assessing the tools of disability outcomes research. Archives of Physical Medicine and Rehabilitation 2000 December;81(12):S15-S20.

(23) McDowell I, Jenkinson C. Development standards for health measures. Journal of Health Services and Research Policy 1996 October;1(4):238-46.

(24) Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010 May;19(4):539-49.

(25) Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research 2011 July 6.

(26) Corrigan P, Kerr A, Knudsen L. The stigma of mental illness: explanatory models and methods for change. Applied & Preventive Psychology 2005;11(3):179-90.

(27) Corrigan PW, Watson AC, Barr L. The self-stigma of mental illness: Implications for self-esteem and self-efficacy. Journal of Social and Clinical Psychology 2006 October;25(8):875-84.

(28) Scambler G. Stigma and disease: changing paradigms. Lancet 1998 September 26;352(9133):1054-5.

(29) von Korff M, Crane PK, Alonso J, Vilagut G, Angermeyer MC, Bruffaerts R et al. Modified WHODAS-II provides valid measure of global disability but filter items increased skewness. Journal of Clinical Epidemiology 2008 November;61(11):1132-43.

(30) van der Zee CH, Priesterbach AR, van der Dussen L, Kap A, Schepers VPM, Visser-Meily JMA et al. Reproducibility of Three Self-Report Participation Measures: the ICF Measure of Participation and Activities Screener, the Participation Scale, and the Utrecht Scale for Evaluation of Rehabilitation-Participation. Journal of Rehabilitation Medicine 2010 September;42(8):752-7.

(31) Kelle U. Combining qualitative and quantitative methods in research practice: purposes and advantages. Qualitative Research in Psychology 2006;3(4):293-311.

(32) Misajon R, Pallant JF, Manderson L, Chirawatkul S. Measuring the impact of health problems among adults with limited mobility in Thailand: further validation of the Perceived Impact of Problem Profile. Health and Quality of Life Outcomes 2008 January 21;6.

(33) Bryman A. Integrating quantitative and qualitative research. Qualitative Research 2006;6(1):97-113.