Assessing health-related stigma and social participation

Stevelink, S.A.M.

2011

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

CHAPTER 6

The cross-cultural equivalence of participation instruments: a systematic review

S.A.M. Stevelink, W.H. van Brakel

Submitted to Health and Quality of Life Outcomes

The cross-cultural equivalence of participation instruments: a systematic review

Purpose Concepts such as health-related quality of life, disability and participation may differ across cultures. Consequently it is important to test the cultural equivalence of a particular measure. This paper reviews the process of cross-cultural equivalence testing of instruments to measure social participation.

Methods An existing cultural equivalence framework was adapted and used to assess the instruments included on five categories of equivalence: conceptual, item, semantic, measurement and operational equivalence. For each category, several aspects were rated, resulting in an overall category rating of 'minimal/none', 'partial' or 'extensive'. The best possible overall study rating was therefore five 'extensive' ratings. Articles were included if the instruments focused explicitly on measuring (social) participation and were theoretically grounded in the ICIDH (-2) or ICF. Cross-validation articles were only included if it concerned an adaptation of an instrument developed in Europe, North America, Australia or New Zealand to an African, Asian or Pacific language version of the instrument or vice versa.

Results Eight cross-cultural validation studies were included in which five participation instruments were tested (Impact on Participation and Autonomy, London Handicap Scale, Perceived Impact and Problem Profile, Craig Handicap Assessment Reporting Technique, Participation Scale). Of these eight studies, only three received at least two 'extensive' ratings for the different categories of equivalence. Five studies scored no 'extensive' rating at all. Two 'extensive' ratings were assigned to conceptual, item and semantic equivalence, whereas measurement equivalence received no 'extensive' rating at all. The majority of the 'none/minimal' ratings were given for item and measurement equivalence.

Conclusion The cross-cultural equivalence testing of participation instruments leaves much to be desired. A detailed checklist is proposed for designing a cross-validation study. Once a study has been conducted, the checklist can be used to ensure comprehensive reporting of the validation (equivalence) testing process and results.

Introduction

In recent years, several instruments have been developed to assess the construct of participation, defined by the WHO as 'involvement in life situations' (1;2). The majority of these instruments have been developed in Western countries and often in an English language version (3-6). As a consequence, researchers wanting to investigate this construct in a different language had two options, either to develop a new instrument or to translate and use an existing one. The development of a new instrument is time consuming and costly, and for that reason the latter option was often chosen (7-10).

This translation and adaptation process, also described as cultural equivalence testing or cross-cultural validation, is recognized as important, because the conceptualization of constructs as participation, disability and health-related quality of life (HRQL) may vary across cultures (11;12). Consequently, not all instruments can be translated directly into another language, but need adaptation. In the worst case, they may not suitable for use in a culture different from that for which they had initially been developed for (12).

The quality of the process of translation and adaptation was criticized in a review published in 2003 (13). During that review, the translation and adaptation process of generic HRQL instruments was assessed, using a cultural equivalence framework developed by Herdman et al., (12;13). This framework defined five categories of cultural equivalence, namely conceptual, item, semantic, operational and measurement equivalence (Table 1). The authors concluded that the assessment of these categories of equivalence was often incomplete or not addressed at all (12;13). In particular conceptual equivalence was poorly assessed, because in many cases only quantitative methods were used, whereas qualitative methods might have provided in-depth insight in the conceptualization of the constructs under investigation. The assessment of item, semantic and functional equivalence was at best only briefly investigated in most studies. The assessment of the psychometric properties of the adapted instruments, defined as measurement equivalence, received most attention. The assessment of internal consistency and construct and convergent validity was reported most frequently. Considering the results of such studies, Bowden & Fox-Rushby recommended that more attention should be paid to using efficient and effective methods of assessing the cultural equivalence of instruments (13).

Current insights in the field of measurement research have led to the development of benchmarks that provide indications for what constitutes good measurement properties (14). The benchmarks may be used to assess the quality of psychometric properties such as content, construct and criterion

validity, internal consistency, reliability, agreement, floor and ceiling effects, interpretability and responsiveness (14). The investigation of these properties can be used to assess measurement equivalence in detail.

To our knowledge, the cultural equivalence framework discussed earlier, has not yet been used to evaluate cultural equivalence testing in participation instruments. The objective of this article is to review the quality of the cultural equivalence testing process for participation instruments using an adapted version of the framework. We aim to provide an overview of the extent to which current insights in cultural equivalence testing have been applied in the field of participation instruments and give examples of good practices. We also provide a checklist for testing cultural validity or equivalence, combining insights from both cultural equivalence research and measurement research.

Table 1: Definitions of categories of equivalence, adapted from Herdman et al. (1998)

Equivalence	**Definition**
Conceptual	'Achieved when the questionnaire has the same relationship to the underlying concept in both cultures, primarily in terms of domains included and the emphasis placed on different domains'.
Item	'Item equivalence exists when items estimate the same parameters on the latent trait being measured and when they are equally relevant and acceptable in both cultures'.
Semantic	'The transfer of meaning across languages, achieving a similar effect on respondents who speak different languages'.
Operational	'The possibility of using a similar questionnaire format, instructions, mode of administration and measurement methods'.
Measurement	'The psychometric properties of the adapted version of the HRQL measures are equivalent'.
Cultural	The extent to which an instrument is suitable for use in a different culture from the one it was initially developed for (defined in this paper).

Methods

A systematic search was conducted to identify all studies describing the development of, or cross-cultural validation of instruments that measure social participation (restriction). Three literature databases were searched: *Pubmed (Medline), PsycINFO* and *Web of Science.* A generic syntax was made that consisted of main key words, presented in the title, abstract or main text. The syntax was a variation of the following: <participation AND (instrument OR measure* OR questionnaire)> AND <(cross-cultur* OR cultur* OR valid* OR equivalence)>. The bibliographies of the articles were scanned to identify other relevant studies. In addition, all the initial developers of the instruments and authors of the cross-validation studies were contacted with the request to inform us if they knew of other possibly relevant studies. The last search was concluded on 10 June 2011.

The systematic literature search was performed by the first author. Titles and abstracts of the articles were reviewed and after the exclusion of non-relevant papers, full-text articles were assessed and checked to ascertain whether they met the two-step inclusion round. During the first step, articles were included only if the instruments focused explicitly on the measurement of (social) participation, were theoretically grounded in the International Classification of Impairments, Disabilities and Handicaps (ICIDH or ICIDH-2) or International Classification of Functioning, Disability and Health (ICF), were written in English, freely available or could be obtained from the authors. This resulted in a comprehensive overview of the current state of participations instruments developed.

During the second step, the selection was further limited to cross-validation articles describing an adaptation of an instrument developed in Europe, North America, Australia or New Zealand to an African, Asian or Pacific language version of the instrument or vice versa.

The process of cultural equivalence testing was assessed by using a modified version of the framework by Herdman (12;13). This framework defines several categories of equivalence (conceptual, item, semantic, operational and measurement) and can be used to assess the extent to which an instrument is suitable for use in a different cultural setting than it was initially developed for (Table 1).

We made several adaptations to this framework, especially regarding measurement equivalence. The benchmarks proposed by Terwee et al. were incorporated in measurement equivalence. Item Response Theory methods (e.g. Rasch analysis) were also added as a subcategory of measurement equivalence, because they provide detailed additional information about the validity of an instrument, relevant in a cross-cultural context. The assessment of the Minimally Important Change

was not considered mandatory as a benchmark of interpretability, because it is a fairly new indicator that is not yet applied widely in the field of measurement. The subcategory interpretability, therefore, was rated positively also when only means and SDs were provided for at least four relevant subgroups of respondents. Other small modifications concerning the remaining categories of equivalence (e.g. addition or removal of certain subcategories), can be found in the Appendices. The following rating procedure was applied. Each category of equivalence consists of several subcategories. For example, conceptual equivalence comprised three subcategories: 1) How were the local populations' conceptualizations of participation assessed? 2) Were any theoretical arguments presented questioning or accepting conceptual equivalence? 3) How was the appropriateness of the instrument judged for use in this particular study population? Some subcategories also comprised underlying sub domains (See Appendix 2). Two independent raters (SS and WvB) assessed the quality of the cultural equivalence testing process, rating each subcategory as 'positive', 'partial', 'negative' or 'no information available'. Based on these ratings, they independently classified each category of cultural equivalence as tested 'extensively', 'partially' or 'none/minimally'. Discrepancies in the ratings were resolved by referring back to the original paper and by discussing the findings. Appendix 1 provides a detailed description of the rating system used. Appendix 2 contains the cultural validity checklist.

Results

The initial search generated 2,084 titles, of which 1,982 titles were excluded because they did not concern instrument development or validation studies. The abstracts of the remaining 102 titles were reviewed. After the exclusion of duplicates (n=22), non-English language articles (n=1) and incomplete versions (n=6), 73 full-text articles were reviewed. In addition, the bibliographies of these articles were searched, resulting in 8 additional papers relevant for this review.

In total, 20 instrument development studies were identified during this process. Of these, 14 concerned the development of a participation instrument for adult populations and 4 studies developed an instrument specifically to measure children's participation. Another 2 studies were identified that adapted an adult participation instrument for use with children (Appendix 3).

During the second step of the selection process, 7 articles were excluded, because they did not meet the selection criteria. Therefore, 13 full-text articles, 5 initial development studies and 8 cross-validation studies, were identified. These were included in the present review. The initial development studies served as background articles to complement the findings of the cross-validation studies.

Context

The majority of the initial instrument development studies took place in high income countries such as the Netherlands, United Kingdom (UK), Australia and the United States of America (USA) (4;6;15;16). The Participation Scale was developed simultaneously in India, Nepal and Brazil, whereas the PIPP was developed in Australia, in close collaboration with researchers from Thailand and Malaysia (16;17). The cross-validation studies were performed mainly in East Asia and the Pacific (Hong Kong, Japan, Malaysia and Indonesia), and the Middle East (Turkey, Iran) (9;18-21). Only one study was performed in Europe (the Netherlands) (22). The cross-validation studies were conducted in 7 different languages. The majority of the study populations consisted of people with non-communicable diseases such as stroke, spinal cord injury and various disabilities. In general, the instruments were intended to be generic in nature. Table 2 presents a summary of the different ratings assigned to the different categories of equivalence.

Impact on Participation and Autonomy

The Impact on Participation and Autonomy (IPA) was developed in the Netherlands and is theoretically grounded in the ICIDH (15). The 32 item scale addresses participation restriction (perceived participation) and autonomy across five domains (19). The five-point response scale of the self-administered IPA ranges from 0 (very good) to 4 (very poor). The higher the score on each domain, the higher the level of perceived handicap (15).

Table 2: Summary rating of the quality of cultural equivalence testing in studies of participation instruments developed in North America, Europe, Australia or New Zealand and translated into an African or Asian language or vice versa.

Equivalence* **Instrument**	**Conceptual**	**Item**	**Semantic**	**Operational**	**Measurement**
IPA-P *(Persian)*	++	++	++	+	+
LHS *(Hong Kong Chinese, 2001)*	+	-	+	-	-
LHS (*Hong Kong Chinese, 2007)*	+	-	+	+	-
PIPP *(Bahasa Malaysia)*	+	+	+	+	-
PIPP *(Thai)*	+	++	++	+	+
CHART *(Japanese)*	-	-	-	-	+
CHART *(Turkish)*	++	-	+	-	-
P-scale *(Dutch)*	+	-	-	+	+

* Extensive: + + Partial: + None/minimal: -
IPA: Impact on Participation and Autonomy, LHS: London Handicap Scale, PIPP: Perceived Impact of Problem Profile, CHART: Craig Handicap Assessment and Reporting Technique, P-scale: Participation Scale

Recently, a Persian version of the IPA (IPA-P) was developed (19). Conceptual equivalence was investigated extensively. The appropriateness of the IPA-P in the cultural setting of Iran was critically evaluated by an expert committee that was asked to check the cross-cultural equivalence and the face validity of the instrument. In addition, several theoretical arguments were provided concerning the conceptualization of participation in Iran and the theoretical foundation of the IPA-P in the ICIDH and ICF. Furthermore, a proper definition of the construct was given and the cultural differences in the construct found in the different dimensions of the scale were discussed extensively. According to the authors, the family-oriented culture in Iran may explain the two different dimensions identified, named 'performance-based participation' and 'social-based participation', instead of the original unidimensional construct of 'perceived participation'. In the first construct, the concept of 'doing/performing' is the most important characteristic of participation, whereas in 'social-based participation', 'being together or with others' and 'social relationships' have a central position. In a family-oriented culture *"family members are expected to take care of members who become ill or disabled"* (19). In this way, restrictions in social participation might become less compared to

performance-based participation. Item equivalence was also rated as 'extensive' as was semantic equivalence. During the pre-testing phase of the IPA-P (n = 25), items were checked for ease of understanding and acceptability. Table 3 shows an example of how the rating was applied. This was done in combination with the results of a Rasch analysis (19). The IPA-P was developed with the close collaboration of the initial instrument developer and extensive translation guidelines were followed. The translation quality was judged primarily by the expert committee. Operational equivalence was rated as 'partial'. No information was found on the handling of missing data. The IPA-P, however, was pre-tested, and a well-explained decision was made to administer the IPA-P using an interview-based method instead of self-administration, because of the low educational level of the respondents (literacy rates were provided). Besides Rasch analysis, also content validity was addressed, resulting in a 'partial' rating for measurement equivalence (19).

Other than the Persian version of the IPA, English, Swedish, Italian and German versions are also available (19;23-27).

Table 3: Example of assessment of item equivalence of the Persian version of the Impact on Participation and Autonomy scale

Item equivalence	**Findings**	**Rating**
None/minimal		
Partial (1 out of 3)		
Extensive (2 out of 3)		x
1. Relevance and acceptability of items	Pre-tested in 25 respondents; checking item meaning, difficulty level, confusing and unclear items. Expert committee; face validity & cross-cultural equivalence.	+
2. Items discussed in the light of quantitative or qualitative analysis results	Qualitative analysis of pre-testing phase. Findings from Rasch analysis discussed into detail.	+
3. Adaptations made based on findings regarding individual items	No information	0

+ positive – negative 0 no information +/- indeterminate

London Handicap Scale

The London Handicap Scale (LHS), also based on the ICIDH, was developed in the UK (4). This instrument measures the severity level of handicap in the six ICIDH dimensions of 'handicap', namely mobility, physical independence, occupation, social integration, orientation and economic self-sufficiency (4). The LHS comprises only one item for each dimension. A six-point disadvantage response scale is used for each item (4).

In 2001, a Hong Kong (HK) Chinese version of the LHS was validated (8). In addition, in 2007, a study was conducted to compare handicap perceptions among mainland Chinese, HK Chinese and UK respondents (English and HK Chinese version) (28). Conceptual equivalence was rated 'partial' for the first study, because only theoretical arguments were provided regarding the suitability of the participation construct in China. A description was provided on how the LHS covered the ICIDH domains. In addition, several possible cultural differences were identified in the scoring of different handicap scenarios. For example, *"Hong Kong subjects rated lower scores (i.e. more handicap) than UK estimates for this scenario [scenario J: "describes an individual who is bed/chair bound with difficulty in keeping occupied, needing help to be available all the time, economically just self-sufficient, but who is fully oriented and gets on well with people"].* It might be that Hong Kong subjects valued mobility, independence and occupation more than orientation and social integration, compared with their UK counterparts (8)". A partial rating was assigned to semantic equivalence, because the only information provided concerned the translation process and the involvement of the initial instrument developer (co-authorship). The other categories of equivalence, however, were all rated 'none/minimal' (8).

Almost the same ratings were provided for the study conducted in 2007 (28). According to the authors, the findings showed that *"Significant differences were also found between age, gender and health status subgroups within the mainland Chinese subjects"* (28). In addition, the ICIDH and ICF were discussed. Despite adequate theoretical argumentation, no assessment of local perceptions of participation was reported and the appropriateness of the measure was not discussed, resulting in a 'partial' rating for conceptual equivalence. Item equivalence was addressed in only one sentence: *"Anecdotal feedback from the research assistants revealed that the LHS was easy to use and well understood by the elderly"*, resulting in a 'none/minimal rating' (28). The Chinese language version used in the 2007 study was developed in the 2001 study (8). For this reason, the information from both articles was combined, resulting in a 'partial' rating for semantic equivalence. Operational equivalence received also a 'partial' rating, based on a description of the educational levels of the study population and the availability of a LHS user manual. Despite this, no information was

presented for the other sub-categories such as the handling of missing data, response options, item or questionnaire format. Measurement equivalence received a 'none/minimal' rating, because the information reported was insufficient (28).

The LHS studies used a different method to assess the validity of the instrument ratings than that used in the other studies. Instead of interviewing a sample of live respondents, raters were asked to score a set of hypothetical vignettes of people with a particular handicap. As a result, therefore, the ratings were only influenced by the raters' understanding of the items. The understanding of respondents was not assessed.

A Swedish version of the LHS is available as is a version for use with children (10;29).

Perceived Impact of Problem Profile

The Perceived Impact of Problem Profile (PIPP) was developed to measure the impact and distress of having a health condition as perceived by the respondent (16). It was originally developed in Australia in collaboration with researchers from Malaysia and Thailand. This self-report instrument has five subscales self-care, mobility, participation, relationships and psychological well-being. The impact and distress level was rated with a 6-point response scale ranging from 'no impact' to 'extreme impact' and 'no distress' to 'extreme distress' (16).

We identified two non-western language studies of the PIPP. One conducted in Malaysia (Bahasa Malaysia) and one in Thailand (Thai) (9;20). The first study received a 'partial' rating for conceptual equivalence. The authors of that study elaborated extensively on possible group and cultural differences. In their study conducted in Malaysia, Misajon et al. aimed to investigate gender differences in the experience of disability. They drew an interesting comparison between findings from previous research in relation to their study (9). They concluded that *"Men also reported greater impact and disability on their ability to assist their family members. This may be related to conventional gender roles, with men feeling a duty to care for the family"* (9). The other two subcategories of conceptual equivalence, 'appropriateness of measure' and 'assessment of local conceptions' of participation, however, were not addressed sufficiently. The result was a 'partial' rating. Item equivalence was assessed as 'partial'. For example, *"In all three countries, wording and content were chosen carefully to ensure that the activities described were suitable across different cultural contexts, for both men and women and across different age groups"* (20). There had been close collaboration with the author of the initial instrument development study, "*the intent and precise meaning of terms were discussed at length and pretested during the training".* This justified

the subcategory rating as 'partial' for semantic equivalence (9). Table 4 shows an example of how the rating was applied.

Table 4: Example of assessment of semantic equivalence of the Malaysian version of the Perceived Impact of Problem Profile

Semantic equivalence	**Findings**	**Rating**
None/ minimal		
Partial (2 out of 5)		x
Extensive (3 out of 5)		
1. Contact with developers	Co-author	+
2. Translation guideline or user manual including translation instructions used	No information	0
3. Details translation procedure		+/-
a) Translators	No information	0
b) Procedure	Translated from English into Malay. No back translation	-
c) Checking translation target populations	"the intent and precise meanings of terms were discussed at length and pretested during the course of training".	+/-
d) Judging translation quality (experts, researchers)	"the intent and precise meanings of terms were discussed at length and pretested during the course of training".	+/-
4. Meaning of key words and phrases	"the intent and precise meanings of terms were discussed at length…".	+
5. Translation problems and difficulties	No information	0

+ positive – negative 0 no information +/- indeterminate

Operational equivalence was also rated as 'partial'. Adequate pre-testing of this language version was performed and well-justified adaptations were made to the response options used. The educational levels of the target population were also given (9). Insufficient information was provided concerning measurement equivalence, however, resulting in a 'none/minimal' rating (9).

Misajon et al. assessed the cultural equivalence of the Thai version of the PIPP more thoroughly. A 'partial' rating was scored for conceptual equivalence. In their article there is also a detailed discussion on possible between-group differences and cultural differences related to 'participation'. For example: *"No DIF [Different Item Functioning] was found for either gender, or age, except for participation in family activities in the Participation subscale. Specifically, men indicated a greater likelihood of endorsing this item than women."* Item equivalence was rated as 'extensive'. Adding to a discussion among researchers about the relevance and acceptability of items was the outcome of the Rasch analysis that showed good results, except for one item. These findings were discussed and

adequate adaptations were made (20). Semantic equivalence was rated also as 'extensive'. The meaning of key words and phrases were discussed in detail. Several translation problems and difficulties were also discussed, e.g., *"One of the difficulties in translating Western developed concepts from English into different languages is ensuring congruent meanings, particularly in the case of abstract nouns. In Thai, the term 'distress' translates to 'took' or 'suffer', while impact is 'pon-kratop', or effect. These two words have a similar meaning in Thai, although distress connotes cause; impact is consequence."* Analysis suggested that participants did not necessarily differentiate between 'impact' and 'distress'. Because of this, the distress subscale was omitted from the Thai version of the instrument. Operational equivalence was rated 'partial'. Based on a careful evaluation of the response patterns, the decision was made to collapse the response categories from six to four. The reason given was the *'…disordered thresholds for many of the items" in the Rasch analysis when using 6 response categories. "...respondents...typically only used 4 response points"* (20). Furthermore, the educational levels of the target population were given. Pilot-testing of the Thai version of the PIPP took place with help of interviews with the target population (20). Finally, measurement equivalence was assessed as 'partial'. Content validity was addressed by a detailed description of the target population under study. Qualitative data and existing measures served as input for item generation, after which items were selected based on discussion with researchers. Construct validity was investigated using the European Quality of Life-5 Dimensions (EQ-5D), but no adequate hypotheses were formulated. Rasch analysis was performed and internal consistency was found to be good (Person Separation Index (PSI) range 0.85 – 0.88) (20).

Craig Handicap Assessment Reporting Technique

The Craig Handicap Assessment and Reporting Technique (CHART) was developed in 1992 in the USA (6). The CHART is used with the aim of measuring handicap among affected persons. It was based on five or six domains of the ICIDH (6;30). The domains were categorised as physical independence, mobility, occupation, social integration and economic self-sufficiency and cognitive independence or orientation. The original version of the CHART consisted of 27 items and was grouped into five domains; the revised version has 32 items, grouped into six domains (6;30).

Two non-western validation studies were identified. These sought to validate a Japanese version and a Turkish version of the CHART (18;21). Conceptual equivalence was rated 'none/minimal', because insufficient information was reported. Information was also insufficient for item, semantic and operational equivalence, resulting in three 'none/minimal ratings'. An article was published in Japanese that may have addressed these categories of equivalence more thoroughly (31). However, the authors did not refer to this. Several subcategories of measurement equivalence were assessed

such as content validity, test-retest reliability and construct validity. However, no adequate prior hypotheses were formulated for the latter. As a result a 'partial' rating was given for measurement equivalence (21).

In 2002, a study was published assessing the validity of the Turkish version of the 20-item CHART Short Form (CHART-S), comparing the results with findings from the USA (18). An extensive comparison of community participation of persons with a disability in Turkey and USA was made, based on the literature and *"the authors' familiarity with the two societies"* (18). The local conceptions of participation and the appropriateness of the CHART were also considered, as illustrated by the following statement: *"Even if we assume that the basic concept of community participation is applicable to Turkish society, and that the significant indicators are the same as those selected for the USA, we cannot always be certain that exactly the same information is collected"* (18). The authors noted that the CHART was modelled directly on the ICIDH's handicap dimensions and a definition of social participation was given. Also several between-group differences and possible cultural differences were described. As a result, conceptual equivalence was rated 'extensive'. Item equivalence was not assessed adequately resulting in a 'none/minimal' rating. Only a back translation procedure was applied and discrepancies between the two translations were discussed. Semantic equivalence was given a 'partial' rating since only collaboration with the initial instrument developer was reported (18). Operational equivalence was rated 'none/minimally', because insufficient information was reported. The same applied for measurement equivalence. Content validity was addressed adequately; however, no adequate hypotheses were formulated for construct validity and interpretability was not assessed thoroughly (only SDs and means of 1 group, instead of 4 subgroups). Consequently, the information provided was insufficient and a 'none/minimal' rating was applied (18).

Participation Scale

The Participation Scale (P-scale) was developed simultaneously in 6 languages in India, Nepal and Brazil (Hindi, Bengali, Telugu, Tamil, Nepali and Portuguese) and aims to measure the level of participation restriction based on the ICF (17). The P-scale consists of 18 items covering eight out of nine domains of participation (major life areas, learning and applying knowledge, communication, domestic life, interpersonal interactions and relationships, mobility, community, social and civic life and self-care). A distinctive feature of the P-scale is the peer concept defined as "people who are similar to the respondent in all aspects (socio-cultural, economic and demographic) except for the health condition and or disability" (17).

A study conducted in the Netherlands assessed the reproducibility of the P-scale and of two other participation measurements in a rehabilitation setting (22). For this study, a self-report version of the P-scale was used. Extensive criteria were considered for the inclusion of participation instruments in the study, such as *"applicable in various diagnostic groups", "feasible" and "having sound psychometric properties"* (22). However, the other subcategories of conceptual equivalence were not addressed. Therefore, conceptual equivalence was rated only 'partial'. No information was found for item and semantic equivalence resulting in a 'none/minimal' rating. For operational equivalence, the subcategories 'missing data' and 'questionnaire format' were addressed and this resulted in a rating of 'partial'. According to the authors *"The proportion of missing item responses [for the P-scale] was somewhat larger than that for the other measures (2.8%)".* Another issue was the format of the P-scale. *"A common comment concerned the layout of the Participation Scale, which was perceived as confusing"* and *"An internet-based version of the questionnaire might solve this problem"* (22). During this study, measurement equivalence was assessed by addressing content validity, test-retest reliability and agreement, showing good results. In addition, no floor or ceiling effects were found. Measurement equivalence was rated as 'partial' (22).

The P-scale has been translated in 18 additional languages, including Arabic (Sudanese and Jordanese), Bahasa Indonesia, Bangla, Chinese, Dutch, Hausa, Hebrew, Khmer, Thai and Vietnamese. However, most of these were non-western languages. In addition, very few formal validation studies have been published. Only one study qualified according to our criterion of 'cross-cultural studies', while two other validation studies concerned two Indian languages and Bahasa Indonesia and were therefore excluded (Kelders et al. accepted) (32).

Discussion

It is our conviction that social participation is a globally relevant construct. The way this is understood in different cultures, however, may well vary. Therefore, it cannot be assumed that instruments to assess such a construct are necessarily cross-culturally valid. Instruments developed in one culture that are to be used in another should be tested for cultural validity first. In situations where results are to be compared between cultures, testing of cultural equivalence between the 'development culture(s)' and the new culture is crucial. Our analysis indicates that cross-cultural equivalence testing of participation instruments leaves much to be desired.

In the present review, we used the cultural equivalence framework designed by Herdman, Fox-Rushby and Badia (1998) to assess the cross-cultural equivalence testing process of participation instruments. Five categories of equivalence were defined in this framework, conceptual, item,

semantic, operational and measurement equivalence (12;33). A sixth umbrella category in the former framework was 'functional equivalence', defined as "...the extent to which an instrument does what it is supposed to do equally well in two or more cultures" (12). A balanced decision concerning cultural equivalence is a major goal of equivalence testing. According to Herdman et al., this decision can be made after addressing the other five categories of equivalence as mentioned above. These findings are summarised under the sixth category, which we suggest should be termed 'cultural equivalence' instead of 'functional equivalence'. We defined 'cultural equivalence' as 'The extent to which an instrument is suitable for use in a different culture from the one it was initially developed for'.

The former framework was used in 2003 to review the process of translation and adaptation of generic HRQL instruments (13). Based on their findings, the authors recommended that the methods for assessing measurement equivalence (e.g. reliability and validity) needed detailed specification. For this reason, we integrated the benchmarks for 'good psychometric properties' proposed by Terwee et al. into the framework. This made it possible to assess the process of testing measurement equivalence in a systematic way, and assess the psychometric properties of the included instruments at the same time (Appendix 2) (14).

Table 2 shows the results of our assessment of the process of cultural equivalence testing applying this modified framework to the 8 different cross-cultural studies included. In the present review only one instruments, namely the IPA-P, received three 'extensive' ratings (19). One 'extensive' rating was provided for the Turkish version of the CHART and the Thai version of the PIPP (18;20). The other five studies scored no 'extensive' ratings (8;9;21;22;28). The Japanese version of the CHART scored four 'none/minimal ratings' (21). Three studies were assigned at least three 'partial' ratings (9;22;28).

The majority of the 'extensive' ratings were given for conceptual (2), item (2) and semantic (2) equivalence, followed by operational equivalence (1). Measurement equivalence received no 'extensive' rating, which is remarkable, since validation studies tended to focus on this category traditionally (13). Conceptual equivalence also received the most 'partial' ratings (5). Studies often did not receive 'extensive' ratings for several reasons. Local perceptions of participation were not informed by the perceptions of the target population, discussion among researchers or other experts was not undertaken, and the assessment of local literature did not take place or was not recorded (Appendix 2). A good exception was the cross-cultural validation of the Turkish version of the CHART, where an extensive literature review was performed, and discussion amongst researchers was recorded (18). Not only assessment of local perceptions, but also the appropriateness of the

instrument was discussed (8;9;13;18;20;21;28). An example of a good practice is the consultation of an expert committee, as done in the IPA-P study (19).

Operational equivalence was often rated as 'partial' (5) or even 'none/minimal' (3), because the revised instrument was not pre-tested, no special attention was paid to missing data or administration format (e.g. literacy or educational levels, questionnaire and item format, response options etc.,). An example of good operational equivalence testing was found in the Thai validation of the PIPP, where the response options were evaluated carefully and pre-testing was performed (20). The authors also provided a good report on translation problems and difficulties experienced during the translation procedure (20). Another good exception is the validation study of the IPA-P, where a good description of the translation procedure was provided (19). In combination with the description of translation difficulties mentioned earlier, these two examples can be seen as good practices for the assessment of semantic equivalence (19;20). Both studies also performed a sufficient assessment of item equivalence using Rasch analysis (19;20). Although quantitative methods were used, operational equivalence can also be assessed with the help of qualitative methods (e.g. in-depth interviews with target population, field experiences from researchers).

Measurement equivalence was often rated as 'none/minimally' (5). In many studies, content validity was addressed sufficiently, in contradiction to the assessment of construct validity. Hypotheses were often not formulated adequately (e.g. *a priori* and stating a magnitude range and the direction of the expected association) (9;18;20;21;28). Furthermore, internal consistency or test-retest reliability were only assessed occasionally, as were floor or ceiling effects (18;20-22). In the case of interpretability, often only one or two means and SDs of respondent subgroups were presented, which is too few to adequately interpret the quantitative findings (9;21). It was noteworthy that studies which applied Rasch analysis commonly neglected other measurement properties. This was unfortunate because some processes, such as test-retest reliability and agreement, and responsiveness, provide valuable additional information to establish the validity of an instrument.

A possible explanation for the poor evaluation of measurement equivalence may be that we used only recently proposed benchmarks, which have not been incorporated extensively in the field of instrument testing (14). A review was recently published wherein the psychometric properties of several participation instruments were assessed in detail (34). Their findings were congruent with ours, in that most participation instruments need further psychometric testing (34). A good example of a participation instrument study that addressed measurement equivalence was published by Rensen et al. (2010) (32).

As far as we are aware, this is the first extensive review of cultural equivalence testing of participation instruments. A similar review was published by Bowden & Fox-Rushby concerning HRQL instruments (13). We identified some interesting differences when comparing the two studies. In the latter, the majority of the instruments included received a 'none/minimal' rating for conceptual equivalence, whereas in our review, often 'partial' and some 'extensive' ratings were ascribed. Furthermore, Bowden & Fox-Rushby stated that attention was focussed on the assessment of measurement equivalence, while many instruments in our review were rated 'none/minimally' or 'partially' on this point. A partial explanation may be the modifications made during our review in several (sub) categories of equivalence, especially in the category of measurement equivalence. However, we agree with their conclusions that more attention has to be paid to adequate and efficient ways of assessing cultural equivalence and that this concept is generally under reported.

In 1993, a literature review was published that reported the findings of an assessment of the cross-cultural adaptations of 17 HRQOL instruments (35). Guidelines were proposed comprising 5 sections: 1) translations, 2) back-translations by qualified people, 3) committee review of translations and back-translations, 4) pretesting for equivalence using adequate techniques and 5) re-examination of the weighting scores, if relevant (35). In that study it was suggested that no standardized approach was used for the cross-cultural adaptation of the instruments (35). Often insufficient information was reported and the methods that were applied varied (35). In contrast to the framework we used, no assessment of validity and reliability was included. A different review in the same year, but also with a focus on the cross-cultural adaptation of HRQOL instruments, concluded that most studies had a particular focus on translation and paid less attention to measurement equivalence (36). Ten years later, Bowden & Fox-Rushby noted that most validation studies paid particular attention to this category of equivalence, allowing us to conclude that the practice of measurement validation is clearly evolving (36).

This review has one limitation. We limited our review to studies where the need for cross-cultural validation would have been very obvious. Therefore we excluded cross-cultural validations of participation instruments between high-income countries. Although cultural differences may be somewhat less in the latter, cultural validation is still necessary. A future research opportunity, therefore, is to complete the overview of the current status of cultural equivalence testing in this particular field.

We suggest three points of caution. The large number of ‘none/minimal’ or ‘partial’ ratings do not indicate directly that the process of cultural equivalence testing was not performed sufficiently, because extended testing procedures may not always have been published. Besides this, we aimed to assess the process of testing and not the cultural equivalence of the instrument per se. The latter was not possible, due to a lack of necessary information and understanding of the cultures under study. Secondly, some examples were given of how certain categories of equivalence were addressed. However, this does not mean that these are ‘best practices’ or that no other methods can be used to assess discrete categories of equivalence in a satisfactory way. With the help of the adapted framework and the accompanying checklist, we hope to encourage improvement of the quality of cultural equivalence testing. This in turn would help ensure the availability of valid and reliable instruments across countries. Finally, we are also well aware of the tension that exists between the theoretical need to use such an extensive framework and the feasibility of using it in terms of available time, resources and manpower.

In conclusion, our findings showed that the cultural equivalence of participation instruments has generally been tested inadequately. Further testing and reporting concerning of the cultural equivalence of participation instruments is recommended. The proposed cultural equivalence framework and accompanying checklist would facilitate this process.

Acknowledgements

We would like to thank Prof. dr. Julia Fox-Rushby for providing the technical report concerning cultural equivalence testing. In addition, we would like to thank Dr. Hugh Cross for his comments on earlier drafts of this review.

Reference List

(1) World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva: WHO; 2001.

(2) World Health Organization. Towards a common language for functioning, disability and health; ICF. Geneva: WHO; 2002.

(3) Bedell GM. Developing a follow-up survey focused on participation of children and youth with acquired brain injuries after discharge from inpatient rehabilitation. Neurorehabilitation 2004;19(3):191-205.

(4) Harwood RH, Rogers A, Dickinson E, Ebrahim S. Measuring handicap: the London Handicap Scale, a new outcome measure for chronic disease. Quality of Health Care 1994 March;3(1):11-6.

(5) Noreau L, Fougeyrollas P, Vincent C. The LIFE-H: Assessment of the quality of social participation. Technology and Disability 2002;14:113-8.

(6) Whiteneck GG, Charlifue SW, Gerhart KA, Overholser JD, Richardson GN. Quantifying handicap: a new measure of long-term rehabilitation outcomes. Archives of Physical and Medical Rehabilitation 1992 June;73(6):519-26.

(7) Lemmens J, van Engelen ISM, Post MW, Beurskens AJ, Wolters PM, de Witte LP. Reproducibility and validity of the Dutch Life Habits Questionnaire (LIFE-H 3.0) in older adults. Clinical Rehabilitation 2007 September;21(9):853-62.

(8) Lo R, Harwood R, Woo J, Yeung F, Ebrahim S. Cross-cultural validation of the London Handicap Scale in Hong Kong Chinese. Clinical Rehabilitation 2001 April;15(2):177-85.

(9) Misajon R, Manderson L, Pallant JF, Omar Z, Bennett E, Rahim RB. Impact, distress and HRQoL among Malaysian men and women with a mobility impairment. Health and Quality of Life Outcomes 2006;4:95.

(10) Westergren A, Hagell P. Initial validation of the Swedish version of the London Handicap Scale. Quality of Life Research 2006 September;15(7):1251-6.

(11) Berry JW, Poortinga YH, Segall MH, Dasen PR. Cross-Cultural Psychology: Research and Applications. Cambridge, UK: Cambridge University Press; 1992.

(12) 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.

(13) Bowden A, Fox-Rushby JA. A systematic and critical review of the process of translation and adaptation of generic health-related quality of life measures in Africa, Asia, Eastern Europe, the Middle East, South America. Social Science and Medicine 2003 October;57(7):1289-306.

(14) 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.

(15) Cardol M, de Haan RJ, van den Bos GAM, de Jong BA, de Groot IJM. The development of a handicap assessment questionnaire: the Impact on Participation and Autonomy (IPA). Clinical Rehabilitation 1999 October;13(5):411-9.

(16) Pallant JF, Misajon R, Bennett E, Manderson L. Measuring the impact and distress of health problems from the individual's perspective: development of the Perceived Impact of Problem Profile (PIPP). Health and Quality of Life Outcomes 2006 June 29;4.

(17) Van Brakel WH, Anderson AM, Mutatkar RK, Bakirtzief Z, Nicholls PG, Raju MS et al. The Participation Scale: Measuring a key concept in public health. Disability and Rehabilitation 2006 February;28(4):193-203.

(18) Dijkers MP, Yavuzer G, Ergin S, Weitzenkamp D, Whiteneck GG. A tale of two countries: environmental impacts on social participation after spinal cord injury. Spinal Cord 2002 July;40(7):351-62.

(19) Fallahpour M, Jonsson H, Joghataei MT, Kottorp A. Impact on Participation and Autonomy (IPA): Psychometric evaluation of the Persian version to use for persons with stroke. Scandinavian Journal of Occupational Therapy 2011 March;18(1):59-71.

(20) 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.

(21) Tozato F, Tobimatsu Y, Wang CW, Iwaya T, Kumamoto K, Ushiyama T. Reliability and validity of the Craig Handicap Assessment and Reporting Technique for Japanese individuals with spinal cord injury. Tohoku Journal of Experimental Medicine 2005 April;205(4):357-66.

(22) 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.

(23) Franchignoni F, Ferriero G, Giordano A, Guglielmi V, Picco D. Rasch psychometric validation of the Impact on Participation and Autonomy questionnaire in people with Parkinson's disease. Eura Medicophys 2007 December;43(4):451-61.

(24) Kersten P, Cardol M, George S, Ward C, Sibley A, White B. Validity of the impact on participation and autonomy questionnaire: A comparison between two countries. Disability and Rehabilitation 2007;29(19):1502-9.

(25) Lund ML, Fisher AG, Lexell J, Bernspang B. Impact on Participation and Autonomy Questionnaire: internal scale validity of the Swedish version for use in people with spinal cord injury. Journal of Rehabilitation Medicine 2007 March;39(2):156-62.

(26) Senn D. Validity of the German version of the Impact on Participation and Autonomy (IPA-G) questionnaire 2005. Dissertation.

(27) Vazirinejad R, Lilley JM, Ward CD. The 'Impact on Participation and Autonomy': acceptability of the English version in a multiple sclerosis outpatient setting. Multiple Sclerosis 2003;9(6):612-5.

(28) Lo RS, Kwok TC, Cheng JO, Yang H, Yuan HJ, Harwood R et al. Cross-cultural validation of the London Handicap Scale and comparison of handicap perception between Chinese and UK populations. Age Ageing 2007 September;36(5):544-8.

(29) Detmar SB, Hosli EJ, Chorus AMJ, van Beekum T, Vogels T, Mourad-Baars PEC et al. The development and validation of a handicap questionnaire for children with a chronic illness. Clinical Rehabilitation 2005 February;19(1):73-80.

(30) Mellick D, Walker N, Brooks CA, et al. Incorporating the domain of cognitive independence into CHART. Journal of Rehabilitation Outcomes Measurements 1999;3:12-21.

(31) Kumamoto K, Iwaya T, Tobimatsu Y, Kumano H, Sonoda K, Tozato F. Japanese version of the Craig Handicap Assessment and Reporting Technique. Sogo Rehabilitation 2002;30:249-56.

(32) 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) Herdman M, FoxRushby J, Badia X. 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Quality of Life Research 1997 April;6(3):237-47.

(34) Magasi S, Post MW. A comparative review of contemporary participation measures' psychometric properties and content coverage. Archives of Physical and Medical Rehabilitation 2010 September;91(9 Suppl):S17-S28.

(35) Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality-of-life measures - literature-review and proposed guidelines. Journal of Clinical Epidemiology 1993 December;46(12):1417-32.

(36) Anderson RT, Aaronson NK, Wilkin D. Critical-review of the international assessments of health-Related quality-of-life. Quality of Life Research 1993 December;2(6):369-95.

Appendix 1: Assessment reporting on cultural equivalence (adapted from Bowden & Fox-Rushby, 2003; Herdman et al. 1998)

	None/minimal adherence	**Partial adherence**	**Extensive adherence**
Conceptual	No mention of any issues related to conceptual equivalence or only a brief statement concerning possible cultural differences.	At least a mention of one of the following: 1) an assessment of the local population's conceptualization of the construct, or 2) an assessment of the appropriateness of the measure in the target setting, or 3) theoretical arguments questioning or accepting conceptual equivalence.	Explicit details about at least two of the three issues listed under "partial".
Item	No mention of any issues related to item equivalence, or a brief mention in the text related to any of the issues related to item equivalence.	Description of the assessment of either 1) the relevance or acceptability of individual items to the target population, or 2) item discussed in the light of any quantitative or quality analyses results, or 3) discussion of adaptations made based on findings regarding individual items.	Explicit details about at least two of the three issues listed under "partial".

Semantic	No mention of any issues related to semantic equivalence, or a brief mention in the text about who was involved in the translation of the HRQL measure.	Description of at least two of the key issues related to semantic equivalence: 1) contact with developers, or 2) reference to the translation guidelines used, or user manual including translation instructions, or 3) details provided on the translation procedure, or 4) meaning of key words and phrases, or 5) a description of any problems or difficulties encountered during the translation.	Extensive details about at least 3 of the key items listed under "partial".
Operational	No mention of any issues related to operational equivalence.	A description of at least one or two of the key issues listed under operational equivalence: 1) an assessment of missing data 2) discussion on administration format 3) pre-testing of the instrument.	An extensive description of all key issues listed under "partial".
Measurement	No mention of any issues relating to measurement equivalence, or a description of only one of the key issues related to measurement equivalence.	An adequate description of two or three of the following: 1) assessment of content validity 2) assessment of construct validity, 3) assessment of test-retest reliability and agreement, 4) assessment of floor and ceiling effects, 5) assessment of interpretability, 6) assessment of responsiveness, 7) application of IRT analysis	A detailed description of at least four of the seven issues listed under "partial".

Appendix 2: Checklist for assessment of reporting on cultural equivalence (based on Bowden & Fox-Rushby, 2003; Herdman et al., 1998; Terwee et al., 2007; and Mokkink et al., 2010)

General information

- Name of instrument
- *Initial study (language)*
    - Authors
    - Journal
    - Article title
    - Location
    - Disease/condition (and intervention) studied
- *Cross-cultural validation study (language)*
    - First author
    - Journal
    - Article title
    - Location
    - Disease/condition (and intervention) studied

Methodological details

- Sample characteristics
- Sample size
- Sampling frame
- Method of selection
- Aim of study
- Other measures used during the study

Conceptual equivalence

- In what ways were the local populations' conceptualizations of participation assessed?
    - Local literature
    - Local questionnaires/instruments
    - Discussion amongst researchers
    - Involvement of anthropologists, sociologists, etc.,
    - Discussion with local people
    - Other

'Local population's' conceptualization was rated positive if 50% of the subcategories received a positive rating (three out of six).

- Were any people of the target population asked to judge the appropriateness of the instrument; was a detailed discussion provided in the article concerning the appropriateness of the instrument or were the domains of importance identified by the local people covered in the instrument?
- Were any theoretical arguments presented questioning or accepting conceptual equivalence?
    - Conceptual framework described in relation to the local concept under investigation
    - Definition of the main construct
    - Discussion of possible between-group differences related to construct
    - Discussion of possible cultural differences related to the construct

'Theoretical arguments' was rated positive if 2 out of 4 subcategories received a positive rating. Conceptual equivalence was rated 'extensive' if two out of three categories were rated positively. The rating 'partial' was assigned if one out of the three categories received a positive rating. If no or minimal information was provided concerning conceptual equivalence a 'none/minimal rating was given.

Item equivalence

- Does the report mention how the authors assessed the relevance and acceptability of the individual items for the target population?
- Are the relevancy and acceptability of items discussed in the light of any quantitative or qualitative analyses?
- Were any adaptations necessary and was this discussed properly regarding individual items?

Item equivalence was rated 'extensive' if two out of three categories were rated positively. The rating 'partial' was assigned if one out of the three categories received a positive rating. If no or minimal information was provided concerning item equivalence, a 'none/minimal' rating was provided.

Semantic equivalence

- Were the initial developers of the scale contacted and what was the nature of the contact?
- Was a translation protocol followed or a user manual including translation instructions?
- Were any details about the translation procedure provided?
    - Description of the translators

    - Was the translation procedure adequate? (translation and back translation, native speakers, with and without knowledge of the particular topic)
    - Was the translation checked with the target population?
    - Was the translation quality judged by experts or researchers?

A positive rating was provided for 'translation procedure' if at least two out of the four subcategories were rated as positive.

- Was the initial meaning of key words and phrases investigated and if yes, how was this done?
- Were there any problems or difficulties reported with the translation?

Semantic equivalence was rated 'extensive' if at least three out of five categories were rated positively. The rating 'partial' was assigned if two out of five categories received a positive rating. If no or minimal information was provided concerning semantic equivalence a 'none/minimal' rating was provided.

Operational equivalence

- What was the percentage missing data and what action was taken if the percentage was too high (>25% per item)?
- Was the same administration format used?
    - Was a description provided about the literacy rates or educational level of the target population?
    - Was the suitability of the questionnaire format discussed?
    - Was the appropriateness of the item format evaluated and discussed?
    - Was the appropriateness of the response options evaluated and discussed?
    - Were instructions for interviewers available?

A positive rating was provided for 'administration format if at least two out of the five subcategories were rated as positive.

- Was the instrument pre-tested before use?

Operational equivalence was rated 'extensive' if all three categories were rated positively. The rating 'partial' was assigned if at least one or two out of three categories received a positive rating. If no or minimal information was provided concerning semantic equivalence a 'none/minimal' rating was provided.

Measurement equivalence

- How was content validity addressed?
    - Is the measurement aim of the instrument described?
    - Is the target population described?
    - Are the concepts described that the instrument intend to measure?
    - Were the target population and researchers or experts involved during item selection and reduction? (Often not applicable during cross-cultural validation)

A positive rating was provided for 'content validity' if at least two out of the four subcategories were rated as positive.

- How was construct validity of the instrument assessed?
    - Were hypotheses formulated *a priori* and was the expected magnitude range and direction of the expected association stated?
    - Was factor analysis applied on an adequate sample size (at least seven times the number of items)?

Construct validity was rated as positive if one out of the two subcategories was addressed.

- Was test-retest reliability and agreement assessed?
    - How was intra or inter-interviewer reliability assessed and were the results found adequate (Intraclass correlation coefficients (ICCs) ≥0.70 or weighted kappa ≥0.70)?
    - Showed the scale adequate internal consistency (Cronbach's alpha at least 0.70)?
    - Were adequate agreement measures provided (e.g. Smallest Detectable Change, Minimal Important Change)

Test-retest reliability was rated positive if one out of the two subcategories was addressed.

- Were any floor or ceiling effects tested (<15%)?
- How was interpretability assessed and were the results found adequate (at least means and standard deviations of four subgroups provided and/or a Minimally Important Change defined)?
- How was responsiveness assessed, were the methods applied adequate, as well as the results found?
- Were any Item Response Theory (IRT) methods applied (e.g. Rasch analysis)?

Measurement equivalence was rated as extensive if at least four out of seven categories were rated positively. The rating partial was assigned if two or three of the seven categories received a positive rating. If no or minimal information was provided concerning measurement equivalence a none/minimal rating was provided.

The subcategories were rated positive (+), negative (-), no information available (0) or indeterminate (+/-) inadequate design or methods used.

Appendix 3: Overview of participation instruments identified

(1) Bedell GM. Developing a follow-up survey focused on participation of children and youth with acquired brain injuries after discharge from inpatient rehabilitation. Neurorehabilitation 2004;19(3):191-205.

(2) Brown M, Dijkers MPJM, Gordor WA, Ashman T, Charatz H, Cheng ZF. Participation objective, participation subjective - A measure of participation combining outsider and insider perspectives. Journal of Head Trauma Rehabilitation 2004 November;19(6):459-81.

(3) Cardol M, de Haan RJ, van den Bos GAM, de Jong BA, de Groot IJM. The development of a handicap assessment questionnaire: the Impact on Participation and Autonomy (IPA). Clinical Rehabilitation 1999 October;13(5):411-9.*

(4) Detmar SB, Hosli EJ, Chorus AMJ, van Beekum T, Vogels T, Mourad-Baars PEC et al. The development and validation of a handicap questionnaire for children with a chronic illness. Clinical Rehabilitation 2005 February;19(1):73-80.

(5) Gandek B, Sinclair SJ, Jette AM, Ware JE, Jr. Development and initial psychometric evaluation of the participation measure for post-acute care (PM-PAC). American Journal of Physical and Medical Rehabilitation 2007 January;86(1):57-71.

(6) Gray DB, Hollingsworth HH, Stark SL, Morgan KA. Participation survey/mobility: psychometric properties of a measure of participation for people with mobility impairments and limitations. Archives of Physical and Medical Rehabilitation 2006 February;87(2):189-97.

(7) Harwood RH, Rogers A, Dickinson E, Ebrahim S. Measuring handicap: the London Handicap Scale, a new outcome measure for chronic disease. Quality of Health Care 1994 March;3(1):11-6.*

(8) Jessen EC, Colver AF, Mackie PC, Jarvis SN. Development and validation of a tool to measure the impact of childhood disabilities on the lives of children and their families. Child Care Health and Development 2003 January;29(1):21-34.

(9) Mulcahey MJ, DiGiovanni N, Calhoun C, Homko E, Riley A, Haley SM. Children's and parents' perspectives about activity performance and participation after spinal cord injury: initial development of a patient-reported outcome measure. American Journal of Occupational Therapy 2010 July;64(4):605-13.

(10) Noreau L, Fougeyrollas P, Vincent C. The LIFE-H: Assessment of the quality of social participation. Technology and Disability 2002;14:113-8.

(11) Noreau L, Lepage C, Boissiere L, Picard R, Fougeyrollas P, Mathieu J et al. Measuring participation in children with disabilities using the assessment of life habits. Developmental Medicine and Child Neurology 2007 September;49(9):666-71.

(12) Ostir GV, Granger CV, Black T, Roberts P, Burgos L, Martinkewiz P et al. Preliminary results for the PAR-PRO: a measure of home and community participation. Archives of Physical and Medical Rehabilitation 2006 August;87(8):1043-51.

(13) Pallant JF, Misajon R, Bennett E, Manderson L. Measuring the impact and distress of health problems from the individual's perspective: development of the Perceived Impact of Problem Profile (PIPP). Health and Quality of Life Outcomes 2006 June 29;4.*

(14) Post MW, de Witte LP, Reichrath E, Verdonschot MM, Wijlhuizen GJ, Perenboom RJ. Development and validation of IMPACT-S, an ICF-based questionnaire to measure activities and participation. Journal of Rehabilitation Medicine 2008 August;40(8):620-7.

(15) Rosenberg L, Jarus T, Bart O. Development and initial validation of the Children Participation Questionnaire (CPQ). Disability and Rehabilitation 2010;32(20):1633-44.

(16) Sandstrom M, Lundin-Olsson L. Development and evaluation of a new questionnaire for rating perceived participation. Clinical Rehabilitation 2007 September;21(9):833-45.

(17) Van Brakel WH, Anderson AM, Mutatkar RK, Bakirtzief Z, Nicholls PG, Raju MS et al. The Participation Scale: Measuring a key concept in public health. Disability and Rehabilitation 2006 February;28(4):193-203.*

(18) Whiteneck GG, Charlifue SW, Gerhart KA, Overholser JD, Richardson GN. Quantifying handicap: a new measure of long-term rehabilitation outcomes. Archives of Physical and Medical Rehabilitation 1992 June;73(6):519-26.*

(19) Whiteneck GG, Dijkers MP, Heinemann AW, Bogner JA, Bushnik T, Cicerone KD et al. Development of the Participation Assessment With Recombined Tools-Objective for Use After Traumatic Brain Injury. Archives of Physical Medicine and Rehabilitation 2011 April;92(4):542-51.

(20) Wilkie R, Peat G, Thomas E, Hooper H, Croft PR. The Keele Assessment of Participation: a new instrument to measure participation restriction in population studies. Combined qualitative and quantitative examination of its psychometric properties. Quality of Life Research 2005 October;14(8):1889-99.

*Initial development studies included in the present review.