Office of Health Economics Research

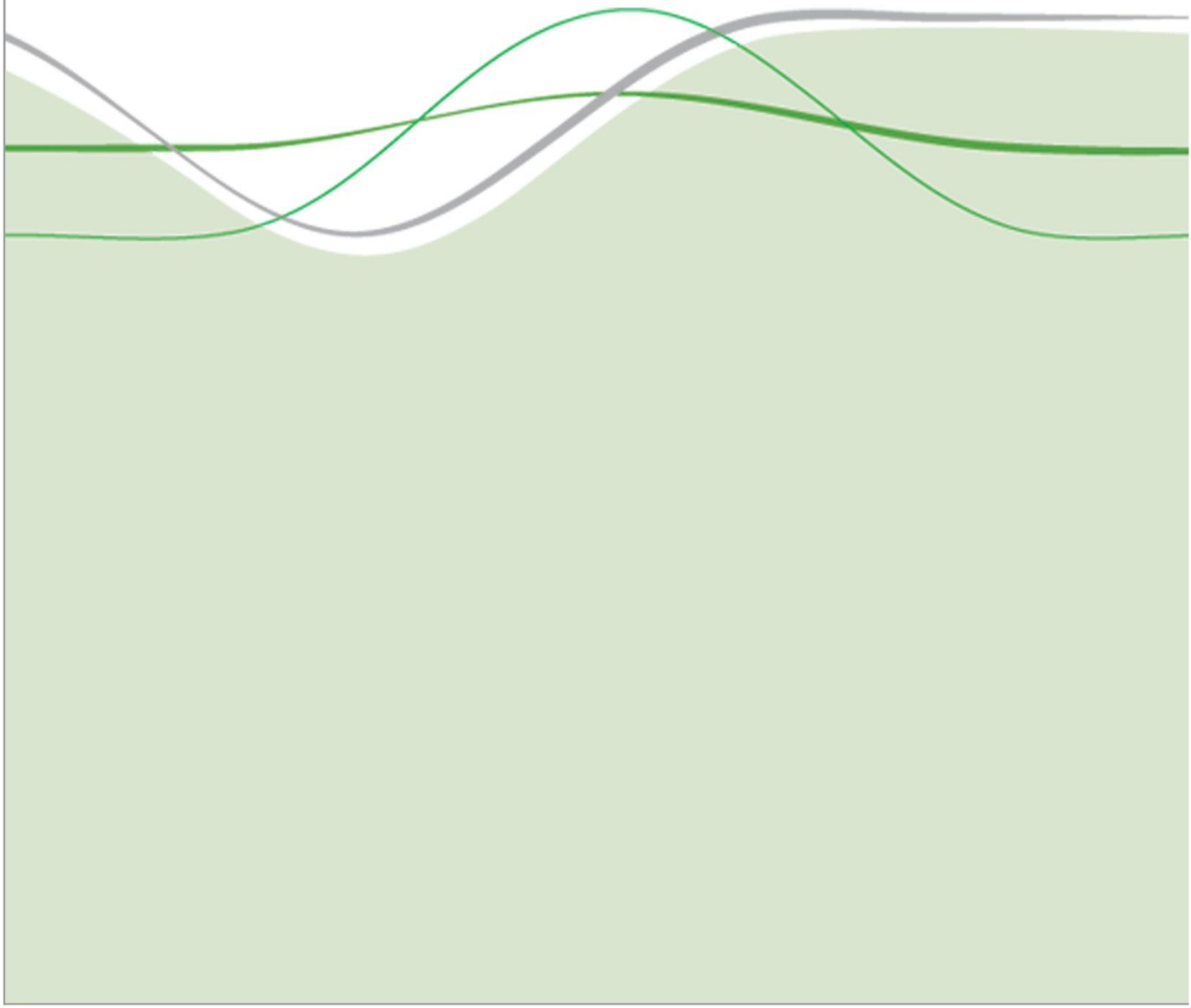
Research Paper

Can We Really Compare and Aggregate PRO Data Between People and Settings?

Implications for Multi-Country Clinical Trials and HTA

April 2019

Nancy Devlin, Paula Lorgelly and Mike Herdman



Can We Really Compare and Aggregate PRO Data Between People and Settings?

Implications for Multi-Country Clinical Trials and HTA

Nancy Devlin[a,b], Paula Lorgelly[c] and Mike Herdman[d]

[a] Health Economics Unit, Melbourne School of Population and Global Health, University of Melbourne, Australia

[b] School of Health and Related Research, University of Sheffield

[c] King's College London

[d] Office of Health Economics, London

April 2019

Please cite this report as:

Devlin, N., Lorgelly, P. and Herdman, M., 2019. Can We Really Compare and Aggregate PRO Data Between People and Settings? Implications for Multi-Country Clinical Trials and HTA. OHE Research Paper, London: Office of Health Economics. Available at: https://www.ohe.org/publications/can-we-really-compare-and-aggregate-pro-data-between-people-and-settings-implications

Corresponding author: Nancy Devlin; nancy.devlin@unimelb.edu.au

The Office of Health Economics (a company limited by guarantee of registered number 09848965; a charity with registered charity number: 1170829)
Southside, 7th Floor
105 Victoria Street
London SW1E 6QT
United Kingdom

For further information please con
Professor Nancy J. Devlin
Head of Health Economics
and Director, Centre for Health Policy
Melbourne School of Population and Global Health
University of Melbourne
Tel: +61 (0)3 9035 6448
nancy.devlin@unimelb.edu.au

About OHE Research Papers

OHE Research Papers are intended to provide information on and encourage discussion about a topic in advance of formal publication. They are subject to internal quality assurance and undergo at least one external peer review, usually by a member of OHE's Editorial Panel. Any views expressed are those of the authors and do not necessarily reflect the views or approval of the OHE's Editorial Panel or Research and Policy Committee, or its sponsors.

Once a version of the Research Paper's content is published in a peer review journal, that supersedes the Research Paper and readers are invited to cite the published version in preference to the original version.

Funding and Acknowledgements

The authors' work on this paper was undertaken in their roles at OHE and was not directly supported by a grant. The authors are grateful to Andrea Manca for useful discussions on this topic, and to Diane Wild for helpful comments received on an earlier draft.

Note that Nancy Devlin and Paula Lorgelly undertook the majority of this work while employed at the Office of Health Economics.

Table of Contents

ABSTRACT

Clinical trials are increasingly being conducted across multiple countries and regions. The collection of patient-reported outcome (PRO) data in these trials, and the practice of pooling data from them in analysis, relies on patients' responses to PRO instruments being strictly comparable. There are two aspects of this assumption: (a) that the PRO instruments themselves generate responses which are strictly comparable when fielded in different contexts; and (b) that the way in which people from different socio-economic or cultural groups self-report their health on those instruments is fundamentally the same. The aim of this paper is to provide an overview of the issues that might limit comparability of PRO data and to highlight some of the evidence that exists on these issues. We note some of the implications for the development and use of PRO instruments, for their application in multi-country clinical trials, and for employing evidence from them in regulatory and reimbursement decisions. Although much progress has been made in this area, there is still scope for further research and improvement. Numerous factors can affect the comparability of PRO data across (and potentially within) countries and cultures. Failure to recognise and account for these differences could lead to incorrect conclusions about the effectiveness and cost effectiveness of new medicines and other health care interventions. We suggest areas where further research and enhanced guidelines for users of PRO instruments and data would be useful.

1. INTRODUCTION

There is growing acceptance of the importance and relevance of patient reported outcomes (PRO) data in informing decisions about pricing and reimbursement of new health care technologies. This reflects increasing recognition that patients' subjective accounts of their health provide an important complement to clinical data.

In parallel, clinical trials are increasingly conducted across multiple countries. Shenoy (2016) notes that it is now common practise to conduct clinical trials across several regions, with an increase in data collection in Asia, Latin America, Middle East and Africa (Shenoy, 2016). Patients are therefore recruited in multiple countries and the data pooled for analysis. Evidence from these trials is subsequently used to inform regulatory and reimbursement decisions in a range of health care systems.

This approach to data collection and analysis assumes that PRO data collected in different countries and cultural contexts is strictly comparable. This same assumption underlies the practice of using PRO data to draw comparisons between the self-reported health of people in different countries (Salomon, Tandon and Murray, 2004; Subramanian, Huijts and Avendano, 2010) or in different regions or sub-populations within a given country (Szende, Janssen and Cabases, 2014).

For the assumption to be correct, it requires: (a) that the PRO instruments themselves generate strictly comparable data when fielded in different contexts; and (b) that the way in which different groups of people self-report their health on those instruments is fundamentally the same.

With respect to (a) conceptual and semantic issues can potentially compromise the comparability of data collected with PRO instruments. The concept of HRQoL, if it exists, may be constructed differently in different cultures and socio-economic groups. For example, Perkins et al (2004) note that the generic preference-based instrument the EQ-5D may not have validity with the indigenous Maori people of New Zealand (Perkins, Devlin and Hansen, 2004), whose people consider biological health to be inextricably linked with mental, spiritual and whanau (family) wellbeing (New Zealand Ministry of Health, 2018). Maher (2008) notes that Australian aboriginal people similarly hold social and spiritual dysfunction as central to beliefs about health and illness(Maher, 1999). In a systematic review of the use of HRQoL instruments with indigenous people globally, Angell et al (2016) notes that there have been limited attempts to develop appropriate instruments and that domains which lie outside traditional PRO measures may be important to the HRQoL of these populations (Angell et al., 2016).

Even where the underlying constructs of a PRO instrument are demonstrated to be valid across settings, semantic issues may mean that PRO questionnaire items and level descriptors do not mean precisely the same thing across different languages, even when translation is undertaken with great care.

With respect to (b), the way individuals interact with PRO instruments when self-reporting their health may differ. Cultural characteristics may influence how scales are used. For example, Feng et al (2017) noted differences both within and between western and Asian countries in the self-reporting of pain. Peoples' self-perception of their health is likely to depend on their expectations about what is 'normal' in terms of health, and levels of knowledge about health and health problems. This, in turn, will be conditioned by social context, education, literacy, income, availability of health care facilities, public health education efforts, and so on.

For example, the visual analogue scale included in the widely used EQ-5D instrument, the EQ-VAS (Van Reenen and Janssen, 2015), asks people to report their overall health today, on a scale where 0 is the 'worst imaginable health' and 100 is 'best imaginable health'. Comparing these data relies on all people sharing the same (unstated) view about what constitutes the best (and worst) possible health they can imagine experiencing. The same issue potentially arises in the 'profile' data captured by the EQ-5D and other PRO instruments: whether a person reports having *no* or *mild* problems on a given question item, for example, will depend on what that person considers to be a 'problem' as opposed to what is 'normal' in their experience.

For the most part, analysis of PRO data from multi-country clinical trials assume that these kinds of issues do not compromise our ability to pool or compare data from different contexts. Is this assumption legitimate?

Sen (2002) expressed similar concerns about self-reported morbidity (e.g. Sen, 1970 and Sen, 1979 see Sen (2017) for collection of work). He presents data showing that the US has higher self-reported morbidity than one of the poorest states of India, Bihar; and similarly, that Bihar has lower levels of self-reported morbidity than Kerala, where education, income and life expectancy are considerably higher. Sen suggests that people in communities where disease is prevalent and where there are few health care facilities may regard their symptoms as normal; whereas people with more education and health care are better positioned to diagnose and perceive themselves as having health problems. He concludes that "the internal [self-reported] view of health deserves attention but relying on it in assessing health care or in evaluating medical strategy can be extremely misleading" (Sen, 2002, p.861).

This is an important challenge to health economists and other health researchers who routinely use PRO data in a manner that assumes that strict comparability holds.

The aim of this paper is to provide an overview of the issues that might limit comparability of PRO data, and to highlight some of the evidence that exists about these issues. We identify gaps in knowledge where further research is required and consider the implications for the development of PROs; their use in multi-centre clinical trials; and the use of that evidence in local health technology appraisal (HTA) and other health care decisions that rely on PRO data. The paper proceeds as follows: we first discuss the challenges as well as the opportunities for using PROs in multi-country studies. We then consider the heterogeneity that can occur in PRO data and specifically that which can impede comparability and transferability. We briefly discuss the use of anchoring vignettes as an example of how to correct for one type of heterogeneity. The paper concludes with a discussion of the implications for the development of PRO instruments; the use of PRO evidence from multi-centre clinical trials in local decisions about regulation and reimbursement and makes recommendations for future research and guideline development.

2. USING PROS IN MULTI-COUNTRY STUDIES: CONCEPTUAL AND PRACTICAL CHALLENGES

The increasing number of multi-country and multi-regional clinical trials has brought to the fore issues and challenges for the PRO field which have also been considered within the clinical trials community. As Komiyama et al. (2013, p.26) state, "Increasing the number of regions in clinical trials results in ethnic and cultural diversity of patients,

which in turn leads to a potential for increased variability in treatment response among patients".

Chen et al. (2010), examining this issue in relation to potential regional effects in multi-regional clinical trials (MRCTs) in schizophrenia, found that the observed treatment effect was generally smaller in the US than in non-US regions, potentially because placebo response had increased more over time in the US region.

Awareness that regional variations of treatment effect can exist, or that there are biases that might create the appearance of such effects, is reflected in the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) guidelines covering MRCTs (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2016) and ethnic factors affecting the acceptability of foreign clinical data (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 1998). The most recent version of the MRCT guidelines states that "Of specific concern in MRCTs are those endpoints that could be understood and/or measured differently across regions." The guidelines go on to note that examples of such endpoints are "psychometric scales, assessment of quality of life, and pain scales" and recommend that to "guarantee that such scales can be properly interpreted, the scales should be validated and their applicability to all relevant regions justified before starting the MRCT" (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 1998).

In the context of PRO data for health economics analysis, a number of guidelines and checklists have been proposed to assess the transferability of results (for a review, see Goeree et al (2011)). However, while some of those guidelines and checklists remark on the importance of using appropriate health state utilities in local estimations of cost effectiveness, none appear to consider the possibility that responses to the questionnaires themselves might be affected by cultural factors that vary by region, country, or ethnicity.

When considering the extent to which PRO data from MRCTs can be compared, aggregated, or transferred, it is useful to bear in mind several characteristics of a PRO measure[1]. The key question is whether or not an instrument is likely to give equivalent results when used in different cultural settings. Herdman et al (1998) argue that there are six types of cross-cultural equivalence that need to be addressed for an instrument to be considered cross-culturally valid. These are: conceptual equivalence, item equivalence, semantic equivalence, operational equivalence, measurement equivalence, and functional equivalence (Herdman, Fox-Rushby and Badia, 1998). The definition of those different types of equivalence, how and when they should be tested, and the risks of not testing them are shown in Table 1. The model offers a theoretical framework for assessing the degree to which data from a specific PRO instrument might be transferable across different countries or different cultural settings.

While all of these equivalence types are relevant when assessing the cross-cultural validity of a questionnaire, three of them - conceptual, semantic, and operational equivalence - are particularly important when considering the transferability of PRO data across cultural settings or decision-making jurisdictions.

[1] Of course, these considerations apply not only to MRCTs but to most or all types of study which are carried out simultaneously in several countries using PRO instruments.

Table 1. Types of equivalence and relevance for comparability and transferability of results

TYPE OF EQUIVALENCE	WHAT IS IT?	WHEN/HOW TESTED	RISKS IF NOT TESTED
Conceptual equivalence	The domains that are important to a concept such as HRQOL and the weight attached to each domain are similar or equivalent across settings	Ideally tested during the development stage, but investigation of suitability of already established content can also be tested later by qualitative research and statistical approaches including factor analysis	Constructs used may be inappropriate in some settings, e.g. relevant domains omitted, and/or irrelevant domains included.
Item equivalence	Appropriateness of items to measure specific domains across settings	Ideally tested during the development stage but can also be investigated during cultural adaptation procedures and using psychometric techniques including factor analysis and item response analysis	Inappropriate items may be used. Items may be unacceptable, irrelevant, and/or assess different levels of the construct in different settings
Semantic equivalence	Possibility of expressing the same item meaning across settings.	Can be tested during the development stage or during cultural adaptation through interviews/focus groups with members of the target population	Interpretations of items differ across settings. May affect comparability/ transferability of responses.
Operational equivalence	Possibility of using the same measurement approach (format, mode of administration, measurement methods and technologies, etc) across settings	Can be tested during the development stage or during cultural adaptation through interviews/focus groups with members of the target population and using quantitative techniques to explore whether operational aspects differentially affect scoring across settings	Risk of using formats, measurement methods, etc which are not equally well understood or appropriate across settings. Possible impact of response style.

TYPE OF EQUIVALENCE	WHAT IS IT?	WHEN/HOW TESTED	RISKS IF NOT TESTED
Measurement equivalence	Does the instrument have equivalent or similar psychometric properties when used in different settings and languages?	Assessed once data has been gathered using the standard array of psychometric methods	Risk of using measures which are less valid, reliable, and/or responsive in some settings
Functional equivalence	An amalgam of the other types of equivalence. Does the instrument do what it is supposed to do equally well in different settings?	Will be shown when all other types of equivalence have been assessed.	A failure to show any of the other types of equivalence could mean that use of the tool is not appropriate and/or will provide non-comparable results across different settings

3. CONCEPTUAL EVIDENCE

The conceptual equivalence of a measure refers to the extent to which "the concept(s) of interest exist, are equally relevant, and share the same structure across different cultures" (Regnault and Herdman, 2015). In the case of a relatively complex concept such as HRQoL, and in the context of preference-based measures such as EQ-5D or HUI, this involves investigating both the way in which different local populations conceptualise HRQoL and the relative weights or values they place on the different dimensions which constitute the concept.

One approach to achieving conceptual equivalence is the simultaneous development of instruments in several different cultural settings. A good example of this approach is the World Health Organization's Quality of Life (WHOQOL) instrument, which was researched and developed from the outset in 15 settings worldwide (Orley and Kuyken, 1994). The Obesity and Weight Loss Quality of Life measure (OWLQOL) was developed across 5 European countries and the US (Niero et al., 2002). Other examples, although on a more limited scale, include the EQ-5D (Kind, Brooks and Rabin, 2015) and the EORTC measures, all of which are developed simultaneously in at least three European countries (Vivat et al., 2013; Velikova et al., 2012). These are, however, rare examples: the vast majority of PRO instruments in use today have been developed in a single country and then 'exported' for use in other countries. In this case, the scope for exploring the concept of interest in other countries is arguably more limited.

While guidelines for the translation or cultural adaptation of PRO measures (Wild et al., 2005; Koller et al., 2012) have helped to set high standards in achieving linguistic comparability between different language versions of a measure, they provide relatively little room for exploring the concept on a more fundamental level. This was noted by Bowden and Fox-Rushby in a review of the translation of nine generic measures of HRQoL (Bowden and Fox-Rushby, 2003). Their conclusion was that there was "a misguided pre-occupation with scales rather than the concepts being scaled and too much reliance on unsubstantiated claims of conceptual equivalence". They also concluded that research practice and translation guidelines needed to change to "facilitate more effective and less biased assessments of equivalence of HRQL measures across countries". The situation has arguably remained relatively unchanged since that review.

This issue can also be framed in terms of content validity. If the content validity of a PRO measure depends on obtaining input from relevant stakeholders (Patrick and Erickson, 1993) including patients, caregivers, and clinical experts, then it is not clear to what extent content validity can be claimed for a measure which is used beyond the cultural setting in which it was originally developed. The question then is how to explore content validity when a measure is used in that way. There might be several ways to do that; one way, however, is to acknowledge that, even when an instrument exists, additional research can and should continue to be carried out into its content validity in other countries or regions, using similar techniques to those used when developing the measure initially, though presumably on a smaller scale.

For example, the PRO guidance from the FDA (U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research and U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health, 2006) indicates that an assessment of content validity of translations should be carried out as

part of the translation and linguistic validation process. However, it did not provide any detailed guidance on the best way to do this, e.g., how many respondents, what type of interviews, and how to deal with the results if they were to suggest a lack of content validity.

4. SEMANTIC EQUIVALENCE

A further critical element when using PROs in multi-country studies is the semantic equivalence of different language versions used. The International Society for Quality of Life Research (ISOQOL) notes that to "be able to compare or combine HRQoL results across those groups, it is critical that the measured HRQoL concept and the wording of the questionnaire used to measure it is interpreted in the same way across translations" (Reeve et al., 2013).

The translation or linguistic validation of PROs has become increasingly standardised in recent years in order to avoid differences in wording affecting responses. Guidelines aim to ensure rigour in the translation process and avoid problems caused by inappropriate wording in the target languages; they generally recommend forward and back translation and cognitive debriefing in a small number of the intended target population (Wild et al., 2005; Koller et al., 2007).

Given the critical importance of PRO translation to the collection of high-quality data in MRCTs, there has arguably been little work into how poor translation might affect results. One example is Regnault et al (2015) who explored the impact of the ''contamination'' of a cultural subgroup by a flawed PRO measurement, such as that stemming from poor translation, on study power (Regnault, Hamel and Patrick, 2015). They observed that this type of poor PRO measurement in a cultural subgroup could lead to a considerable decrease in study power which would reduce the likelihood of showing a treatment effect and emphasised the importance of optimising the conceptual and linguistic equivalence of PRO measures when pooling data from multi-country clinical trials.

Guidelines should also not be considered as static documents but may need updating as more is learnt about processes and procedures. For example, the EuroQol Group has its own guidelines for translation of the EQ-5D which are also employ forward and back translation, and cognitive testing (Rabin et al., 2014). However, after several years of experience in translating the instrument, it was observed that responses would often just briefly paraphrase the existing item wording, which was not very helpful when judging the suitability of the translated wording. In an effort to improve the richness and depth of the feedback received from cognitive testing, the guidelines were therefore modified to include examples of the type of responses that were required and which are considered useful when deciding whether respondents interpret items as intended. Given that EQ-5D is a preference-based measure, the accurate and appropriate translation of the labels used to represent levels of problem severity in each dimension ('slight', 'moderate', 'severe', etc) is also of paramount importance. For that reason, a further modification to the standard translation procedure was introduced, whereby participants in the cognitive debriefing exercise are asked to score the severity of each label on a visual analogue scale like the one used in the instrument itself. This is a useful check on how these labels are interpreted in different languages.

With preference-based measures the importance of the labels is paramount not just because of the role they play in allowing respondents to describe their health but also

because they are critical to the *valuation* of health states. Craig et al (2017) report a novel approach to investigating the interpretation of severity labels and their possible impact in health state valuation (Craig et al., 2017). The researchers used a paired comparisons approach to assess preference inversions between the fourth (severe problems) and fifth (extreme problems/unable to) levels of each of the five domains in the new EQ-5D-5L. This is believed to occur because it is not sufficiently clear to respondents in stated preference exercises, where the ordinal structure of the underlying descriptive system is not obvious, whether 'severe' or 'extreme' is meant to be the worse level of problem. The study was performed in the US and Brazil and showed that, in the English-speaking respondents, preference inversion (i.e. the propensity of respondents to interpret the theoretically less severe label as representing a higher level of severity than the, theoretically, more extreme label) was only present to any extent in the anxiety/depression dimension, whereas in the Portuguese-speaking respondents, preference inversions were more common overall, and particularly in the pain/discomfort and anxiety/depression dimensions. Although this sort of issue is likely to be minimised when responding to the questionnaire itself, because in that instance respondents see all response options in context of the overall descriptive system, so the ordinal nature of the labels and problem levels is obvious. However, the issue is arguably greater in valuation exercises, as the items will be seen in isolation, in the context of a health state profile. Cole et al (2018) report an attempt to overcome this issue in stated preference studies to value HRQoL by presenting the health states to be valued 'in the context' of the descriptive system, in the same way patients see it when self-reporting their health (Cole et al., 2018).

5. OPERATIONAL EQUIVALENCE: THE IMPACT OF RESPONSE STYLE

A third factor which can affect responses across cultures and therefore the comparability or transferability of data is response style, which refers to a respondent's tendency to systematically respond to questionnaire items in a given way regardless of item content (Baumgartner and Steenkamp, 2001). An understanding of response style effects in different settings and using different measurement approaches, such as different types of response scale, is part of the assessment of operational equivalence.

Examples of this include extreme response style (ERS) and acquiescent response style (ARS). ERS refers to the tendency to use the endpoints of a scale, such as the rating categories 0 and 4 on a five-point scale from 0 (no problems) to 4 (unable to). ARS, on the other hand, refers to a tendency to "respond to descriptions of conceptually distinct attributes or attitudes with agreement/affirmation (agreement acquiescence) or disagreement/opposition (counter-acquiescence) regardless of their content" (Rammstedt, Danner and Bosnjak, 2017). Likewise, mild response style (MRS) is reflected by a tendency to avoid extreme response categories and a preference for middle categories of response options. The presence, characteristics, and sources of such response styles have been widely studied and reported in the social sciences and marketing literature. One example of a large-scale study of response styles in 26 countries found major differences in response styles between countries with, for example, students from Spanish-speaking countries showing high ERS and acquiescence and East Asian (Japanese and Chinese) respondents showing a relatively high level of MRS. Within Europe, German respondents appeared to show higher acquiescence than British respondents (Harzing, 2006). Even within certain regions, such as Eastern

Europe, the authors report the presence of two clear patterns of responding with Russia and Poland showing high disacquiescence, low MRS and low positive ERS, and Bulgaria and Lithuania showing the reverse pattern. The authors report that country-level characteristics such as power distance, collectivism, uncertainty avoidance, and extraversion all influenced response styles and that English-language questionnaires elicited a higher level of middle responses, while questionnaires in a respondent's native language generated more extreme response styles. There has been arguably little research on these areas within the PRO field.

Even in the case of a universal symptom such as pain, studies have shown that Japanese subjects provide lower pain ratings for equivalent 'objective' levels of pain than European subjects (Komiyama et al., 2009), possibly because of different views related to being expressive about pain (Hobara, 2005; Abe et al., 2008). While there has been little investigation into these differences between Japanese and European respondents, their impact could be important. For example, Gerlinger et al (2012) reported differences between European and Japanese populations in pain relief from endometriosis medication when measured using a VAS scale (Gerlinger et al., 2012) (see Figure 1). Although the authors indicated that there was very little difference in effectiveness between the two treatments tested in the two populations, the considerable difference between Japanese and European populations in terms of the pain relief obtained was striking (mean VAS changes of −47.5 mm and −30.2 mm in European and Japanese women, respectively).

Figure 1. Mean change in VAS (mm) between baseline and week 24 in European and Japanese trials of DNG vs GnRH analogue

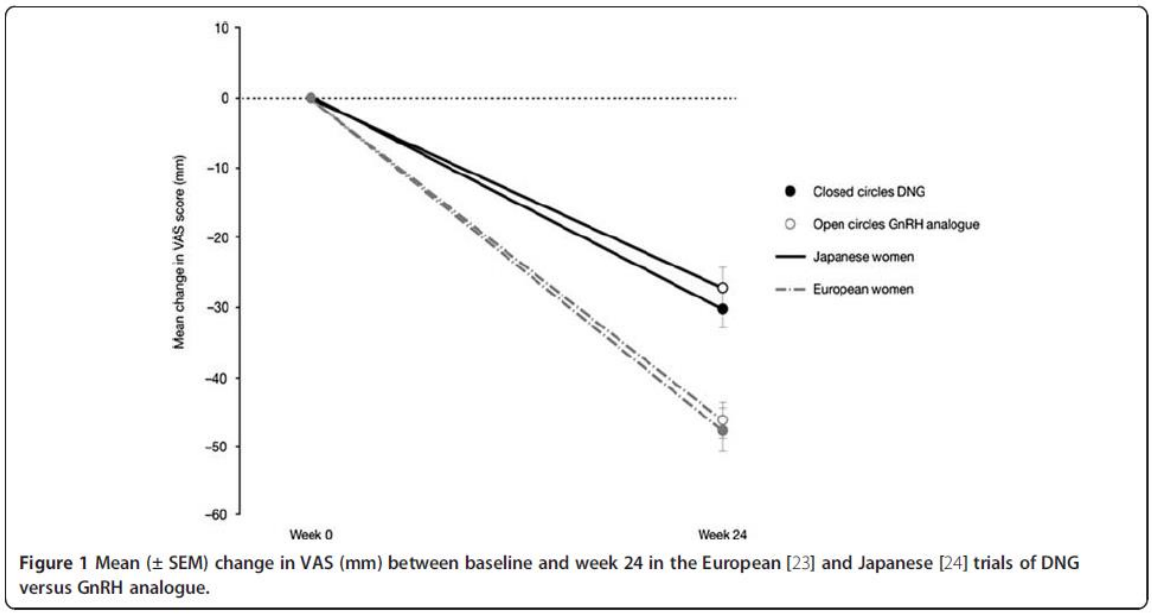


Figure 1 Mean (± SEM) change in VAS (mm) between baseline and week 24 in the European [23] and Japanese [24] trials of DNG versus GnRH analogue.

Source: Reproduced with permission from Gerlinger et al. (2012)

The situation with regard to the use of PROs in MRCTs or other studies involving patients from different regions is perhaps best summed up by Salomon et al (2011). They examined the comparability of EQ-5D data from a multicentre clinical trial in diabetes performed in 20 countries. Patients were grouped into 3 regions defined by geography and levels of economic development (Asia, Established Market Economies, Eastern Europe). Substantial regional reporting differences in presence of problems on EQ-5D were found even after controlling for demographics, common risk factors, and history of

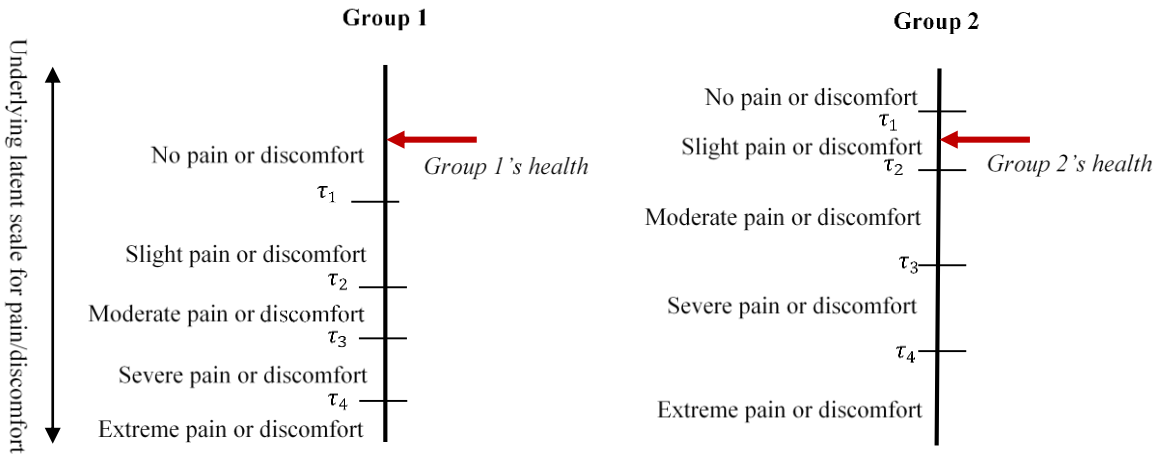
major disease complications. Compared to Established Market Economies, adjusted odds ratios for reporting problems in at least 1 health domain were 1.79 (1.55 to 2.06) in Eastern Europe and 0.76 (0.67 to 0.86) in Asia. They concluded that commonly used health status instruments may have important problems in comparability across settings.

6. RESPONSE SCALE-HETEROGENEITY: METHODS FOR IDENTIFYING AND ADJUSTING FOR IT

Systematic variation in the use of response categories by groups of individuals (either across regions, culture, gender or age groups) arising from differences in response style is referred to as response-scale heterogeneity (RSH) (Angelini et al., 2014; Knott et al., 2017a)[2]. Several PROs have been found to suffer from RSH (Salomon et al., 2011; Whynes et al., 2013; Grol-Prokopczyk, Freese and Hauser, 2011) suggesting that comparing results between groups of people could be misleading if they systematically differ in the use of the PRO response categories.

RSH is illustrated in Figure 2, using the example of the pain/discomfort domain of the EQ-5D-5L. The underlying true, but unobserved (latent) scale for health is represented by the vertical line. Assume that we wish to compare the health of two groups (say, Europeans and non-Europeans) where respondents are asked to rate their level of pain or discomfort using the response categories *no*, *slight*, *moderate*, *severe* or *extreme* pain/discomfort. How each group divides the latent scale into the five response categories is represented by the placement of the inter-category thresholds τ1, τ2, τ3 and τ4. Despite having identical levels of true health (with respect to pain/discomfort) as illustrated by the red arrows, Group 2 reports slight pain/discomfort, whereas Group 1, who may be more stoic compared to Group 2, report no pain/discomfort. Researchers are unaware of the groups' latent health and would typically be unaware of the location of each group's inter-category thresholds – so will incorrectly conclude that Group 1 is in better health than Group 2. However, if the placement of the thresholds were "observable", the presence of RSH would be evident.

Figure 2 Response-scale heterogeneity in the EQ-5D-5L pain/discomfort domain (example)



[2] RSH is also referred to as reporting heterogeneity (Bago d'Uva et al., 2011a, 2008) and differential item functioning (DIF) (King et al., 2004; Hopkins and King, 2010; Van Soest et al., 2011).

6.1. Addressing response-scale heterogeneity: anchoring vignettes

To make a meaningful comparison between the self-reported health of Groups 1 and 2 it is essential to adjust for RSH. One approach to doing so is through the use of anchoring vignettes (King et al., 2004), which allow us to observe the thresholds noted above. Vignettes have been used to address RSH in self-reported measures of political efficacy, job/income/life satisfaction, and general/specific health measures, (Grol-Prokopczyk, Freese and Hauser, 2011; Bago d'Uva et al., 2011b; Kapteyn, Smith and van Soest, 2007; Kapteyn et al., 2011; Salomon, Tandon and Murray, 2004; Bago d'Uva et al., 2011a; King et al., 2004). The anchoring vignette approach involves the inclusion of at least one, but typically several, brief health descriptions of hypothetical individuals (vignettes) that respondents are asked to rate using the PRO of interest (King et al., 2004) (see Box 1). As the health state described in the vignettes is the same for all respondents, variation in ratings of it can be used to identify and correct for RSH.

Box 1. Vignette example, self-care dimension in the EQ-5D-5L

Tom takes twice as long as others to put on and take off clothes but needs no help with this. Although it requires an effort, he is able to bathe and groom himself, though less frequently than before.

Select the ONE option that best describes TOM'S SELF-CARE:

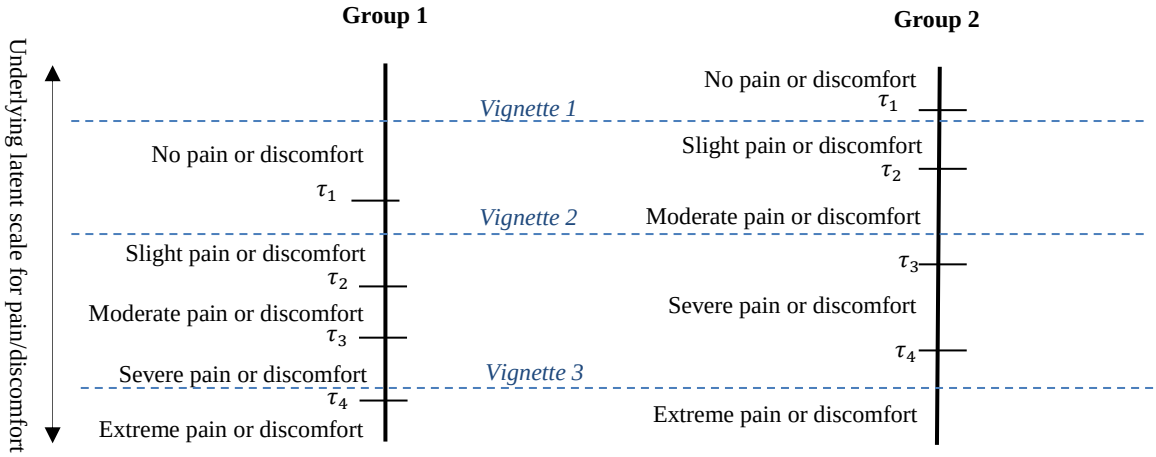
- ☐ He has no problems washing or dressing himself
- ☐ He has slight problems washing or dressing himself
- ☐ He has moderate problems washing or dressing himself
- ☐ He has severe problems washing or dressing himself
- ☐ He is unable to wash or dress himself

The intuition behind this approach is illustrated in Figure 3, which extends the example in Figure 2. Groups 1 and 2 are assumed to divide the underlying latent scale for pain/discomfort as before; but to allow us to identify their response thresholds, respondents from both groups are given several (in this example, three) vignettes describing differing levels of pain or discomfort, for a hypothetical individual, which they are asked to rate using the same underlying response scale they use to rate their own pain/discomfort. An example of a vignette in this instance may be "Alex suffers from back pain every day and is unable to stand or sit for more than half an hour at a time". Respondents are asked to rate the health of Alex using the same response categories they use to rate their own health (i.e. the levels of the EQ-5D-5L domain for pain/discomfort). In the diagram the fixed health of each vignette is represented by the dotted horizontal lines. Aggregation of the individual vignette responses and analysis identifies the various thresholds (cut-points) and appropriately compares Group 1 and Group 2; in this example it is evident that Group 2 considers all vignettes to be describing situations of more pain/discomfort.

It is important to note that the anchoring vignette method relies heavily on two assumptions holding true (King et al., 2004). First, it requires that all respondents interpret the health states described by the vignettes in the same way and on the same uni-dimensional scale, aside from random error, i.e. vignette equivalence (VE). VE is demonstrated in the example above by the horizontal dotted lines. Second, it requires response consistency (RC), whereby respondents rate the health of the hypothetical

people described in the vignettes in the same way (i.e., using the same underlying scale) that they would use to rate their *own* health.

Figure 3 Anchoring vignettes in the EQ-5D-5L pain/discomfort domain (example)



Various studies have explored VE and RC. Grol-Prokopczyk et al. (2011) used a series of vignettes with no disease, heart disease and diabetes examples with different levels of health functioning and disability to test whether RC is affected by the inclusion of medical diagnoses and by personal experiences of medical conditions in vignettes. Kapteyn et al. (2007) considered RC in vignettes specific to the domains of sleep, mobility, concentration, breathing and affect/depression. They found that RC was satisfied for the sleep domain only. They concluded that there is a need for a more systematic approach to the design of anchoring vignettes. Au and Lorgelly (2014) addressed this by using qualitative research to develop and design vignettes, with a focus on ensuring questionnaire design satisfied VE and RC. This included a number of design features, such as ensuring that the health problems in the vignettes could be experienced by any age group (i.e. were not associated with old age). This resulted in both domain-specific vignettes (like those above) and more holistic vignettes that cover the whole construct of health (see Box 2).

Box 2. Vignette example, EQ-5D-5L as a whole

Brian walks for one to 3 kilometres every day without tiring, but he cannot run anymore due to an injured knee. He keeps himself neat and tidy. He showers and dresses himself each morning in under 15 minutes. He works in the public sector. He misses work 1 or 2 days per year due to illness. He has a headache once a month that is relieved 1 hour after taking a pill. Brian remains happy and cheerful most of the time, but once a week feels worried about things at work. He feels very sad once a year but is able to come out of this mood within a few hours.

In a subsequent study, vignettes were used to identify RSH in the EQ-5D-5L in a sample of Australian respondents 55-65 years old (Knott et al., 2017b). The EQ-5D-5L index for the sample was compared to an RSH-adjusted index. The average difference in the EQ-5D-5L index between people born in Australia and other English-speaking countries was 0.072 prior to adjustment; following adjustment for RSH it was 0.155. Using a minimal important difference (MID) of 0.074 (Walters and Brazier, 2005), this would suggest that

a difference exists between the groups that was not evident previously. Other researchers have drawn similar conclusions when examining RSH in general and domain-specific self-assessments of health (Grol-Prokopczyk, Freese and Hauser, 2011; Molina, 2016). The presence of RSH in PRO data across countries and regions could affect the conclusions drawn from multi-country trials, as well as potentially limiting the transferability and generalisability of findings from one (or more countries) to another. This has important implications for the use of PRO evidence in local regulatory and reimbursement decisions, particularly if a technology or service is considered cost effective in a subgroup, and that subgroup systematically uses the response scales differently.

Although the vignettes approach has promise, its use in correcting for RSH relies heavily on the assumptions of VE and RC. Recall that RC holds if respondents use the same scales when evaluating themselves and when evaluating the vignette individuals, while VE holds if different respondents interpret the same vignette in the same way. Further research is required on the use of vignettes and their value particularly as their elicitation is not costless; it is also important to consider other methods for identifying and correcting for RSH.

7. CONCLUSIONS: IMPLICATIONS FOR RESEARCHERS AND USES OF PRO DATA

Concerns about the potential non-comparability of PRO data between countries and regions have been noted before; in this paper we have highlighted a number of examples of these issues in practise. It is notable that there appears to have been relatively little methods development to address these issues. Our aim here was to try to heighten awareness of these concerns in the hope that some of the examples and ideas might spur more research and thinking in that direction.

There are lessons here for instrument development: despite a recognition that content validity is likely to be specific to a particular setting, most PRO instruments are still developed in one country/region and then 'exported'. In reality most PRO instruments are intended for broader use across different countries or settings, so simultaneous development across different cultural settings should be encouraged as good practice. More generally, a greater focus on content validity is needed. This links to the rise in interest in 'patient-centricity' and patient-relevant outcomes: standardised PROs may need to make greater recognition of the differences *between* patients, in different settings, in terms of what matters to them about their health.

For those analysing PRO data obtained from multiple countries or regions, analysis and reporting of results by country/region should be encouraged as good practice and reported alongside overall sample averages where those data are submitted as part of evidence to support regulatory or reimbursement decisions. Decision makers need to be aware of any potential limitations of the transferability of PRO results obtained across multiple countries/regions to the local settings. Statistical techniques for addressing heterogeneity, such as clustering methods, have an important role here (Regnault and Herdman, 2015). These issues are equally relevant to condition-specific as well as generic PROs. In the case of preference-based generic measures, there may be differences between groups not only in terms of how patients self-report their health but also in the weights assigned to health states (Elbarazi et al., 2017). Whether these two elements combine to modify or amplify differences in the use and interpretation of

preference-weighted PROs between countries or regions is unknown and is under-researched.

It is worth noting that while this paper has focused on issues arising from the use of PROs in multi-country and multi-region clinical trials, these same issues could also arise within a country and affect (for example) the conclusions drawn from population health surveys, especially as these relate to inequalities in self-reported health.

In conclusion, we think there are two questions about the use of PROs in multi-country clinical trials which merit wider attention. First, given that multi-country, multi-region trials have become so common, (how) can we be sure we're measuring HRQOL - and related concepts - in an appropriate way in all of the different settings? Second, although it appears to be widely accepted that it is important to take into account differences in preferences between countries when analysing clinical trial data (for example, by using local 'utilities' to preference-weight PRO data in estimating QALYs), what about the patient reported outcomes themselves? (How) can we be sure that results (health gains and losses) seen in multi-country trials are an adequate reflection of what would happen in any one country if the trial was just carried out there?

REFERENCES

Abe, Y., Miyashita, M., Ito, N., Shirai, Y., Momose, Y., Ichikawa, Y., Tsuji, S. and Kazuma, K., 2008. Attitude of outpatients with neuromuscular diseases in Japan to pain and use of analgesics. *Journal of the Neurological Sciences*, 267(1–2), pp.22–27. 10.1016/j.jns.2007.09.027.

Angelini, V., Cavapozzi, D., Corazzini, L. and Paccagnella, O., 2014. Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases. *Oxford Bulletin of Economics and Statistics*, 76(5), pp.643–666. 10/gfxhzf.

Angell, B., Muhunthan, J., Eades, A.-M., Cunningham, J., Garvey, G., Cass, A., Howard, K., Ratcliffe, J., Eades, S. and Jan, S., 2016. The health-related quality of life of Indigenous populations: a global systematic review. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 25(9), pp.2161–2178. 10/f83w8k.

Au, N. and Lorgelly, P.K., 2014. Anchoring vignettes for health comparisons: an analysis of response consistency. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 23(6), pp.1721–1731. 10/f59xmt.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and van Doorslaer, E., 2011a. Education-related inequity in healthcare with heterogeneous reporting of health. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 174(3), pp.639–664. 10/dbg6d4.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and van Doorslaer, E., 2011b. Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *The Journal of human resources*, 46(4), pp.875–906.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M. and O'Donnell, O., 2008. Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, 17(3), pp.351–375. 10/fgtcvd.

Baumgartner, H. and Steenkamp, J.-B.E.M., 2001. Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, 38(2), pp.143–156.

Bowden, A. and Fox-Rushby, J.A., 2003. 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 & Medicine (1982)*, 57(7), pp.1289–1306.

Chen, Y.-F., Wang, S.-J., Khin, N.A., Hung, H.M.J. and Laughren, T.P., 2010. Trial design issues and treatment effect modeling in multi-regional schizophrenia trials. *Pharmaceutical Statistics*, 9(3), pp.217–229. 10/dsnfv5.

Cole, A., Shah, K., Mulhern, B., Feng, Y. and Devlin, N., 2018. Valuing EQ-5D-5L health states 'in context' using a discrete choice experiment. *The European Journal of Health Economics*, 19(4), pp.595–605. 10/gdtkc5.

Craig, B.M., Monteiro, A.L., Herdman, M. and Santos, M., 2017. Further evidence on EQ-5D-5L preference inversion: a Brazil/U.S. collaboration. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 26(9), pp.2489–2496. 10/gbsc8h.

Elbarazi, I., Devlin, N.J., Katsaiti, M.-S., Papadimitropoulos, E.A., Shah, K.K. and Blair, I., 2017. The effect of religion on the perception of health states among adults in the United Arab Emirates: a qualitative study. *BMJ Open*, 7(10), p.e016969. 10/gb24m6.

Feng, Y., Herdman, M., van Nooten, F., Cleeland, C., Parkin, D., Ikeda, S., Igarashi, A. and Devlin, N.J., 2017. An exploration of differences between Japan and two European countries in the self-reporting and valuation of pain and discomfort on the EQ-5D. *Quality of Life Research*, 26(8), pp.2067–2078. 10/gdtkdb.

Gerlinger, C., Faustmann, T., Hassall, J.J. and Seitz, C., 2012. Treatment of endometriosis in different ethnic populations: a meta-analysis of two clinical trials. *BMC women's health*, 12, p.9. 10/gb3rpv.

Goeree, R., He, J., O'Reilly, D., Tarride, J.-E., Xie, F., Lim, M. and Burke, N., 2011. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. *ClinicoEconomics and Outcomes Research: CEOR*, 3, pp.89–104. 10.2147/CEOR.S14404.

Grol-Prokopczyk, H., Freese, J. and Hauser, R.M., 2011. Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health. *Journal of health and social behavior*, 52(2), pp.246–261. 10/cqjw8r.

Harzing, A.-W., 2006. Response Styles in Cross-national Survey Research: A 26-country Study. *International Journal of Cross Cultural Management*, 6(2), pp.243–266. 10/djp83r.

Herdman, M., Fox-Rushby, J. and Badia, X., 1998. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 7(4), pp.323–335.

Hobara, M., 2005. Beliefs about appropriate pain behavior: cross-cultural and sex differences between Japanese and Euro-Americans. *European Journal of Pain (London, England)*, 9(4), pp.389–393. 10/fdbqwp.

Hopkins, D. and King, G., 2010. Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. *Public Opinion Quarterly*, pp.1–22.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 1998. *Ethnic Factors in the Acceptability of Foreign Clinical Data - E5(1)*. ICH Harmonised Tripartite Guideline.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2016. *General Principles for Planning and Design of Multi-regional Clinical Trials – E17*. ICH Harmonised Guideline.

Kapteyn, A., Smith, J.P. and van Soest, A., 2007. Vignettes and Self-Reports of Work Disability in the United States and the Netherlands. *The American Economic Review*, 97(1), pp.461–473.

Kapteyn, A., Smith, J.P., Van Soest, A. and Vonkova, H., 2011. *Anchoring Vignettes and Response Consistency*. [Product Page] Available at: https://www.rand.org/pubs/working_papers/WR840.html [Accessed 27 Mar. 2019].

Kind, P., Brooks, R. and Rabin, R., 2015. *EQ-5D concepts and methods: - a developmental history*. [online] Netherlands: Springer. Available at: https://www.springer.com/gb/book/9781402037115 [Accessed 27 Mar. 2019].

King, G., Murray, C.J.L., Salomon, J.A. and Tandon, A., 2004. Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98, pp.191–207.

Knott, R.J., Black, N., Hollingsworth, B. and Lorgelly, P.K., 2017a. Response-Scale Heterogeneity in the EQ-5D. *Health Economics*, 26(3), pp.387–394. 10/f9pc8s.

Knott, R.J., Lorgelly, P.K., Black, N. and Hollingsworth, B., 2017b. Differential item functioning in quality of life measurement: An analysis using anchoring vignettes. *Social Science & Medicine (1982)*, 190, pp.247–255. 10/gchhvk.

Koller, M., Aaronson, N.K., Blazeby, J., Bottomley, A., Dewolf, L., Fayers, P., Johnson, C., Ramage, J., Scott, N., West, K. and EORTC Quality of Life Group, 2007. Translation procedures for standardised quality of life questionnaires: The European Organisation for Research and Treatment of Cancer (EORTC) approach. *European Journal of Cancer (Oxford, England: 1990)*, 43(12), pp.1810–1820. 10.1016/j.ejca.2007.05.029.

Koller, M., Kantzer, V., Mear, I., Zarzar, K., Martin, M., Greimel, E., Bottomley, A., Arnott, M., Kuliś, D. and ISOQOL TCA-SIG, 2012. The process of reconciliation: evaluation of guidelines for translating quality-of-life questionnaires. *Expert Review of Pharmacoeconomics & Outcomes Research*, 12(2), pp.189–197. 10/gfxhzg.

Komiyama, O., Hiro, S., Isogawa, N., Toyoizumi, S., Matsuoka, N., Hashigaki, S., Yoshiyama, T. and Maruyama, N., 2013. Evaluation of Data for Multi-Regional Trials: A Three-Layer Approach. *Applied Clinical Trials*, 22(11), pp.25–29.

Komiyama, O., Wang, K., Svensson, P., Arendt-Nielsen, L., Kawara, M. and De Laat, A., 2009. Ethnic differences regarding sensory, pain, and reflex responses in the trigeminal region. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 120(2), pp.384–389. 10/bt7cns.

Maher, P., 1999. A review of 'traditional' aboriginal health beliefs. *The Australian Journal of Rural Health*, 7(4), pp.229–236.

Molina, T., 2016. Reporting Heterogeneity and Health Disparities Across Gender and Education Levels: Evidence From Four Countries. *Demography*, 53(2), pp.295–323. 10/gfxhz9.

New Zealand Ministry of Health, 2018. *Māori health models – Te Whare Tapa Whā*. Available at: https://www.health.govt.nz/our-work/populations/maori-health/maori-health-models/maori-health-models-te-whare-tapa-wha [Accessed 27 Mar. 2019].

Niero, M., Martin, M., Finger, T., Lucas, R., Mear, I., Wild, D., Glauda, L. and Patrick, D.L., 2002. A new approach to multicultural item generation in the development of two obesity-specific measures: the Obesity and Weight Loss Quality of Life (OWLQOL) questionnaire and the Weight-Related Symptom Measure (WRSM). *Clinical Therapeutics*, 24(4), pp.690–700.

Orley, J. and Kuyken, W. eds., 1994. The Development of the World Health Organization Quality of Life Assessment Instrument (the WHOQOL). In: *Quality of Life Assessment: International Perspectives*. Springer Berlin Heidelberg, pp.41–57.

Patrick, D.L. and Erickson, P., 1993. Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation. In: *Health Care Evaluation and Resource Allocation*. New York: Oxford University Press.

Perkins, M.R.V., Devlin, N.J. and Hansen, P., 2004. The validity and reliability of EQ-5D health state valuations in a survey of Māori. *Quality of Life Research*, 13(1), pp.271–274. 10/bqjs99.

Rabin, R., Gudex, C., Selai, C. and Herdman, M., 2014. From translation to version management: a history and review of methods for the cultural adaptation of the EuroQol five-dimensional questionnaire. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 17(1), pp.70–76. 10/f2w78k.

Rammstedt, B., Danner, D. and Bosnjak, M., 2017. Acquiescence response styles: A multilevel model explaining individual-level and country-level differences. *Personality and Individual Differences*, 107, pp.190–194. 10/f9pgvd.

Reeve, B.B., Wyrwich, K.W., Wu, A.W., Velikova, G., Terwee, C.B., Snyder, C.F., Schwartz, C., Revicki, D.A., Moinpour, C.M., McLeod, L.D., Lyons, J.C., Lenderking, W.R., Hinds, P.S., Hays, R.D., Greenhalgh, J., Gershon, R., Feeny, D., Fayers, P.M., Cella, D., Brundage, M., Ahmed, S., Aaronson, N.K. and Butt, Z., 2013. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 22(8), pp.1889–1905. 10/f5g36w.

Regnault, A., Hamel, J.-F. and Patrick, D.L., 2015. Pooling of cross-cultural PRO data in multinational clinical trials: how much can poor measurement affect statistical power? *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 24(2), pp.273–277. 10/f6x23n.

Regnault, A. and Herdman, M., 2015. Using quantitative methods within the Universalist model framework to explore the cross-cultural equivalence of patient-reported outcome instruments. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 24(1), pp.115–124. 10/f6vxr9.

Salomon, J.A., Patel, A., Neal, B., Glasziou, P., Grobbee, D.E., Chalmers, J. and Clarke, P.M., 2011. Comparability of patient-reported health status: multicountry analysis of EQ-5D responses in patients with type 2 diabetes. *Medical Care*, 49(10), pp.962–970. 10/fpcvcm.

Salomon, J.A., Tandon, A. and Murray, C.J.L., 2004. Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ*, 328(7434), p.258. 10/btqfq7.

Sen, A., 2002. Health: perception versus observation: Self reported morbidity has severe limitations and can be extremely misleading. *BMJ*, 324(7342), pp.860–861. 10/csz8q2.

Sen, A., 2017. *Collective choice and social welfare*. Expanded edition ed. London: Penguin Books.

Shenoy, P., 2016. Multi-regional clinical trials and global drug development. *Perspectives in clinical research*, 7(2), pp.62–67. 10/gfxhzj.

Subramanian, S., Huijts, T. and Avendano, M., 2010. Self-reported health assessments in the 2002 World Health Survey: how do they correlate with education? *Bulletin of the World Health Organization*, 88(2), pp.131–138. 10/bkbr7z.

Szende, A., Janssen, B. and Cabases, J. eds., 2014. *Self-Reported Population Health: An International Perspective based on EQ-5D*. [online] Dordrecht: Springer. Available at: http://www.ncbi.nlm.nih.gov/books/NBK500356/ [Accessed 27 Mar. 2019].

U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research and U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health, 2006. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes*, 4, p.79. 10.1186/1477-7525-4-79.

Van Reenen, M. and Janssen, B., 2015. *EQ-5D-5L User Guide: Basic Information on How to Use the EQ-5D-5L Instrument – Version 2.1. Study guide.* Available at: https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf. [Accessed 13 Dec. 2018].

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A. and Smith, J.P., 2011. Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjctive Questions. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 174(3), pp.575–595. 10/fft8h5.

Velikova, G., Coens, C., Efficace, F., Greimel, E., Groenvold, M., Johnson, C., Singer, S., van de Poll-Franse, L., Young, T. and Bottomley, A., 2012. Health-Related Quality of Life in EORTC clinical trials — 30 years of progress from methodological developments to making a real impact on oncology practice. *European Journal of Cancer Supplements*, 10(1), pp.141–149. 10/gfxhzh.

Vivat, B., Young, T., Efficace, F., Sigurđadóttir, V., Arraras, J.I., Åsgeirsdóttir, G.H., Brédart, A., Costantini, A., Kobayashi, K., Singer, S. and EORTC Quality of Life Group, 2013. Cross-cultural development of the EORTC QLQ-SWB36: a stand-alone measure of spiritual wellbeing for palliative care patients with cancer. *Palliative Medicine*, 27(5), pp.457–469. 10/f4wk5b.

Walters, S.J. and Brazier, J.E., 2005. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 14(6), pp.1523–1532.

Whynes, D.K., Sprigg, N., Selby, J., Berge, E., Bath, P.M. and ENOS Investigators, 2013. Testing for differential item functioning within the EQ-5D. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 33(2), pp.252–260. 10/f4pz5w.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., Erikson, P. and ISPOR Task Force for Translation and Cultural Adaptation, 2005. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 8(2), pp.94–104. 10/c2dxrt.