"Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project."

Tennant, Alan ; Penta, Massimo ; Tesio, Luigi ; Grimby, Gunnar ; Thonnard, Jean-Louis ; Slade, Anita ; Lawton, Gemma ; Simone, Anna ; Carter, Jane ; Lundgren-Nilsson, Asa ; Tripolski, Maria ; Ring, Haim ; Biering-Sørensen, Fin ; Marinek, Crt ; Burger, Helena ; Phillips, Suzanne

ABSTRACT

INTRODUCTION: In Europe it is common for outcome measures to be translated for use in other languages. This adaptation may be complicated by culturally specific approaches to certain tasks; for example, bathing. In this context the issue of cross-cultural validity becomes paramount. OBJECTIVE: To facilitate the pooling of data in international studies, a project set out to evaluate the cross-cultural validity of impairment and activity limitation measures used in rehabilitation from the perspective of the Rasch measurement model. METHODS: Cross-cultural validity is assessed through an analysis of Differential Item Functioning (DIF) within the context of additive conjoint measurement expressed through the Rasch model. Data from patients undergoing rehabilitation for stroke was provided from 62 centers across Europe. Two commonly used outcome measures, the Mini-Mental State Examination (MMSE) and the Functional Independence Measure (FIM) motor scale are used to illustrate the approach....

CITE THIS VERSION

Tennant, Alan ; Penta, Massimo ; Tesio, Luigi ; Grimby, Gunnar ; Thonnard, Jean-Louis ; et. al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. In: Medical care, Vol. 42, no. 1 Suppl, p. i37-48 (2004) http://hdl.handle.net/2078.1/28110 -- DOI : 10.1097/01.mlr.0000103529.63132.77

Le dépôt institutionnel DIAL est destiné au dépôt et à la diffusion de documents scientifiques émanants des membres de l'UCLouvain. Toute utilisation de ce document à des fin lucratives ou commerciales est strictement interdite. L'utilisateur s'engage à respecter les droits d'auteur lié à ce document, principalement le droit à l'intégrité de l'oeuvre et le droit à la paternité. La politique complète de copyright est disponible sur la page Copyright policy

DIAL is an institutional repository for the deposit and dissemination of scientific documents from UCLouvain members. Usage of this document for profit or commercial purposes is strictly prohibited. User agrees to respect copyright about this document, mainly text integrity and source mention. Full content of copyright policy is available at Copyright policy
Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales Through Differential Item Functioning Within the Framework of the Rasch Model

The PRO-ESOR Project

Alan Tennant,* Massimo Penta,† Luigi Tesio,‡ Gunnar Grimby,§ Jean-Louis Thonnard,† Anita Slade,* Gemma Lawton,* Anna Simone,‡ Jane Carter,|| Asa Lundgren-Nilsson,§ Maria Tripolski,§ Haim Ring,¶ Fin Biering-Sørensen,** Črt Marincek,‖ Helena Burger, ‖ Helena Burger, ‖ ‖ and Suzanne Phillips||

Introduction: In Europe it is common for outcome measures to be translated for use in other languages. This adaptation may be complicated by culturally specific approaches to certain tasks; for example, bathing. In this context the issue of cross-cultural validity becomes paramount.

Objective: To facilitate the pooling of data in international studies, a project was set out to evaluate the cross-cultural validity of impairment and activity limitation scales used in rehabilitation from the perspective of the Rasch measurement model.

Methods: Cross-cultural validity is assessed through an analysis of Differential Item Functioning (DIF) within the context of additive conjoint measurement expressed through the Rasch model. Data from patients undergoing rehabilitation for stroke was provided from 62 centers across Europe. Two commonly used outcome measures, the Mini-Mental State Examination (MMSE) and the Functional Independence Measure (FIM) motor scale are used to illustrate the approach.

Results: Pooled data from 3 countries for the MMSE were shown to fit the Rasch model with only 1 item displaying DIF by country. In contrast, many items from the FIM expressed DIF and misfit to the model. Consequently they were allowed to be unique across countries, so resolving the lack of fit to the model.

Conclusions: Where data are to be pooled for international studies, analysis of DIF by culture is essential. Where DIF is observed, adjustments can be made to allow for cultural differences in outcome measurement.

Key Words: Rasch, DIF, outcomes, cross-cultural, rehabilitation
(Med Care 2004;42: 1-37-1-48)

Throughout Europe, a wide range of therapeutic interventions is used within rehabilitation, often country-specific and reflecting the different traditions of health care. To match this diversity of treatment modalities, there is also a broad range of outcome measures used to evaluate the efficacy of treatment. These measures focus mostly on aspects of impairment and activity limitation (disability) as defined by the International Classification of Functioning, Disability and Health (ICF). Their use, which attempts to quantify the results of rehabilitation, is seen as an increasingly important part of good clinical practice. However, if we wish to identify the most effective and efficient treatment modalities at the European level, it follows that we require outcome measures that serve this purpose.

Some attempt at standardization of outcome measurement in rehabilitation has been made in North America. The Functional Independence Measure (FIM), a measure of disability, is used across a wide range of conditions and in a wide range of situations in rehabilitation. This scale has also

From the *Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds, 36 Clarendon Road, Leeds, LS2 9NZ, United Kingdom; †Unité de Réadaptation et de Médecine Physique, Université Catholique de Louvain, B-1200 Brussels, Belgium; ‡Divisione di Recupero e Rieducazione Funzionale, Salvatore Maugeri Foundation, IRCCS, Pavia, Italy; §Department of Rehabilitation Medicine, Sahlgrenska Academy at Göteborg University, Sweden; ||The Bath Head Injury/Neuro-Rehabilitation Unit, Royal National Hospital for Rheumatic Diseases, National Health Services Trust, Bath, United Kingdom; and ¶Loewenstein Hospital, Rehabilitation Centre, Raanan, Tel Aviv University, School of Medicine, Israel.

Reprints: Alan Tennant BA, PhD, Professor of Rehabilitation Studies, Academic Unit of Musculoskeletal & Rehabilitation Medicine, The University of Leeds, 36 Clarendon Road, Leeds LS2 9NZ, UK. E-mail: a.tennant@leeds.ac.uk.

Copyright © 2003 by Lippincott Williams & Wilkins
ISSN: 0025-7079/04/4200-0037
DOI: 10.1097/01.mlr.0000103529.63132.77

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
been widely adopted throughout Europe. Despite this, across Europe, each diagnostic group tends to have its own set of commonly used outcome measures. For example, for the diagnosis of stroke, as well as the FIM, measures such as the Barthel Index and the Mini-Mental State Examination (MMSE) are also widely used.

Once outcome measures are introduced, it is common for them to be translated for use in other languages. In this context, the issue of cross-cultural validity becomes paramount. Although guidelines, both old and new, exist for adapting instruments into different languages, these tend to focus on self-complete instruments, whereas those measures in use in routine clinical practice tend to involve assessment of impairment and disability by professionals. Although in principle the adaptation procedures should be similar, it is often unclear as to exactly how (and by whom) these clinical measures were first translated. For purposes of comparing the rehabilitation services across Europe, it thus becomes essential to determine whether or not different language versions of such measures do work in the same way.

The objective of this paper is to demonstrate the use of Differential Item Functioning (DIF) as a mechanism to evaluate the cross-cultural validity of outcome measures used in rehabilitation and so facilitate the pooling of data in international studies. The approach is set within the framework of the Rasch unidimensional measurement model. Within Item Response Theory [IRT] this is known as the 1-parameter model. IRT is a general statistical theory about item (question or task) and scale performance and how that performance relates to the abilities measured by the items in the scale. Although early work in IRT was contiguous with that of Rasch (and it was known as Latent Trait Theory in the 1950s) it was Lord and Novick’s 1968 publication of their seminal work on Statistical Theories of Mental Test Scores, and particularly the contribution of Birnbaum that formally established IRT. Today, IRT encompasses many different types of parametric and nonparametric models, each with various extensions of the original dichotomous model. However, the Rasch model has unique properties which are crucial to attaining additive conjoint measurement, a prerequisite for the calculation of change scores, effect sizes and drawing correct inferences from various statistical tests which are common in health outcome studies.

The Rasch model is a unidimensional model which asserts that the easier the item the more likely it will be passed, and the more able the person, the more likely they will pass an item compared with a less able person. The dichotomous Rasch model assumes that the probability of a given respondent to give a “correct” answer to a particular item is a logistic function of the relative distance between the item location and the respondent location on a linear scale. In other words the probability that a person will affirm an item is a logistic function of the difference between the person’s ability (θ) and the difficulty of the item (b) (ie, the ability required to affirm item i), and only a function of that difference.

\[ P_{ni} = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \]  

where \( P_{ni} \) is the probability that person \( n \) will answer item \( i \) correctly (or be able to do the task specified by that item), \( \theta \) is person ability, and \( b \) is the item difficulty parameter.

From this, the expected pattern of responses to an item set is determined given the estimated \( \theta \) and \( b \). When the observed response pattern coincides with or does not deviate too much from the expected response pattern, then the items constitute a true Rasch scale. Taken with confirmation of local independence of items, that is, no residual associations in the data after the Rasch trait has been removed, this supports unidimensionality.

The model can be extended to cope with items with more than 2 categories and this involves an explicit “threshold” parameter (τ), where the threshold represents the equal probability point between any 2 adjacent categories within an item. In the logit form this is:

\[ \ln \left( \frac{P_{nk}}{1 - P_{nk}} \right) = \theta_k - b_k - \tau_k \]  

where \( P \) is the probability of person \( n \) affirming category \( k \) in item \( i \); \( \theta \) is person ability, \( b \) is the item difficulty parameter, and \( \tau \) is the difficulty of the \( k \) threshold. The model used in the present analysis is a further derivation, the Partial Credit Model:

\[ \ln \left( \frac{P_{nk}}{1 - P_{nk}} \right) = \theta_k - b_k \]  

where no assumptions are made about the equality of threshold locations relative to each item.

It is easy to see how the approach was readily adopted in rehabilitation in the late 1980s. Patients undergoing rehabilitation have a given ability level, and they are presented with a range of tasks with differing degrees of difficulty. The language of ability and difficulty easily transferred from education to rehabilitation. From this early work the approach quickly moved into the mainstream of health status measurement. The early published work on Rasch analysis in rehabilitation explored issues of unidimensionality and scaling properties and this has remained a central theme to date.

However, the approach can also contribute to the issue of cross-cultural validity. Essentially the scale should work in the same way, irrespective of which group is assessed. Thus, the location of items along the measurement construct should
remain the same across cultures. Originally, tests for such
differences came under the rubric “item bias.” Here, individ-
uals with the same score on a test are expected to have the
same probability of affirming an item, irrespective of any
group membership. This type of analysis has latterly been
given the name DIF. The basis of the DIF approach lies in
the item response logistic function, the proportion of individ-
uals at the same ability level who answer a given item
correctly (or can do a particular task). Under the requirement
that the ability under consideration is unidimensional, if the
item measures the same ability across groups then, except for
random variations, the same proportion is found irrespective
of the nature of the group for whom a function is plotted.
Items that do not yield the same item response function for 2
or more groups display DIF and are violating the requirement
of unidimensionality. Consequently, it is possible to exam-
ine whether or not a scale works in the same way by
contrasting the response function for each item across cul-
tures. It was thus the potential of the IRT approach, and
specifically the application of the Rasch model with its
attendant measurement properties, which led to the project to
examine cross-cultural validity, funded under the European
Commission’s BIOMED2 program. This project was the
European Standardization of Outcome Measurement in Phys-
cal Medicine and Rehabilitation which, as with all European
funded projects, is given a mnemonic, in this case PRO-
ESOR.

MATERIALS AND METHODS

Rasch Analysis
Thus the project sought to explore the potential of the
IRT Rasch model for evaluating the cross-cultural validity of
outcome scales used in Rehabilitation in Europe. A survey of
the use of outcome measures across Europe (1) identified a
set of measures for evaluation. Two scales identified in this
survey have been chosen to illustrate the approach, both
commonly used with patients who have had a stroke. These
are the MMSE8 and the FIM motor subscale. The former is
a measure of cognitive impairment and the latter, a measure
of activity limitation, looking at the independence in activi-
ties of daily living.

The approach, utilizing the Rasch model described above,
follows a sequence of analyses. All analysis is under-
taken on pooled data with age, gender, and country entered as
“person factors” for subsequent DIF analysis. In the first
instance, where data are polytomous, an analysis is under-
taken of the ordering of each category. An “ordered category”
implies that the response categories (eg, 1, 2, 3...) reflect an
increasing amount of the latent variable under investigation.
The boundaries between categories, as we have seen above,
are called “thresholds” and it is possible to observe “disor-
dered thresholds” in the data, where, for example, with an
increasing category number representing higher disability,
the threshold between categories 2 and 3 is found to represent
a lower level of disability than the threshold between cate-
gories 1 and 2. Where this occurs, it will be necessary to
collapse adjacent categories which can be undertaken as part
of the ongoing Rasch analysis.

Following this, the data are then (re)fit to the Rasch
model to determine overall fit, and how well each item fits the
model. These statistics indicate how far the observed data
match that expected by the model. Three overall fit statistics
are considered. Two are item-person interaction statistics
distributed as a Z statistic with mean of zero and SD of 1
(which indicates perfect fit to the model). A third is an
item-trait interaction statistic reported as a $\chi^2$, reflecting
the property of invariance across the trait. This means that
the hierarchical ordering of the items remains the same at differ-
ent levels of the underlying trait, indicated by a nonsignificant
$\chi^2$. In addition, individual item-fit statistics are presented,
both as residuals (a summation of individual person and item
deviations—usually acceptable within the range $\pm 2$) and as
a $\chi^2$ statistic (deviation from the model by groups of people
defined by their ability level—requiring a nonsignificant $\chi^2$;
 ie, a $P$ value of 0.05 and above, with appropriate adjustment
for repeated tests). Misfit of items indicates a lack of the
expected probabilistic relationship between the item and
other items in the scale. This may derive from DIF (see
below) or may indicate that the item does not contribute to the
trait under consideration.

Where data fit the Rasch model, an estimate of both
person ability and item difficulty is provided on the same
linear scale. The logit scale is an interval scale, where each
unit represents an increase in the odds of a person succeeding
on an activity by 2.716 times. It is customary to set the mean
difficulty of items as 0.0 logits to solve the indeterminacy of
the scale. The items are situated along the interval scale
according to their difficulty. The average person ability and
spread (ie, the SD) will indicate how well the scale is targeted
at the sample, with less well-targeted scales having an aver-
age person ability well away from the central zero logit. A
person separation index can be used to compute the number
of discernible strata in the distribution of patient measures.48
The index gives a practical indication of how precisely
patients have been spread out along the measurement con-
struct defined by the items.

As we have seen, within the framework of Rasch
measurement, the scale should work in the same way, irre-
spective of which group is being assessed. Thus, in the case
of disability, the probability of a person affirming an item (or
category) at a given level of disability should be the same for
younger or older people, men and women, and so on. Items
that do not yield the same item response function for 2 or
more groups display DIF and are violating the requirement
of unidimensionality. Consequently, every item is checked for
DIF by age and gender and, for the current analysis, by country.

Briefly, for the response of each person to each item, the standardized residual $Z_{mi}$ of the observed score $X_{mi}$ from that predicted by the model, $E[X_{mi}]$ is calculated according to

$$Z_{mi} = \frac{X_{mi} - E[X_{mi}]}{\sqrt{V[X_{mi}]}}$$  \hspace{1cm} (4)

Then each person is classified according to one of G class intervals and for example, in the case of gender, giving a set of residuals suitable for a 2-way analysis of variance (ANOVA) design.\(^{49}\) Thus, the statistical test used for detecting DIF is an ANOVA of the person-item deviation residuals with person factors (eg, gender) and class intervals (eg, Group along the trait) as factors. Two types of DIF can be identified: uniform and nonuniform DIF. With the former, there is a constant difference between groups in the probability of affirming an item (or category) across the trait (ANOVA main effect), and with the latter the difference varies across the trait (ANOVA interaction effect).

Where some but not all items display DIF, adjustments can be made to allow items with DIF to vary by group. The approach adopted has been described as an iterative “top-down purification” approach in that a requirement for identifying DIF is a baseline set of “pure” items.\(^{50}\) Consequently, a biased item with the poorest fit to the Rasch model is removed first and the procedure repeated until an unbiased and scalable subset of items is identified.\(^{51}\) The rejected items are then reintroduced to test the results. When an item displays DIF, it is rendered unique to the group(s) which display DIF. For example, suppose that Italy displayed DIF for a bathing item, compared with England and Sweden (post hoc tests for the ANOVA identify where the difference(s) lie). In this case, the item would be split into 2, with 1 item for Italy (and the responses for patients from England and Sweden entered as structural missing values) and 1 item for England and Sweden together (with the responses for Italian patients entered as structural missing values). Consequently, there would now be 2 bathing items, an Italian item and an Anglo-Swedish item. The unsplit (pure) items act as links in the calibration. In this way, item difficulty for some items is allowed to vary across countries. Fit is again reassessed, and if items still display misfit to the model, then they are removed from the scale and do not contribute to the person estimate.

Finally, person-item deviation residuals are examined by principal components analysis (PCA) for associations which may be indicative of the breach of the assumptions of local independence. The absence of such associations, taken with adequate fit to the Rasch model, support unidimensionality and indicate that the attributes of conjoint additive measurement are retained in the data.

Sample Size and Statistical Software

Sixty-two rehabilitation facilities contributed anonymous data from consecutively admitted patients. It was known that a sample size of 150 patients would estimate the item difficulty, with $\alpha$ of 0.01, to within $\pm 0.5$ logits.\(^{52}\) This sample size is also sufficient to test for DIF where (a) at $\alpha$ of 0.01 a difference of 0.5 SD within the residuals can be detected for any 2 groups with $\beta$ of 0.20; or (b) with identical $\alpha$ and $\beta$, a difference in SDs within the residuals can be detected where this difference varies by 0.1 logits across each of 6 groups.\(^{53}\) Bonferroni corrections are applied to both fit and DIF statistics due to the number of tests undertaken for any given scale.\(^{54}\)

In any event, 3 countries, Italy ($n = 108$), Slovenia ($n = 78$) and Sweden ($n = 149$) contributed 335 cases for the analysis of the MMSE, and data were received from 6 countries (Belgium 143, France 157, Italy 1046, Israel 319, Sweden 642, and UK 239) for the FIM motor scale. Due to the unequal sample size for the latter scale, a random sample of cases was taken from Italy, Sweden, and Israel to provide approximately 150 cases from each country for the analysis. This was necessary to conform to equal group size requirements for the ANOVA of residuals.\(^{55}\)

Rasch analysis was undertaken using the RUMM2010 package.\(^{56}\)

The Outcome Measures

Briefly, the FIM motor scale comprises 13 items with a 1–7 category response. Items are tasks, ranging from bladder management, to climbing stairs. A category response of 1 reflects total dependence, and 7, total independence. Thus a total score can range from 13 to 91. The MMSE covers 6 domains, including orientation and recall. The domains have a variable number of items, and the items have a different number of categories (ranging from dichotomous to a 6-category response). In practice, 11 items are summed with a maximum score of 30. A low score reflects cognitive dysfunction, with a score of 23 considered nonnormal; 24 and above, normal.

RESULTS

The MMSE

Overall, 15 items were found to fit the model (Table 1; Item Fit Mean $-0.038$, SD 1.300) and the Person Separation Index was satisfactory (0.899). However, Item-Trait Interaction showed marginal significance, indicating a variable hierarchical ordering of items across the trait ($\chi^2$ (df = 33) 58.203, $P = 0.004357$). Also, 5 items displayed disordered thresholds (Fig. 1). For example, while the item “disorientation in time” displayed ordered thresholds, in that each category demonstrates an increasing level of the trait of

I-40

© 2003 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
TABLE 1. Fit of MMSE to Rasch model

<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>SE</th>
<th>Residual</th>
<th>( \chi^2 )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation time</td>
<td>0.61</td>
<td>0.07</td>
<td>0.28</td>
<td>0.24</td>
<td>0.6221</td>
</tr>
<tr>
<td>Orientation place</td>
<td>-0.09</td>
<td>0.08</td>
<td>-1.76</td>
<td>1.15</td>
<td>0.2843</td>
</tr>
<tr>
<td>Registration</td>
<td>-1.39</td>
<td>0.18</td>
<td>-0.40</td>
<td>0.81</td>
<td>0.3670</td>
</tr>
<tr>
<td>Attention and calculation</td>
<td>1.14</td>
<td>0.11</td>
<td>-0.41</td>
<td>0.81</td>
<td>0.3686</td>
</tr>
<tr>
<td>Recall</td>
<td>0.44</td>
<td>0.16</td>
<td>0.99</td>
<td>0.33</td>
<td>0.5669</td>
</tr>
<tr>
<td>Language naming</td>
<td>-3.09</td>
<td>0.44</td>
<td>-1.46</td>
<td>1.22</td>
<td>0.2690</td>
</tr>
<tr>
<td>Language repeating</td>
<td>0.40</td>
<td>0.16</td>
<td>1.08</td>
<td>1.29</td>
<td>0.2567</td>
</tr>
<tr>
<td>Language command—verbal</td>
<td>-0.48</td>
<td>0.14</td>
<td>0.56</td>
<td>0.14</td>
<td>0.7042</td>
</tr>
<tr>
<td>Language command—written</td>
<td>-0.24</td>
<td>0.18</td>
<td>-1.27</td>
<td>3.04</td>
<td>0.0811</td>
</tr>
<tr>
<td>Language write</td>
<td>0.58</td>
<td>0.16</td>
<td>0.41</td>
<td>0.04</td>
<td>0.8406</td>
</tr>
<tr>
<td>Copying</td>
<td>2.13</td>
<td>0.14</td>
<td>1.61</td>
<td>2.72</td>
<td>0.0988</td>
</tr>
</tbody>
</table>

mental state, the item “registration” failed to display an ordering over the trait and could thus not be presented in a graphical fashion as the categories are out of order.

These items were rescored to see if fit to the model was improved. During rescorer, 2 items had to be reduced to dichotomous responses. These were “Language—naming” and “Recall”. This produced an improvement in overall fit to model (Item Fit Mean = -0.034, SD 1.113; Person Fit Mean = -0.144, SD 0.464) and all individual item fit statistics were satisfactory, both for residuals and \( \chi^2 \) (Table 2). The Item-Trait Interaction statistic became nonsignificant (\( \chi^2 \) (11) 11.799, \( P = 0.37862 \)), demonstrating invariance across the trait. The Person Separation Index remained satisfactory (0.897), indicating the scale was able to differentiate almost 4 separate groups of patients across the scale.\(^8\)

Although the overall fit of the data to the model was good, 2 items were found to display uniform DIF by age, and 1 item displayed nonuniform DIF by country. The “language—naming” and “language—copying” displayed uniform DIF by age. The “language command—written” item displayed DIF by country, and the comparative response functions for this item are shown in Figure 2. A Bonferroni corrected probability of 0.0045 would indicate a significant difference in these circumstances.

In Figure 2, the model expectation curve for the item is given, and also for each country. In practice, for each country,

---

**FIGURE 1.** Threshold ordering of MMSE.

© 2003 Lippincott Williams & Wilkins
TABLE 2. Fit of MMSE (adjusted for DIF) to Rasch model

<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>SE</th>
<th>Residual</th>
<th>$\chi^2$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation time</td>
<td>0.58</td>
<td>0.07</td>
<td>0.31</td>
<td>3.49</td>
<td>0.4793</td>
</tr>
<tr>
<td>Orientation place</td>
<td>-0.13</td>
<td>0.08</td>
<td>-1.76</td>
<td>4.60</td>
<td>0.3313</td>
</tr>
<tr>
<td>Registration</td>
<td>-1.42</td>
<td>0.18</td>
<td>-0.41</td>
<td>2.87</td>
<td>0.5798</td>
</tr>
<tr>
<td>Attention and calculation</td>
<td>1.11</td>
<td>0.11</td>
<td>-0.36</td>
<td>3.27</td>
<td>0.5143</td>
</tr>
<tr>
<td>Recall</td>
<td>0.42</td>
<td>0.16</td>
<td>0.98</td>
<td>3.50</td>
<td>0.4778</td>
</tr>
<tr>
<td>Language naming</td>
<td>-3.12</td>
<td>0.44</td>
<td>-1.44</td>
<td>3.99</td>
<td>0.4069</td>
</tr>
<tr>
<td>Language repeating</td>
<td>0.37</td>
<td>0.16</td>
<td>1.09</td>
<td>7.03</td>
<td>0.1344</td>
</tr>
<tr>
<td>Language command—verbal</td>
<td>-0.52</td>
<td>0.14</td>
<td>0.56</td>
<td>2.96</td>
<td>0.5640</td>
</tr>
<tr>
<td>Language command—written</td>
<td>-0.56</td>
<td>0.22</td>
<td>-1.34</td>
<td>3.83</td>
<td>0.4294</td>
</tr>
<tr>
<td>Language command—written, Slovenia</td>
<td>0.63</td>
<td>0.34</td>
<td>-0.86</td>
<td>2.41</td>
<td>0.6610</td>
</tr>
<tr>
<td>Language write</td>
<td>0.55</td>
<td>0.16</td>
<td>0.46</td>
<td>1.80</td>
<td>0.7726</td>
</tr>
<tr>
<td>Copying</td>
<td>2.10</td>
<td>0.14</td>
<td>1.63</td>
<td>2.21</td>
<td>0.6973</td>
</tr>
</tbody>
</table>

The expected score is given for 2 groups of people at different levels of the trait (the number of groups is determined by sample size). It becomes obvious from Figure 2 (as well as post hoc tests) that the difference in response function can be attributed to the Slovenia data. Consequently, the item was split in 2; 1 for Italy and Sweden together and 1 for Slovenia. Following this, fit of the resulting 12-item scale to the model was found to be good (Item Fit Mean −0.095; SD 1.099; Person Fit Mean −0.146; SD 0.462; Item-Trait Interaction $\chi^2$ (48) 41.957, $P = 0.717653$; Table 3), and DIF for country was absent. PCA analysis of residuals showed no discernable pattern, with the first factor taking 13% of the variation among the residuals, so supporting unidimensionality.

The FIM Motor Subscale

Overall fit to the model of the FIM motor scale was poor (Item Fit Mean −0.360, SD 4.642). The Item-Trait Interaction statistic was significant ($\chi^2$ (117) 504.309, $P < 0.0001$) showing a lack of invariance across the trait. Individual item fit is shown in Table 3.

Only 5 of the 13 items were found to have ordered thresholds. Consequently, 8 disordered items had to be re-scored, 3 of which had to be dichotomized (Fig. 3).

Following recoring overall fit to the model remained poor (Item Fit Mean −0.444, SD 2.450) and the Item-Trait Interaction Statistic remained significant ($\chi^2$ (117) 229.288, $P < 0.0001$) although the Person Separation Index was satisfactory (0.969). DIF was tested with respect to age, gender, and country. No items were found to have DIF by gender or age, but 8 items were found to have DIF by Country (Table 4). Consequently, the items that displayed DIF were split and made unique across Countries. A new analysis was run on the resulting 53 items (5 original “pure” items and 8 split across 6 countries). Items with disordered thresholds were once again

[Graph showing comparative response functions for the item “language command—written” in the MMSE.]

FIGURE 2. Comparative response functions for the item “language command—written” in the MMSE.
TABLE 3. FIM individual item fit

<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>SE</th>
<th>Residual</th>
<th>$\chi^2$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eating</td>
<td>-1.11</td>
<td>0.04</td>
<td>6.39</td>
<td>80.46</td>
<td>0.0000</td>
</tr>
<tr>
<td>Grooming</td>
<td>-0.41</td>
<td>0.03</td>
<td>0.34</td>
<td>8.95</td>
<td>0.4423</td>
</tr>
<tr>
<td>Bathing</td>
<td>0.39</td>
<td>0.03</td>
<td>-3.37</td>
<td>27.06</td>
<td>0.0014</td>
</tr>
<tr>
<td>Dressing—upper body</td>
<td>0.03</td>
<td>0.03</td>
<td>-0.87</td>
<td>23.67</td>
<td>0.0049</td>
</tr>
<tr>
<td>Dressing—lower body</td>
<td>0.47</td>
<td>0.03</td>
<td>-6.00</td>
<td>41.53</td>
<td>0.0000</td>
</tr>
<tr>
<td>Toileting</td>
<td>0.11</td>
<td>0.03</td>
<td>-4.14</td>
<td>31.97</td>
<td>0.0002</td>
</tr>
<tr>
<td>Bladder management</td>
<td>-0.64</td>
<td>0.03</td>
<td>5.74</td>
<td>49.70</td>
<td>0.0000</td>
</tr>
<tr>
<td>Bowel management</td>
<td>-0.88</td>
<td>0.03</td>
<td>4.70</td>
<td>42.88</td>
<td>0.0000</td>
</tr>
<tr>
<td>Transfer bed</td>
<td>-0.15</td>
<td>0.03</td>
<td>-5.79</td>
<td>39.26</td>
<td>0.0000</td>
</tr>
<tr>
<td>Transfer toilet</td>
<td>-0.04</td>
<td>0.03</td>
<td>-7.39</td>
<td>57.19</td>
<td>0.0000</td>
</tr>
<tr>
<td>Transfer tub</td>
<td>0.80</td>
<td>0.03</td>
<td>1.88</td>
<td>24.29</td>
<td>0.0039</td>
</tr>
<tr>
<td>Walk/wheelchair</td>
<td>0.24</td>
<td>0.03</td>
<td>1.57</td>
<td>14.17</td>
<td>0.1164</td>
</tr>
<tr>
<td>Stairs</td>
<td>1.19</td>
<td>0.03</td>
<td>2.27</td>
<td>63.20</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

*Bonferroni corrected significance level $P = 0.0008$.

rescored, and misfitting items were deleted, reducing the scale to 50 items (all 5 original items—Eating, Dressing—Upper Body, Bladder Management, Walk/Wheelchair and Stairs—to serve as “link” items, and 45 split items). The unique “transfer-toilet” items for Sweden and Israel were deleted, along with the “grooming” item for Italy. The final fit of this scale was good at the individual level, and at the Item Fit level (mean $-0.462$, SD 1.550), although the Item-Trait Interaction statistic remained borderline nonsignificant ($\chi^2 (50) = 89.889, P = 0.000462$). Final fit of the individual 50 items is shown in Table 5. Once again, PCA of residuals showed no discernable pattern among the data, with a first factor accounting for only 19% of the total variation.

DISCUSSION

A DIF approach within the framework of the Rasch measurement model offers a sophisticated way of identifying and, under certain circumstances, of adjusting for problems associated with cross-cultural differences in outcome scales.

![FIGURE 3. Threshold Map of FIM motor scale following rescorning.](image)

© 2003 Lippincott Williams & Wilkins
The 2 scales chosen for illustration in this paper, despite their widespread use as international standards, both demonstrated initial problems with basic measurement requirements. At a given level of the trait, the probability of response to various categories was shown to be different across Countries. Through adjusting for DIF, these problems were resolved for the MMSE, and largely for the FIM. Given this adjustment, unidimensionality was supported, although the solution for the FIM required the omission of 3 country-specific items.

Fit of the data to the Rasch model, including a PCA of the residuals to test the local independence assumption, supports the unidimensionality of the scale. Although factor analysis is a traditional approach to determine unidimensionality, there are considerable problems associated with using ordinal data for this type of analysis and, for example, spurious factors can be caused by DIF.57

MMSE scores have previously been found to be influenced by both culture and educational level58 and Teresi and colleagues59 have identified DIF for culture and education. The latter has not been considered in the present analysis but, particularly with an expanding Europe, may introduce another important source of DIF which will need to be considered. It has also been suggested that for cognitive screening tests, shorter tests composed of easier items which rely more on memory recall, rather than performance related to language and literacy, would perform better.60 The interaction between culture and educational level, and its effect upon DIF, remains unknown.

The FIM is widely used in rehabilitation throughout Europe and North America and poses particular challenges for cross-cultural studies. This is explicitly acknowledged by proponents of the FIM who have established formal procedures for translation and training across countries. It became clear during the analysis that the original scoring function of 7 categories is not working well. This is true both across and within countries, such as Italy, where formal training programs help therapists and others assign grades according to strict criteria. The rescoring undertaken in the analysis above suggests that a common scoring function across items may be difficult to achieve, in that some items had to be dichotomized. This indicates that some tasks are perceived by therapists as an “all or nothing” situation for patients, compared with other tasks which can be graded. However, it is possible that such a situation may vary by diagnostic group.

Many items displayed DIF by Country, but through making those items unique to each Country, a solution was found that included most items for most Countries. However, problems remain. The “grooming” item in Italy for example, would not fit the model as a Country-specific item. Analysis of the Italian data only (not shown) showed this item to lack fit to the model. Is this because the adaptation of the scale to the Italian culture brought about a fundamental change in its relationship with other items? The FIM motor scale works in other Countries, for example the UK, which suggests that the adaptation may have changed the item. We must also stress here that it is not the wording of the item that necessarily changes, but rather the way in which the task is evaluated in the clinical setting. Once again, different traditions across Europe may influence this, for example, in the way in which patients are bathed in a hospital setting.61 This may result in subtle differences in the degree of difficulty of the tasks.

However, other issues may be relevant here; for example, the nature of the underlying lesion causing the stroke (DIF by case mix, based on lesion type). If the case mix is not homogeneous across countries, this may cause the discrepancy and should be investigated before definitive statements are made about the quality of the scale. Even on homogeneous case mix, with an apparent paradox, differential training may be a source of DIF, as long as it introduces systematic differences in interpretation of the items across countries.

There are several methodological issues which are raised by proposing this approach as a standard for assessing and adjusting for the cross-cultural validity of outcome measures across Europe. The detection of DIF by ANOVA techniques is vulnerable to sample size issues. With large samples, DIF can usually be found. Therefore this issue becomes one of substantive difference as opposed to statistical significance, yet there is little literature to guide what might constitute a substantive difference. Sample size also affects interpretation of χ²-based fit statistics62 and, again, with large enough sample sizes, misfit to the model can always be found. In the current study, we took samples of patients from some countries which returned large numbers of cases, to satisfy the equality of group requirements in the ANOVA analysis. We know, in the case of the Rasch model, that certain sample sizes can deliver given levels of precision for the parameter esti-
<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>SE</th>
<th>Residual</th>
<th>$\chi^2$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eating</td>
<td>-5.11</td>
<td>0.19</td>
<td>0.09</td>
<td>5.53</td>
<td>0.0187</td>
</tr>
<tr>
<td>Grooming, UK</td>
<td>-0.41</td>
<td>0.12</td>
<td>3.28</td>
<td>4.07</td>
<td>0.0437</td>
</tr>
<tr>
<td>Grooming, Sweden</td>
<td>-1.16</td>
<td>0.14</td>
<td>-0.14</td>
<td>0.77</td>
<td>0.3810</td>
</tr>
<tr>
<td>Grooming, Israel</td>
<td>-0.40</td>
<td>0.11</td>
<td>0.15</td>
<td>2.10</td>
<td>0.1474</td>
</tr>
<tr>
<td>Grooming, France</td>
<td>-0.12</td>
<td>0.13</td>
<td>-0.37</td>
<td>0.15</td>
<td>0.7000</td>
</tr>
<tr>
<td>Grooming, Belgium</td>
<td>-1.16</td>
<td>0.09</td>
<td>1.66</td>
<td>1.00</td>
<td>0.3183</td>
</tr>
<tr>
<td>Bathing, UK</td>
<td>0.43</td>
<td>0.10</td>
<td>1.03</td>
<td>0.33</td>
<td>0.5664</td>
</tr>
<tr>
<td>Bathing, Sweden</td>
<td>0.80</td>
<td>0.11</td>
<td>0.57</td>
<td>0.20</td>
<td>0.6576</td>
</tr>
<tr>
<td>Bathing, Italy</td>
<td>0.39</td>
<td>0.10</td>
<td>-1.31</td>
<td>0.43</td>
<td>0.5141</td>
</tr>
<tr>
<td>Bathing, Israel</td>
<td>0.02</td>
<td>0.11</td>
<td>-1.36</td>
<td>0.02</td>
<td>0.8848</td>
</tr>
<tr>
<td>Bathing, France</td>
<td>0.74</td>
<td>0.18</td>
<td>-2.92</td>
<td>5.80</td>
<td>0.0160</td>
</tr>
<tr>
<td>Bathing, Belgium</td>
<td>0.20</td>
<td>0.11</td>
<td>-0.89</td>
<td>1.01</td>
<td>0.3144</td>
</tr>
<tr>
<td>Dressing—upper body</td>
<td>-0.10</td>
<td>0.04</td>
<td>2.80</td>
<td>2.62</td>
<td>0.1057</td>
</tr>
<tr>
<td>Dressing—lower body, UK</td>
<td>0.67</td>
<td>0.09</td>
<td>-1.23</td>
<td>0.56</td>
<td>0.4531</td>
</tr>
<tr>
<td>Dressing—lower body, Sweden</td>
<td>0.41</td>
<td>0.10</td>
<td>-1.58</td>
<td>0.26</td>
<td>0.6090</td>
</tr>
<tr>
<td>Dressing—lower body, Italy</td>
<td>1.00</td>
<td>0.10</td>
<td>-0.65</td>
<td>0.43</td>
<td>0.5102</td>
</tr>
<tr>
<td>Dressing—lower body, Israel</td>
<td>0.82</td>
<td>0.10</td>
<td>-2.17</td>
<td>0.32</td>
<td>0.5730</td>
</tr>
<tr>
<td>Dressing—lower body, France</td>
<td>0.16</td>
<td>0.14</td>
<td>-1.36</td>
<td>6.05</td>
<td>0.0139</td>
</tr>
<tr>
<td>Dressing—lower body, Belgium</td>
<td>0.86</td>
<td>0.10</td>
<td>-0.97</td>
<td>0.07</td>
<td>0.7971</td>
</tr>
<tr>
<td>Toileting, UK</td>
<td>0.10</td>
<td>0.19</td>
<td>-0.58</td>
<td>3.99</td>
<td>0.0458</td>
</tr>
<tr>
<td>Toileting, Sweden</td>
<td>-0.04</td>
<td>0.15</td>
<td>-1.69</td>
<td>1.38</td>
<td>0.2403</td>
</tr>
<tr>
<td>Toileting, Italy</td>
<td>1.00</td>
<td>0.19</td>
<td>-1.54</td>
<td>0.93</td>
<td>0.3349</td>
</tr>
<tr>
<td>Toileting, Israel</td>
<td>-0.02</td>
<td>0.17</td>
<td>-1.22</td>
<td>2.18</td>
<td>0.1399</td>
</tr>
<tr>
<td>Toileting, France</td>
<td>0.29</td>
<td>0.14</td>
<td>-3.17</td>
<td>1.82</td>
<td>0.1769</td>
</tr>
<tr>
<td>Toileting, Belgium</td>
<td>0.37</td>
<td>0.18</td>
<td>-1.27</td>
<td>2.49</td>
<td>0.1145</td>
</tr>
<tr>
<td>Bladder management</td>
<td>-1.25</td>
<td>0.07</td>
<td>2.49</td>
<td>5.06</td>
<td>0.0245</td>
</tr>
<tr>
<td>Bowel management, UK</td>
<td>-1.00</td>
<td>0.17</td>
<td>1.21</td>
<td>2.65</td>
<td>0.1034</td>
</tr>
<tr>
<td>Bowel management, Sweden</td>
<td>-1.53</td>
<td>0.21</td>
<td>2.88</td>
<td>2.68</td>
<td>0.1018</td>
</tr>
<tr>
<td>Bowel management, Italy</td>
<td>-1.42</td>
<td>0.17</td>
<td>0.30</td>
<td>0.54</td>
<td>0.4606</td>
</tr>
<tr>
<td>Bowel management, Israel</td>
<td>-1.64</td>
<td>0.18</td>
<td>0.45</td>
<td>0.82</td>
<td>0.3644</td>
</tr>
<tr>
<td>Bowel management, France</td>
<td>-2.49</td>
<td>0.17</td>
<td>1.41</td>
<td>1.50</td>
<td>0.2204</td>
</tr>
<tr>
<td>Bowel management, Belgium</td>
<td>-1.79</td>
<td>0.16</td>
<td>-0.56</td>
<td>1.22</td>
<td>0.2695</td>
</tr>
<tr>
<td>Transfer bed, UK</td>
<td>-0.08</td>
<td>0.14</td>
<td>-2.05</td>
<td>6.19</td>
<td>0.0129</td>
</tr>
<tr>
<td>Transfer bed, Sweden</td>
<td>-0.23</td>
<td>0.10</td>
<td>-2.22</td>
<td>0.16</td>
<td>0.6895</td>
</tr>
<tr>
<td>Transfer bed, Italy</td>
<td>-0.07</td>
<td>0.09</td>
<td>-0.52</td>
<td>0.57</td>
<td>0.4503</td>
</tr>
<tr>
<td>Transfer bed, Israel</td>
<td>-0.20</td>
<td>0.10</td>
<td>-2.63</td>
<td>0.15</td>
<td>0.6962</td>
</tr>
<tr>
<td>Transfer bed, France</td>
<td>-1.15</td>
<td>0.09</td>
<td>-0.29</td>
<td>0.47</td>
<td>0.4952</td>
</tr>
<tr>
<td>Transfer bed, Belgium</td>
<td>-0.42</td>
<td>0.10</td>
<td>-1.13</td>
<td>0.85</td>
<td>0.3558</td>
</tr>
<tr>
<td>Transfer toilet, UK</td>
<td>-0.20</td>
<td>0.12</td>
<td>-1.45</td>
<td>3.42</td>
<td>0.0644</td>
</tr>
<tr>
<td>Transfer toilet, Italy</td>
<td>0.91</td>
<td>0.16</td>
<td>-2.85</td>
<td>3.58</td>
<td>0.0584</td>
</tr>
<tr>
<td>Transfer toilet, France</td>
<td>-0.64</td>
<td>0.11</td>
<td>-2.47</td>
<td>0.22</td>
<td>0.6416</td>
</tr>
<tr>
<td>Transfer toilet, Belgium</td>
<td>-0.34</td>
<td>0.10</td>
<td>-2.13</td>
<td>0.63</td>
<td>0.4279</td>
</tr>
<tr>
<td>Transfer tub, UK</td>
<td>0.87</td>
<td>0.09</td>
<td>1.04</td>
<td>1.34</td>
<td>0.2468</td>
</tr>
<tr>
<td>Transfer tub, Sweden</td>
<td>1.03</td>
<td>0.14</td>
<td>-1.52</td>
<td>0.07</td>
<td>0.7928</td>
</tr>
<tr>
<td>Transfer tub, Italy</td>
<td>2.10</td>
<td>0.19</td>
<td>-0.08</td>
<td>0.39</td>
<td>0.5343</td>
</tr>
<tr>
<td>Transfer tub, France</td>
<td>2.95</td>
<td>0.27</td>
<td>0.39</td>
<td>8.16</td>
<td>0.0043</td>
</tr>
<tr>
<td>Transfer tub, Belgium</td>
<td>2.63</td>
<td>0.24</td>
<td>-0.24</td>
<td>0.21</td>
<td>0.6506</td>
</tr>
<tr>
<td>Walk/wheelchair</td>
<td>0.82</td>
<td>0.06</td>
<td>0.96</td>
<td>0.66</td>
<td>0.4165</td>
</tr>
<tr>
<td>Stairs</td>
<td>2.49</td>
<td>0.07</td>
<td>0.19</td>
<td>1.87</td>
<td>0.1716</td>
</tr>
</tbody>
</table>

*Benferroni corrected significance level $P = 0.0002$. 

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
mates. However, where patient databases are much larger, is it appropriate to take a sample so that the analysis can be done with known (and perhaps commonly agreed) levels of power and significance?

It has been argued that DIF is evaluated once the conditions for fit to the Rasch model have been satisfied. However, it has also been argued that DIF is a breach of the assumptions of unidimensionality. It is possible to have data which fit the model but which also display DIF. We have adopted the approach whereby we consider DIF as a violation to unidimensionality and thus one possible contribution to misfit, and thus deliberately refitted data to the model after adjustment for DIF. This adjustment involves a physical separation of the item for those countries displaying DIF and almost always results in an improvement of fit. An alternative approach would be to allow for that variation statistically by using the 2PL model (which adds a discrimination parameter). Unfortunately, the latter has problems with convergence when dealing with relatively small item sets, missing values, and small sample sizes typically found in medical outcome studies.

It is true that the approach used here, the 1-parameter Rasch model, places great demands upon the quality of measurement derived from the outcome measures. The Rasch measurement model is chosen because it conforms to the axioms of additive conjoint measurement and, given adequate fit to the model and local independence, can support the claim for an invariant unidimensional scale and provide a linear transformation of the ordinal data into interval measures. This class of measurement is a requirement for carrying out arithmetic operations such as the calculation of valid change scores. Otherwise, analysis should be restricted to appropriate nonparametric statistics. Some analysis has used 2-parameter models as more appropriate for examining DIF in already developed measures. However, in addition to the much larger sample size requirements for the 2PL and 3PL models (the latter which adds a guessing parameter), it is known that almost 100% of the time their parameters violate interval scaling. Thus, these models do not provide the quality of measurement which is required for most clinical outcome studies. Despite this, 2PL models have entered the mainstream of health status measurement.

While this paper has addressed the issue of cross-cultural validity from a DIF perspective within the Rasch measurement model framework, there are many other non-IRT approaches, all well catalogued by Angoff. A commonly used approach is the Mantel Haenzel procedure which estimates a common odds ratio across groups. Newly developed packages such as SIBTEST facilitate analysis of DIF using this method. However, it must be stressed that analysis of DIF within the Rasch framework is an analysis within a theoretical measurement framework which we believe to be crucial to fundamental measurement in health sciences.

For the future, developments based upon item banking, may bring the prospect of having different sets of items for different countries. As long as these items are calibrated on the same underlying metric, any subset of items may be chosen, based on Rasch’s notion of specific objectivity. The strategy of allowing items to be unique by country, used in the analysis above, essentially establishes an item bank post hoc. In the future, new scales will use the power of Rasch analysis to allow for cultural variation by combining unique and common items in the development phase of the instrument.

In conclusion, where data are to be pooled for international clinical trials, it is essential to test for, and if necessary to adjust for DIF. Rasch’s notion of specific objectivity means that, in practice, estimates of patient ability can be obtained when some (country-specific) items are omitted because of misfit to the model. We would argue that the requirements of outcome measurement in rehabilitation, specifically the calculation of change scores and other arithmetic operations on the data, demand that such analysis be undertaken within a framework that supports additive conjoint measurement. This can be operationalized, in a probabilistic framework, through the Rasch measurement model.

ACKNOWLEDGMENTS

This work was funded by the European Commission within its BIOMED 2 program under contract BMH4-CT98-3642.

We would like to thank the anonymous reviewers of this paper who made several valuable suggestions which contributed to our revision.

REFERENCES


