The AMC Linear Disability Score (ALDS) : measuring disability in clinical studies

Weisscher, N.

Citation for published version (APA):
Chapter 9

General discussion
could also include an item like ‘running a mile’, but one can doubt if these types of items still belong to the disability continuum and are clinically useful. Another interpretation could be that patients who are able to ride a bike for over two hours report no disability at all, which implies it is not a ceiling effect of the ALDS item bank but the ‘ceiling’ of the disability construct. Even though the item bank showed no floor effect in our studies, one can imagine there are easier items than ‘put on and take off a T-shirt’, therefore the lower limit of the scale will be extended in further calibration stages.

Linearity

An advantage of constructing item banks using IRT models is the placement of the items on a common hierarchical difficulty measure with linear measurement properties. This, in combination with the possibility to span the entire range of the disability continuum, produces a sensitive outcome scale to measure subtle functional health differences between patients and within patients in time. It should be mentioned that the central part of the scale is very well filled, with a high density of items. Items at the lower and upper end are more sparsely distributed with large gaps between adjacent items (Figure). This means that it is possible to estimate the functional status of patients in the middle of the scale very accurately, but that at both extremities of the scale less accurate estimates of disability status will be obtained.

Main findings

In Chapter 2 we examined the appropriateness of the disability concept in clinical efficacy studies. Subsequently, in Chapter 3, the ALDS was introduced and the psychometric properties of the items in a mixed population were presented. We concluded the ALDS item bank has promising measurement characteristics in this patient population, but further clinimetric studies in other populations are needed. The latter recommendation is not only important from a psychometric point of view, but also from the perspective of further implementing the ALDS in clinical research. Although measures constructed with modern psychometric methods theoretically are superior to traditional measures, it is essential to investigate this advantage and its implications in practice. Therefore, Chapters 4 to 6 focused on the measurement properties of the item bank in populations with rheumatoid arthritis, Parkinson’s disease and stroke. In addition, Chapter 7 examined the clinical meaning of dichotomizing an often used stroke outcome scale (reflecting ‘good’ versus ‘poor’ outcome), by calculating the patients’ average chance to perform certain ALDS items. Finally, if an item bank is to be used for various groups of patients, it is important to compare the metric behavior of the items in different groups of patients. In Chapter 8, we investigated the generic character of the ALDS item bank by comparing its measurement properties in subgroups of patients suffering from neurological diseases and diseases of internal medicine.

Floor & Ceiling effects

Most traditional instruments are either long and assess the whole disability continuum, or are short and assess only the lower end or the upper end of the continuum. There are also instruments which measure a wide range of the disability construct but these measures are less detailed. One goal of the ALDS project was to construct a detailed measure containing both basic activities of daily life as well as instrumental activities of daily life. As a result the item bank hardly showed a floor effect, however a small percentage patients reached the ceiling of the item bank. The most difficult item in the current item bank is ‘ride a bike for more than two hours’. To further reduce the ceiling effect we
Another possibility to construct an even more sensitive item bank could be accomplished by using categorical response options, which are known to be more informative and responsive. At present the response options to the ALDS items are; ‘can carry out the activity’, ‘can carry out, but with difficulty’, and ‘cannot carry out the activity’. The option ‘with difficulty’ has been added since dichotomous rating scales are disliked by some respondents as they are perceived as too restrictive. Nevertheless, using a dichotomous IRT model, the present scoring of the ALDS is based on two response categories: (the responses ‘can’ and ‘can carry out, but with difficulty’ are analyzed as one response category). It has been shown that this dichotomized scoring approach maximizes the reproducibility of scoring between two time points and interviewers, increases clinical interpretability and is easier to manage and comprehend in adaptive testing. However, inter-item correlations are lower and skewed score distributions are more common. There is no clear consensus which score method is best, especially in patients at the ceiling and floor of the ability continuum, the sensitivity of the item bank might be improved by calibrating the items using the graded response model. This model is an extension of the two-parameter logistic model used to calibrate the current item bank, and is appropriate to use when item responses can be characterized as ordered categorical responses.

Sample size, number of items and statistical power

Since linear measures are more sensitive compared to the classic ordinal or quasi-interval instruments, it is argued that IRT based item banks have more statistical power to detect treatment effects than traditional outcome scales. Although this argument seems reasonable, it is not per definition necessarily true. For example, in general, to detect an effect size of 0.20, 0.50 or 0.80 using a t-test with a 5% two-sided significance level and a power of 80%, it is required to include 394, 64 or 26 patients respectively in each treatment arm of a clinical trial. A simulation study of our research group, however, showed that 450, 90 and 40 patients are required in each arm to detect the same effects using 50 items and IRT (Table). The main reason for these differences between these figures is that when a t-test is used, it is assumed that the outcome parameter of interest can be measured directly and without error, whereas IRT takes into account that a latent variable (e.g., disability status) cannot be measured without error. On the other hand, compared to traditional outcome scales, IRT instruments make a statistical power analysis more flexible as the number of patients required in a trial to demonstrate a given effect size also depends on the number if items used. Moreover, we demonstrated that the power of a study can be increased further when a selection of items is tailored by clinicians to the target population.

Differential item functioning

A potential problem using item banks, including the ALDS, in outcome assessment is differential item functioning (DIF). When comparing outcomes of the ALDS between patient groups, item characteristics of the ALDS should be consistent across groups. DIF occurs when people from different groups (commonly gender or age) with the same ability have a different probability to be able to carry out an activity. If DIF is ignored, inaccurate measurements are obtained and true differences between patient groups may be obscured and non-existent differences ‘created’.

Due to the generic nature of the ALDS item bank and consequently, its application in a broad range of disorders it is important to examine DIF carefully. For example, one might argue that indoor mobility activities might be relatively difficult for patients with primarily motor impairment, and relatively easy for patients without motor impairment. Although, DIF probably exist in all outcome measures, IRT, in contrast to the classical test theory, gives the statistical opportunity to test the presence of DIF, to evaluate the impact, and to model DIF.

In the current ALDS item bank 11 items with different measurement properties for men and women as well as older and younger respondents have been removed from the instrument. An alternative approach could have been to include these differences in the ALDS item bank by using gender and age dependent ‘weights’ or difficulty parameters for such items. For clarity, gender and age specific items were not preferred. In the latter part of this thesis we showed a new DIF analysis by contrasting patients suffering from (mainly) neurological diseases and patients with diseases of internal medicine. Eighteen of the 72 ALDS items had statistical significant DIF. Nevertheless, the DIF could

<table>
<thead>
<tr>
<th>Effect size</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>394</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>950</td>
<td>680</td>
<td>500</td>
<td>450</td>
<td>440</td>
<td>160</td>
</tr>
<tr>
<td>0.5</td>
<td>160</td>
<td>125</td>
<td>96</td>
<td>90</td>
<td>87</td>
<td>64</td>
</tr>
<tr>
<td>0.8</td>
<td>70</td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>39</td>
<td>26</td>
</tr>
</tbody>
</table>

*Theoretically measuring without error.
be accounted for so that DIF did not affect the disability estimates of both groups. Therefore, no adjustment in terms of removing items is required.

Although, our analysis across a wide range of diseases showed some DIF, the question remains how much DIF is acceptable? If we compare the estimated disability level between the two groups studied not accounting for DIF, small differences in the estimated disability level exist (a difference of 0.12 logit, which corresponds with less than two points on the 0 to 100 linear transformed ALDS scale). If we want to use the ALDS in epidemiologic studies explicitly focusing on comparison of diseases groups between medical specialties it may be advisable to model DIF using disease-specific item parameters to increase the measurement precision. Further, in Chapter 8 it can be seen that the direction of the differences of the item difficulty parameters of the two groups were evenly distributed. That is, some of the items are more difficult for one group and some of the items are more difficult for the other group. Further, the differences in the discrimination parameters were usually not significant. Therefore, the items for an outcome measure can be selected is such a way that the effects of DIF may cancel (this is in line with the approach chosen by many large-scale educational assessments). At present, in clinical research within patients groups, there is no need to modify the item parameters presented in this thesis as the relatively small DIF identified in these subgroups has no clinical significance.

Adaptive testing

On several occasions we have stressed that IRT is superior to classical test theory because it can eliminate test dependency and achieve more precise measurement through adaptive testing. Adaptive testing reduces test administration times and allows varied and precise estimates of ability, because each ‘test’ is tailored to the unique ability level of each respondent. If a respondent cannot walk one block, it is not essential to ask whether he or she can walk a mile. Instead, it is asked whether the respondent can walk across a room. The test burden of patients can thus be highly reduced. Additionally, pre-test, post-tests and follow-up tests no longer have to consist of the same items, challenging respondents at their ability level instead of having an annoying or discouraging effect. With adaptive algorithms, where every patient answers a different set of questions with different severity, it is showed that a selection of a small set of items is sufficient to have unbiased estimates of the scores obtained using the total set of items.1

A large incentive for the construction of IRT based item banks in clinical research is the possibility of computer adaptive testing (CAT).4,10 In the further development phase of the ALDS high priority should therefore be given to the development of an ALDS website. On this website an easy to use CAT program should be implemented. With CAT, items are selected from an item bank on the basis of patients’ responses to previously administered items. CAT targets the difficulty of items to the level of disability of the patient that is assessed and patients need only answer a limited number of items in order to obtain a measure that accurately estimates what would have been obtained had the entire item set been administered. This practical approach also gives physicians and nurses the opportunity to monitor their patient’s level of disability in their daily clinical practice.

Future research

At present data of approximately 6000 patients with a wide variety of diseases of 35 different clinical studies is available in the ALDS database. All studies are using three response options and many of those studies included additional (to be calibrated) items as well. In the coming calibration phase of the ALDS the lower limit of the scale will be extended and the item density around the extremities will be increased.

Furthermore, additional analyses will be performed using the graded response model to decide if scoring categorical response options gives a surplus value in terms of clinical information and score interpretation. Subsequently, using such a model, the floor and ceiling effects can be investigated in more detail. Although, the ALDS showed less of a ceiling effect compared to other often used disability scales in the described disorders, these results must be interpreted with caution. Floor and ceiling effects were demonstrated by presenting the percentage patients that reached a minimum or maximum score, which is a quite simple approach. An alternative and more sensitive approach could be, for example, the use of Shannon’s indices. These indices incorporate the frequency distribution across all categories of an instrument and not just the highest and lowest categories. In an item where a response option has a very high (or low) endorsement, e.g., more than 95% of the patients are able to carry out a certain activity, there is very little information being transmitted.11

Further implementation of the ALDS should proceed by comparing the clinimetric properties of the item bank with the most often used disability scales in peripheral vascular surgery patients, neuromuscular disorders and pulmonary diseases. Finally, new DIF analysis should also focus on cross-cultural differences by comparing the present ALDS language versions (Dutch, English, French, Italian, and Spanish), and by contrasting measurements on base of interview by doctors, nurses or paramedics on the one hand and patient self-reports on the other.

To conclude

This thesis showed promising results using the IRT based ALDS item bank in clinical research. The
The challenge of the studies included in this thesis was bringing IRT and adaptive testing to the clinical setting. No matter how superior IRT is compared to the well-known classical test theory, it is important IRT does not become a ‘black box’ for clinicians, since they have to interpret the provided outcomes in clinical research and practice. The ALDS item bank will be improved further as new items are added from studies and assessments are fully (web based) computer adaptive. Combined with other merits of IRT based item banks it is expected that this technique gradually will suppress the very large number old classical constructed paper and pencil questionnaires.

References


(8)  Holman R. How does item selection procedure affect power and sample size when using an item bank to measure health status? Quality of Life Newsletter. 2004;9:11.

