Rasch Analysis of the ADL Scale of the A-ONE

Guðrún Árnadóttir, Anne G. Fisher

The ADL-focused Occupation-based Neurobehavioral Evaluation (A-ONE; Árnadóttir, 1990) can be used to evaluate both performance of activities of daily living (ADL) tasks and neurobehavioral problems that interfere with ADL task performance among clients with neurological disorders. This study examined the rating scale structure and aspects of validity and reliability of the A-ONE’s ordinal ADL scale by applying Rasch analysis methods (Bond & Fox, 2001). Rasch analysis of 209 clients’ A-ONE assessments indicated that misfit of items to the ADL scale could be reduced by removing the two communication items. Threshold disordering could be corrected by combining two adjacent scoring categories (supervision and verbal assistance), thus supporting four response categories. Separation reliability for item calibrations (.98) was high and acceptable for people (.90). Finally, principal components analysis of the residuals supported unidimensionality. The study provided support for converting the ordinal ADL scale to an interval scale that has potential to be used to measure changes in ADL task performance over time.


This study used Rasch analysis methods to explore the rating scale structure and aspects of scale validity and reliability of the Activities of Daily Living (ADL) scale of the A-ONE (ADL-focused Occupation-based Neurobehavioral Evaluation), originally developed by traditional psychometric methods and previously called the Árnadóttir Occupational Therapy–Activities of Daily Living (OT–ADL) Neurobehavioral Evaluation (Árnadóttir, 1990). Thus, this study is an effort to potentially strengthen an existing assessment by determining whether it has a potential to be used for a reason other than the one for which it was originally designed—that is, as an outcome measure, not merely as a guide to intervention planning. Rasch analysis of the Neurobehavioral scale will be addressed in a future article.

The A-ONE is intended for use by occupational therapists to evaluate clients with neurological disorders. It is based on the idea that the occupational therapist is able to identify not only the level of ADL assistance (ADL ability) needed but also the nature of underlying neurobehavioral problems that interfere with ADL task performance. That is, the A-ONE is unique in that it also is used to evaluate the underlying reason for the lack of independence (Árnadóttir, 1990, 1999, 2004).

The A-ONE comprises two scales representing two different hypothetical constructs. The two scales are the Functional Independence scale (commonly referred to as the ADL scale because most of the items pertain to the construct of ADLs) and the Neurobehavioral scale. Both the A-ONE ADL and Neurobehavioral scales were developed as criterion-referenced rating scales of the ordinal type. The ADL scale comprises five domains (dressing, grooming and hygiene, transfers and mobility, feeding, and communication). Totaling scores within domains was discouraged because of the ordinal nature of the raw scores (Árnadóttir, 1990). For similar reasons, adding scores across domains to generate an overall total score

KEY WORDS
- activities of daily living (ADLs)
- assessment
- cerebrovascular disorders
- dementia
was strongly discouraged. Using inappropriate mathematical manipulations (e.g., adding ordinal scores as though they were equal interval to form a total score) has been criticized (Merbitz, Morris, & Grip, 1989) because the results lack meaning and can result in serious misinterpretation of the data. To prevent this problem, Rasch measurement methods have been recommended to transform ordinal data into equal interval measures expressed in linear log-odds probability units (logits; Rasch, 1960/1980; Wright & Linacre, 1989). Such data can be subjected to mathematical manipulation without risk of generating invalid results (Bond & Fox, 2001; Wright & Linacre, 1989).

One reason for implementing this study was that we wanted to determine whether the items from the five domains of the ADL scale could be combined and shown to work together to define a single unidimensional construct. The original purpose for developing the A-ONE was to provide a tool to gather useful information for intervention planning; it was not designed to evaluate change in ADL task performance over time (e.g., effectiveness of interventions). If, however, all the items in the ADL scale can be shown to be unidimensional (i.e., to evaluate the same latent variable), the potential also exists to use Rasch measurement methods (Bond & Fox, 2001; Wright & Linacre, 1989) to construct linear measures that also can be used to evaluate the effectiveness of interventions.

More specifically, Rasch analysis computer programs generate goodness-of-fit statistics for items, indicating how well the data fit the Rasch model. When the items demonstrate statistical goodness of fit to the Rasch model, the scale can be said to be unidimensional. Unidimensionality indicates that the items are ordered along a single unidimensional construct or line on which items range from those that are easily performed to those that are hard to perform, thus supporting the scale’s internal validity (Bond & Fox, 2001; Wright & Linacre, 1989). Unidimensionality also can be evaluated by using principal components analysis of the residuals (Smith, 2000).

The Rasch rating scale model is based on the following two assertions: (1) The easier the items, the more likely it is for a person to obtain higher ratings, and (2) the more able the client, the more likely he or she is to obtain higher ratings on hard items than are less able clients. Should Rasch analysis reveal that the ADL items of the A-ONE demonstrate goodness of fit to those assertions, it will be possible to develop a table in the test manual in which therapists can look up the Rasch linearized equivalent for the client’s raw total ordinal ADL score, thus allowing them to use the A-ONE as an outcome measure.

Additional evidence for scale validity can be provided by verification of logical hierarchical ordering (based on the item difficulty calibration values) of the items along the linear scale and the targeting of the items to the abilities of the persons tested. Targeting refers to how well the item difficulties are aimed at the performance level of the target population (Wright & Stone, 1979). For the ADL scale, targeting would be indicated by a match between (1) the average performance and ability dispersion of persons with neurological dysfunctions and (2) the mean item difficulty calibration and dispersion of ADL item calibration values, respectively.

One method used to verify the logical ordering of the items is to examine the hierarchical order of the item difficulties in relation to other scales. Examples of easy ADL items reported in the literature include the “Eating” item on the Functional Independence Measure (FIM; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994) and the “Feeding” item on the Barthel Index (Tennant, Geddes, & Chamberlain, 1996). At the more difficult ends of these scales are the “Tub and shower transfer” item on the FIM and the “Bathing” item on the Barthel Index (Linacre et al., 1994; Tennant et al., 1996).

Rasch analysis computer programs also estimate reliability for both persons and items. This information is revealed by two indexes in the form of a separation reliability coefficient and a separation index. The reliability coefficient indicates replicability of person or item placements on the line. The person separation index indicates how well the ADL items separate the sample into statistically distinct levels of ability. Similarly, item separation is an index of how well the sample separates the items into different levels of difficulty. Higher separation indicates a scale that covers a wider range of the construct being measured (Bond & Fox, 2001).

Finally, a high-quality measure requires a statistical assessment of the psychometric properties of the rating scale (Linacre, 2002). Examination of the psychometric properties of the rating scale should be performed before exploration of other forms of validity (Lopez, 1996).

Preliminary reliability and validity studies of the A-ONE ADL scale performed with traditional psychometric methods indicated that the instrument has the potential to be converted into a linear measure. These studies have included examination of item correlations both within and between domains, exploratory factor analysis (Árnadóttir, 1990), and item analysis based on Cronbach’s alpha (Steuiljens, 1998). Relatively high (≥ .80) Cronbach’s alpha coefficients suggested homogeneity of all but the communication items. Homogeneity, however, does not confirm unidimensionality (Anastasi & Urbina, 1997). Moreover, factor analysis suggested the existence of three
underlying factors (with the communication items representing one factor), indicating that the scale items may not be unidimensional. Considered together, the results from these studies suggest that although the communication items likely belong to a construct different from that of the items in the four ADL domains, unidimensionality might be achieved if the communication items are omitted. That communication is a different construct from ADL has been supported by research findings related to the FIM (Heinemann, Linacre, Wright, Hamilton, & Granger, 1993).

In view of the above background information, four research questions were posed. The first three questions address validity and the last one addresses reliability:

1. Does the rating scale of the A-ONE demonstrate sound psychometric properties as evidenced by ordering of category measures, acceptable goodness of fit of the rating scale categories to the Rasch model, and ordering of the calibration thresholds between the rating scale categories?

2. Do the items on the ADL scale of the A-ONE define a single unidimensional construct, as evidenced by goodness of fit and principal components analysis?

3. Do the item difficulties on the A-ONE ADL scale show logical hierarchical ordering, and are they appropriately targeted to people with cerebral vascular accidents (CVAs) or dementia?

4. Do the items on the A-ONE ADL scale separate people into different levels of ability when evaluating those with CVA or dementia, and do the people tested separate the items into different levels of difficulty?

Method

Participants

This study used a retrospective design in which A-ONE evaluation records from 209 clients (all available records between 2000 and 2004) of the rehabilitation and geriatric wards at Landspítali University Hospital in Iceland were reviewed and analyzed. According to Linacre (1994), a sample size of 150 is acceptable for most purposes (99% confidence interval for estimated item difficulty calibrations remaining stable within the absolute value of 0.5 logit). Heterogeneity of both ADL ability (i.e., level of assistance needed to perform ADL tasks), as required by Rasch analysis (Bond & Fox, 2001; Wright & Masters, 1982), and diagnosis, reflecting the group of clients to which the A-ONE is intended to be applied, was ensured in the selection of the participants. One hundred eleven participants (53.1%) had been diagnosed with dementia of different types, 95 with CVA (45.5%), and 3 (1.4%) with other neurological diagnoses. See Table 1 for more detailed participant demographic information.

Instrumentation

The ADL scale of the A-ONE includes 22 items representative of the five domains: dressing (“Put on shirt,” “Put on pants,” “Put on socks,” “Put on shoes,” “Manipulate fastenings”), grooming and hygiene (“Wash face,” “Comb hair,” “Brush teeth,” “Shave beard/apply cosmetics,” “Perform toilet hygiene,” “Bathe”), transfers and mobility (“Sit up in bed,” “Transfer from sitting,” “Maneuver around,” “Transfer to toilet,” “Transfer to tub”), feeding (“Drink from glass/cup,” “Use fingers to bring food to mouth,” “Bring food to mouth by fork or spoon,” “Use knife to cut and spread”), and communication (“Expression” and “Comprehension”). Each item is rated using a five-category ordinal scale: 0 = full assistance needed, 1 = minimum to considerable physical assistance needed, 2 = verbal assistance needed, 3 = supervision needed, and 4 = independent. People are scored on the basis of the observed level of assistance needed for the ADL task performance and are not penalized for using assistive devices. The A-ONE manual includes conceptual and operational definitions for all items and detailed criteria for administration and scoring of the instrument (Árnadóttir, 1990).

Table 1. Age and Gender of Participants by Diagnostic Group

<table>
<thead>
<tr>
<th>Age and Gender</th>
<th>Dementia</th>
<th>CVA</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>80.7</td>
<td>66.4</td>
<td>49.0</td>
<td>73.7</td>
</tr>
<tr>
<td><em>SD</em></td>
<td>8.0</td>
<td>13.0</td>
<td>5.2</td>
<td>13.0</td>
</tr>
<tr>
<td>Range</td>
<td>61–99</td>
<td>22–84</td>
<td>43–52</td>
<td>22–99</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>24 (28.6)</td>
<td>59 (70.2)</td>
<td>1 (1.2)</td>
<td>84 (40.2)</td>
</tr>
<tr>
<td>Female</td>
<td>87 (69.6)</td>
<td>36 (28.8)</td>
<td>2 (1.6)</td>
<td>125 (59.8)</td>
</tr>
<tr>
<td><strong>Total, n (%)</strong></td>
<td>111 (53.1)</td>
<td>95 (45.5)</td>
<td>3 (1.4)</td>
<td>209 (100)</td>
</tr>
</tbody>
</table>

Note. CVA = cardiovascular accident.
Procedures and Data Analysis

All participants had been evaluated as part of routine occupational therapy services at the Landspítali University Hospital. The 11 therapists who performed the evaluations had completed a 5-day A-ONE training course and administered the evaluations according to the standardized procedures described in the A-ONE manual. Because this study was retrospective, it was not possible to obtain information on rater reliability of the therapists involved. However, studies of the intrarater reliability of therapists working at the same hospital have indicated a kappa of .84 for the ADL scale (Árnadóttir, 1990) and, more recently, a weighted kappa of .90 (Árnadóttir, 2005). Before collection of raw scores and participant demographic information from the available A-ONE forms in the hospital records, written approval for the study was obtained from the Ethical Committee of Landspítali University Hospital. The first author extracted the raw data from the participants’ records.

The raw scores were analyzed using the WINSTEPS Rasch computer software program (Version 3.63.0; Linacre, 2006). Rasch analysis procedures have been described elsewhere in detail (Bond & Fox, 2001; Wright & Masters, 1982). The data analysis was divided into two phases, as described in more detail in the following section. It was anticipated that both phases might involve several analyses. Although the two communication items were expected to misfit, they were retained in Phase 1 with the intention of verifying their expected lack of fit in Phase 2.

Phase 1: Rating Scale Analysis

The first phase involved evaluation of the rating scale’s psychometric properties using Linacre’s (2002) guidelines. Thus, a minimum of 10 observations was desired for precise estimation of category measures, and the average measures were required to advance with scoring category. The conservative criterion of 1.5 was chosen for outfit mean square ($MnSq$) values (Smith, 1996). We were prepared to collapse nonadvancing categories with the ones below if the average measures did not advance with category (Linacre, 2002), provided doing so made sense theoretically. We would attempt to extract maximum information, however, by retaining the highest possible number of distinct categories between which raters could clearly differentiate (Stone & Wright, 1994).

The literature reflects different opinions about the necessity of ordered thresholds (cf. Andrich, 1996; Linacre, 2002). Disordering of thresholds can result when a category is undersused (i.e., has low frequency counts), its definition is not clear, or the number of categories exceeds the number of levels the raters can distinguish between (Andrich, 1996; Linacre, 2002). Thus, if threshold disordering appeared in the rating scale, we intended to interpret its source and attempt to resolve the disordering (Wright & Linacre, 1992). Our goal was to ensure the best person separation along the variable when determining optimal categorization (Lopez, 1996). Moreover, we planned to consider whether each of the threshold calibrations of the rating scale advanced by at least 1.0 logit (Linacre, 2002).

Phase 2: Internal Scale Validity and Reliability Analysis

Phase 2 of the analysis began by exploring for misfitting items. If items failed to meet the assertions of the Rasch model for the A-ONE ADL scale, the plan was to omit them one at a time (except for communication, for which both items would be removed even if only one misfit was found). At each step, the criteria for failure to meet the assertions were based on the combined consideration of $MnSq$ and standardized $z$ goodness-of-fit statistics (Wilson, 2005). The first item targeted for removal was that with the largest $MnSq$, provided infit $MnSq > 1.4$ (Wright & Linacre, 1994) and $z \geq 2$ (Wilson, 2005). We based our decision for item removal on infit because of its sensitivity to item performance (Bond & Fox, 2001; Wright & Masters, 1982), and we focused on high $MnSq$ values because they signal a particular threat to validity (Wilson, 2005; Wright, 1995).

Several criteria were considered for when to stop the item removal process: If all items had infit $MnSq$ and $z$ infit statistics that met the criteria for inclusion, the item removal process would be stopped. If, however, removal of a misfitting item resulted in the scale’s diminished ability to separate people into different levels of ability (as indicated by a decrease in the person separation index), the step-by-step removal of items also would be stopped. Because no criteria for a significant change have been published, we set our criterion for a decreased separation index at 0.10. Finally, because of our focus on potential for revision of the ADL scale, we also considered the theoretical importance of items before item removal (Bohlig, Fisher, Masters, & Bond, 1998). That is, because up to 5% of items on a scale are expected to misfit by chance (Smith, 1991), if a single theoretically important misfitting item remained, we would stop the removal process and consider the item’s theoretical importance, its future potential to be revised (e.g., split into two or more new items), or both, rather than remove it (Linacre, 1995). Moreover, if any misfitting items remained, we would attempt to identify the source of disruption (e.g., rater scoring error). We also planned to evaluate unidimensionality by means of principal components analysis. If the proportion of variance explained by the measures (Rasch dimension) was >60% and the proportion of unexplained...
variance accounted for by the first contrast (the largest secondary dimension) was <5%, the results would be considered to support unidimensionality (Linacre, 1991–2006).

Once item analysis and removal were complete, we proceeded to examine aspects of reliability. For this purpose, the person and item separation indices were examined. The separation index should be at least 2.0 to obtain the desired reliability coefficient of .80. A person separation index of 2.0 indicates that the sample can be separated into at least three distinct groups (Fisher, 1992), and an item separation index of 2.0 indicates that the items on the scale define at least three levels of ability.

Results

**Phase 1: Rating Scale Analysis**

The frequencies for all five categories of the ADL scale exceeded 10 ratings (minimum = 130, maximum = 1,696). When examining for measure disordering, the category measures did not signal a problem, but the thresholds were disordered (Table 2). Finally, the highest category outfit \( MnSq \) (1.48) was within acceptable limits.

In a subsequent analysis, verbal assistance (score = 2) and supervision (score = 3) were successfully combined such that threshold disordering was eliminated. The revised modeled category probability curves are shown in Figure 1. Thresholds now advanced more than 1 logit, and the person separation index increased by 0.14 to 2.87.

**Phase 2: Scale Validity and Reliability Analysis**

As we began Phase 2, we retained the collapsed rating scale and proceeded to examine item goodness of fit to the Rasch rating scale model. Several analyses were implemented during this procedure. When data from all 22 items of the five ADL domains were included, the two communication items failed to demonstrate acceptable goodness of fit to the model (“Expression”: infit \( MnSq = 1.74, z = 5.9 \) and outfit \( MnSq = 2.62, z = 7.3 \); “Comprehension”: infit \( MnSq = 1.48, z = 3.7 \) and outfit \( MnSq = 2.79, z = 5.9 \)). The “Use knife” item from the feeding domain also demonstrated misfit, resulting in a total of 13.6% item misfit to the model.

In the next analysis, we removed the two communication items according to our predetermined plan. The “Use knife” item continued to fail to show acceptable goodness of fit (infit \( MnSq = 1.63, z = 4.6 \); outfit \( MnSq = 1.9, z = 5.3 \)). All other items (95%) continued to demonstrate acceptable infit (see Table 3). The person separation index increased by 0.06 to 2.93.

Because the “Use knife” item remained the only misfitting item, we proceeded to examine the data for a potential source of the disruption. Among the 32 people who had unexpected ratings (misfitting responses) on the “Use knife” item, 20 had unexpectedly low ratings; among those, 75% were persons with CVA. The 12 people with unexpectedly high ratings on the “Use knife” item were equally distributed between those with CVA and dementia. Otherwise, no pattern was discernable among the misfitting responses by gender, ADL ability level, age, or diagnostic subgroup (e.g., right CVA vs. left CVA). Content analysis of the item revealed

![Figure 1. Modeled category probability curves after combining categories “2” (verbal assistance) and “3” (supervision). A-ONE ADL scale = ADL-focused Occupation-based Neurobehavioral Evaluation Activities of Daily Living scale.](image)

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency (%)</th>
<th>Outfit ( MnSq )</th>
<th>Calibration Threshold</th>
<th>Category Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>130 (3.4)</td>
<td>1.30</td>
<td></td>
<td>−3.29</td>
</tr>
<tr>
<td>1</td>
<td>862 (22.7)</td>
<td>1.02</td>
<td>−2.16</td>
<td>−0.84</td>
</tr>
<tr>
<td>2</td>
<td>442 (11.7)</td>
<td>1.48</td>
<td>1.01</td>
<td>0.37</td>
</tr>
<tr>
<td>3</td>
<td>660 (17.4)</td>
<td>0.92</td>
<td>0.50</td>
<td>1.11</td>
</tr>
<tr>
<td>4</td>
<td>1696 (44.8)</td>
<td>1.02</td>
<td>0.66</td>
<td>2.23</td>
</tr>
</tbody>
</table>

*Table 2. Rating Scale Category Statistics (22 Items, 5 Categories)*

*Note. Disordered thresholds appear in bold.*
that it includes two different behaviors: cutting food and spreading butter. Because the rate of item misfit was within acceptable limits (≤5%) after removal of the two communication items, the item removal process was stopped and the “Use knife” item was tentatively retained. Principal components analysis revealed that 84% of the total variance was explained by the measures and that only 3.6% of the unexplained variance was accounted for by the first contrast.

Our next step was to examine logical hierarchical ordering of item difficulties along the A-ONE ADL scale. The hierarchical ordering of item difficulties appeared to be logical, with three of the four feeding items being the easiest items and items such as “Transfers to tub” and “Bathe” being the hardest items (Table 3).

The distribution of items along the linear ADL scale and their targeting to the sample are shown in the item–person map in Figure 2. The item difficulty calibrations ranged from −2.83 to 1.61 and spanned 4.44 logits (Table 3 and Figure 2, columns labeled “Item Difficulty Calibration”). The spread of the participants’ abilities ranged over 8.03 logits (Figure 2), thus exceeding the range of the items; 9 participants had maximum scores. The three easiest items were targeted to the few least able persons, but no items were targeted to the most able persons. The mean person ability measure of 1.61 logits, as opposed to 0.00 for item difficulty calibration, also indicated that the items might not be targeted well to the most able participants. The only gap in the item distribution exceeding 0.50 logit was between −4.75 and −3.00 logits, at the lowest performance categories of these items (item distribution column labeled “Items: Bottom”), and no people in the sample were located across from this gap.

Our final step was to examine reliability. Our final analysis revealed a person separation index of 2.93 and separation reliability coefficient of .90. These results indicate that we can reliably differentiate the sample into at least three statistically distinct strata of ADL ability. The item separation index was 8.02, revealing that the items defining the ADL variable were well separated into at least nine strata of difficulty by the people in the sample (Fisher, 1992). The associated reliability coefficient was .98.

Discussion
The purpose of the study was to apply Rasch analysis methods to explore the rating scale’s psychometric properties and to evaluate aspects of the A-ONE ADL scale’s validity and reliability to determine whether valid measures could be generated from the ordinal ratings so that the scale might be used as an outcome measure. A key objective was to identify need for revisions of the instrument to strengthen its psychometric qualities. Potentially useful refinements were revealed, including correcting threshold disorder by collapsing two of the rating scale scoring categories and reduction of item misfit from 13.6% to an acceptable level by removal of the two communication items. Supporting these revisions, an

---

Table 3. Item Measurement Report With Communication Items Omitted and Four Rating Scale Categories

<table>
<thead>
<tr>
<th>Items</th>
<th>Item Difficulty Calibration</th>
<th>SE</th>
<th>MnSq</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer to tub</td>
<td>1.61</td>
<td>0.18</td>
<td>0.90</td>
<td>−0.6</td>
</tr>
<tr>
<td>Bathe</td>
<td>1.59</td>
<td>0.19</td>
<td>0.79</td>
<td>−1.4</td>
</tr>
<tr>
<td>Manipulate fastenings</td>
<td>1.11</td>
<td>0.12</td>
<td>1.38</td>
<td>3.3</td>
</tr>
<tr>
<td>Put on socks</td>
<td>1.08</td>
<td>0.12</td>
<td>0.85</td>
<td>−1.6</td>
</tr>
<tr>
<td>Put on pants</td>
<td>1.03</td>
<td>0.11</td>
<td>0.56</td>
<td>−5.4</td>
</tr>
<tr>
<td>Put on shirt</td>
<td>0.83</td>
<td>0.11</td>
<td>0.60</td>
<td>−4.8</td>
</tr>
<tr>
<td>Put on shoes</td>
<td>0.57</td>
<td>0.12</td>
<td>0.81</td>
<td>−2.0</td>
</tr>
<tr>
<td>Brush teeth</td>
<td>0.33</td>
<td>0.12</td>
<td>1.04</td>
<td>0.5</td>
</tr>
<tr>
<td>Wash face</td>
<td>0.25</td>
<td>0.11</td>
<td>0.81</td>
<td>−2.1</td>
</tr>
<tr>
<td>Use knife to cut and spread</td>
<td>0.19</td>
<td>0.14</td>
<td>1.63</td>
<td>4.6</td>
</tr>
<tr>
<td>Transfer to toilet</td>
<td>0.00</td>
<td>0.15</td>
<td>1.25</td>
<td>1.8</td>
</tr>
<tr>
<td>Maneuver around</td>
<td>0.06</td>
<td>0.12</td>
<td>1.16</td>
<td>1.6</td>
</tr>
<tr>
<td>Perform toilet hygiene</td>
<td>0.13</td>
<td>0.15</td>
<td>0.99</td>
<td>−0.1</td>
</tr>
<tr>
<td>Shave beard or apply cosmetics</td>
<td>−0.02</td>
<td>0.14</td>
<td>1.04</td>
<td>0.4</td>
</tr>
<tr>
<td>Sit up in bed</td>
<td>−0.32</td>
<td>0.12</td>
<td>1.13</td>
<td>1.2</td>
</tr>
<tr>
<td>Transfer from sitting</td>
<td>−0.36</td>
<td>0.12</td>
<td>0.93</td>
<td>−0.6</td>
</tr>
<tr>
<td>Comb hair</td>
<td>−0.39</td>
<td>0.12</td>
<td>1.04</td>
<td>0.4</td>
</tr>
<tr>
<td>Bring food to mouth by fork or spoon</td>
<td>−2.12</td>
<td>0.17</td>
<td>1.10</td>
<td>0.7</td>
</tr>
<tr>
<td>Use fingers to bring food to mouth</td>
<td>−2.56</td>
<td>0.19</td>
<td>1.10</td>
<td>0.7</td>
</tr>
<tr>
<td>Drink from glass or cup</td>
<td>−2.83</td>
<td>0.20</td>
<td>1.00</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*Outfit statistics are not reported but revealed similar results.

*Item misfit.
Figure 2. Item–person map indicating spread of items and people along the logistic Activities of Daily Living (ADL) scale. Each item appears in three columns indicating different rating scale measures. Item Difficulty Calibration = the mean item difficulty calibration; Items: Bottom = item measure in which a person has a 50:50 chance of being rated 0 or 1 (threshold between the lowest rating and the next highest rating); Items: Top = item measure in which a person has a 50:50 chance of being rated 2 or 3 (threshold between the next to highest rating and the highest rating); : = 1 person; # = 2 persons.
improved and acceptable person separation index (2.93) was obtained in the final analysis. Finally, the results suggested a need to revise and reevaluate the “Use knife” item.

As mentioned earlier, we found threshold disordered for scoring categories of “2” (verbal assistance) and “3” (supervision). This could indicate that therapists have a problem discriminating between these two categories (Andrich, 1996; Linacre, 2002; Stone & Wright, 1994). This finding was not surprising because both ratings may be assigned when some verbal assistance is given (Arnadóttir, 1990). Although opinions regarding the importance of threshold disordering differ (cf. Andrich, 1996; Linacre, 2002), we were able to demonstrate that the threshold disordering of the ADL scale could easily be resolved and, in this case, improve the scale’s psychometric properties. That is, although sensitivity of evaluations may increase with more scoring categories, in our case, sensitivity (as indicated by the separation index) slightly improved after categories “2” and “3” were combined and the number of categories was reduced from five to four. The increase in person separation indicated that the items may be slightly more able to differentiate persons into different ability levels and subsequently better able to detect changes in performance over time. Moreover, combining these two categories seemed logical from a theoretical perspective (Linacre, 2002; Lopez, 1996). Thus, because the category combination will likely make scoring of the A-ONE easier for therapists and thereby reduce error, this is a desirable solution. Before a final decision is made about the optimal number of categories for the ADL rating scale, however, rater use of the reduced number of categories should be evaluated.

The findings indicating acceptable goodness of fit for items obtained after removal of the two communication items are in agreement with previous indications from item analysis applying Cronbach’s alpha coefficients (Steultjens, 1998). Results from the principal components analysis confirmed unidimensionality. The “Use knife” item, accounting for the remaining misfit to the model after removal of the two communication items, was tentatively retained for two reasons. First, the item misfit was within the acceptable statistical misfit limits expected by chance when analyzing data from a 20-item scale (Type I error; Smith, 1991). Second, and more important, the item content was considered to be of both clinical and conceptual importance as a part of the feeding domain construct of the scale. Therefore, reasons for the item’s misfit should be explored and an attempt made to revise the item before a final decision is made regarding retention or removal. It is already clear that this item is composed of two tasks: cutting and buttering. Thus, the item may seem unclear or diffuse when it comes to scoring. Rather than omit the item outright, we preferred the more conservative approach of tentatively retaining it (Bohlig et al., 1998) and revising it by dividing it into two separate items (Linacre, 1995). In future research, we plan to reevaluate the scale’s unidimensionality by evaluating the goodness of fit of the items “Cutting” and “Buttering” after the division of the item “Use knife” into two separate items and performing principal components analysis of item residuals.

Scale validity was further supported by the logical ordering of items, in terms of agreement with findings from other studies, such as Rasch studies performed on the FIM motor scale (Linacre et al., 1994) and the Barthel Index (Tennant et al., 1996). Finally, the results revealed an unexpected situation in which 9 participants reached maximum scores on all items on the ADL scale (this outcome is referred to as a ceiling effect when it occurs using traditional psychometric methods) and people of lesser ability were properly targeted to easier items. Ceiling effects are in agreement with the findings from studies of other ADL scales (e.g., Tennant et al., 1996) and may not be a clinically relevant problem as long as the use of the A-ONE is restricted to people who are not independent in ADLs; a person who receives a maximum score is independent. Unless harder items are added to the instrument, however, this effect will limit the use of the evaluation as an outcome measure for more able clients.

Items sometimes misfit because they behave differently with different groups of people (Bond & Fox, 2001). That almost 50% of the misfitting ratings on the “Use knife” item were for people with CVA who had lower-than-expected ratings may suggest that this item, in its current form, displays differential item functioning. The participants who misfit on the item were otherwise proportionately distributed by gender, ADL ability, age, and diagnostic subgroup. Thus, no other systematic pattern was detected that could explain the misfit of this item. One reason may be that people with CVA and dementia have different patterns of motor impairment that might affect their use of knives. That is, it is common clinical knowledge that people with CVA often have unilateral weakness or paralysis, whereas those with dementia develop bilateral motor impairments that increase in the later stages of the disease process. A future study with larger sample sizes for each diagnostic group should examine for differences, including differential item functioning. Similarly, the performance of other diagnostic groups and cultures should be studied, as this study was limited by the fact that the analysis included mainly two diagnostic groups from Icelandic rehabilitation and geriatric wards. Thus, the generalizability of the results to performance of persons with neurological conditions may be limited.
Research results also are of limited use if they cannot be incorporated into clinical practice. Therefore, to enable clinicians to more easily use the A-ONE ADL scale as an outcome measure, conversion tables will need to be developed that enable clinicians to convert raw ordinal total scores to Rasch linearized measures without the need to use a Rasch computer program like WINSTEPS. More specifically, such tables would enable the occupational therapist to look up the linearized ADL ability measure expressed in logits (and its associated standard error) for the raw total ordinal ADL scores from a client’s evaluation. Subsequently, changes in logit values could be used to measure occupational therapy outcomes in a valid way. Development of conversion tables is considerably less expensive and less cumbersome than the development and maintenance of computer programs used for score conversion.

Conclusion and Clinical Implications

The present study provides an important first step in revising the ordinal ADL domains of the A-ONE into a single-interval ADL scale suitable for measuring outcomes, an event that would create a clinically more useful instrument. It is evident that several additional studies must be undertaken. The present study, however, provided evidence that item unidimensionality can be achieved by removal of two communication items and the possible future revision of the “Use knife” item. Scale structure can be improved by collapsing two adjacent scoring categories. Finally, conversion tables should be developed for therapists interested in using the A-ONE ADL scale as a linear outcome measure.

Acknowledgments

We would like to acknowledge the occupational therapists in the rehabilitation and geriatric wards of the Landspítali University Hospital who used the A-ONE to evaluate their clients. The study was supported by grants from the University of Akureyri and Landspítali University Hospital in Iceland.

References


