Refining an IADL measure by determining category functioning

Lindy Clemson
Anita Bundy
Lynnette Kay
Tim Luckett
Faculty of Health Sciences, The University of Sydney
L.Clemson@usyd.edu.au

Introduction

This paper provides an example of using category function analysis to refine the rating scale of an instrumental activities of daily living (IADL) assessment designed for use with older people. The scale, the *Assessment of Living Skills and Resources* (*ALSAR*), is unique in that it considers risk of not being able to perform instrumental daily living tasks as a function of both a person's skill and the adequacy of available resources. We adapted the original scoring to create a 9-point scale that reflected both skill and resources without summing ordinal values. We then investigated the rating scale's effectiveness using guidelines recommended by Linacre (2002). We ultimately settled on a 6-point scale. This paper details our process and describes challenges that face test developers in presenting scales that are useful and accessible to both clinicians and researchers.

Developing a scale to measure instrumental activities of daily living

Test developers seek to create measures that rate peoples' performance on a particular construct with the greatest accuracy and precision possible (Linacre, 2004). Defining the rating scale is a critical step in test development. Both Lopez (1996: 482) and Linacre (2004: 89) stated that effective category rating scales should be clearly defined, mutually exclusive, substantively relevant and perform as a conceptually exhaustive ordered sequence.

Previous studies have not shown that defining multiple scoring categories in IADL assessments leads to improved differentiation of people. In fact, after examining

category effectiveness, previous researchers have reduced their IADL rating scales from five- to three- (Doble & Fisher 1998) or two-categories (Finlayson et al. 2005; Hsueh et al. 2003).

The *Assessment of Living Skills and Resources* (*ALSAR*) has been designed to rate accomplishment of IADL tasks by elderly people living in the community (Williams et al. 1991). Both skills and resources are considered in the *ALSAR*. Skill is defined as accomplishment of the task by the person. Resources are any supports for task accomplishment extrinsic to the person.

The original *ALSAR* rating scale contained 5-categories (0 to 4) determined by adding scores on each item of 0 to 2 for both skills and resources (See Table 1.) All scores were then summed to provide a total risk score for IADL.

Table 1 about here

We created an alternative schema that considered both skills and resources without summing the values (See Table 2). The resulting rating incrementally increases across 9 hierarchical levels reflecting risk of inability to perform IADLs.

Table 2 about here

Method

We investigated the functioning and meaningfulness of our adapted category rating scale using Winsteps 3.61 computer software (Linacre 2006). We explored different category rating structures, guided by Linacre (2004) and by quality (fit) and robustness (reliability) (Wright & Masters 1982). Other aspects of this investigation have been detailed elsewhere (Clemson et al. 2006).

Category function analysis was used to check that the average measures for the rating scale categories advanced along a logical continuum, an essential dimension. We expect that observations in higher categories are produced by higher measures. Categories with average measures that do not advance monotonically or that show minimal advancement should be combined. Outfit mean-squares for each category are

expected to be less than 2.0 to indicate a reasonably uniform level of randomness and not too much noise in the data.

Linacre (2002: 97) also stated that it is helpful if step calibrations advance monotonically with each category so that clear interpretations can be made. Step calibrations correspond to probability characteristic curves. As the scores for an individual increase, these curves should reflect that each advancing category is the most likely to be chosen (Andrich 1996; Linacre 2004). If categories do not advance, they are considered to be "step disordered." In addition, the degree of advancement, though not essential for validity, can indicate the number of categories the scale can best support. For example, a dichotomous scale should aim to advance in step difficulty by 1.4 logits and a five category rating scale by 1 logit.

Another helpful dimension is the coherence of "ratings that imply measures" with that of "measures that imply ratings". That is, relationships between measures and average expected ratings are modelled along an expected item score ogive with preference to 40% or more fitting within the model expectation used as a guide to coherence.

Data

The sample included 260 ALSAR ratings from 160 people aged 55-101 years (mean 79 years). Diagnosis included stroke, hip fracture, visual dysfunction, and a small sample of general aged care in-patients. Data collection has been described more fully in Clemson at al. (2006)

Results: Developing the 6- Level Solution

Category structure effectiveness

In the original 9-level scale, the category observed average measures for each item followed the intended pattern of advancement. The outfit mean scores were below 2 suggesting that the measure did not include too much "noise." That is, no segments of unexplained outliers or unpredictable responses that would hinder the usefulness of the test. However, there were some irregularities. While the observed average measures increased monotonically, the degree of advancement was minimal, indicating that some categories needed to be combined. In addition, the step difficulty calibrations or thresholds were disordered and were at erratic intervals ranging from

9.11 to 0.06 logits. Thus, we set out to achieve a more workable number of categories and improve step ordering.

Closer examination of the advance of category measures and disordered thresholds indicated that scale categories "01" (2%, n=66), "02'' (0%, n=2), and "12" (1%, n=20) contained the lowest number of observations. Few observations in a category suggest that the category represents a very narrow segment of the latent variable or that the category is poorly defined or not understood by the raters (Linacre 2004).

Category scale "02" where "skill accomplishment is independent and consistent despite a lack of consistent resources" seems to be a meaningless rating; this was supported by its lack of use by raters (n=2). Scale "12" (n=20) where "partial skill accomplishment is accompanied by a total lack of consistent resources" is possible but the boundaries between this and "11," or possibly even "21," seem to be somewhat blurred, and the rating was not helpful. Thus, we deleted the "02" ratings and combined the insufficient resources ("12") with the partially supported resources ("11"). We recommend that future uses of the ALSAR also eliminate "02" and "12".

In contrast, a person accomplishing a task independently even though a resource is at risk in some way (category "01") might be a clinically meaningful category suggestive of an impending need for intervention. While it seems worthwhile to report this category for clinical purposes, it had extremely low coherence and was rarely awarded. Thus, we decided, for outcome measurement research purposes, to combine "01" with "00." This reduced the 9-level category ratings to a 6-level scale. Combining these categories retained the intent of the scale and improved its psychometric properties.

After these adjustments, the observation distribution and step calibration advance reflected more acceptable intervals. Nonetheless, both "reading" and "telephoning" had unacceptably high fit statistics, necessitating closer scrutiny of these items.

Examining item fit for "reading" and "telephoning"

Twenty scores for "reading" (19 persons) and 5 for "telephoning" had residuals of 3 or greater. A close look at the people behind these erratic scores revealed specific circumstances in which these very easy items became the most difficult.

Reading was limited for people with severe visual problems and for one person who was illiterate. Further, while vision did not change much over time for most people, other skills improved as a result of intervention. For other people, getting new glasses influenced reading but did not affect other IADL skills. With regard to telephoning, a few people were unable to get to the phone in time to answer it and one person with poor mobility used a public phone across the road. All of these accounted for the failure of "reading" and "telephoning" to conform to the expectations of the Rasch model. Because there were valid explanations for these poorly fitting scores, there was little compelling reason to drop "reading" or "telephoning" from the scale.

Further, deleting either or both of the items did not improve the construct. There was very minimal gain to person reliability or separation and deletion of these very easy items resulted in a decreased potential for discriminating the least capable persons. However, the category rating structure revealed that deleting the reading score for those persons with a known severe visual deficit or illiteracy was beneficial. Severe visual loss included visual acuity of 6/36 or worse (n = 47) and stroke where comments indicated severe visual dysfunction (n = 5).

Determining the final solution

The final solution was a 6-item scale with the reading score deleted for specified persons. This scale possessed average category measures that advanced monotonically at acceptable intervals — ranging from 0.59 to 1.09 logits. Tables 3 and 4 contrast the category structures of the 9-point scale with that of the 6-point solution. Figures 1 and 2 demonstrate the improvements to the ordering of the category rating structure. Step disordering was improved although the "11" category remained at a lower threshold than the "10" category (See Figures 3 and 4). Possibly the "11" category of "partial task accomplishment and partial support" is less clear to raters or, more likely, it measures a less distinctive level or narrower band of the construct.

Tables 3 and 4 about here

Figures 1 and 2 about here

A 5-category scale eliminated step disordering for all items; however, it produced 2 items (18%) with data that failed to conform to the fit expectations of the Rasch model.

A 3-category scale produced the most effective category functioning for all the category features and demonstrated item fit. The person reliability was good at 0.87 though separation was reduced at 2.62. This scale, however, effectively considered only skill levels and not the adequacy of available resources, thus violating a basic premise of the ALSAR.

As the concept of "11" (partial skill accomplishment, inconsistent resources) is in keeping with the intent of the scale and the six-level scale demonstrated those category functioning features that are deemed essential (Linacre 2004) we chose to maintain the 6 level rating scale. In addition, this scale has only one item (9%) with both misfitting infit and outfit statistics which is acceptable. Reliability is excellent at 0.90 and it has a separation of 2.98 (Clemson et al. 2006). Unidimensionality was also supported by conducting an unrotated principal components factor analysis, which is also reported elsewhere (Clemson et al. 2006).

In all category scales examined, the 50% cumulative probability was consistently at the "20" level (see Table 1), meaning that the probability of observing the categories below this level equals the probability of observing the categories equal or above (Linacre 2004; Winsteps Help 2006). This supports maintaining the full range of scoring options for resources (dependent on others for doing or taking full responsibility) across the "2" skill accomplishment level.

Discussion

The approach to category scaling presented in this study offers a way of enhancing differentiation among people at risk for not being able to perform IADLs. Achieving the most coherent, meaningful and psychometrically sound assessment necessitates

establishing a balance between fine differentiation, robustness and category functioning (Linacre, 2004).

The preferred solution for the *ALSAR* was a 6-point scale representing combinations of ratings of individual task accomplishment and extrinsic resource adequacy rather than simply adding scores for each item. The hierarchical scoring, although more complicated than simply adding skill and risk scores, seems to make more sense than the original additive scoring.

We have asserted that the validity of this scoring approach depends as much or more on the substantive relevance of the proposed hierarchy as on the psychometric properties. Despite the one disordered item, the preferred scale meets all the essential features of category functioning and considered along with other attributes indicates the scale is useful and valid for use with elders across a range of diagnosis. It was the best solution that also allowed maximum differentiation of people.

Presenting alternate scoring solutions for a scale can generate challenges when it comes to ensuring that the scale remains accessible to both clinicians and researchers. We want measures to provide as much information as possible but we need them to remain simple or they will be misunderstood or not used. In this case, different scoring may be needed for clinical and research purposes.

The present work has provided further validation of the *ALSAR* as a useful tool. In addition, we generated a new category rating scale and suggested that different scoring be used depending on whether the purpose is to assess clients for intervention planning or to measure the outcome of research. The next challenge is to find the best way of representing the new category rating scale, the 6-level solution, in the *ALSAR* tool format so that it can be easily scored by raters. Practicalities such as this will influence future acceptability and usage.

References

Andrich, D. A. (1996). Chapter 1 Measurement criteria for choosing among models for graded responses. In A. Von Eye and C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp. 3-35). Orlando, FL: Academic Press.

Clemson, L., Bundy, A., Unsworth, C., and Fiatarone Singh, M. (2006). Rasch analysis of the Assessment of Living Skills and Resources (ALSAR), an IADL measure for older people. *Manuscript submitted for publication*.

Doble, S. E., and Fisher, A. G. (1998). The dimensionality and validity of the Older Americans Resources and Services (OARS) Activities of Daily Living (ADL) Scale. *Journal of Outcome Measurement., 2*(1), 4-24.

Finlayson, M., Mallinson, T., and Barbosa, V. M. (2005). Activities of daily living (ADL) and instrumental activities of daily living (IADL) items were stable over time in a longitudinal study on aging. *Journal of Clinical Epidemiology., 58*(4), 338-349.

Hsueh, I. P., Wang, W. C., Sheu, C. F., and Hsieh, C. L. (2003). Rasch analysis of combining two indices to assess comprehensive ADL function in stroke patients. *Stroke., 35*(3), 721-726.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement., 3*(1), 85-106.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. J. Smith and R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 258-278). Maple Grove, Minnesota: JAM Press.

Linacre, J. M. (2006). *WINSTEPS Rasch measurement computer program Version 3.61*. Chicago: Winsteps.com.

Lopez, W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions, 10*, 482.

Williams, J. H., Drinka, T. J. K., Greenberg, J. R., Farrell-Holtan, J., Euhardy, R., and Schram, M. (1991). Development and testing of the assessment of living skills and resources (ALSAR) in elderly community-dwelling veterans. *The Gerontologist, 31*(1), 84-91.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Table 1. Original ALSAR scoring

SKILL **Individual task accomplishment is:**	**Scoring Options:** SKILL + RESOUCES	**RESOURCES** **Support for task completion extrinsic to individual is**:
0 Independent & consistent	0 1 2	0 Consistently available 1 Inconsistently available, unstable or unreliable 2 Insufficient or not used
1 Partial accomplishment	1 2 3	0 Consistently available 1 Inconsistently available, unstable or unreliable 2 Insufficient or not used
2 Not accomplished or no responsibility for doing	2 3 4	0 Consistently available 1 Inconsistently available, unstable or unreliable 2 Insufficient or not used

Table 2. New ALSAR scoring

SKILL **Individual task accomplishment is:**	**Scoring Options:** SKILL RESOURCES INCREMENTAL SCORE	**RESOURCES** **Support for task completion extrinsic to individual is**:
0 Independent & consistent	"00" = 0 "01" = 1 "02" = 2	0 Consistently available 1 Inconsistently available, unstable or unreliable 2 Insufficient or not used
1 Partial accomplishment	"10" = 3 "11" = 4 "12" = 5	0 Consistently available 1 Inconsistently available, unstable or unreliable 2 Insufficient or not used
2 Not accomplished or no responsibility for doing	"20" = 6 "21" = 7 "22" = 8	0 Consistently available 1 Inconsistently available, unstable or unreliable 2 Insufficient or not used

Table 3. Summary of category rating results for 9-category scale

Category		Observed count	Average measure	Outfit MeanSq	Structure calibration	Coherence	
Code	Score					M→C	C→M
00	0	728	1.12	1.06	None	83%	27%
01	1	66	.89	1.71	1.38	5%	27%
02	2	2	.59	0.15	2.71	0%	0%
10	3	686	.50	0.97	-6.40	46%	25%
11	4	321	.25	1.96	.41	16%	21%
12	5	20	.04	0.52	2.66	1%	40%
20	6	889	-.30	0.83	-3.67	58%	41%
21	7	280	-.56	1.19	1.58	27%	22%
22	8	182	-1.24	1.12	1.32	81%	19%

Table 4. Summary of category rating results for 6-category scale

Category		Observed count	Average measure	Outfit MeanSq	Structure calibration	Coherence	
Code	Score					M→C	C→M
00	0	783	2.37	0.93	None	81%	54%
10	1	675	1.28	0.78	-1.65	42%	50%
11	2	331	0.67	1.44	-0.27	16%	32%
20	3	884	-0.10	0.83	-1.28	58%	51%
21	4	269	-0.69	1.18	1.57	30%	32%
22	5	179	-1.66	1.20	1.63	81%	27%

Figure 1. Ordered category rating structure for 9-category scale

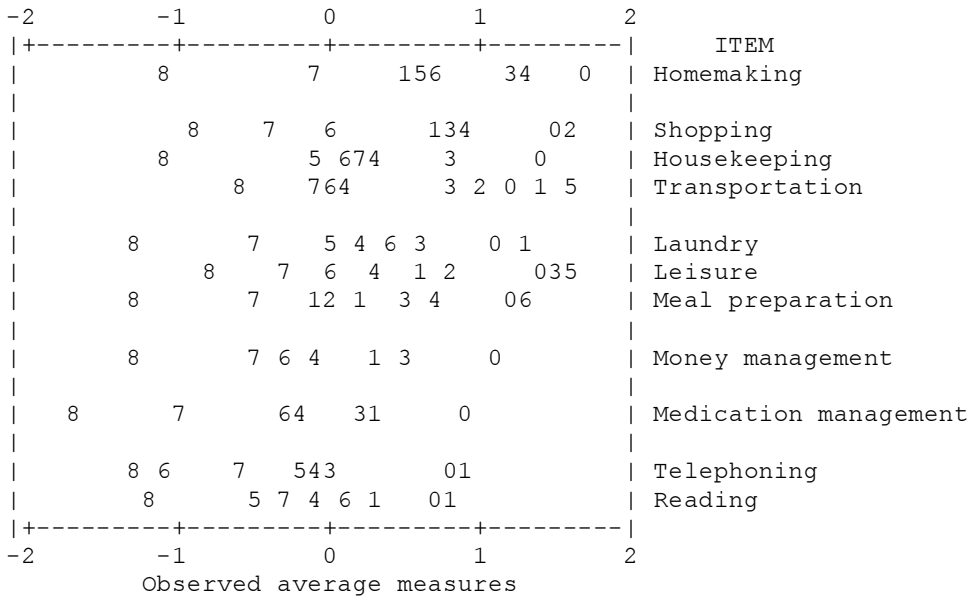

Figure 2 Ordered category rating structure for 6-category scale

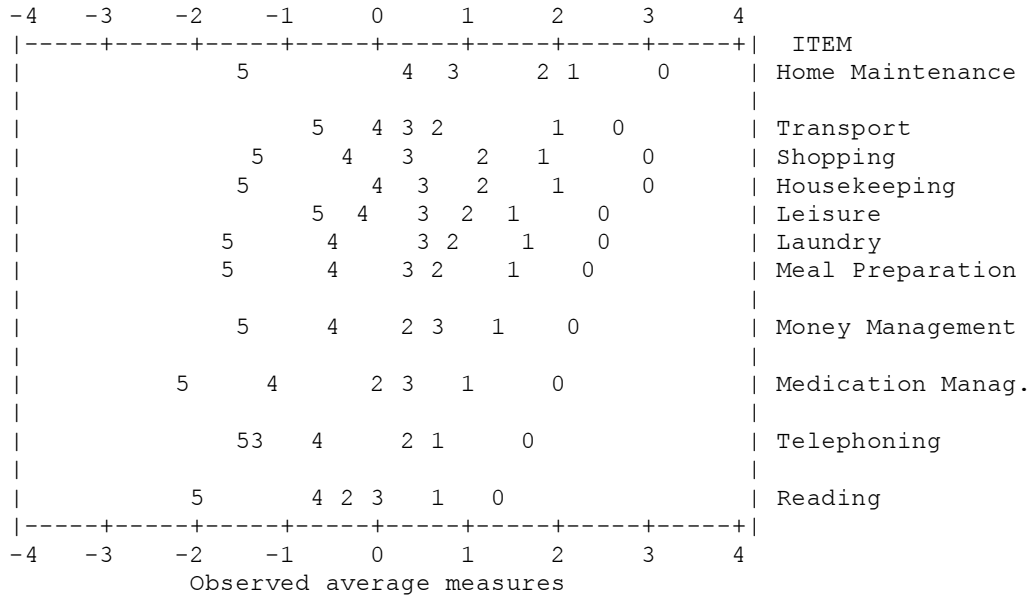

Figure 3: Category probability curves for the 9-category rating scale

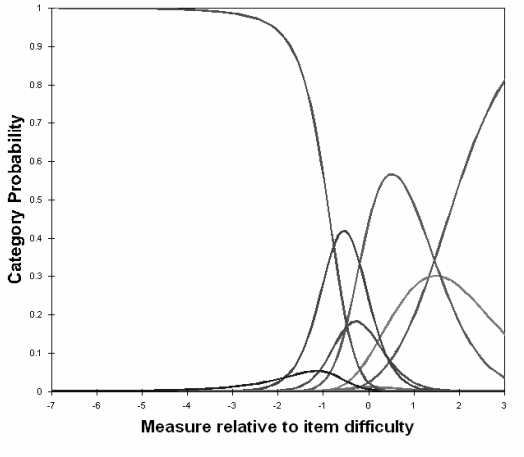

Figure 4: Category probability curves for the 6-category rating scale