Article

of Occupational Therapy

Hong Kong Journal of Occupational Therapy
2019, Vol. 32(1) 32–40

DOI: 10.1177/1569186119825885
journals.sagepub.com/home/hjo
SAGE

Reliability and validity of the Japanese version of the ADL-focused Occupation-based Neurobehavioural Evaluation (A-ONE J): Applying Rasch analysis methods

Yasuhiro Higashi[1,2], Shinichi Takabatake[2], Asako Matsubara[3], Koji Nishikawa[4], Hiroto Shigeta[5] and Guðrún Árnadóttir[6,7]

Abstract

Background/objective: The ADL-focused Occupation-based Neurobehavioral Evaluation (A-ONE) can be used to evaluate both performances of activities of daily living (ADL) tasks and neurobehavioural problems that interfere with ADL task performance among clients with neurological disorders. Research studies have demonstrated acceptable psychometric properties of the original version of the A-ONE as well as the Rasch analysed version. The aim of this study was to examine the reliability and validity of the Japanese version of the A-ONE (A-ONE J).

Methods: Rasch analysis was performed on data obtained from eight different hospitals in Japan on performances of 150 individuals diagnosed with a stroke based on the functional independence (FI) scale items. The rating scale structure was investigated and internal validity and reliability were examined. Unidimensionality of the items was examined by mean square infit values and principal component analysis of residuals. The targeting between person ability and item difficulty was explored, as well as the separation reliability. Finally, psychometric values and item difficulty hierarchies obtained in this study were compared to the original Rasch analysis of the A-ONE.

Results: The rating scale structure might be improved by collapsing two categories twice (from five categories to three categories). Unidimensionality of the items was obtained for 20 items. Targeting was acceptable, and separation reliability for item calibrations was high and acceptable for people.

Conclusion/limitations: This study provides important information regarding the possibilities for revising the ordinal A-ONE J FI Scale, converting it into a unidimensional scale. Further study with increased and more diverse sample is needed.

Keywords

Activities of daily living, assessment, cerebral vascular accidents, Rasch analysis

Received 6 March 2018; accepted 31 December 2018

Introduction

The evaluation and intervention of activities of daily living (ADL) are among the most common rehabilitation services provided by occupational therapists for people presenting with neurological disorders (Árnadóttir, 2010; Gillen, 2013a; Steultjens et al., 2003). Thus, the importance of evaluating neurobehavioural impairments (NBIs) that impact ADL performance in naturalistic contexts has gained increased support in the rehabilitation literature (Árnadóttir,

[1]Kansai Rehabilitation Hospital, Japan
[2]Graduate School of Comprehensive Rehabilitation, Osaka Prefecture University, Japan
[3]Hiroshima City Rehabilitation Hospital, Japan
[4]Ishikawa Prefectural Takamatsu Hospital, Japan
[5]Kita-Osaka Police Hospital, Japan
[6]The National University Hospital, Iceland
[7]University of Iceland, Iceland

Corresponding author:
Yasuhiro Higashi, Department of Rehabilitation, Kansai Rehabilitation Hospital, 3-11-1, Sakuranocho, Toyonaka, Osaka 560-0054, Japan.
Email: yasuhirohigashi2000@gmail.com

Löfgren, & Fisher, 2012). The ADL-focused Occupation-based Neurobehavioral Evaluation (A-ONE) was the first instrument developed within the discipline of occupational therapy to evaluate the impact of neurobehavioural impairments on task performance (Gillen, 2013b). Later, other instruments were designed to evaluate NBIs in naturalistic contexts. For example, these include the ADL test for those with apraxia (Goldenberg, Daumüller, & Hagmann, 2001), the Catherine Bergego Scale (Azouvi et al., 2003) for unilateral spatial neglect and the Moss Attention Rating Scale (Whyte, Hart, Bode, & Malec, 2003) for attention. However, in our review, no instruments other than the A-ONE were identified that could be used to evaluate the impact of a wide range of NBIs in a natural context on ADL task performance.

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 can identify the level of assistance required for ADL performance and the nature of underlying NBIs that interfere with ADL task performances. The A-ONE comprises two scales representing two different hypothetical constructs. Both scales, the Functional Independence Scale (FI scale) and the Neurobehavioural Scale (NB scale), were developed as criterion-referenced ordinal rating scales (Árnadóttir, 1990; Árnadóttir & Fisher, 2008).

The original purpose of the A-ONE was to use the ordinal scales to provide useful information for planning intervention, not to evaluate outcomes. Ordinal scale scores should not be used to measure outcomes, as these scores cannot be added as if they were interval scores needed to generate an overall total score (Árnadóttir & Fisher, 2008; Merbitz, Morris, & Grip, 1989). Treating ordinal data as interval data can result in incorrect conclusions about whether the outcomes of two treatment approaches or facilities are different (Velozo, Kielhofner, & Lai, 1999). To prevent this problem, the Rasch analysis has been used to evaluate the measurement properties of existing ordinal-level instruments. Thus, Árnadóttir and Fisher (2008) implemented the Rasch analysis of the FI scale of the A-ONE to examine the potential for using the evaluation as an outcome measure. Their results based on several different types of Rasch analyses provided support for that the ordinal ADL items could be converted to an interval scale, thus indicating potential for using the scale to measure change.

In Japan, the Japanese A-ONE study group developed the Japanese version of the A-ONE (A-ONE J) through translation of original items and definitions taking aim at cultural differences. This process is described in more detail under methods. The FI scale of A-ONE J was examined by a pilot study performed on a trial basis using the Rasch analysis with a small sample size ($n = 65$) and a limited variety of analyses. Five misfit items were detected (Higashi, Takabatake, Matsubara, Nishikawa, & Shigeta, 2016), indicating a need for further studies.

The Rasch analysis quantifies the interaction between a person's ability and a scale's individual item level of difficulty. It examines the extent to which observed scores fit with the expected scores under the Rasch model. A fundamental assumption of the Rasch model is that items follow an ordered hierarchy on a unidimensional scale. Furthermore, the model can be used to test whether ordinal-level scores approximate interval-level measurements on a linear scale after raw (ordinal) scores have been converted into equal units, termed "logits" (Tennant & Conaghan, 2007). Unidimensionality refers to the idea that items included on a scale must define a single construct that is represented by a hierarchy of items arranged from those easily performed to those that are hard to perform; thus, it supports the scale's internal validity (Bond & Fox, 2015). Unidimensionality can be examined by goodness-of-fit analyses based on the Rasch model (Linacre, 2016). Principal component analysis (PCA) of the residuals is also used to test unidimensionality and its underlying assumption that all data can be explained by the latent variable measured (Dickens, Rudd, Hallett, Ion, & Hardie, 2017; Linacre, 2016; Sick, 2011).

Goodness-of-fit analysis can also be used to examine the psychometric properties of the rating scale. It enables exploration of the categorisations of responses yielding higher quality measures than other categorisations (Bond & Fox, 2015).

Additional evidence for scale validity can be provided by targeting the items' difficulty to the abilities of the persons (Bond & Fox, 2015). Finally, reliability is evaluated in terms of separation defined as the ratio of the person (or item) true standard deviation to the error standard deviation (Bond & Fox, 2015). Item separation is used to verify the item hierarchy and reflects the number of strata of measures that are statistically discernible (Fisher, 1992; Morrone et al., 2017).

Five research questions were posed in accordance with questions in the preceding study of the Rasch analysis of the original FI scale of the A-ONE (Árnadóttir & Fisher, 2008). The first three questions address validity and the fourth addresses reliability. The last question relates to the comparison of the results obtained from the two studies:

1. Does the rating scale of the A-ONE J 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 FI scale of the A-ONE J define a single unidimensional construct, as evidenced by goodness-of-fit and PCA?
3. Are the items of the FI scale appropriately targeted to participants who experienced cerebral vascular accidents (CVAs)?
4. Do the items on the FI scale separate participants into different levels of abilities when evaluating those with CVA, and do the participants tested separate the items into different levels of difficulty?
5. Are the psychometric qualities and the item hierarchies of the original Rasch analysed A-ONE version and the A-ONE J comparable?

Methods

Participants

This was a multicentre study conducted between October 2015 and March 2017. The participants were recruited from eight different acute and rehabilitation hospitals in Japan where A-ONE trained occupational therapists worked. Ten therapists and their attending physicians selected 150 participants for inclusion based on: (a) the presence or suspicion of cognitive or perceptual dysfunction as a result of CVA, revealed by a medical examination, and (b) the person's medical readiness for an ADL evaluation. The exclusion criteria pertained to those that were not medically stable or not able to perform any ADL task. The detailed participant demographic information is presented in Table 1. 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) (Linacre, 1994; Morrone et al., 2017).

Table 1. Demographic information of participants.

	RCVA (*n*=68)	LCVA (*n*=64)	Both (*n*=18)	Total (*n*=150)
Age				
M	74.1	72.8	71.8	73.3
SD	12.2	12.4	10.4	12
Range	41–96	39–93	49–87	39–96
Gender				
Male	43	37	11	91
Female	25	27	7	59
Diagnosis				
Infarction	45	37	12	94
Hemorrhage	23	27	6	56
Days after onset				
M	77.7	66	85.7	73.7
SD	58.9	52.2	56.6	55.9
Range	2–298	4–215	25–187	2–298
Br. Stage (arm)				
Median[a]	3.5	5	5	4.5
Range[a]	1–6	1–6	1–6	1–6
Br. Stage (hand)				
Median[a]	4	5	5	5
Range[a]	1–6	2–6	2–6	1–6
Br. Stage (leg)				
Median[a]	4	5	5	5
Range[a]	1–6	2–6	2–6	1–6

RCVA: right cerebral vascular accident; LCVA: left cerebral vascular accident; Both: both right and left cerebral vascular accident; Br. stage: Brunnstrom stage.
[a]Converted Roman numerals to Arabic numbers.

Instrumentation

The A-ONE is commonly used for adults that have acquired central nervous system dysfunction. The FI scale consists of 20 ADL items and two communication items. It is a 5-category ordinal rating scale ranging from 0 to 4 used to score the observed level of assistance needed to overcome the impact of impairment on ADL performance (0 = full assistance needed, 1 = minimum to considerable physical assistance needed, 2 = verbal assistance needed, 3 = supervision needed and 4 = independent). The NB scale is used to evaluate the impact of NBIs that interfere with ADL task performances. To use the A-ONE reliably in clinical practice and research, a five-day training course emphasising clinical reasoning is essential (Árnadóttir, 2010; Árnadóttir & Fisher, 2008; Árnadóttir, Fisher, & Löfgren 2009).

In Japan, the Japanese A-ONE study group developed the Japanese version of the A-ONE through translation of original items and definitions taking aim at cultural differences. The translation process included four steps. These included: (1) Forward translation from English to Japanese performed separately by two translators, who both are members of the study group (both had prior experience in translating educational materials and expertise in neurological occupational therapy); (2) Then 10 Japan A-ONE Study Group members (some were educated in English speaking countries with prior experience in translating educational materials) compared and discussed the translations and any discrepancies in focus-group meetings; (3) Subsequently the content of A-ONE J and cultural differences were discussed with five Japan A-ONE Study Group members and the author of the A-ONE. Back translation was not considered necessary as the instrument is an observation tool (not a questionnaire) intended for use by specifically trained therapists that have completed a 40-h training course consisting of instruction and supervised scoring

practice; (4) Finally, the FI scale of A-ONE J was examined by a pilot study performed on a trial basis as mentioned before (Higashi et al., 2016).

Two changes were made to the FI scale of the original A-ONE to accommodate the Japanese culture. The item of 'Wash face and upper body' was changed to (simplified) 'Wash face and hands' and the item of 'Use knife to cut and spread' was changed to 'Use chopsticks to manipulate and carry'. There were no other changes from the original A-ONE (Higashi et al., 2016). All items of the FI scale in the A-ONE J are listed in Appendix 1.

Procedures and data analysis

The participants were evaluated in the different hospitals by therapists who had completed a five-day A-ONE training course and administered the evaluations according to the standardised procedures described in the manual. The study was approved by the School of Comprehensive Rehabilitation Osaka Prefecture University as well as the Research Ethics Committees of each hospital.

The raw scores were analysed using the WINSTEPS Rasch computer software program (Version 4.0.0) (Linacre, 2017). The analysis was divided into two phases in accordance with the previous Rasch analytic study of the A-ONE (Árnadóttir & Fisher, 2008). As the Rasch analysis of the original A-ONE FI scale demonstrated that the 20 ADL items formed one construct but not the two communication items, and the preliminary study of the A-ONE J supported that finding, only the 20 ADL items excluding the two communication items from the FI scale were included in this study.

Phase 1: Rating scale analysis

The psychometric properties of the 5-category rating scale were examined as suggested by Lopez (1996) as well as Bond and Fox (2015) before investigating the validity and reliability of the scaled items. We used the criteria of the outfit mean square (MnSq) value within 2.0 and advancement by at least 1.4 logits of the calibration thresholds between the rating scale categories (Bond & Fox, 2015; Linacre, 2002). If these set criteria were not met, we were prepared to collapse non-advancing categories, and subsequently, reanalyse the data.

Phase 2: Internal scale validity and reliability analysis

Step 1: Construct validation. The range of fit statistics differs according to test characteristics. In the clinical observation type of assessment, MnSq > 1.7 associated with Zstd > 2.0 indicates a misfit, that is, a problem with the internal consistency of test items, the ability pattern of participants, or the measurement pattern (Bond & Fox, 2015).The Rasch analysis includes two types of MnSq values, infit MnSq and outfit MnSq. We based our decision of item removal on the infit MnSq values. Infit is an information-weighted indicator of misfit. The outfit statistic is not weighted and therefore remains relatively more sensitive to the influence of outlying scores: the performances of persons distant from the item's location. Aberrant infit statistics usually cause more concern than do large outfit statistics (Bond & Fox, 2015). We also focused on high MnSq values, as low MnSq are not likely to have practical implications (Bond & Fox, 2015).

We evaluated unidimensionality by means of the PCA. We used the Fisher's (2007) five-level quality criteria for rating scale instruments and aimed for good quality. Thus, the required proportion of variance explained by the measures (Rasch dimension) needed to be >50% and the proportion of unexplained variance accounted for by the first contrast (the largest secondary dimension) needed to be <5%–10% for the results to support unidimensionality.

Step 2: Targeting. We also examined whether the items were located at the targeted difficulty levels to capture the range of participant abilities in the sample. Targeting refers to the difference between the average item measure and the average person measure, using average error as the unit of comparison (Fisher, 2007). In addition, comments on the quality of the rating scales were made based on the ceiling effect (percentage of scores at the maximum possible scores) and floor effect (percentage of scores at the minimum possible scores) according to the rating scale instrument quality criteria (Fisher, 2007; Lim, Rodger, & Brown, 2010).

Step 3: Reliability. The person reliability index refers to the replicability of the person ability logit score. The item reliability index indicated the replicability of the item along the pathway when these same items were given to another sample of a similar size that behaved in the same manner (Bond & Fox, 2015). A reliability value of >0.8 is usually the goal, as it allows for distinguishing at least three strata (i.e. a reliability value that allows discerning of three significantly different levels of measures in the sample), and a separation index of 2.0 indicates that the persons or items on the scale define at least three levels of difficulties or abilities (Caronni, Sciumè, Donzelli, Zaina, & Negrini, 2017; Fisher, 1992).

Finally, obtained values of psychometric properties in both phases of the present study were compared to the findings of the original Rasch analysis of the

A-ONE FI scale. Sequence of item difficulty on the hierarchies from both studies was further examined.

Results

Phase 1: Rating scale analysis

Investigation of the 5-category rating scale revealed an interval of <1.4 logits for the category measures from 2 to 3 and 3 to 4 (see Table 2). Subsequently, several possibilities and combinations of collapsing categories were attempted. As a 4-category scale did not solve the rating scale problem (outfit MnSq was over 2.0), other possibilities were tested, which resulted in a 3-category scale. As planned, we collapsed the non-advancing categories by combining categories 0 and 1, and 2 and 3 (Table 2, Figure 1). This resulted in an increased person separation index from 3.49 to 3.58. The modelled category probability curves and distribution of persons and items for each rating category on the scales are presented in Figure 1.

Phase 2: Internal validity and reliability analysis

Step 1: Construct validation. We used the new 3-category rating scale and examined the item goodness-of-fit to the Rasch rating scale model. When data from all 20 ADL items of the FI scale were included, no misfit items were detected. PCA results from analysing the unidimensionality of the 20 items revealed that 70.9% of the total variance was explained by the measures and that 4.4% of the unexplained variance was accounted for by the first contrast. See Table 3 for the item measurement report.

Table 2. Rating scale category statistics (20 items, 5 categories, 4 categories, 3 categories).

Score	Frequency (%)	Outfit MnSq	Calibration threshold	Category measure
5-Category scale				
0	50 (2)	**2.38**	NONE	−5.16
1	737 (25)	1.26	−4.06	−1.49
2	351 (12)	1.34	1.19	0.93
3	356 (12)	0.61	1.48	1.80
4	1506 (50)	0.97	1.39	2.99
4-Category scale				
0	50 (2)	**2.53**	NONE	−5.79
1	737 (25)	1.29	−4.69	−1.66
2	707 (24)	0.80	1.38	2.35
3	1506 (50)	0.96	3.31	4.51
3-Category scale				
1	787 (26)	1.70	NONE	−2.22
2	707 (25)	0.85	−1.03	0.00
3	1506 (52)	1.04	1.03	2.22

Note: Misfit categories appear in bold.

Step 2: Targeting. The item difficulty logits of the 20 items ranged from −4.27 to 3.97 and the person ability logits ranged from −7.32 to 6.98. The mean person ability measure was 1.26 logits and the targeting error was 1.83. The floor effect was 2% (3 participants received the minimum score), and the ceiling effect was 6.7% (10 participants achieved the maximum score).

Step 3: Reliability. Reliability analysis revealed a person separation index of 3.58 and separation reliability coefficient of 0.93. The item separation index was 9.69 and separation reliability coefficient was 0.99.

Table 4 summarises the psychometric qualities of both the original four category Rasch version of the A-ONE and the three-category version of A-ONE J. Item order on the A-ONE J scale as presented in Table 3 was like the original Rasch A-ONE hierarchy for most items. However, 'Wash Face and Hands' and "Use chopsticks" became considerably easier on the A-ONE J, and all the transfer items became considerably harder to perform.

Discussion

Rating scale analysis

In the preceding study (Árnadóttir & Fisher, 2008), categories "2" (verbal assistance) and "3" (supervision) were collapsed, resulting in a 4-category rating scale. However, in our study, the category of "0" (full assistance) and "1" (physical assistance) were collapsed, and subsequently, the categories of "2" (verbal assistance) and "3" (supervision) were also collapsed, resulting in a 3-category rating scale. While in the preceding study, 95 participants with CVA were evaluated, in addition to 111 participants with dementia, all 150 participants in our study were diagnosed with CVA. This difference in sample composition may have contributed to the different category structures obtained from the rating scale. According to literature (Bond & Fox, 2015; Linacre, 2002), misfitting category value may be caused by several reasons. For example, therapists may have a problem discriminating between specific categories or there may be a rarely used category. In the A-ONE J, it may be difficult to differentiate between "0" (full assistance) and "1" (minimal to considerable physical assistance) because both categories involve physical assistance. Further, the discrepancy may also be due to that the category of zero was seldom used (50 times vs. 130 times in the original study), which could relate to limitations in sample size. Thus, although Figure 1 looks promising for a 4-category rating scale of the A-ONE J, outfit misfit was detected for the "0" category in Table 2.

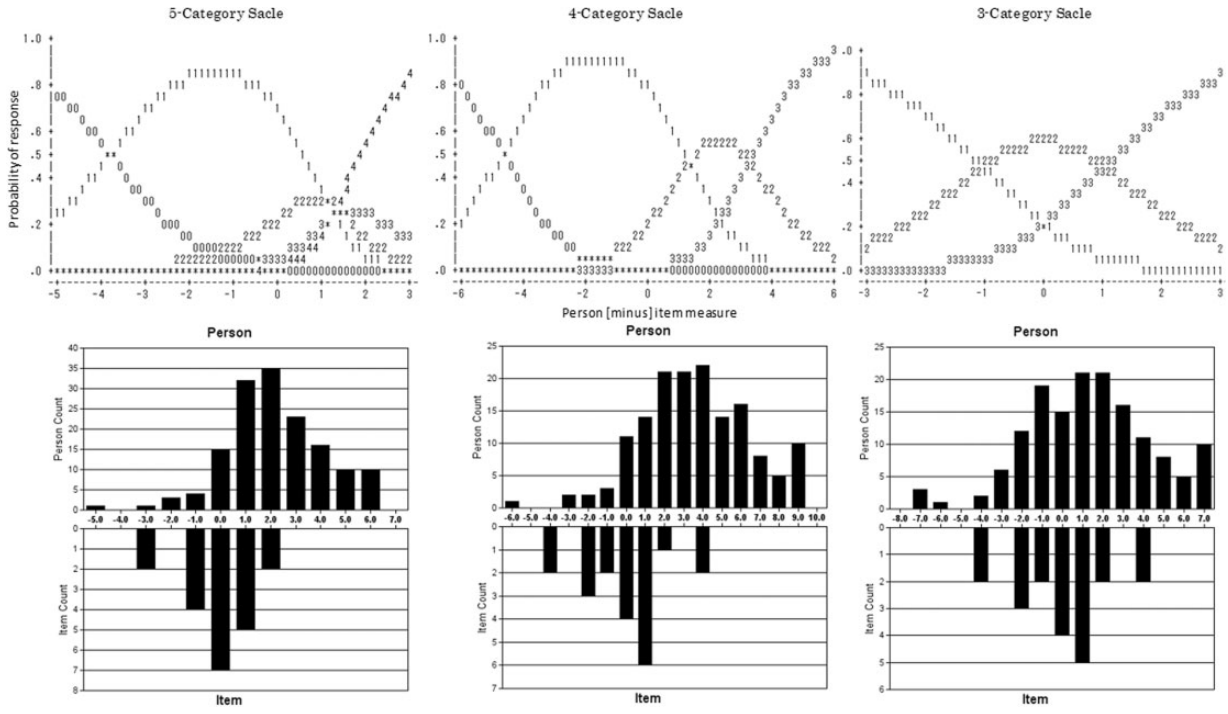

Figure 1. Modelled category probability curves and distribution of persons and items for each rating category scale.

Table 3. Item measurement report of all items and the three rating scale categories.

Item	Item difficulty (logits)	SE	infit		Pt measure-A	
			MnSq	Zst	Corr.	Exp.
Bathe	3.97	0.21	0.73	−1.8	0.75	0.72
Transfer to tub	3.72	0.20	1.35	2.1	0.68	0.73
Put on pants	1.69	0.17	0.71	−2.5	0.81	0.78
Perform toilet hygiene	1.55	0.17	0.80	−1.6	0.81	0.78
Transfer to toilet	1.28	0.17	0.59	−3.7	0.84	0.78
Manoeuvre around	1.11	0.17	1.00	0.0	0.77	0.78
Put on shirt	1.05	0.17	1.02	0.2	0.78	0.77
Put on socks	0.79	0.17	0.98	−0.1	0.77	0.77
Transfer from sitting	0.73	0.17	0.54	−4.4	0.84	0.77
Manipulate fastenings	0.46	0.17	1.68	4.5	0.69	0.77
Put on shoes	0.11	0.17	0.72	−2.4	0.80	0.76
Brush teeth	−0.33	0.18	0.88	−0.9	0.76	0.75
Shave beard or apply cosmetics	−0.33	0.18	1.49	3.3	0.68	0.75
Wash face and hands	−0.72	0.18	1.10	0.8	0.71	0.73
Sit up in bed	−0.75	0.18	0.89	−0.7	0.75	0.73
Comb hair	−1.72	0.20	1.18	1.2	0.66	0.69
Use chopsticks	−1.76	0.21	1.57	3.2	0.60	0.69
Bring food to mouth by spoon	−2.31	0.22	0.87	−0.7	0.68	0.66
Drink from glass or cup	−4.27	0.35	0.98	0.0	0.55	0.55
Use fingers to bring food to mouth	−4.27	0.35	0.85	−0.5	0.56	0.55

The fact that the evaluation was performed many weeks (M = 10.5 weeks) after the onset of stroke may also have diminished the need for use of the "0" category on the scale. Moreover, we detected that the ability of the participants was relatively high according to Figure 1. Thus, the distribution of the participants' ability might be another reason for category misfit. Problems with differentiating between the categories

Table 4. Comparison of the A-ONE original and A-ONE J scales.

	FI scale of A-ONE J	A-ONE original FI scale	Criteria	Comparison of results
Diagnostic groups	CVA (150)	CVA (95), Dementia (111)	–	Less diversity for A-ONE J
Persons (n)	150	209	150	Both samples met the criteria
Items (i)	20	20	–	One item fewer on A-ONE J
Omitted items	0 ADL	0 ADL	–	None
Misfitting items	0	1 = 5%	≤5%	All final items met the criteria
Number of categories	3	4	–	Reduced by 1 from original
Item model fit MnSq	≤1.7	≤1.4		Good/very good[a]
PCA: Rasch factor	70.9%	84%	>50%	Very good/excellent[a]
PCA: First contrast	4.4%	3.6%	<5%–10%	Very good/very good[a]
Person separation index	3.58 = 4–5 strata	2.93 = 4 strata	>2	Very good/very good[a]
R (persons)	0.93	0.9	>0.8	Very good/good[a]
Item separation index	9.69	8.02	>2	Excellent/excellent[a]
R (items)	0.99	0.98	>0.8	Excellent/excellent[a]
Targeting error	1.83	3.22	<2	Fair/poor[a]
Ceiling effect	10 (6.7%)	9 (4.3%)	<5%	Poor/fair[a]
Floor effect	3 (2%)	0 (0%)	<5%	Good/excellent[a]

[a]Interpretation based on Fisher's (2007) quality criteria for rating scale instrument.
Note: Comparison of psychometric qualities of Rasch analysed versions of the ADL construct on the A-ONE original and A-ONE J scales.

of "2" (verbal assistance) referring to a verbal assistance due to item performance, and "3" (supervision), which may include a non-item specific general verbal cue, may be due to the possible presence of a verbal component in both categories, according to Árnadóttir and Fisher (2008).

Internal validity

The 20 ADL items fitted with the expected scores under the Rasch model. And PCA confirmed unidimensionality of the 20-item scale. Therefore, it can be concluded that the scale can be used as a psychometrically sound outcome measure by using the total score of the 20 items on a 3-category rating scale.

The identified error figure in targeting was 1.83. Ten participants (6.7%) achieved the maximum score and three participants (2%) received the minimum score. According to the criteria (Fisher, 2007), the targeting error can be classified as fair; the ceiling effect was poor and the floor effect was good. These findings were consistent with the preceding study (Árnadóttir & Fisher, 2008). As long as the A-ONE J targets participants who cannot perform ADL independently, the ceiling effect may not result in any clinical problems. However, it might be possible to add more difficult items such as Instrumental ADL items to reduce the ceiling effect, as suggested by Árnadóttir and Fisher (2008).

When comparing the item hierarchies on the A-ONE J and the original A-ONE Rash analysed version of the FI scale, it is noted that most items remain in a similar location on the hierarchy. However, "Washing face and hands" became less difficult, which could be explained by the simplification of the item. "Use chopsticks" also became considerably less difficult than "Use knife", which could be related to the different characteristics and requirements of objects used. Knife requires bilateral hand use, but chopsticks can be manipulated with only one hand, thus using knife could be more difficult than chopsticks for persons with CVA. Further, all transfer items became more difficult on the A-ONE J scale, which probably relates to differences in the composition of the sample and resulting impairments with a higher percentage of the Japanese sample having paralyses than on the original sample where people with dementia were also included.

In this study, interrater reliability of the A-ONE J was not examined. All evaluators completed the A-ONE training course and were certified. Thus, they had learned how to administer and score the A-ONE J. However, it is necessary to examine interrater reliability of the A-ONE J in the future.

A potential limitation of this study influencing the misfitting rating scale category on the four category version of the scale could be that the sample size may have been too small. Thus, although a sample size of 150 should be sufficient for an exploratory study, this may have contributed to infrequent use of categories as larger samples are likely to produce more stable results (Bond & Fox, 2015). Another limitation possibly affecting item hierarchies might be as mentioned earlier that the sample only included people with CVA.

Therefore, our recommendations for future studies include expansion of sample size and composition including more diverse diagnostic categories, aiming for results that would allow same number of categories and items as included on the Rasch analysed version of the original A-ONE.

Conclusion

This study provides an important first step in exploring the revision potential of the ordinal FI scale of the A-ONE J into a single interval scale as required for an outcome measure. Construct validity of a reliable 20-item measure with a 3-category rating scale was obtained, although the results were not identical to the original 4-category Rasch version of the A-ONE. It would be worthwhile to study further the effect of the different composition of the diagnostic sample groups on the original A-ONE version and the A-ONE J with an increased sample size and cross validation of results.

Acknowledgements

Clients who participated as research participants are thanked for their valuable contribution to this study. We would also like to acknowledge the occupational therapists that used the A-ONE J to evaluate their clients.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was supported by grants from the Japanese Association of Occupational Therapists in 2016.

References

Árnadóttir, G. (1990). *The brain and behavior: Assessing cortical dysfunction through activities of daily living*. St. Louis: Mosby.

Árnadóttir, G. (2010). *Measuring the impact of body functions on occupational performance*: *Validation of the ADL-focused Occupation-based Neurobehavioral Evaluation (A-ONE)*. Doctoral dissertation, Umeå University, Sweden. [Medical Dissertations, New series no. 1322; 2010].

Árnadóttir, G., & Fisher, A. G. (2008). Rasch analysis of the ADL scale of the A-ONE. *American Journal of Occupational Therapy*, *62*, 51–60.

Árnadóttir, G., Fisher, A. G., & Löfgren B. (2009). Dimensionality of nonmotor neurobehavioral impairments when observed in the natural contexts of ADL task performance. *Neurorehabilitation and Neural Repair*, *23*, 579–586.

Árnadóttir, G., Löfgren, B., & Fisher, A. G. (2012). Neurobehavioral functions evaluated in naturalistic contexts: Rasch analysis of the A-ONE neurobehavioral impact scale. *Scandinavian Journal of Occupational Therapy*, *19*, 439–449.

Azouvi, P., Olivier, S., de Montety, G., Samuel, C., Louis-Dreyfus, A., & Tesio, L. (2003). Behavioral assessment of unilateral neglect: Study of the psychometric properties of the Catherine Bergego Scale. *Archives of Physical Medicine and Rehabilitation*, *84*, 51–57.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Caronni, A., Sciumè, L., Donzelli, S., Zaina, F., & Negrini, S. (2017). ISYQOL: A Rasch-consistent questionnaire for measuring health-related quality of life in adolescents with spinal deformities. *Spine Journal*, *17*, 1364–1372.

Dickens, G. L., Rudd, B., Hallett, N., Ion, R. M., & Hardie, S. M. (2017). Factor validation and Rasch analysis of the individual recovery outcomes counter. *Disability and Rehabilitation*, *11*, 1–12.

Fisher, W. P. J. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*, 238. Retrieved from www.rasch.org/rmt/rmt63i.htm

Fisher, W. P. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, 21, 1095. Retrieved from www.rasch.org/rmt/rmt211m.htm

Gillen, G. (2013a). *Cerebrovascular accident/stroke*. In H. M. Pendleton & W. Schultz-Krohn (Eds), *Pedretti's occupational therapy: Practice skills for physical dysfunction* (7th ed., pp. 844–880). St. Louis, MO: Mosby Elsevier.

Gillen, G. (2013b). A fork in the road: An occupational hazard? (Eleanor Clarke Slagle Lecture). *American Journal of Occupational Therapy*, *67*, 641–652.

Goldenberg, G., Daumüller, M., & Hagmann, S. (2001). Assessment and therapy of complex activities of daily living in apraxia. *Neuropsychological Rehabilitation*, *11*, 147–169

Higashi, Y., Takabatake, S., Matsubara, A., Nishikawa, K., & Shigeta, H. (2016, December). *A study of internal validity and item difficulty of the Japanese version of the ADL-focused occupation-based neurobehavioral evaluation*. Poster session presented at the 41st Annual Scientific Meeting of Japan Society for Higher Brain Dysfunction, Nagano, Japan

Lim, S. M., Rodger, S., & Brown, G. (2010). Validation of child behavior rating scale in Singapore (Part 1): Rasch analysis. *Hong Kong Journal of Occupational Therapy*, *20*, 52–62.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328.

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

Linacre, J. M. (2016). A user's guide to WINSTEPS MINISTEP Rasch model computer programs: Program manual 3.92.0., 599-606. Retrieved from www.winsteps.com/manuals.htm

Linacre, J. M. (2017). *Winsteps (Version 4.0.0) [Computer software]*. Chicago, IL: Winsteps.com.

Lopez, W. (1996). Communication, validity and rating scales. *Rasch Measurement Transactions*, *10*, 482–483.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, *70*, 308–312.

Morrone, E., Sguazzin, C., Bertolotti, G., Giordano, A., Braghiroli, A., Balestroni, G. L., ... Fanfulla F. (2017). Development and validation of the Maugeri Sleep Quality and Distress Inventory (MaSQuDI- 17). *PLoS One*, *12*, e0180743.

Sick, J. (2011). Rasch measurement in language education part 6: Rasch measurement and factor analysis. *Shiken*, *15*, 15–17.

Steultjens, E. M. J., Dekker, J., Bouter, L. M., van de Nes, J. C. M., Cup, E. H. C., & van den Ende, C. H. M. (2003). Occupational therapy for stroke patients: A systematic review. *Stroke*, *34*, 676–687.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism*, *57*, 1358–1362.

Velozo, C. A., Kielhofner, G., & Lai, J.-S. (1999). Quantitative research series – The use of Rasch analysis to produce scale-free measurement of functional ability. *American Journal of Occupational Therapy*, *53*, 83–90.

Whyte, J., Hart, T., Bode, R. K., & Malec, J. F. (2003). The moss attention rating scale for traumatic brain injury: Initial psychometric assessment. *Archives of Physical Medicine and Rehabilitation*, *84*, 268–276.

Appendix 1. FI Scale Items on the A-ONE J.

ADL CONSTRUCT ITEMS	
DRESSING DOMAIN	TRANSFER AND MOBILITY DOMAIN
Put on shirt	Sit up in bed
Put on pants	Transfer from sitting
Put on socks	Maneuver around
Put on shoes	Transfer to toilet
Manipulate fastenings	Transfer to tub
GROOMING AND HYGIENE DOMAIN	FEEDING DOMAIN
Wash face and hands	Drink from a glass/cup
Comb hair	Use fingers to bring food to mouth
Brush teeth	Bring food to mouth by spoon
Shave beard/apply cosmetics	Use chopsticks
Perform toilet hygiene	
Bathe	
COMMUNICATION CONSTRUCT ITEMS[a]	
Comprehension	
Expression	

[a]The communication items on the FI scale did not fit the ADL construct in the original Rasch analysed version of the A-ONE.