Quality of Life Research (2019) 28:1627–1639
https://doi.org/10.1007/s11136-019-02146-4

Supporting construct validity of the Evaluation of Daily Activity Questionnaire using Linear Logistic Test Models

Núria Duran Adroher[1,2] · **Alan Tennant**[1,2]

Accepted: 21 February 2019 / Published online: 9 March 2019

Abstract

Purpose Construct validity is commonly assessed by applying statistical methods to data. However, purely empirical methods cannot explain what happens between the attribute and the instrument scores, which is the core of construct validity. Linear Logistic Test Models (LLTMs) can provide such explanation by decomposing item difficulties into a weighted sum of theoretical item properties. In this study, we aim to support construct validity of the Evaluation of Daily Activity Questionnaire (EDAQ) by using item properties accounting for item difficulties.

Methods Dichotomized responses to the EDAQ were analyzed with (1) the Rasch model (to estimate item difficulties), and (2) LLTMs (to predict item difficulties). Seven properties of the items were identified and rated in ordinal scales by 39 Occupational Therapists worldwide. Aggregated metric estimates—the weights used to predict item difficulties in LLTMs—were derived from the ratings using seven cumulative link mixed models. Estimated and predicted item difficulties were compared.

Results The Rasch model showed acceptable fit and unidimensionality for a sample of 42 locally independent EDAQ items. The LLTM plus error showed significantly better fit than the LLTM. In the former, three of the seven properties were not significant, and the corresponding model including only the significant properties was used to predict item difficulties; they explained 77.5% of the variance in estimated item difficulties.

Conclusion A satisfactory theoretical explanation of what makes an activity of daily living task more difficult than another has been provided by a LLTM plus error model, therefore supporting construct validity of the EDAQ.

Keywords Construct validity · Activities of daily living · Linear logistic test model · Cumulative link mixed model · Rasch measurement

Introduction

Although the common understanding of construct validity refers to the extent to which a test measures what it purports to measure [1, 2], there are several opinions on the notion of construct validity and on the ways to be assessed. The concept of 'construct validity' was conceived in 1954 by the American Psychological Association [3], and it was further elaborated by Cronbach and Meehl [4]. Their definition of construct validity based on nomological networks—a system of laws relating theoretical concepts to each other and to observations—is still prevailing, but some authors consider that assessing validity in this way is insufficient to conclude that a test is valid [5]: Despite recognizing the importance of assessing relations between the measured attribute and other attributes, they claim that the primary concern of validation should be on construct representation [6].

Construct representation consists of analyzing the psychological processes and theoretical mechanisms underlying task performance [2, 7]. Hence, it is concerned with task variability rather than subject variability [6]. Little attention has been given to this kind of analysis in health sciences [2]. To validate a test, instead of developing formal theories which could explain the processes between the attribute and the test scores, most of the studies do it empirically by applying statistical methods to data. However, purely empirical methods or nomological networks cannot provide an explanation of how these processes work [5, 8]; such explanation must be given by substantive theory.

✉ Núria Duran Adroher
nuria.duranadroher@paraplegie.ch

[1] Swiss Paraplegic Research, Guido A. Zäch Strasse 4, 6207 Nottwil, Switzerland

[2] Department of Health Sciences and Health Policy, University of Lucerne, Frohburgstrasse 3, 6002 Lucerne, Switzerland

Rasch Measurement [9] consists of a series of models determining the probability of endorsing an item in terms of a person parameter (denoted as 'ability') and an item parameter (denoted as 'difficulty'). After fitting a Rasch model, estimates of person abilities and item difficulties are obtained so that persons are sorted from the 'least able' to the 'most able,' and items from the 'easiest' to the 'most difficult.' Thus, Rasch Measurement postulates a probabilistic relationship between person ability and item difficulty. However, the model is descriptive, i.e., it does not embody a construct theory, providing no explanation on the mechanisms determining item difficulties—i.e., which underlying mechanisms make an item be more difficult than another.

Two different modeling approaches provided an explanation of the mechanisms determining item difficulties. Stenner et al. developed a methodology specifying such mechanisms in two particular constructs, short-term memory [10] and receptive vocabulary [11]. Basically, they built a construct specification equation by regressing item difficulties (derived from the Rasch model) on selected characteristics of the items. For short-term memory, they used data from the Knox Cube Test [12], and they concluded that two item characteristics—distance covered and number of taps—govern item difficulties.

Earlier, Fischer had developed an extension of the Rasch model, the Linear Logistic Test Model (LLTM) [13], in which the item difficulties are determined theoretically by means of a linear combination of several basic parameters. These basic parameters usually represent processes or cognitive operations defined by theory, and they are given weights. In many cases, these weights are specified either dichotomously (present or not) or by the number of times a certain operation is present [14]. There exists an interesting extension of the LLTM model, LLTM plus error (LLTM + ε) [15], and we refer to both LLTM and LLTM + ε as LLTMs. Embretson [6] shows that the LLTM fulfills the four criteria she defined for assessing construct representation. Moreover, LLTMs are considered as a method of construct validation [14, 16, 17]. Several applications of LLTMs are found in education [13, 18] or cognitive psychology [19], but to our knowledge, there are no applications of LLTMs to explain the item difficulty variability of Activities of Daily Living (ADL) instruments.

Soon et al. [20] considered a number of ADL activities (previously determined in an ADL taxonomy) and established an ordered structure of actions within each activity. Although there is evidence for such a hierarchy, we did not find any study trying to explain what makes a task more difficult to perform than another. On the other hand, Fong et al. [21] argued that understanding the contributions of physical and cognitive demands to be able to carry out a task could provide more accurate knowledge for predicting independence, detecting early functioning decline, addressing patient safety, and tailoring treatment; they gave numerical ratings to the cognitive demand and to the physical demand of some tasks, and they argued that this information could help in clinical decision making and planning in rehabilitation. However, they did not assess if these potential predictors (cognitive and physical demand) could help to explain the variability in task difficulty, choosing also not to sort the tasks from easy to difficult.

In this study, we wanted to not only establish a hierarchy from easier to more demanding ADL task, but try to explain why the tasks are sorted in this way considering intrinsic properties of the tasks (such as cognitive and physical demand) as predictors of the item difficulties in the LLTM and LLTM + ε models. These intrinsic properties would constitute the theoretical framework necessary to assess the validity of an instrument assessing ADL such as the Evaluation of Daily Activity Questionnaire (EDAQ). Hence, the objective of this study is to support construct validity of the EDAQ by using item properties accounting for item difficulties. Specifically, we aim (1) to identify potential predictors of the hierarchical order in difficulty of the items that make up the EDAQ, (2) to derive a reliable weight for each predictor and item, and (3) to explain the hierarchical order from a theoretical perspective.

Methods

Instrument

The EDAQ is a Swedish Patient-Reported Outcome Measure consisting of ADL items. The linguistically validated English version contains 138 items of 14 domains [22]: eating and drinking; in the bathroom and personal care; dressing; bathing and showering; cooking; moving indoors; cleaning the house; laundry and clothes care; moving and transfers; communication; moving outdoors and shopping; gardening and house maintenance; caring; leisure and social activities. The last two domains (consisting of 18 items) were not considered in the LLTMs because the activities described are very broad and do not correspond to tasks as such. The remaining 12 domains (120 items) relate to chapters d4 (Mobility), d5 (Self-care), and d6 (Domestic Life) of the International Classification of Functioning, Disability, and Health (ICF) [23].

The EDAQ was administered to 383 adults with Rheumatoid Arthritis. They had to indicate their ability to carry out each of the activities. The answers to all items were in an ordinal scale '0 = without difficulty', '1 = with some difficulty', '2 = with much difficulty', '3 = unable to do'. For each item, they had to fill in two sections, (A) 'How do you do the activity without using an aid/gadget, alternate method or help?' and (B) 'How else do you do it with an aid/gadget or

alternate method?'. In ICF's terminology, A corresponds to capacity and B to performance. For this study, we analyzed the capacity (A) answers.

Construct validity

To assess the construct validity of the EDAQ, we analyzed how close the estimated item difficulties derived from the Rasch model were to the predicted ones using LLTMs in a three-step process. The first step consisted of estimating the Rasch item difficulties. In the second step, if the Rasch model showed acceptable fit, LLTM and LLTM + ε were built to predict item difficulties and the best model was selected. To do that, a set of intrinsic properties of the items were identified and the weights (the values given to these properties) were determined. Finally, in the last step, the correlation among estimated and predicted item difficulties, as well as the proportion of variance explained by predicted item difficulties, was assessed.

The ordinal responses of the EDAQ were dichotomized so that the LLTM + ε could be applied in the R software. The categories were grouped as follows: '0 = without difficulty or with some difficulty', '1 = with much difficulty or unable to do'. This dichotomization was done conceptually, so that the two lower levels of the EDAQ responses—representing either no or a small amount of difficulty, and the two higher levels—representing either a large amount of difficulty or an inability to perform a task, were compared. This seemed to us the most natural split to distinguish easy from difficult tasks.

Step 1: Estimating Rasch item difficulties from the EDAQ

The Rasch model for dichotomous data is expressed as follows:

$$P\left(X_{iv}=1|\theta_v,\beta_i\right)=\frac{e^{\theta_v-\beta_i}}{1+e^{\theta_v-\beta_i}}, \tag{1}$$

where the probability P that person v has much difficulty or is unable to do task i depends on the difference between the person ability θ and item difficulty β.

There was a large amount of Local Dependency (LD) among the 120 EDAQ items. Therefore, we selected $I = 42$ items free of LD from which we derived the Rasch item difficulty estimates using the RUMM2030 software, where the estimation is based on conditional maximum likelihood [24]. Fit of the model was evaluated via the item–trait interaction Chi-square test. Unidimensionality was evaluated following Smith's Principal Component Analysis (PCA) approach [25]: Two item sets from the residuals PCA analysis were identified, the person estimates from the two sets were compared using a t test, and the number of cases differing significantly at the 5% level was considered. If this number (or the lower bound of a binomial 95% confidence interval) was below 5%, unidimensionality was accepted.

Step 2: Predicting item difficulties with LLTMs

The Linear Logistic Test Model [26] is defined as a Rasch model with a restriction on the item difficulties β_i

$$\beta_i=\beta_0+\sum_{j=1}^{J}q_{ij}\alpha_j, \tag{2}$$

where q_{ij} are the weights of item i on property j and α_j are basic parameters of the LLTM to be estimated. Hence, β_i are not estimated in the Rasch model as in (1) but decomposed as a weighted sum of theoretical item properties, what reduces the number of parameters to estimate, because in general the number of properties J is smaller than the number of items I.

The LLTM assumes that all the variance in item difficulty is explained by the item properties, which is rarely the case. This assumption can be relaxed by applying the LLTM + ε [15] which incorporates an error term to (2)

$$\beta_i=\beta_0+\sum_{j=1}^{J}q_{ij}\alpha_j+\varepsilon_i, \tag{3}$$

where $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$.

A design matrix containing the weights q_{ij}, denoted as Q-matrix, was defined for both LLTMs. The items were located in rows, and the properties in columns. The correlations among the properties were computed to detect potential multicollinearity. Having highly correlated properties (above 0.9 [27]) in (2) and (3) should be avoided.

Identifying a set of intrinsic properties of the ADL items

After performing a pilot study, reviewing the literature, and having discussions with clinical experts of ADL tasks, namely, Occupational Therapists (OTs), we selected $J = 7$ intrinsic properties of ADL items as potential predictors of the item difficulties: (1) Overall physical demand, (2) Bilateral hand involvement, (3) Fine hand use, (4) Physical endurance, (5) Overall cognitive demand, (6) Sequence complexity, and (7) Concentration.

Ordinal ratings of the properties

The properties had to be assigned a value to put in the Q-matrix. We needed clinical judgment for that, and therefore we sent out a questionnaire to professional OTs where they were asked to rate each ADL property for the

42 ADL items. In Table 1, the range of the ordinal ratings for the seven properties is specified. Bilateral hand involvement was rated in a 0–3 ordinal scale, and the other properties in a 0–10 ordinal scale.

The ratings were mostly collected using a snowball technique. NDA contacted OTs worldwide whom she knew directly or through a colleague. She or the colleague sent an email to the OT inviting them to participate, and also asking them to forward the email to OT colleagues. In the email of invitation, OTs were asked to fill in an attached Excel file. Instructions on how to complete the form were in the Excel file itself. Once completed, they sent back the Excel file to NDA attached to an email.

NDA also contacted several OT associations including the Council of Occupational Therapists for the European Countries (COTEC) and World Federation of Occupational Therapists (WFOT). Only one OT from the COTEC participated.

All the professionals contributed to the study on a voluntary basis, giving explicit consent to participate by returning the completed Excel file via email. A confirmation that ethics approval was not required for this study was granted from the Ethics Committee Northwest and Central Switzerland (EKNZ) on the 9th of May 2018.

Deriving the weights of the Q-matrix from the ordinal ratings

Thirty-nine OTs participated in the project, so we had 39 ratings for each combination of ADL item and property. These ratings were on an ordinal scale, and we required an aggregate estimate, ideally at an interval scale metric, for each combination of task and property. A Cumulative Link Mixed Model (CLMM), implemented in the *ordinal* package [28] of the R software, allowed us to satisfy that purpose. Cumulative link models are designed to model ordinal responses [29]: For each level l of the ordinal response (0, 1, 2, or 3 for Bilateral hand involvement, or 0–10 for the other properties), the cumulative probability of being in level l or lower is modeled. As the OTs might have used the response scale differently, we added random effects for OTs so that the different scale perceptions were taken into account.

We applied seven CLMMs, one for each property. In each model, the ordinal ratings to the property (integers from 0 to 3 for Bilateral hand involvement, or from 0 to 10 otherwise) were the response variable. A categorical factor with the ADL items as categories was the single explanatory variable of the model. We took the OT effects to be random (random intercept), assuming that they followed a normal distribution.

The formula of each CLMM is as follows:

$$\log \mathrm{it}(P(Y_{ijk} \leq l)) = z_{jl} - q_{ij}ADL - u_{ijk}, \quad (4)$$

Table 1 Explanation of the ordinal ratings

Property	Range of the ordinal ratings	Explanation
Overall physical demand	0–10	0 = no physical demand required to 10 = extreme physical demand required
Bilateral hand involvement	0–3	0 = one hand suffices 1 = one hand in a sequence 2 = keeping one steady hand 3 = two active hands
Fine hand use	0–10	0 = no fine hand use required to 10 = extreme fine hand use required
Physical endurance	0–10	0 = no physical endurance required to 10 = extreme physical endurance required
Overall cognitive demand	0–10	0 = no cognitive demand required to 10 = extreme cognitive demand required
Sequence complexity	0–10	0 = no complexity to 10 = extreme complexity
Concentration	0–10	0 = no concentration to 10 = extreme concentration

where $i = 1,\ldots,I = 42$ items (ADL tasks), $j = 1,\ldots,J = 7$ item properties, $k = 1,\ldots,K = 39$ OTs, $l = 0,\ldots,L$ levels of ordinal response, $L = 3$ if the property is Bilateral hand involvement, $L = 10$ otherwise.

Y_{ijk} were the ordinal ratings of OT k to the property j of item i; z_{jl} were threshold parameters; q_{ij} were the coefficients for the categories of the ADL factor (the 'weights' of the Q-matrix, an aggregated measure in an interval scale metric for j); and u_{ijk} were the random effects (OT variable) following a normal distribution, $u_{ijk} \sim N(0, \sigma_u^2)$.

A sum zero parameterization was employed so that an estimate for each category of the ADL factor could be obtained. Maximum likelihood estimates of the parameters were provided using the adaptive Gauss–Hermite quadrature method to compute the likelihood function. Each of the seven CLMMs was compared to the respective cumulative link model without random effects, all with unstructured thresholds. The distribution of the random effects estimates was evaluated.

LLTM estimation

Both LLTM and LLTM + ε can be formulated as generalized linear mixed models [30, 31]. We used the *glmer* function from the *lme4* package [32] from the R software (version 3.5.0) to estimate both LLTM and LLTM + ε. In *lme4*, a Laplace approximation of the likelihood is maximized [33]. In both models, the person parameter θ was treated as a random effect and the basic parameters α_j as fixed effects. To include the error term ε in LLTM + ε, the items were also treated as random effects [34].

The goodness-of-fit indices Akaike information criterion (AIC) [35] and Bayesian information criterion (BIC) [36] were used to compare the LLTMs.

Predicted item difficulties were obtained from the best fitting LLTM model, and their standard error was computed via bootstrapping with 1000 iterations [37].

Step 3: Comparing estimated and predicted item difficulties

The predicted item difficulties were correlated with the estimated Rasch item difficulties from Eq. (1). The higher the correlation, the more plausible the fact that variation in ADL difficulties was explained by the properties in the LLTM model, and therefore, the higher the support for construct validity. The proportion of variance explained in the estimated item difficulties by the predicted item difficulties, or the R^2 value [38], was also computed.

Results

Step 1: Rasch item difficulties

The 42 EDAQ dichotomous items fitted a unidimensional Rasch model (p value of the item–trait interaction Chi-square test: 0.14). The relative number of significant t tests from the residuals PCA was 0.041. The estimated Rasch difficulties of the 42 items are shown in the penultimate column of Table 2. They range from −3.18 logits (S5A13: Open fridge door) to 4.57 logits (S12A3: Heavy gardening).

Step 2: Predicting item difficulties with LLTMs

Deriving the weights of the Q-matrix

Figure 1 shows a barplot of the OTs who participated in the study by country. Seventeen OTs from Europe, 16 from Asia, five from North and South America, and one from Oceania participated. Thailand, Japan, Spain, and Switzerland were the countries with most representation.

Seven CLMMs were fitted considering each property as the response variable, the factor of $I = 42$ categories corresponding to the ADL items as the single explanatory variable, and the OT ratings as random effects, as explained in the "Methods" section. The coefficients of the ADL factor were the weights of the Q-matrix. Each model was also compared to the respective model without random effects, and in the seven cases adding a random effect for the intercept considerably improved the fit.

All the models converged with 20 quadrature nodes. The standard deviations of the random effects ranged from 1.20 (Bilateral hand involvement) to 1.96 (Concentration). We computed a 95% confidence interval for the random effects based on the estimated random effects standard deviation for each model. From 3 to 4 OTs (10% of the 39) in each model had a random effect outside the 95% Confidence Interval.

The weights of the Q-matrix are found in Table 2, and the most demanding tasks for each property are shown in Table 3.

LLTM estimation

Four correlations among properties were above 0.9: (1) Overall physical demand and Physical endurance (0.97), (2) Overall cognitive demand and Sequence complexity (0.95), (3) Overall cognitive demand and Concentration (0.95), and (4) Sequence complexity and Concentration (0.91).

Under the LLTM, the seven properties had a significant contribution to predicting the difficulty of ADL tasks (Table 4). Under the LLTM + ε, three properties were no

Table 2 Q-Matrix and estimated and predicted item difficulties

Item	Q-Matrix							Item difficulties	
	Overall physical demand	Bilateral hand involvement	Fine hand use	Physical Endurance	Overall cognitive demand	Sequence Complexity	Concentration	Rasch model	Predicted from LLTM + ε sig
	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)	Estimate (SE)[b]
S1A1: Lift a glass	− 2.78 (0.3)	− 4.38 (0.46)	− 1.93 (0.3)	− 2.78 (0.3)	− 3.15 (0.32)	− 3.12 (0.31)	− 2.85 (0.31)	− 1.48 (0.23)	− 1.9 (0.27)
S1A4: Slice food (e.g., bread, cheese)	− 0.26 (0.28)	0.59 (0.3)	0.4 (0.28)	− 0.2 (0.28)	0.35 (0.3)	0.39 (0.3)	1.08 (0.31)	0.55 (0.17)	− 0.18 (0.05)
S1A8: Open a screw top jar or bottle	0.31 (0.29)	0.8 (0.29)	0.18 (0.27)	− 0.06 (0.28)	− 0.94 (0.28)	− 1.38 (0.28)	− 1.44 (0.27)	3.54 (0.16)	1.95 (0.28)
S2A2: Wipe yourself with toilet paper/ clean self below	− 0.48 (0.29)	− 0.92 (0.29)	− 0.33 (0.27)	− 0.78 (0.29)	− 0.44 (0.29)	0.21 (0.28)	− 0.41 (0.27)	− 1.14 (0.22)	− 1.1 (0.13)
S2A4: Flush the toilet	− 3.14 (0.31)	− 4.18 (0.43)	− 2.85 (0.3)	− 3.6 (0.31)	− 3.34 (0.32)	− 3.42 (0.31)	− 4.46 (0.34)	− 2.89 (0.32)	− 2.69 (0.31)
S2A6: Wash your hands	− 2.1 (0.3)	3.73 (0.61)	− 1.33 (0.28)	− 2.2 (0.29)	− 1.59 (0.29)	− 0.43 (0.29)	− 1.61 (0.28)	− 3.06 (0.33)	− 2 (0.46)
S2A9: Use a tube of toothpaste	− 2.21 (0.28)	0.61 (0.3)	0.39 (0.27)	− 1.81 (0.29)	− 0.85 (0.28)	− 0.21 (0.27)	− 1.02 (0.28)	− 1.52 (0.24)	− 1.4 (0.2)
S2A11: Do your make up or shave	0.03 (0.29)	1.19 (0.35)	2.33 (0.29)	0.06 (0.27)	1.94 (0.28)	2.3 (0.28)	2.46 (0.3)	− 1.74 (0.26)	− 0.05 (0.19)
S2A12: Put on jewelery/watch	− 1.94 (0.3)	0.99 (0.31)	2.8 (0.31)	− 1.64 (0.3)	0.32 (0.28)	0.83 (0.27)	0.84 (0.29)	0.28 (0.18)	− 0.03 (0.27)
S3A1: Put on/take off a coat	− 0.47 (0.27)	1.53 (0.38)	− 1.42 (0.31)	− 0.74 (0.29)	− 0.18 (0.28)	0.23 (0.27)	− 0.72 (0.28)	− 1.03 (0.21)	− 1.62 (0.27)
S3A5: Pull clothes over your feet	0.09 (0.27)	1.77 (0.36)	− 0.6 (0.27)	− 0.24 (0.29)	− 0.43 (0.27)	0.2 (0.26)	− 0.9 (0.28)	− 0.33 (0.19)	− 0.32 (0.16)
S3A6: Do up/undo zips	− 2.14 (0.28)	0.63 (0.3)	2.1 (0.29)	− 1.93 (0.28)	− 0.38 (0.29)	0 (0.27)	− 0.14 (0.28)	− 0.64 (0.2)	− 0.06 (0.26)
S3A11: Fasten clothes at the back	0.54 (0.27)	2.06 (0.38)	1.55 (0.3)	0.17 (0.27)	1.17 (0.28)	0.58 (0.27)	0.76 (0.28)	3 (0.17)	1.62 (0.2)
S4A4: Turn taps (any in home)	− 1.37 (0.28)	− 3.54 (0.37)	− 0.99 (0.29)	− 2.03 (0.3)	− 1.96 (0.28)	− 2.46 (0.3)	− 2.8 (0.29)	− 0.19 (0.19)	− 0.31 (0.24)
S4A5: Wash your back and neck	0.73 (0.27)	− 0.01 (0.31)	− 1.29 (0.27)	0.56 (0.28)	− 0.49 (0.26)	− 0.28 (0.28)	− 0.4 (0.27)	1.01 (0.17)	− 0.16 (0.12)
S4A9: Style/ blow-dry your hair	0.9 (0.28)	2.11 (0.41)	0.26 (0.28)	1.43 (0.28)	0.81 (0.27)	0.98 (0.27)	1.05 (0.28)	0.39 (0.2)	0.45 (0.13)

Table 2 (continued)

Item	Q-Matrix							Item difficulties	
	Overall physical demand	Bilateral hand involvement	Fine hand use	Physical Endurance	Overall cognitive demand	Sequence Complexity	Concentration	Rasch model	Predicted from LLTM + ε sig
	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)	Estimate (SE)[b]
S4A10: Cut/file your finger nails	−0.91 (0.28)	0.49 (0.32)	2.72 (0.29)	−0.82 (0.28)	1.22 (0.26)	0.64 (0.27)	2.62 (0.3)	0.6 (0.18)	0.96 (0.26)
S5A1: Stand while working in the kitchen	1.86 (0.29)	−1.26 (0.38)	−3.08 (0.46)	2.48 (0.31)	−0.64 (0.37)	−1.11 (0.37)	−0.27 (0.36)	−0.55 (0.2)	−0.06 (0.28)
S5A4: Carry a full pan to/from the cooker	1.91 (0.27)	1.69 (0.36)	−0.73 (0.28)	1.75 (0.28)	0.45 (0.28)	−0.1 (0.27)	1.28 (0.28)	2.75 (0.16)	1.52 (0.16)
S5A10: Put crocker/pans etc into kitchen cupboards	1.1 (0.27)	1.41 (0.35)	−0.39 (0.26)	0.85 (0.28)	0.39 (0.26)	0.2 (0.27)	0.53 (0.26)	−0.3 (0.19)	0.73 (0.1)
S5A11: Use a kettle (e.g., fill, pour)	0.21 (0.28)	0.18 (0.29)	−0.36 (0.25)	0.36 (0.26)	0.7 (0.27)	0.28 (0.25)	1.14 (0.27)	0.22 (0.18)	−0.37 (0.08)
S5A13: Open fridge door	−1.41 (0.28)	−3.17 (0.36)	−2.49 (0.29)	−1.71 (0.29)	−2.74 (0.31)	−2.96 (0.31)	−3.15 (0.31)	−3.18 (0.34)	−1.07 (0.23)
S6A3: Lock and unlock doors	−1.61 (0.28)	−1.58 (0.29)	0.9 (0.28)	−2.01 (0.28)	0.1 (0.29)	0.03 (0.28)	−0.38 (0.29)	−0.87 (0.2)	−1.02 (0.15)
S6A9: Reach up	−0.49 (0.28)	−3.17 (0.35)	−3.47 (0.33)	−0.63 (0.29)	−3.03 (0.33)	−3.52 (0.34)	−2.1 (0.32)	0.85 (0.17)	−0.51 (0.28)
S6A12: Manage heating (e.g., controls, wood-burner, multifuel stove, open fire)	−0.43 (0.33)	−0.36 (0.32)	0.62 (0.27)	−0.22 (0.3)	2.44 (0.29)	1.39 (0.29)	1.89 (0.29)	−1.43 (0.24)	−1.31 (0.19)
S7A1: Make the bed	2.67 (0.29)	2.67 (0.43)	0.31 (0.28)	2.07 (0.29)	1.34 (0.27)	2.19 (0.27)	1.16 (0.27)	0.6 (0.17)	1.02 (0.23)
S7A4: Wring out a cloth	0.98 (0.28)	3.39 (0.54)	0.14 (0.27)	1.64 (0.29)	−1.02 (0.27)	−0.71 (0.28)	−0.93 (0.28)	1.56 (0.16)	2.34 (0.34)
S7A5: Vacuum clean	2.62 (0.28)	1.01 (0.33)	−0.78 (0.27)	2.98 (0.28)	0.93 (0.27)	1.11 (0.26)	0.93 (0.27)	1.44 (0.17)	0.81 (0.22)
S7A6: Open a window	−0.75 (0.27)	−1.5 (0.32)	−1.22 (0.27)	−1.02 (0.27)	−1.57 (0.27)	−1.34 (0.28)	−1.81 (0.27)	−0.34 (0.19)	−0.67 (0.11)
S7A7: Clean windows	2.52 (0.28)	−0.01 (0.29)	−0.46 (0.27)	2.56 (0.29)	0.6 (0.27)	0.94 (0.27)	0.99 (0.28)	2.47 (0.17)	0.99 (0.22)

Table 2 (continued)

Item	Q-Matrix							Item difficulties	
	Overall physical demand	Bilateral hand involvement	Fine hand use	Physical Endurance	Overall cognitive demand	Sequence Complexity	Concentration	Rasch model	Predicted from LLTM + ε sig
	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)[a]	Estimate (SE)	Estimate (SE)[b]
S8A1: Do the hand washing	1.52 (0.3)	3.04 (0.49)	0.31 (0.27)	1.53 (0.3)	0.62 (0.29)	0.88 (0.29)	0.5 (0.27)	1.52 (0.18)	1.34 (0.18)
S8A3: Hang out washing	2.21 (0.28)	1.82 (0.36)	0.55 (0.27)	2.19 (0.28)	0.54 (0.27)	1.32 (0.27)	0.74 (0.26)	0.51 (0.18)	1.52 (0.15)
S8A4: Plug in and pull out a plug (any in home)	− 1.12 (0.28)	− 2.67 (0.33)	0.22 (0.28)	− 1.43 (0.29)	− 0.91 (0.28)	− 1.48 (0.28)	− 0.96 (0.3)	− 0.12 (0.18)	0.16 (0.22)
S8A5: Put up an ironing board	0.8 (0.27)	3.11 (0.5)	− 0.22 (0.25)	0.37 (0.27)	0.36 (0.26)	0.73 (0.27)	0.47 (0.27)	0.51 (0.18)	0.38 (0.21)
S8A8: Use scissors (any in home)	− 0.99 (0.28)	− 0.12 (0.3)	2.74 (0.29)	− 0.24 (0.28)	1.48 (0.28)	0.92 (0.28)	2.24 (0.28)	0.81 (0.17)	0.51 (0.24)
S10A1: Use a phone/mobile/smartphone	− 2.49 (0.3)	− 0.73 (0.3)	2.22 (0.28)	− 1.91 (0.29)	3.46 (0.31)	2.91 (0.3)	2.64 (0.28)	− 2.56 (0.29)	− 3.41 (0.42)
S10A3: Write	− 1.23 (0.31)	− 1.49 (0.31)	3.84 (0.31)	− 0.12 (0.29)	3.72 (0.32)	2.76 (0.32)	3.46 (0.31)	− 0.66 (0.2)	− 0.83 (0.37)
S10A6: Use remote controls (e.g., TV)	− 2.98 (0.33)	− 3.46 (0.35)	1.41 (0.31)	− 2.83 (0.3)	1.39 (0.3)	0.47 (0.31)	0.32 (0.3)	− 3.18 (0.34)	− 2.61 (0.32)
S11A5: Get in and out of a car and open car door	1.77 (0.27)	− 1.27 (0.31)	− 1.2 (0.26)	1.21 (0.28)	− 0.16 (0.29)	0.22 (0.31)	− 0.55 (0.29)	− 0.92 (0.21)	0.16 (0.23)
S11A8: Open a heavy (e.g., shop) door	2.83 (0.29)	− 0.99 (0.32)	− 1.57 (0.29)	2.16 (0.31)	− 1.7 (0.31)	− 1.66 (0.3)	− 1.57 (0.3)	1.69 (0.16)	2.74 (0.32)
S1112: Hold a walking stick	− 0.73 (0.29)	− 3.68 (0.38)	− 1.25 (0.28)	0.33 (0.32)	− 1.81 (0.31)	− 1.95 (0.32)	− 1.73 (0.32)	− 0.76 (0.23)	− 0.45 (0.22)
S12A3: Heavy gardening (e.g., dig, mow)	6.45 (0.39)	3.66 (0.61)	1.98 (0.31)	6.22 (0.37)	2.99 (0.3)	3.38 (0.31)	3.08 (0.3)	4.57 (0.2)	4.94 (0.47)

LLTM + ε sig linear logistic test model plus error with significant predictors, *SE* standard error

[a]Estimates from cumulative link mixed models

[b]Bootstrap standard error

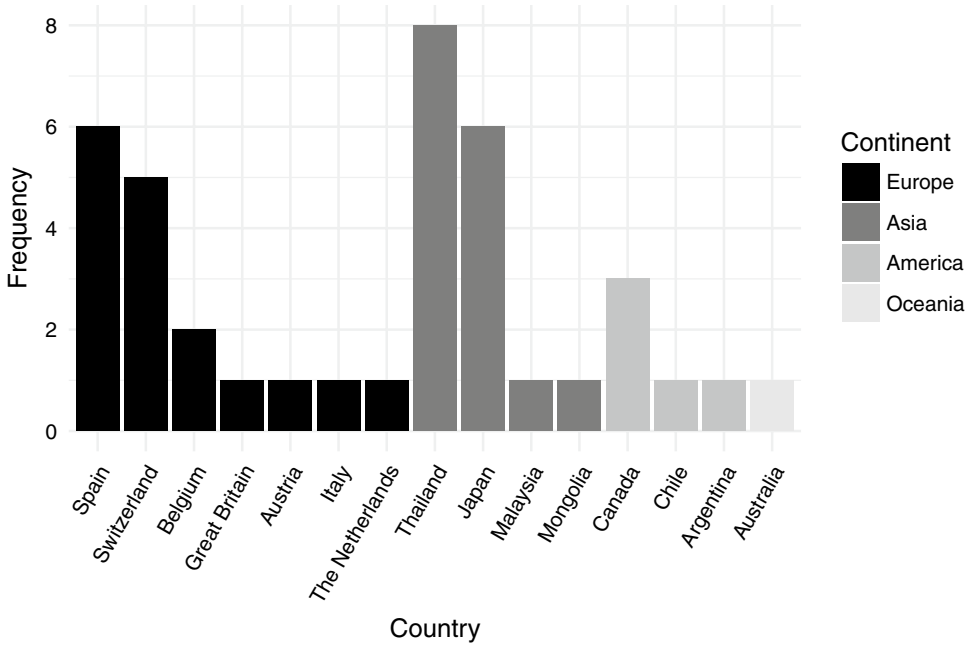

Fig. 1 Participation of occupational therapists by country

Table 3 Most demanding ADL tasks by property

ADL property	Most demanding task 1	Most demanding task 2	Most demanding task 3	Most demanding task 4	Most demanding task 5
Overall physical demand	S12A3: Heavy gardening (e.g., dig, mow)	S11A8: Open a heavy (e.g., shop) door	S7A1: Make the bed	S7A5: Vacuum clean	S7A7: Clean windows
Bilateral hand involvement	S2A6: Wash your hands	S12A3: Heavy gardening (e.g., dig, mow)	S7A4: Wring out a cloth	S8A5: Put up an ironing board	S8A1: Do the hand washing
Fine hand use	S10A3: Write	S2A12: Put on jewelry/watch	S8A8: Use scissors (any in home)	S4A10: Cut/file your finger nails	S2A11: Do your make up or shave
Physical endurance	S12A3: Heavy gardening (e.g., dig, mow)	S7A5: Vacuum clean	S7A7: Clean windows	S5A1: Stand while working in the kitchen	S8A3: Hang out washing
Overall cognitive demand	S10A3: Write	S10A1: Use a phone/mobile/smartphone	S12A3: Heavy gardening (e.g., dig, mow)	S6A12: Manage heating (e.g., controls, woodburner, multi-fuel stove, open fire)	S2A11: Do your make up or shave
Sequence complexity	S12A3: Heavy gardening (e.g., dig, mow)	S10A1: Use a phone/mobile/smartphone	S10A3: Write	S2A11: Do your make up or shave	S7A1: Make the bed
Concentration	S10A3: Write	S12A3: Heavy gardening (e.g., dig, mow)	S10A1: Use a phone/mobile/smartphone	S4A10: Cut/file your finger nails	S2A11: Do your make up or shave

ADL activities of daily living

longer significant (Physical endurance, Overall cognitive demand, and Concentration). The LLTM+ε fitted significantly better than the LLTM ($\chi^2(1)=489$, p value <0.001). A model with the four significant properties 'LLTM+ε sig' was also considered, and a Chi-square test revealed that LLTM+ε did not fit significantly better than LLTM+ε sig ($\chi^2(3)=5.7$, p value $=0.13$). Besides, under LLTM+ε sig all the correlations among the properties were below 0.9, so that the strong collinearity present in LLTM and LLTM+ε was no longer there in LLTM+ε sig. Hence, LLTM+ε sig was the preferred choice.

Under LLTM+ε sig, Overall physical demand, Bilateral hand involvement, and Fine hand use had positive coefficients, while the coefficient for Sequence

Table 4 Estimates of the item properties under the LLTM and LLTM + ε

Effects	Parameter	LLTM Estimate (SE)	LLTM + ε Estimate (SE)	LLTM + ε sig Estimate (SE)
Fixed effects	Intercept	− 2.93* (0.15)	− 3.19* (0.2)	− 3.19* (0.2)
	Overall physical demand	0.95* (0.08)	0.99* (0.31)	0.92* (0.09)
	Bilateral hand involvement	0.27* (0.02)	0.27* (0.09)	0.18* (0.08)
	Fine hand use	0.65* (0.03)	0.72* (0.14)	0.86* (0.14)
	Physical endurance	− 0.17* (0.08)	− 0.14 (0.33)	
	Overall cognitive demand	0.56* (0.09)	0.48 (0.38)	
	Sequence complexity	− 1.57* (0.07)	− 1.59* (0.31)	− 0.99* (0.16)
	Concentration	0.15* (0.06)	0.22 (0.26)	
Random effects	θ_v	7.50 (2.74)[a]	8.89 (2.98)[a]	8.90 (2.98)[a]
	ε_i		0.54 (0.73)[a]	0.62 (0.79)[a]

LLTM, linear logistic test model, *LLTM* + ε linear logistic test model plus error, *LLTM* + ε sig linear logistic test model plus error with significant predictors, *SE* standard error

*p value < 0.05

[a]Variance (standard deviation)

Table 5 Goodness of fit

Model	Number of parameters	AIC	BIC
LLTM	9	8940.45	9009.08
LLTM + ε	10	8453.82	8530.08
LLTM + ε sig	7	8453.53	8506.92

AIC Akaike information criterion, *BIC* Bayesian information criterion, *LLTM* linear logistic test model, *LLTM* + ε linear logistic test model plus error, *LLTM* + ε sig linear logistic test model plus error with significant predictors

complexity was negative. A positive coefficient indicated that the more of the property, the harder the ADL task. Bilateral hand involvement had the lowest positive impact to the item difficulty, and Overall physical demand the highest positive impact. The negative impact of Sequence complexity translates into the fact that the higher the number of steps to be performed, the easier the task.

Table 5 shows goodness-of-fit indies (AIC and BIC) for the three models. Lower values of these indices indicate better fit. LLTM + ε sig shows the lowest values in both AIC and BIC, being again the preferred choice. Thus, the predicted item difficulties were computed via LLTM + ε sig, and they are presented in the last column of Table 2 together with a bootstrap standard error. The estimated Rasch standard errors were similar to the bootstrap standard errors for the predicted difficulties (the differences among them ranged between − 2.78 and 0.12, with a median of − 0.03 and an interquartile range of 0.12).

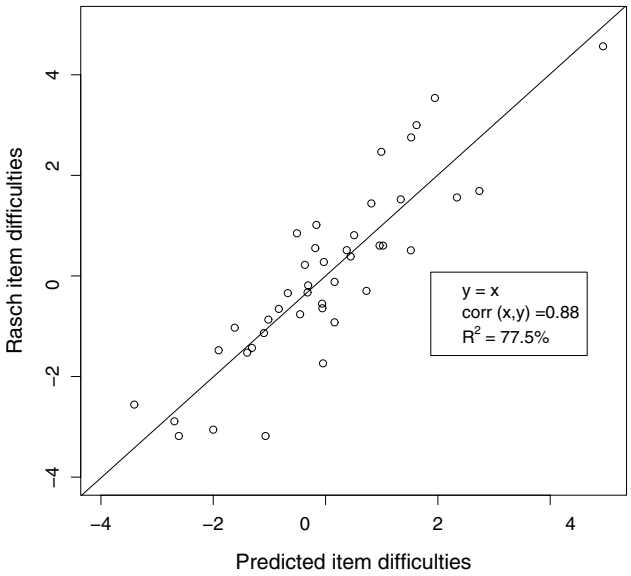

Fig. 2 Rasch item difficulties versus predicted item difficulties

Step 3: Comparing estimated and predicted item difficulties

The correlation among the estimated item difficulties from the Rasch model (penultimate column of Table 2) and the predicted item difficulties under LLTM + ε sig (last column of Table 2) was 0.88 (p value < 0.001). The variance of the predicted item difficulties, 2.44, was lower than the variance of the Rasch difficulties, 3.17. Figure 2 shows a scatterplot of the estimated versus predicted item difficulties. The proportion of variance in estimated item difficulties explained by the predicted item difficulties was 77.5%.

Discussion

In this study, we aimed to support construct validity of the EDAQ by using item properties accounting for item difficulties. We identified seven properties of ADL items rated in a 0–10 scale (except Bilateral hand involvement, in a 0–3 scale) by 39 occupational therapists. Seven CLMMs were employed to derive the weights of the properties to 42 locally independent EDAQ items. Three LLTMs were fitted, and the LLTM + ε model including only the four significant predictors of LLTM + ε (LLTM + ε sig) was the preferred choice to predict item difficulties. These four predictors—Overall physical demand, Bilateral hand involvement, Fine hand use, and Sequence complexity—accounted for 77.5% of the variability in the Rasch estimated item difficulties. Thus, a substantial amount of item variability has been explained from a theoretical perspective.

The three non-significant properties—Physical endurance, Overall cognitive demand, and Concentration—were highly correlated to Overall physical demand (the first) and to Sequence complexity (the latter two), and therefore a subset of the seven properties (not the whole set) with acceptable correlations should be considered when predicting item difficulties. A plausible subset was the one consisting of the significant properties of LLTM + ε. Having a non-significant Chi-square value in comparing the LLTM + ε models with and without the non-significant predictors indicates that the contribution of the non-significant predictors in explaining the variance in item difficulty was negligible. Hence, excluding them when it comes to predict item difficulties should not impact the evaluation of the validity of the EDAQ.

As expected, the higher the amount of Overall physical demand, Bilateral hand involvement, and Fine hand use, the more difficult is the task to perform. On the contrary, the higher the Sequence complexity, the easier the task; this could be explained due to the fact that, after the other properties are accounted for, if a task can be broken down into several smaller tasks done in a sequence, this should make the process easier.

It seems that Sequence complexity acted as a 'suppressor variable' [39], because in a univariate context, it was significantly correlated with the predictors Overall physical demand and Fine hand use, and it was not correlated with the outcome (Rasch estimated item difficulties). In a model containing the predictors Overall physical demand and Fine hand use, after adding Sequence complexity as a predictor, the latter became significant and the positive predictive effects of the former two were enhanced. These former two on the one hand and Sequence complexity on the other hand had significant opposite effects on the outcome, being the former two positive and the latter negative. However, it is important to note that while Sequence complexity has the hallmarks of a suppressor variable, it also has a conceptual validity in making tasks easier if they can be broken down into discrete sub-components, conditioned on the physical attributes.

The LLTM + ε sig model derived in this study provides a theoretical explanation of the processes involved in performing ADL tasks, therefore supporting construct validity of the EDAQ. The EDAQ had already been validated using empirical and statistical methods [22], as it is done in the majority of health instruments. However, in such empirically oriented methods, there is no theoretical explanation on what happens between the attribute and the instrument scores. And without such an explanation, the validation of an instrument is incomplete [5]. Hence, LLTMs are a valuable tool for providing validity evidence. Ideally, they should be part of the process of construct validation, not making use of just purely empirical and statistical methods.

This study presents some limitations. First, it is impossible to include all the possible properties of the items that predict item difficulty. Although we carefully thought about the selected ADL properties, we could have missed relevant predictors. Secondly, the weights of the Q-matrix were estimates derived from CLMMs, therefore they had measurement error. The potential impact of this error in assessing construct validity using item properties to predict item difficulties could be the topic of future studies. Thirdly, the item difficulties derived in both original Rasch analysis and LLTMs were based on a dataset from those with rheumatoid arthritis. It is possible that other diagnostic groups, such as stroke, may deliver a different hierarchical ordering, and as a consequence a different emphasis upon the intrinsic properties of each item. This will require further investigation.

Finally, we had to dichotomize the EDAQ responses to apply the LLTM + ε using the *glmer* function from the *lme4* R package. To have an idea of whether dichotomizing had a considerable impact on the results, we estimated the linear equivalent of the LLTM + ε model with the original EDAQ responses. In the linear equivalent, the ordinal EDAQ responses were treated as an interval scale, which is not correct, but in this way we could somehow assess the impact of dichotomizing. Interestingly, we obtained the same significant predictors in the linear as in the dichotomized version, namely, Overall physical demand, Bilateral hand involvement, Fine hand use, and Sequence complexity. The coefficients of both versions were obviously not the same, but they had the same sign and kept the same hierarchy. Hence, it seems that dichotomizing the EDAQ responses did not have a considerable impact on the results.

The main strength of this study is that it is the first application of the LLTM in trying to explain the variability in item difficulties of ADL. In addition, deriving an aggregated

measure from ordinal ratings using CLMMs, although containing some error, has the advantage of obtaining interval scale values and taking into account the variability of the OT ratings, which is not possible in a summary measure such as the median. As far as we know, it is the first time that such a model is used for this purpose. We believe that this model can open up the opportunity to assess construct validity by using item properties accounting for item difficulties across a wider range of health outcomes using ordinal clinical ratings of properties of the task.

Conclusion

We strongly believe that using LLTMs adds significant value in assessing construct validity. The mere fact of formulating the basic parameters of the LLTM leads to a clearer understanding of the task [40]. We hope that the four ADL properties identified in this study can help in predicting performance of other ADL tasks.

Acknowledgements We are extremely thankful to the 39 Occupational Therapists who participated and made possible this project. We would also like to thank Armin Gemperli and Cristina Ehrmann for their comments in a previous version of this paper. This paper is part of the cumulative PhD thesis of NDA.

Funding This work was supported by Swiss Paraplegic Research and University of Lucerne.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval A confirmation that ethics approval was not required for this study was granted from the Ethics Committee Northwest and Central Switzerland (EKNZ) on May 9, 2018.

Informed consent All the occupational therapists contributed to the study on a voluntary basis, giving explicit consent to participate by returning the completed Excel file via email.

References

1. Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill
2. Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639.
3. American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington DC: American Psychological Association
4. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957.
5. Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295x.111.4.1061.
6. Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179.
7. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066x.50.9.741.
8. Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x.
9. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. København: Danmarks Paedagogiske Institut.
10. Stenner, J. A., & Smith M. III (1982). Testing construct theories. *Perceptual and Motor Skills, 55*, 415–426.
11. Stenner, J. A., Smith M. A. III, Burdick D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20*(4), 305–316.
12. Arthur, G. (1947). *A point scale of performance tests*. New York: Psychological Corp.
13. Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6.
14. Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research and Evaluation, 20*, 1–11.
15. Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models*. New York: Springer.
16. Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*(4), 369–381. https://doi.org/10.2307/1165055.
17. Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*(4), 665–686. https://doi.org/10.1177/0013164411430707.
18. Kubinger, K. D. (1979). Das Problemlöseverhalten bei der statistischen Auswertung psychologischer Experimente. Ein Beispiel hochschuldidaktischer Forschung. *Zeitschrift für Experimentelle und Angewandte Psychologie, 26*, 467–496.
19. Zeuch, N., Holling, H., & Kuhn, J. T. (2011). Analysis of the Latin square task with linear logistic test models. *Learning and Individual Differences, 21*, 629–632.
20. Sonn, U., Törnquist, K., & Svensson, E. (1999). The ADL taxonomy—From individual categorical data to ordinal categorical data. *Scandinavian Journal of Occupational Therapy, 6*(1), 11–20. https://doi.org/10.1080/110381299443807.
21. Fong, T. G., Gleason, L. J., Wong, B., Habtemariam, D., Jones, R. N., Schmitt, E. M., et al. (2015). Cognitive and physical demands of activities of daily living in older adults: Validation of expert panel ratings. *PM R, 7*(7), 727–735. https://doi.org/10.1016/j.pmrj.2015.01.018.
22. Hammond, A., Tennant, A., Tyson, S. F., Nordenskiold, U., Hawkins, R., & Prior, Y. (2015). The reliability and validity of the English version of the evaluation of daily activity questionnaire for people with rheumatoid arthritis. *Rheumatology, 54*(9), 1605–1615. https://doi.org/10.1093/rheumatology/kev008.
23. WHO. (2001). *International classification of functioning, disability and health*. Geneva: World Health Organization (WHO). Retrieved from http://nla.gov.au/nla.cat-vn1515102.

24. Andrich, D., Sheridan, B., & Luo, G. (2010). *Rasch models for measurement: RUMM2030*. Perth: RUMM Laboratory Pty, Ltd.
25. Smith, E. V. Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2), 205–231.
26. Fischer, G. H. (1995). *Rasch models: foundations, recent developments, and applications*. New York: Springer.
27. Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: Sage Publications, Inc.
28. Christensen, R. (2015). A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package. https://cran.rproject.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf. Accessed 23 May 2018.
29. Agresti, A. (2012). *Analysis of ordinal categorical data* (2nd ed.). Hoboken: Wiley.
30. De Boeck, P., Bakker, M., Zwitser, R., & Nivard, M. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software. 39*(12), 1–28.
31. De Boeck, P., Cho, S. J., & Wilson, M. (2016). Explanatory item response models. In A. A. Rupp, & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*. New Jersey: Wiley.
32. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software; 67*(1), 1–48.
33. Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical software, 20*(2), 18. https://doi.org/10.18637/jss.v020.i02.
34. De Boeck, P. (2008). Random item IRT models. *Psycometrika, 73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x.
35. Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.
36. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.
37. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
38. Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 1). New York: Wiley.
39. Ludlow, L., & Klein, K. (2014). Suppressor variables: The difference between 'is' versus 'acting as'. *Journal of Statistics Education*. https://doi.org/10.1080/10691898.2014.11889703.
40. Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6*(4), 397–416. https://doi.org/10.1177/014662168200600403.