



An assessment of the content and properties of extended and instrumental activities of daily living scales: a systematic review

Eline Kelbling, David Ferreira Prescott, Mary Shearer & Terence J. Quinn

To cite this article: Eline Kelbling, David Ferreira Prescott, Mary Shearer & Terence J. Quinn (2023): An assessment of the content and properties of extended and instrumental activities of daily living scales: a systematic review, *Disability and Rehabilitation*, DOI: [10.1080/09638288.2023.2224082](https://doi.org/10.1080/09638288.2023.2224082)

To link to this article: <https://doi.org/10.1080/09638288.2023.2224082>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 06 Jul 2023.



[Submit your article to this journal](#)




[View related articles](#)



[View Crossmark data](#)

An assessment of the content and properties of extended and instrumental activities of daily living scales: a systematic review

Eline Kelbling, David Ferreira Prescott, Mary Shearer and Terence J. Quinn 

Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK

ABSTRACT

Purpose: We performed a systematic review to assess the psychometric properties of extended Activities of Daily Living (eADL) scales.

Materials and Methods: Articles assessing eADL scales' properties were retrieved by searching multidisciplinary databases, and reference screening. Data on the following properties were extracted: validity, reliability, responsiveness, and internal consistency. The COSMIN (Consensus-based Standards for the selection of health status Measurement Instruments) risk of bias checklists are used to assess the quality of included articles. All aspects were performed by two independent researchers.

Results: Of 245 titles, 26 articles were eligible, comprising 15 different eADL scales. The Lawton scale had the most papers describing properties, while the Performance-based Instrumental Activities of Daily Living received the highest COSMIN rating. Properties most often assessed were convergent validity and reliability, no articles assessed all COSMIN properties. The COSMIN assessment rated 43% of the properties as 'positive', 31% 'doubtful' and 26% 'inadequate'. Only Lawton was assessed in more than one paper, available data suggest that this scale has excellent reliability, construct validity, internal consistency, and medium criterion validity.

Conclusion: Despite their common use, there are limited data on the properties of eADL scales. Where data are available there are potential methodological issues in the studies.

ARTICLE HISTORY

Received 9 January 2023

Revised 26 May 2023

Accepted 27 May 2023

KEYWORDS

Extended activities of daily living; eADL; instrumental Activities of Daily Living; iADL; properties; validity; reliability; responsiveness

> IMPLICATIONS FOR REHABILITATION

- The functional abilities of older adults are most commonly measured using extended activities of daily living scales (eADL).
- There are many eADL scales available to clinicians and no guidance on a preferred tool.
- Despite the frequent use of eADL scales in research and practice, there is limited published literature on their psychometric properties (for example validity, reliability and responsiveness).
- The Lawton Scale has the most supporting evidence and its properties are generally acceptable, more research is needed on other eADL scales.


Introduction

Functional decline is a common and important feature of ageing. It reflects how an individual's limitations interact with the demands of the environment [1]. Independence and participation in everyday activities are essential to older adults, and maintaining their autonomy plays a considerable role in ageing 'successfully' [2]. The inability to achieve everyday activities without assistance may suggest unsafe conditions and inferior quality of life [3]. Therefore, it is important to assess the functional abilities of older adults. A function is usually assessed by using activities of daily living (ADL) measurement instruments. These activities can be divided into 'basic activities of daily living' (bADL) and 'extended activities of daily living' (eADL), sometimes also called instrumental ADL (iADL). The current article will adopt the term eADL, referring to both eADL and iADL. bADL tasks include mobility and basic self-care such as bathing and eating, whereas eADL includes higher-level complex activities such as using the telephone or public transportation [4].

Assessment of function is integral to research and clinical practice. From charting the natural history of disease through assessing the efficacy of a novel treatment to resource allocation and policy – all require a standardised assessment of functional ability. For example, as eADL are cognitively more demanding than bADL, loss of ability in eADL is used to distinguish mild cognitive impairment from dementia [5]. Using accessible, standardized scales to assess daily activities in older adults may help healthcare professionals in making diagnoses, describing prognoses and monitoring recovery from disease or some other functional insult [6]. As the number of people over 65 years of age continues to increase, so does the importance of the eADL scale [7].

There are many different eADL scales available and no consensus on the optimal assessment for older adults. Some eADL scales have been developed for a specific purpose, for example, assessment of dementia, while others are more generic. With so many scales available, we need a framework to help us choose the best scale for a particular situation. Knowledge of the psychometric properties

CONTACT Eline Kelbling  e.kelbling@outlook.com  Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/09638288.2023.2224082>.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

of scales could help in decision-making about which tool to use. Psychometric properties describe a test's appropriateness, usefulness, and meaningfulness [8]. It provides a distinct insight into whether the test measures what it is supposed to measure, its stability over time, and the ability to detect a change in conditions – in other words, validity, reliability, and responsiveness.

Despite the extensive use of eADL scales, their psychometric properties have been questioned in previous research and commentary [9]. A comprehensive, objective assessment of the eADL scales available and their properties would be a useful addition to the literature. In this systematic review, we aim to provide an overview of available eADL scales and systematically assess the properties of those eADL scales.

Methods

We conducted this systematic review according to the 'Consensus-based Standards for the selection of health status Measurement Instruments (COSMIN) methodology for systematic reviews of Patient-Reported Outcome measures (PROMs)' [10]. Where relevant we followed Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidance for reporting (see [supplementary appendix](#) for reporting checklist). This systematic review protocol was registered on the Open Science Framework (protocol provided in the [supplementary appendix](#)). All aspects of title selection, data extraction and quality assessment were performed by two researchers who worked independently and compared results. Where differences could not be resolved through discussion a third reviewer was consulted.

Literature search

We created a search, using validated search syntax where possible. Our search was developed around concepts of eADL and psychometric properties. The primary search was complemented by a purposive search based on the names of commonly used eADL scales ([Table 1 in the supplementary appendix](#)). We reviewed the following databases from inception to May 2022: EMBASE (OVID), HaPI (OVID), MEDLINE (OVID), PsychINFO (EBSCO), and CINAHL (EBSCO). Searches were combined in Mendeley reference management software and de-duplicated. Reference lists of relevant articles were screened for potentially eligible articles.

Inclusion and exclusion

This systematic review was focused on eADL assessments. We defined eADL scales based on their content and purpose. These scales should include items and tasks beyond basic self-care, such as using public transport and managing finances. The scales provide information on a person's functional abilities and the capability to maintain an independent lifestyle. A questionnaire was selected when it aimed to assess iADL, eADL, or complex ADL. The scales did not have to be disease-specific. No distinction was made based on the structure (i.e. informed-based, self-reported, etc.) or country of development of the questionnaire. We included full papers, written in English describing the psychometric properties of one or more than one eADL scale.

Studies that evaluated bADL scales or scales with a purpose other than ADL assessment, such as cognitive scales, were excluded. Scales that assessed both bADL and eADL were included and assessed in the same way the other articles were assessed. Our evaluation majored on psychometric properties. To be eligible,

a paper had to describe one of the properties as included in the COSMIN guidance.

Quality assessment

The methodological quality of each paper was evaluated using the COSMIN Risk of Bias Checklist [11]. Differing checklists are available for content validity, structural validity, internal consistency, cross-cultural validity, reliability, measurement error, criterion validity, construct validity, and responsiveness (defined below). Each checklist contained a number of question items which together generated an overall score on a five-point rating system: 'very good', 'adequate', 'doubtful', 'inadequate', or 'not applicable'. Scoring was based on the 'worst score counts' principle, where the lowest rating of any item in the checklist determines the final rating. In general, a design requirement is rated as 'very good' when there are convincing arguments that the standard is met; 'adequate' when it is reasonable to assume that the standard is met, but it is not explicitly described; 'doubtful' when it is unclear whether the standard is met; 'inadequate' when there is evidence the standard is not met; and 'not applicable' when information regarding the criteria was lacking.

Data extraction

For each paper, we extracted the details of the eADL scale including the individual tasks included in the scale. We also extracted data on the application of the scale and the context of the assessment. We used a bespoke data extraction tool that we piloted on two exemplar studies. We created a data visualisation describing the components of each scale. These components were chosen after thoroughly evaluating the scales and identifying the most common items. This process was repeated by a second reviewer. We used a narrative approach to give an overview of each COSMIN-defined property (see below) and an individual assessment of any scale with three or more articles assessing the scale.

Psychometric properties

Reliability

Reliability refers to the ability of a scale to produce consistent results with repeated measurements. It includes both consistencies among scale items and reproducibility among observers [12]. There are four subtypes of reliability. Test-retest reliability involves achieving consistent results over time, showing that the research methods are reliable and not influenced by external factors. Internal consistency tests the homogeneity of the scale and whether the different scale items correlate with each other. It is usually calculated using Cronbach's alpha, with values ranging from 0 to 1. A Cronbach's alpha between 0.8–0.95 is considered good while an alpha coefficient is very high (i.e. >0.95) it is at risk for redundancy. Inter-rater reliability measures the consistency of a scale when administered by different reviewers on the same occasion and can be calculated by using Kappa statistics [13]. Intra-rater reliability measures the scale by using the same reviewer on the same subjects, with assessments separated by time.

Validity

Validity is the extent to which a scale measures a factor accurately and whether it measures the concept it is supposed to measure. Validity can be divided into criterion, content, construct, and face validity [12]. Criterion validity includes a correlation with the 'gold

standard' and it measures to what extent one scale predicts an outcome for another scale that theoretically represents the construct. Content validity assesses if the scale items are fully representative of the concept of interest, this usually involves gathering feedback from healthcare professionals and people with an index condition. Construct validity evaluates if the measurement tool represents the construct of interest. Other than the 'golden standard' in criterion validity, construct validity is usually assessed by demonstrating a relationship between the novel scale and other established measures of a relevant concept. Within the COSMIN checklist, construct validity is divided into convergent and known-group validity, where the scale is compared to a comparison instrument measure or different subgroups. Face validity assesses the appropriateness of the scale at face value.

Responsiveness

Responsiveness reflects the ability to measure changes over time in the construct [3]. For measurements, it is important to detect changes related to time or interventions at all levels of the scale. Sometimes it may not be possible to detect subjects near the bottom or top of the scale, also known as the 'floor' or 'ceiling' effects. This means that when a person scores near the possible upper or lower limit, it can be difficult to identify changes over time [14].

Results

Selection process

The initial literature search produced 245 articles. After title and abstract screening, 30 articles were left for full-text screening. Six of those articles did not assess the properties of eADL scales, two included a cognitive scale rather than an eADL scale, and five were duplicates, leading to 17 articles for inclusion (Table 2 in the supplementary appendix). Hand searching and purposive searches identified an additional 9 eligible articles (Figure 1). The 26 papers included 15 eADL scales for evaluation. Table 1 provides an overview of the included eADL scales. The main items included in each scale are found in Table 2. Additional items are shown in Table 5 in the supplementary appendix.

Overview of questionnaires

Only six of the 15 scales exclusively assessed eADL, the remaining questionnaires assessed a combination of eADL and bADL. All the included scales were first described in the English language and were developed from 1969 through to 2018.

The Disability Assessment for Dementia scale (DAD), Blessed Dementia Rating Scale (Blessed-DS) and the Bristol Activities of

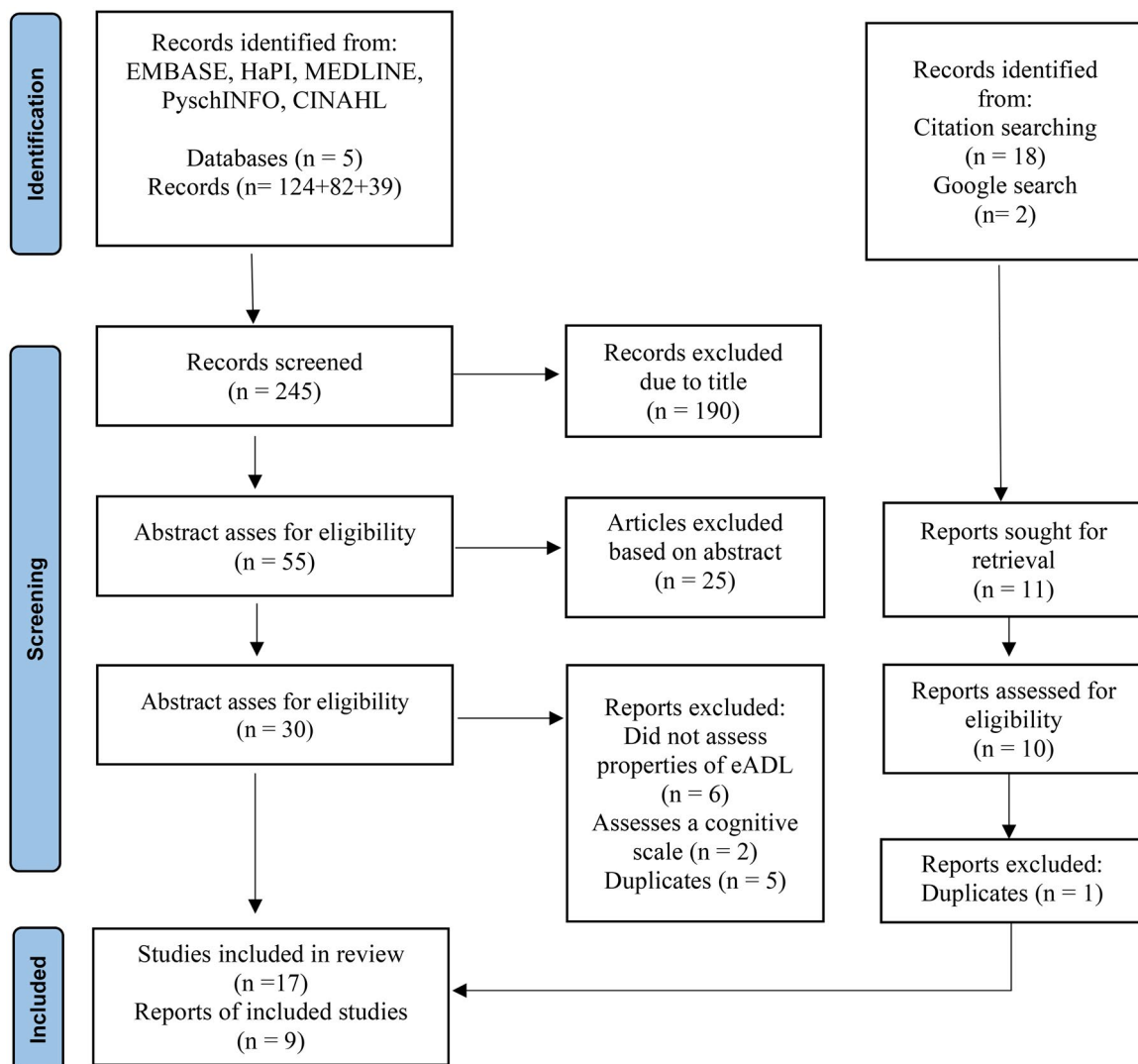


Figure 1. Prisma flowchart.

Table 1. Description of the included questionnaires.

Scale	Item no.	Assessment	Goal	Function measured	Year
Activities of daily living questionnaire (adlq) [16]	28	Questionnaire	Assessment of functional abilities in patients with ad and dementia	BADL + EADL	2004
Bayer activities of daily living scale (badl) [17]	25	Questionnaire	Assess deficits in patients with mci/mild-to-moderate dementia	BADL, EADL + COGNITION	1998
Bblessed dementia rating scale (blessed ds) [18]	22	Interview-based questionnaire	To quantify the degree of intellectual and personality deterioration in elderly	BADL, EADL AND BEHAVIOUR	1968
Bristol activities of daily living (bristol adl) [19]	20	Questionnaire	To assess the ability of patient with dementia in daily activities	BADL + EADL	1996
Cleveland scale for activities of daily living (csadl) [20]	47	Interview-based questionnaire	To assess functional difficulties of patients with dementia	BADL + EADL	2001
Disability assessment for dementia (dad) [21]	40	Interview-based questionnaire	To assess basic and instrumental daily activities in patients with dementia	BADL, EADL	1999
Interview for deterioration in daily living activities in dementia (idd) [22]	33	Interview-based questionnaire	To assess (e)adl in dementia	BADL + EADL	1997
Lawton & Brody instrumental activities of daily living scale (lawton iadl) [1,23–30]	8	Interview-based questionnaire	To assess eadl necessary for older people	EADL	1969
Performance-based instrumental activities of daily living (tpiadl) [31]	5	Performance-based assessments	To assess eadl necessary for older people	EADL	2014
International classification of functioning, disability and health (icf-iadl) [32]	8	Questionnaire	To assess eadl necessary for older people	EADL	2017
Self-care ability scale for the elderly (t-sase) [33]	17	Questionnaire	Assess self-care ability reviewed from cognitive, affective and behavioural component	BADL + EADL	1996
Amsterdam iadl questionnaire (a-iadl-q) [34,35]	70	Questionnaire	To assess complex adls necessary for older people	EADL	2012
The Nottingham extended activities of daily living (neadl) [36]	22	Questionnaire	Assess of iadl for use with patients recovering from stroke	EADL	1987
Functional independence measure (fim) [26]	18	Questionnaire	Assess disability in a variety of populations	BADL + EADL	1994
Performance based assessment of instrumental activities of daily living (pa-iadl) [37]	14	Performance-based assessments	Assess disability in a variety of populations	EADL	2018

* Basic activities of daily living (BADL) are basic selfcare activities to live independently. **Extended activities of daily living (eADL) are everyday tasks beyond BADL for a better quality of life.

Daily Living scale (Bristol ADL) are dementia-specific scales. The Activities of Daily Living Questionnaire (ADLQ) is specific for measuring Alzheimer's disease, and the Bayer Activities of Daily Living Scale (BADL) and Cleveland Scale for Activities of Daily Living (CSADL) were developed for use in people living with cognitive impairment. The NEADL was developed to assess recovery after a stroke. The Lawton IADL was the only generic, older adult assessment scale.

Assessment of psychometric properties

Table 3 provides an overview of the COSMIN assessment of the included articles. Convergent validity and reliability were the properties most often assessed, while structural validity was the least assessed. The properties that received a positive rating are $N=34$ (43%) (i.e. 'very good'/'adequate'), a doubtful rating was assigned to 25 properties (31%), and 21 properties (26%) received an inadequate rating. Overall, internal consistency received the best rating. The Lawton iADL scale was the most assessed scale and the TPIADL received the best ratings for the quality of the assessment of psychometric properties.

Table 4 provides a summary of the outcomes of the psychometric properties per scale, where data were available. Taking the most commonly used eADL scales of Lawton, we can see that Lawton has

an overall excellent internal consistency, a doubtful to inadequate reliability and a considerable doubtful convergent validity.

Except for the study of Patterson et al. [17] (*post hoc* analysis) and Stringer et al. [34] which studied the construct validity in the CSADL and the A-IADL-Q, all the outcomes were significant ($p < 0.05$). Most of the effect sizes and correlations were found to be 'large' or 'medium' by conventional criteria. However, two articles found a weak correlation when assessing the responsiveness and the criterion validity for the ICF-IADL and the Lawton iADL.

Construct validity

In Table 3, construct validity was divided by convergent and known-group validity. Of the 26 reviewed papers, $n=23$ (88%) assessed construct validity. Six assessments were labelled as 'very good' in the convergent validity group and five in the known-group validity group. The comparison measurement most used to assess construct validity were the Mini-Mental State Examination (MMSE) (9 articles) and the Lawton iADL scales (4 articles).

The paper by Tong and Man [23] assessed the construct validity of the Lawton iADL by assessing structural validation using factor analysis rather than a comparison measurement. However, since this does not follow the standard of the COSMIN checklist, the convergent validity is labelled as 'inadequate'.

Table 2. An overview of the main tasks included in eADL scales.

	Finances	Meal prep	Medication	Transportation	Reading	Shopping	Social behaviour	House Keeping	Laundry	Telephone/ Computer	Communication	Hobbies	Employment
ADLQ													
B-ADL													
Blessed-DS													
Bristol-ADL													
CSADL													
DAD													
IDD													
Lawton ADL													
NEADL													
TRIADL													
ICF-IADL													
T-SASE													
A-IADL-Q													
FIM													
PA-IADL													

Grey boxes signify that the item is part of the scale.

Table 3. Overview COSMIN reviewed properties.

Scale	Convergent validity	Known-group validity	Content Validity	Structural validity	Responsiveness	Reliability	Cross-cultural validity	Criterion validity	Internal consistency
Lawton IADL	Red	Grey	Red	Green		Yellow	Red	Green	Green
Siriwardhana et al. [24]									
Lawton IADL	Green	Green	Green	Green				Green	Green
Chen et al. [38]									
TPIADL	Red					Yellow			Grey
Ozkeskin et al. [33]									
T-SASE	Green					Yellow			
DAD	Green								
FIM	Green								
Ottenbacher et al. [38]									
Isik et al. [25]	Grey								
Lawton IADL	Grey								Green
Mehraban et al. [26]			Red						
Lawton IADL	Green								Green
Vergara et al. [27]	Red					Red			Red
Chen et al. [31]									
Hokoishi et al. [39]	Red					Grey			Green
Pérez et al. [40]	Red					Grey			Green
Stringer et al. [34]	Red					Grey			Green
Ng et al. [28]	Yellow	Yellow							Green
Chuang et al. [32]	Yellow								Green
Wang et al. [29]	Green								Green
Sikkies et al. [9]	Green								Green
Kadar et al. [30]	Red								Green
Mirzadeh et al. [1]	Green								Green
Cole et al. [18]	Red								Green
Johnson et al. [16]	Red								Green
Erzigkeit et al. [17]	Yellow								Green

Green = very good; Yellow = adequate; grey = doubtful; Red = inadequate.

Most articles assessed the convergent validity by calculating Pearson's correlation coefficient. However, the study by Stringer et al. [34] calculated the construct with Kendall's Tau-B, rather than Spearman's correlation for the comparison of the A-IADL-Q and the Addenbrooke's Cognitive Examination-III (0.11; $p=0.4$), the Digit Span Backwards Task (0.38; $p=0.46$), the Trails Making Test B (-0.04 ; $p=0.77$), and the Measurement of Everyday Cognitive Function (-0.46 ; $p=0.00$). Patterson and Mack [20] investigated the known-group validity among healthy older adults, physically impaired participants, and three groups of people with Alzheimer's disease and different levels of cognitive impairment. Their analysis reported that all between-group differences were significant except that between the healthy and physically impaired older adult groups ($p<0.064$).

Criterion validity

Six articles assessed the criterion validity for the PIADL, ICF-IADL BADL and the Lawton IADL scale. Only one article assessing the Lawton IADL received an inadequate rating. Since eADL scales lack an agreed gold standard, many other measurement scales were used that were assumed to have a positive or negative correlation with the scale of interest. Hence, Chuang et al. [32] used five measurement scales to investigate the criterion validity for the ICF-IADL. They showed that the ICF-IADL significantly correlated with the Lawton IADL scale ($r=-0.574$ to -0.804 , $p<0.01$), Montreal Cognitive Assessment (MoCA) ($r=-0.517$, $p<0.0$), Digit Symbol Substitution Test (DSS) ($r=-0.380$, $p<0.01$), Words Lists Test (WLT) ($r=-0.290$ to -0.437 , $p<0.01$), and Time Up and Go Test (TUG) ($r=0.404$ to 0.606 , $p<0.01$).

In Table 4, the study of Chen et al. [31] shows a range of accuracy (based on ROC analyses) between 0.53 and 0.91 ($p<0.001$) for the TPIADL scale. They compared outcomes for different groups; healthy participants, participants with mild cognitive impairment (MCI), and participants with dementia. This showed a higher ROC outcome for the dementia group and healthy subjects (0.91). However, when looking at healthy subjects and the MCI group, the results showed the ROC curve to be 0.53 ($p=0.62$). Mirzadeh et al. [1] and Chuang et al. [32] found a weak correlation between the Lawton IADL and the 36-item Short Form Mental Component Summary (SF36-MCS) ($r=0.09$, $p<0.01$) and between the ICF-IADL and the MoCA ($r=-0.247$, $p<0.05$).

Content validity

Three articles assessed content validity. Two of them were labelled 'inadequate' and one 'doubtful'. These low ratings were due to the limited scope of the consultation. The COSMIN checklist requires that both the opinion of healthcare professionals and participants are taken into account, while the included articles only consulted healthcare professionals. All three articles assessed the Lawton iADL scale.

There is no clear pattern in the content validity of the different articles assessing the Lawton IADL. For example, Mehraban et al. [26] found that the lowest homogeneity and agreement was for shopping and housekeeping, this was rated as average. Agreement was highest in the subheading medication management. Conversely, Tong and Man [23] found the lowest agreement among healthcare professionals in the subcategory handiwork and not in the same items as the other articles.

Responsiveness

Three articles assessed the responsiveness of the Lawton iADL, NEADL and the ICF-IADL scale. Only one article assessing the Lawton iADL covered all three subcategories included in the COSMIN assessment; a comparison between other outcome measures, between subgroups, and before and after an intervention [21]. The other articles reviewing the ICF-ADL and NEADL were labelled 'inadequate' as only one of the three subcategories was covered.

To assess responsiveness, Vergara et al. [27] excluded patients who had an improved Barthel Index, due to the small sample size ($n=7$). This study used the standardized effect size (SES) and the standardized response mean (SRM) to measure change. These are both effect size indexes to measure the responsiveness of outcome measures, calculated by dividing the mean change scores by the standard deviation of the baseline scores for the SES and the standard deviation of the change scores for the SRM. They found an SES of 0.79 and an SRM of .84 among the group classified as worsened after the intervention, indicating a moderate to large change. Otherwise, among the unchanged group, the SES and SRM were 0.31 and 0.38, indicating a small change.

Harwood et al. [36], initially found a small effect size of 0.1–0.3 of the improvements in EADL total. However, after adopting Likert-type scoring, the responsiveness improved considerably. The subscale mobility improved with an effect size of 0.7 at six months, and the other subscales reached a total score effect of 0.4–0.5.

Reliability

Reliability was investigated for the Lawton iADL, FIM, ADLQ, B-ADL, Bristol iADL, IDDD, and NEADL. None of the assessments was labelled as 'very good'. This is mainly because it was not clearly stated whether the test conditions were similar in the re-test after the time interval. Mostly, the re-test interval was around 7 days, and it could be assumed that the participants were stable in that period. In Table 4, reliability is divided into test-retest reliability and interrater reliability, while some articles did not assess both. Most articles used the Interclass Correlation Coefficient (ICC) to test the reliability, however, Spearman's rank coefficient and Kappa were also used in some articles.

The study by Hokoishi et al. [39] described the interrater reliability for the physical self-maintenance scale (PSMS) as well as for the Lawton iADL and found a variety of correlations between different healthcare professionals (i.e. neuropsychiatrists, public health nurse, clinical psychologist, neurologist, occupational therapist). For the PSMS they found an interclass correlation coefficient between 0.847 and 0.962 and for the Lawton IADL between 0.901 and 0.95 ($p<0.001$).

Table 4 states the overall interrater reliability of the Blessed Dementia scale, however, Cole et al. [18] used three additional methods to investigate interrater validity. They found a low (i.e. $r<0.7$) interrater reliability for the total Blessed Dementia Scale ($r.59$), for the intra-class scores ($r 0.297$), and different correlations for the item scores ranging from 0.04 for 'increased rigidity' and 0.64 for the 'inability to interpret surrounding' item.

The study by Johnson et al. [16], also tested the test-retest reliability in multiple ways. They provided the concordance coefficient (0.65 $<r<0.96$), the correlation coefficient (0.65 $<r<0.96$), the means, standard errors, and ranges for each subscale (SD 14.7–47.9), and the kappa (0.42 $<k<1.00$) to calculate the reliability for the ADLQ.

Internal consistency

The property assessment for internal consistency received a 'very good' rating in 64% of the articles ($N=9$). The most common reason for an article to be rated 'doubtful' or 'inadequate' was the lack of an internal consistency statistic calculation for each uni-dimensional (sub)scale separately.

Articles assessing the internal consistency reported a Cronbach's alpha for the TPIALD, T-SASE, DAD, Lawton iADL, Bristol iADL, NEADL, and CSADL. The value ranged from .82 for the TPIALD to 0.98 for the IDDD. Although some studies also reported the factor analysis, this calculation is not a part of the COSMIN checklist and will not be discussed here.

Other properties

Two of the least investigated properties are cross-cultural- and structural validity. Structural validity was only assessed once for the Lawton iADL scale [19]. It was calculated by using confirmatory factor analysis and ranged from 0.660 to 0.958 ($p < 0.001$). The cross-cultural validity was investigated in two articles for the Lawton iADL (inadequate) and Amsterdam iADL scale (doubtful) [19,28].

The Lawton scale

The Lawton iADL scale was the only scale assessed by more than three articles. Ten articles assessed the Lawton iADL scale and overall it showed a strong effect size. However, a moderate effect size was found in the study by Siriwardhana et al. [24] in the inter-rater reliability for investigators B and E, the other investigators scored a large effect size. Next to that, in the study by Wang et al. [29] the Lawton iADL scale received a moderate agreement ($\kappa = 0.51-0.66$) for the test-retest reliability. Furthermore, the Lawton iADL scale in the study of Mirzadeh et al. [1], showed a low correlation with the Short Form-36 Mental Component Summary (SF36-MCS) ($r=1.82$), which is logical as the Lawton iADL scale is not a mental measurement scale. However, it showed only a moderate correlation ($r=.563$) with the Barthel index.

Discussion

This systematic review shows that comprehensive assessments of psychometric properties for eADL scales are lacking and that none of the included articles provides a comprehensive description. Our review assessed 26 articles consisting of 15 different eADL scales and despite all the available analyses, it is not clear that one scale is superior to any other. The majority of the scales showed a large effect size regarding internal consistency and reliability. Convergent validity received the lowest effect size.

Our results should be considered in the context of existing literature. A previous review considered eADL scales designed for use in dementia settings [9]. This review found results in line with our review. Although the number of articles rated positive for methodological quality in this study is higher, still more than half of the articles were rated negative and no article included the assessment of all properties.

Similar findings were found in the study of Hopman-Rock et al., who investigated the psychometric properties in bADL scales. After a thorough literature search, they concluded that information about psychometric properties is mostly not sufficiently included [41]. In line with their findings, the included articles in the current

review showed either a lack of information about the properties or very detailed information on specific properties only, which makes the comparison between articles and scales a difficult task.

Our review found a difference in the frequency of the psychometric properties assessed. The reason construct validity may be assessed more frequently is that it involves a relatively accessible research method. To assess this property, the researcher adds a questionnaire or assessment and compares it with contemporaneous eADL scales. It does not require additional participants or substantial extra time. Within the COSMIN assessment, many properties were labelled negative, including construct validity. This might be considered strict, as a paper would be negatively assessed even if the rest of the method was conducted robustly. However, when a comparison is made without a true understanding of the properties, it does not provide a valuable evaluation.

In terms of eADL assessments, there is no agreed gold standard. Therefore, it seems more appropriate to assess construct validity rather than criterion validity in the found articles. However, the current article also distinguished between the two properties to provide a clear overview. Assessing factors such as cognition seems reasonable but is not a perfect comparator. As mentioned in the criterion validity results section, many articles use different measurement instruments to compare the eADL scale and this makes the interpretation of the results difficult.

Reliability is another important property when using an eADL scale and this property was also commonly assessed. When evaluating the functional abilities of an older adult or monitoring disease outcomes such as dementia or stroke, it is essential to repeat the questionnaire over time and intra-observer variability should be minimal. Equally in contemporary healthcare, eADL assessments may be performed by differing healthcare team members, emphasising the need for interobserver reliability. When assessing the intra-observer reliability, most articles selected a time interval of 7 days. In the context of eADL, 7–10 days are seen as an appropriate time interval while there is enough time in between the two tests to rule out the recall of the questions, but not too much time for health conditions to change significantly. The overall effect size of the reliability was large, which means that the scales can produce similar results under consistent conditions. However, with around two-thirds of the assessment quality scores labelled as inadequate, it remains unclear if the change in outcome is due to the functional shift in the participants or to external factors.

Another aspect of reliability is the internal consistency of the items in the eADL scale. For an eADL scale to assess the true functional ability, it is useful to know correlations between the different items, ensuring that the whole scale assesses functional ability and results are not biased by an item measuring a different quality. In the current review, internal consistency was rated positively with around half of the scales having an assessment labelled 'very good'. A possible explanation for this might be that the only requirement for assessing this property is to have sufficient individual participant data to calculate Cronbach's alpha for every subcategory. This is a quick and convenient assessment, and it does not involve additional participants or resources. It could be asked why only 14 articles assessed this important property. For internal consistency, the articles used the same outcome measurement. This makes it possible to have a valuable comparison. No clear explanation is given in the article why the internal consistency of the Lawton iADL scale is lower in comparison to other articles studying the same scale. The IDDD and the BADL received the highest score regarding internal consistency. However, this high internal consistency could implicate that the different items correlate too much and there may be redundancy in the items included in the scale.

The remaining psychometric properties described in COSMIN are other aspects of validity, namely content-, structural-, and cross-cultural validity. These properties have been studied the least. This may relate to the complexity of the method. For content validity research, the researcher must organize focus groups with healthcare professionals and people with experience in the condition of interest for a valuable content evaluation.

Cross-cultural validation was evaluated in two articles and received the labels 'inadequate' and 'doubtful'. Generally, it may not be relevant to measure cross-cultural validity, however, within this systematic review, most included articles adopted the eADL scale for a different culture. For a proper integration of the eADL scale in a different culture, it is essential to assess the cross-cultural validity. The poor evaluation of this property might be because it is a complex and time-consuming task. However, without this assessment, it is unclear whether the new scale is meaningful, applicable, and thus equivalent in the new culture [42].

Our review had several strengths and limitations. To begin with, this study did not distinguish between scales for different healthcare purposes and included every eADL scale. Furthermore, this is the first systematic review assessing the psychometric properties of eADL scales using the standardized methodological COSMIN checklist. This checklist provides a systematic method for the reviewer to investigate the value of the properties and compare different eADL scales even with different outcome measures. However, the use of the COSMIN Checklist to assess the articles could be criticised. The checklist used the 'worst-score-counts method', which means that the lowest-scored rating in a box determines the final rating of the property. This method caused more ratings to be inadequate, even when the mean assessment was 'very good'. Next to that, the COSMIN Checklist leaves space open for the researcher's judgement. Statements such as 'is a clear description provided' or 'are there any other important flaws' have a degree of subjectivity. To limit differences in interpretations, our study used two individual reviewers to assess the methodological quality and when no consensus was formed, a third reviewer was consulted. Furthermore, most articles assessed the measure properties after a cross-cultural adaptation. Therefore, the measurement properties of the adapted version were assessed, rather than the original scale. Although the assessment will still be valuable, it would be preferable to assess the original measure. Lastly, the COSMIN checklist does not include an assessment of the costs or time it takes to perform the eADL evaluation. eADL scales should be an efficient and low-cost tool to help assess older adults' functional abilities, which can be used in healthcare settings and research. Both of these items could be incorporated into the content validity, while patients and healthcare professionals already provide their opinion about the accessibility of the scale in this property.

Taking all the information above into consideration, it is a difficult task to compare the different scales and choose a superior scale. The most educated choice would be the Lawton iADL scale, simply because it has the largest supporting evidence base and it provides the most information to make an informed choice. However, per psychometric properties, the Lawton iADL scale does not always have the largest effect size. Within the convergent validity and internal consistency, the BADL has the largest effect size. Even though most of the outcome measures are high for reliability, the DAD scale is among the largest effect size for inter-rater reliability, as well as for re-test reliability. While no article assessed every property and only one article received positive ratings for every assessment (only construct validity was tested) [35], it is impossible to make a well-advised recommendation. In order to select a proper eADL scale, additional research

consisting of a more comprehensive and adequate methodological quality assessment per eADL scale should be performed.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Terence J. Quinn  <http://orcid.org/0000-0003-1401-0181>

References

- [1] Mirzadeh FS, Alizadeh-Khoei M, Sharifi F, et al. Validity and reliability: the Iranian version of Lawton IADL in elderly community dwellers. *JPMH*. 2020;19(3):241–250. doi: [10.1108/JPMH-05-2020-0036](https://doi.org/10.1108/JPMH-05-2020-0036).
- [2] Fried TR, Tinetti ME, Iannone L, et al. Health outcome prioritization as a tool for decision making among older persons with multiple chronic conditions. *Arch Intern Med*. 2011; 171(20):1856–1858. doi: [10.1001/archinternmed.2011.424](https://doi.org/10.1001/archinternmed.2011.424).
- [3] Edemekong PF, Bomgaars DL, Sukumaran S, et al. Activities of daily living. *Encycl Neurol Sci [Internet]*. 2022. 47–8.
- [4] Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. *Gerontologist*. 1969;9(3):179–186.
- [5] Wicklund AH, Johnson N, Rademaker A, et al. Profiles of decline in activities of daily living in non-Alzheimer dementia. *Alzheimer Dis Assoc Disord*. 2007;21(1):8–13. doi: [10.1097/WAD.0b013e3180324549](https://doi.org/10.1097/WAD.0b013e3180324549).
- [6] Tozlu M, Cankurtaran M, Yavuz BB, et al. Functional disability in Alzheimer disease: a validation study of the Turkish version of the disability assessment for dementia scale. *J Geriatr Psychiatry Neurol*. 2014;(4):237–246. doi: [10.1177/0891988714532014](https://doi.org/10.1177/0891988714532014).
- [7] Department of Economic and Social Affairs of the United Nations. *World population ageing*. 2019. P. 1–64.
- [8] Asunta P, Viholainen H, Ahonen T, et al. Psychometric properties of observational tools for identifying motor difficulties – a systematic review. *BMC Pediatr*. 2019;19(1):1–13.
- [9] Sikkes SAM, De Lange-De Klerk ESM, Pijnenburg YAL, et al. A systematic review of instrumental activities of daily living scales in dementia: room for improvement. *J Neurol Neurosurg Psychiatry*. 2009;80(1):7–12. doi: [10.1136/jnnp.2008.155838](https://doi.org/10.1136/jnnp.2008.155838).
- [10] Morkink Cecilia AC, Prinsen Donald L, Patrick Jordi Alonso Lex M, et al. COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of patient-reported Outcome Measures (PROMs) user manual. 2018. [cited 2022 Aug 3]. Available from: www.cosmin.nl.
- [11] Morkink LB, De Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1171–1179. doi: [10.1007/s11136-017-1765-4](https://doi.org/10.1007/s11136-017-1765-4).
- [12] Johnson G. Measurement scales used in elderly care abhaya gupta radcliffe measurement scales used in elderly care £24.95 168pp 9781846192661 1846192668. *Nurs Older People*. 2009;21(10):10–10. doi: [10.7748/nop.21.10.10.s9](https://doi.org/10.7748/nop.21.10.10.s9).

- [13] McHugh ML. Interrater reliability: the Kappa statistic. *Biochem Medica*. 2012;22(3):276.
- [14] Garin O. Ceiling effect. *Encycl Qual Life Well-Being Res*. 2014;31:631–633.
- [15] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;3:n71.
- [16] Johnson N, Barion A, Rademaker A, et al. The activities of daily living questionnaire: a validation study in patients with dementia. *Alzheimer Dis Assoc Disord*. 2004;18(4):223–230.
- [17] Erzigkeit H, Lehfeld H, Peña-Casanova J, et al. The Bayer-activities of daily living scale (B-ADL): results from a validation study in three European countries. *Dement Geriatr Cogn Disord*. 2001;12(5):348–358. doi: [10.1159/000051280](https://doi.org/10.1159/000051280).
- [18] Cole MG. Interrater reliability of the blessed dementia scale. *Can J Psychiatry*. 1990;35(4):328–330.
- [19] Bucks RS, Ashworth DL, Wilcock GK, et al. Assessment of activities of daily living in dementia: development of the Bristol activities of daily living scale. *Age Ageing*. 1996;25(2):113–120. doi: [10.1093/ageing/25.2.113](https://doi.org/10.1093/ageing/25.2.113).
- [20] Patterson MB, Mack JL. Cleveland scale for activities of daily living (CSADL): its reliability and validity. *J Clin Geropsychology*. 2001;7(1):15–28.
- [21] Gelinas I, Gauthier L, McIntyre M. Development of a functional measure for persons with Alzheimer's disease: the disability assessment for dementia. *Am J Occup Ther*. 1999;53(2):471–481.
- [22] Böhm P, Peña-Casanova J, Aguilar M, et al. Clinical validity and utility of the interview for deterioration of daily living in dementia for Spanish-speaking communities NORMACODEM group. *Int Psychogeriatr*. 1998;10(3):261–270.
- [23] Tong AYC, Man DWK. The validation of the Hong Kong Chinese version of the Lawton instrumental activities of daily living scale for institutionalized elderly persons OTJR: occupation, participation and health. 2002
- [24] Siriwardhana DD, Walters K, Rait G, et al. Cross-cultural adaptation and psychometric evaluation of the sinhala version of lawton instrumental activities of daily living scale. *PLoS One*. 2018;13(6):e0199820.
- [25] Isik EI, Yilmaz S, Uysal I, et al. Adaptation of the lawton instrumental activities of daily living scale to turkish: validity and reliability study. *Ann Geriatr Med Res*. 2020;24(1):35–40. doi: [10.4235/agmr.19.0051](https://doi.org/10.4235/agmr.19.0051).
- [26] Hassani Mehraban A, Soltanmohamadi Y, Akbarfahimi MT. Validity and reliability of the persian version of Lawton instrumental activities of daily living scale in patients with dementia. *J Islam Repub Iran*. 2014;3(28):25.
- [27] Vergara I, Bilbao A, Orive M, et al. Validation of the Spanish version of the Lawton IADL scale for its application in elderly people. *Health Qual Life Outcomes*. 2012. doi: [10.1186/1477-7525-10-130](https://doi.org/10.1186/1477-7525-10-130).
- [28] Ng TP, Niti M, Chiam PC, et al. Physical and cognitive domains of the instrumental activities of daily living: validation in a multiethnic population of asian older adults. *J Gerontol A Biol Sci Med Sci*. 2006;61(7):726–735. doi: [10.1093/gerona/61.7.726](https://doi.org/10.1093/gerona/61.7.726).
- [29] Wang CY, Hu MH, Chen HY, et al. Self-reported mobility and instrumental activities of daily living: test-retest reliability and criterion validity. *J Aging Phys Act*. 2012;20(2):186–197. doi: [10.1123/japa.20.2.186](https://doi.org/10.1123/japa.20.2.186).
- [30] Kadar M, Ibrahim S, Razaob NA, et al. Validity and reliability of a Malay version of the Lawton instrumental activities of daily living scale among the Malay speaking elderly in Malaysia. *Aust Occup Ther J*. 2018;65(1):63–68. doi: [10.1111/1440-1630.12441](https://doi.org/10.1111/1440-1630.12441).
- [31] Chen HM, Lin HF, Huang MF, et al. Validation of Taiwan performance-based instrumental activities of daily living (TPIADL), a performance – based measurement of instrumental activities of daily living for patients with vascular cognitive impairment. *PLoS One* 2016;11(11):e0166546. doi: [10.1371/journal.pone.0166546](https://doi.org/10.1371/journal.pone.0166546).
- [32] Chuang IC, Hsu WC, Chen CL, et al. Psychometric evaluation of an ICF-Based instrumental activities of daily living assessment with older adults with cognitive decline. *Am J Occup Ther*. 2020;74(6):7406205050p1–7406205050p8.
- [33] Özkeskin M, Özden F, Şahin S. Translation, cross-cultural adaptation, and psychometric properties of the turkish version of the self-care ability scale for the elderly. *Ann Geriatr Med Res*. 2021;25(2):122–128. doi: [10.4235/agmr.21.0046](https://doi.org/10.4235/agmr.21.0046).
- [34] Stringer G, Leroi I, Sikkes SAM, et al. Enhancing “meaningfulness” of functional assessments: UK adaptation of the Amsterdam IADL questionnaire. *Int Psychogeriatr*. 2021;33(1):39–50. doi: [10.1017/S1041610219001881](https://doi.org/10.1017/S1041610219001881).
- [35] Sikkes SAM, Knol DL, Pijnenburg YAL, et al. Validation of the Amsterdam IADL questionnaire, a new tool to measure instrumental activities of daily living in dementia. *Neuroepidemiology*. 2013;41(1):35–41. doi: [10.1159/000346277](https://doi.org/10.1159/000346277).
- [36] Harwood RH, Ebrahim S, Harwood RH, et al. Disability and rehabilitation the validity, reliability and responsiveness of the Nottingham extended activities of daily living scale in patients undergoing total hip replacement the validity, reliability and responsiveness of the nottingham extended activities of daily living scale in patients undergoing total hip replacement. 2009. [cited 2022 Sep 2]; Available from: <https://www.tandfonline.com/action/journalInformation?journalCode=idre20>.
- [37] Chen HM, Yeh YC, Su WL, et al. Development and validation of a new performance-based measurement of instrumental activities of daily living in Taiwan. *Psychogeriatrics*. 2015;15(4):227–234. doi: [10.1111/psyg.12096](https://doi.org/10.1111/psyg.12096).
- [38] Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Arch Phys Med Rehabil*. 1996;77(12):1226–1232.
- [39] Hokoishi K, Ikeda M, Maki N, Nomura M, Torikawa S, Fujimoto N, Fukuhara R, Komori K, Tanabe H. Interrater reliability of the physical self-maintenance scale and the instrumental activities of daily living scale in a variety of health professional representatives. *Aging Ment Health*. 2001;5(1):38–40.
- [40] Pérez L, Antonio J, Menor J. Development and validation of a performance-based test to assess instrumental activities of daily living in Spanish older adults. *Eur J Psychol Assess*. 2018;34(6):386–398. doi: [10.1027/1015-5759/a000352](https://doi.org/10.1027/1015-5759/a000352).
- [41] Hopman-Rock M, van Hirtum H, de Vreede P, et al. Activities of daily living in older community-dwelling persons: a systematic review of psychometric properties of instruments. *Aging Clin Exp Res*. 2019;31(7):917–925. doi: [10.1007/s40520-018-1034-6](https://doi.org/10.1007/s40520-018-1034-6).
- [42] A guide for cross-cultural validation of measurement instruments in mental health [Internet]. [cited 2022 Sep 6]. Available from: https://www.researchgate.net/publication/285334108_A_guide_for_cross-cultural_validation_of_measurement_instruments_in_mental_health.