Applied Neuropsychology: Adult

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hapn21

Ecological validity in neurocognitive assessment: Systematized review, content analysis, and proposal of an instrument

Joana O. Pinto, Artemisa R. Dores, Bruno Peixoto & Fernando Barbosa

To cite this article: Joana O. Pinto, Artemisa R. Dores, Bruno Peixoto & Fernando Barbosa (08 Feb 2023): Ecological validity in neurocognitive assessment: Systematized review, content analysis, and proposal of an instrument, Applied Neuropsychology: Adult, DOI: 10.1080/23279095.2023.2170800

To link to this article: https://doi.org/10.1080/23279095.2023.2170800

APPLIED NEUROPSYCHOLOGY: ADULT
https://doi.org/10.1080/23279095.2023.2170800

Ecological validity in neurocognitive assessment: Systematized review, content analysis, and proposal of an instrument

Joana O. Pinto[a,b,c], Artemisa R. Dores[a,b,d], Bruno Peixoto[c,e,f], and Fernando Barbosa[a]

[a]Laboratory of Neuropsychophysiology, Faculty of Psychology and Education Sciences, University of Porto, Porto, Portugal; [b]ESS, Polytechnic of Porto, Porto, Portugal; [c]CESPU, University Institute of Health Sciences, Gandra, Portugal; [d]Center for Rehabilitation Research, ESS, Polytechnic of Porto, Porto, Portugal; [e]NeuroGen, Center for Health Technology and Services Research (CINTESIS), Porto, Portugal; [f]TOXRUN – Toxicology Research Unit, University Institute of Health Sciences, CESPU, Gandra, Portugal

ABSTRACT

Objectives: The main objectives of this study are to identify the dimensions of Ecological Validity (EV) within the definitions of this concept, understand how they are operationalized in neurocognitive tests, and propose a checklist for EV attributes in neurocognitive tests.

Method: A systematized review was combined with content analysis of the selected papers, using the inductive method. We analyzed 82 studies on the EV of neurocognitive tests, 19 literature reviews and 63 empirical studies. Based on this review, we identified the relevant criteria for evaluating EV.

Results: EV is a multidimensional concept with two main dimensions: representativeness and generalization. Representativeness involves the subdimensions simplicity-complexity and artificial-natural and several criteria organized on a continuum from low EV to high EV. Generalization is dependent on representativeness and is influenced by different cognitive and non-cognitive factors. We propose six stages for operationalizing EV, from defining the objectives of the neurocognitive assessment to the methodology for scoring and interpreting the results.

Conclusion: This systematized review helps to operationalize the concept of EV by providing a tool for evaluating and improving EV while developing new tests. Further studies with a longitudinal design can compare the predictive value of tests with higher versus lower EV-checklist scores.

KEY POINTS

- Question: Understand the definition of EV, its dimensions and subdimensions, how EV is operationalized in neurocognitive tests and propose a checklist for the EV attributes of neurocognitive tests.
- Findings: The primary findings were that representativeness and generalization are the main dimensions of EV. Representativeness involves several subdimensions, whereas generalization is dependent on representativeness and is influenced by cognitive and non-cognitive factors. We provided an EV-checklist organized into six parts.
- Importance: The EV-checklist can be used to guide the development of ecologically valid neurocognitive tests and/or assess the EV of existing ones.
- Next steps: Examine the predictive value of tests that have higher EV-checklist scores.

KEYWORDS
Checklist; content analysis; ecological validity; neurocognitive assessment; systematized review

Introduction

The primary goal of neurocognitive assessment is to improve the lives of people with neurocognitive disorders (Woods, 2021), regardless of whether the request is to identify cognitive decline (Silverberg & Millis, 2009), plan neuropsychological rehabilitation (Zgaljardic et al., 2011), or predict functionality in daily living, occupational, and community activities (Holleman et al., 2020; Kibby et al., 1998; Silverberg & Millis, 2009; Tang et al., 2018). However, only about 1% of the studies published in the main neuropsychology journals over the past 35 years, have focused on the relationship between cognition and activities of daily living (Woods, 2021). A reason for this could be that cognitive functioning in everyday life is still assessed and recorded in an unstructured manner (Domensino & van Heugten, 2020). Another important factor that may lead to less research in this field is the controversy surrounding the ecological validity (EV) of neurocognitive tests, starting with the lack of consensus regarding the definition of EV and its operationalization (Aubin et al., 2018; Holleman et al., 2020; Lewkowicz, 2001; Weber et al., 2019).

CONTACT Joana O. Pinto joanafftopinto@gmail.com Laboratório de Neuropsicofisiologia, Faculdade de Psicologia e Ciências da Educação, Universidade do Porto, Rua Alfredo Allen, Porto 4200-135, Portugal.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/23279095.2023.2170800.

The first definition of the term EV was proposed by Brunswik (1955), to refer to the conditions that favor the generalization of results obtained in experimentally controlled situations to natural environments (Tupper & Cicerone, 1990). Later, the EV concept was defined as the functional and predictive relationship between individuals' performance in neurocognitive assessment and their behavior in real-world environments (Sbordone, 1996). Even though the argument of the need for an ecological model of neuropsychology dates back to 1990 (Tupper & Cicerone, 1990), researchers rarely explain what they mean by the term and what criteria are used to establish higher EV (Holleman et al., 2020). Among the most cited approaches to establishing the EV of neurocognitive assessment measures are *verisimilitude* and/or *veridicality*.

According to Franzen and Wilhelm (1996), verisimilitude refers to the degree to which the cognitive requirements of a neurocognitive test resemble the requirements found in the person's daily living environment (Chaytor & Schmitter-Edgecombe, 2003; Spooner & Pachana, 2006). In turn, veridicality refers to the degree to which existing neurocognitive tests empirically relate to measures of daily functionality, i.e., to their ability to predict the individuals' functionality in daily living (Chaytor & Schmitter-Edgecombe, 2003; Franzen & Wilhelm, 1996; Spooner & Pachana, 2006). In addition, Parsons (2011) proposed four criteria to ensure the EV of neurocognitive tests: (a) correspondence—tasks should correspond to relevant aspects of real-world activities and environments; (b) representativeness—tasks should be representative of the population for whom they were developed, considering their cultural knowledge; (c) expedience—the domains assessed should have practical implications for functioning in daily activities; (d) relevance—the outcome measures should be relevant to the domain being assessed.

Given the relevance of EV in neuropsychology, several literature reviews have been developed to systematize the available knowledge in terms of the: (a) concept of EV (Lewkowicz, 2001) and its dimensions (Schmuckler, 2001); (b) criticisms regarding the concept of EV (Holleman et al., 2020); (c) EV of neuropsychological measures (Barkley, 1991; Chaytor & Schmitter-Edgecombe, 2003; Olson et al., 2013; Poletti, 2010; Romero-Ayuso et al., 2019; Silver, 2000; Spooner & Pachana, 2006; Wallisch et al., 2018; Weber et al., 2019; Wilson, 1993); (d) processes and approaches used in the development of tests with good EV for the evaluation of executive functions (Burgess et al., 2006; Parsons et al., 2015); (e) role of the representativeness of the tasks in EV (Dhami et al., 2004); (g) potential of virtual reality to increase EV in cognitive, clinical, affective, and social neurosciences (Parsons, 2011, 2015); and (f) role of multisensory integration in EV (De Gelder & Bertelson, 2003).

Previous studies have highlighted the need for an objective and operational definition of the EV concept, thus enabling the evaluation of tests and their stimuli or tasks as more or less ecologically valid (Lewkowicz, 2001) and serving as a guide for developing, conducting, and interpreting research in the field of neuropsychology (Schmuckler, 2001). To our knowledge, this is the first article that combines a systematized review, aimed at identifying studies that focus on the analysis of EV in neurocognitive assessment, with a qualitative content analysis of the identified studies. The content analysis was intended to: (a) understand how the concept of EV is defined; (b) identify its underlying dimensions and subdimensions; and (c) understand how it is operationalized in neurocognitive tests. Based on the results of the systematized review and content analysis, this study also aimed to propose a checklist for assessing the EV of neurocognitive tests. The following questions guided this systematized review:

Question 1: How is EV defined in neuropsychology?

Question 2: What are the dimensions and subdimensions underlying the concept of EV?

Question 3: Which outcome measures are more commonly used to assess EV?

Question 4: What are the factors with a potential impact on the EV of a neurocognitive test?

Question 5: What are the stages and procedures involved in the development of neurocognitive tests with higher EV?

Method

A systematized review was performed to provide a basis for further analyses, as suggested by Grant and Booth (2009), and to find responses to the above-mentioned questions. It is worth noting that systematized reviews include several elements of a systematic review (see *Preferred Systematic Review and Meta-analysis*, PRISMA; Moher et al., 2009; Page et al., 2021), but not all requirements of the latter are met (Grant & Booth, 2009).

Research strategy

This systematized review involved a systematic search of papers on the topic of EV in *PubMed* and in the databases included in *EBSCO*, namely *PsycInfo*, *PsycArticles*, and *Psychology and Behavioral Sciences Collection*, followed by coding and analysis of the results in a systematic manner (Grant & Booth, 2009). The search expression included the term "ecological valid*" in titles.

Study selection

The criteria for selecting the studies were: (a) empirical studies on the EV of neurocognitive tests; and (b) reviews on the EV of neurocognitive tests. We applied no constraints regarding date of publication, but excluded studies (a) written in a language other than English, (b) outside the field of neuropsychology, (c) assessing social cognition, (d) not presenting relevance to the topic, (e) consisting of secondary literature, such as books, book chapters, dissertations, or thesis, and (f) involving animal models.

After eliminating duplicates, the studies were selected by a reviewer based on the reading of the abstracts.

Content analysis

A content analysis was performed on the selected studies, following a data-driven approach, according to the inductive model. The unit of analysis was the phrase. As proposed by Mayring (2014), two reviewers performed the stages of the inductive model for the development of categories, namely: (1) formulation of the research questions and description of the theoretical framework; (2) definition of categories and level of abstraction; (3) coding the entire text line-by-line, except for the results, as we did not intend to report quantitative data or the effect of different variables; (4) review of the categories formed after coding 50% of the texts; (5) final coding of all material according to the same rules (defined categories and level of abstraction); (6) organization of the main categories by grouping the previous ones or building new categories to answer the research questions; (7) assessment of intra-coding agreement; and (8) presentation of results. A peer agreement was reached after the fifth stage.

Results

A total of 455 articles were identified in the selected databases (*EBSCO* $n = 317$; *PubMed* $n = 138$). Seven articles were found by manual search in the databases described above and from the references of the selected publications. After excluding duplicates, we analyzed the titles and abstracts of 344 papers. A total of 258 articles were excluded for the following reasons: (a) a field of study other than neuropsychology ($n = 129$); (b) dissertations/thesis ($n = 43$); (c) books or book chapters ($n = 35$); (d) unrelated to the topic ($n = 33$); (e) type of paper other than empirical study or review ($n = 15$); (f) did not involve human samples ($n = 1$); (g) articles were not found ($n = 2$).

After reading the full text, ten articles were excluded for the following reasons: (a) type of paper other than empirical study or review ($n = 1$); (b) unrelated to the topic ($n = 8$); (c) study of a neurocognitive assessment instrument without evaluation of its EV ($n = 1$). Thus, 83 articles remained for analysis. The text of these articles was exported to *webQDA—Qualitative Data Analysis Software* (Sousa et al., 2019) and subjected to content analysis.

Study characteristics

The studies were published in 45 different journals. The journals with the highest number of records were *Archives of Clinical Neuropsychology* ($n = 10$) and the *Journal of the International Neuropsychological Society* ($n = 9$).

Of the 83 articles reviewed, 63 were empirical studies (76%), 19 literature reviews (23%), and one was a theoretical article presenting a methodology for assessing the EV of neurocognitive tests (1%).

The participants were adults in most studies ($n = 45$, 71%). A smaller percentage focused solely on the elderly ($n = 13$, 21%) and children ($n = 4$, 6%). The authors of one study did not specify the age of the participants. Most studies focused on clinical samples, especially with neurological pathology ($n = 44$, 70%), with acquired brain injury being the most studied pathology ($n = 20$, 32%). Only 18 (29%) studies recruited healthy and/or community samples (see Supplemental Material).

Response to the research questions

How is EV defined in neuropsychology?

Through the inductive content analysis, we identified four key concepts used to define EV: (a) representativeness (Alderman et al., 2003; Aubin et al., 2015; Barkley, 1991; Burgess et al., 2006; Campbell et al., 2009; Chaytor and Schmitter-Edgecombe, 2003; Schmuckler, 2001; Solanto et al., 2001; Wallisch et al., 2018; Wilson, 1993); (b) generalization (Aubin et al., 2015; Burgess et al., 2006; De Gelder and Bertelson, 2003; Hoc, 2001; Krishna et al., 2016; Owen et al., 2004; Parsons, 2015; Poletti, 2010; Schmuckler, 2001; Wallisch et al., 2018; Weis & Totten, 2004; Yantz et al., 2010); (c) prediction (Alderman et al., 2003; Azouvi et al., 2014; Barkley, 1991; Bowman, 1996; Bromley et al., 2012; Burgess et al., 2006; Campbell et al., 2009; Chaytor et al., 2007; Cuberos-Urbano et al., 2013; Davies et al., 2011; Drozdick & Cullum, 2011; Farias et al., 2003; Farley et al., 2011; Gioia, 2009; Gioia & Brekke, 2009; Groth-Marnat & Teal, 2000; Groth-Marnat & Baker, 2003; Higginson et al., 2000; Holleman et al., 2020; Kieffaber et al., 2007; Maeir et al., 2011; Mitchell & Miller, 2008; Odhuba et al., 2005; Owen et al., 2004; Pezzuti et al., 2013; Phillips et al., 2006; Poletti, 2010; Possin et al., 2014; Price et al., 2003; Ready et al., 2001; Renison et al., 2012; Silverberg and Millis, 2009; Solanto et al., 2001; Spooner and Pachana, 2006; Temple et al., 2009; Thornton et al., 2010; van der Elst et al., 2008; Verdejo-Garcia et al., 2006; Weis & Totten, 2004; Wilson, 1993; Wood and Liossi, 2006; Yantz et al., 2010; Ziemnik & Suchy, 2019); and (d) transfer (Hoc, 2001) (see Supplemental Material).

While *representativeness* refers to the correspondence between the form and context of the assessment and the situations in natural contexts (Burgess et al., 2006), *generalization* concerns the degree to which performance on a neurocognitive test can be predictive of the behavior in natural environments (Burgess et al., 2006). Thus, the term *prediction* can be conceived as representing the same phenomenon as generalization (Burgess et al., 2006). The concept of *transfer* seems to be intrinsically associated with the goals of traditional neurocognitive tests, assuming that such tests represent artificial situations, and their results need to be transferred to real-word situations, defined as natural (Hoc, 2001).

Summing-up, given that *prediction* and *transfer* can be dismissed, it seems reasonable to conclude that the main dimensions of EV are *representativeness* and *generalization*.

What are the dimensions and subdimensions underlying the concept of EV?

According to the studies reviewed, EV varies along a continuum from weak to high representativeness of the behaviour in the real world (Barkley, 1991). EV is presumed to be

weaker when artificial situations and simple environments are used, and to increase when natural situations and complex environments are used (Holleman et al., 2020). Therefore, at least two subdimensions may be considered in the dimensional analysis of representativeness: (a) simplicity-complexity (Holleman et al., 2020); (b) artificial-natural (Gioia, 2009; Helmstaedter et al., 1998; Holleman et al., 2020; Parsons, 2015). Generalization is influenced by representativeness and has no subdimensions, although it depends on cognitive and non-cognitive factors, as explained later in this review.

Which outcome measures are more commonly used to assess EV?

Regarding outcome measures, the authors mostly used self- and/or other-report measures to evaluate EV (Adjorlolo, 2016; Alderman et al., 2003; Benge et al., 2020; Burgess et al., 1998; Chaytor et al., 2006, 2007, 2020; Cuberos-Urbano et al., 2013; Davies et al., 2011; Dawson et al., 2009; Drozdick & Cullum, 2011; Farias et al., 2003; Gioia & Brekke, 2009; Goodman & Zarit, 1995; Higginson et al., 2000; Kibby et al., 1998; Kieffaber et al., 2007; Norris and Tate, 2000; Odhuba et al., 2005; Renison et al., 2012; van der Elst et al., 2008; Wen et al., 2006; Wood and Liossi, 2006). The most frequently reported were derived from the Dysexecutive Questionnaire (Wilson et al., 1996), although self-report and/or informant report variations were used (see Supplemental Material).

The statistical methods used to assess EV were not consensual. While most studies combined correlations and regressions (n = 26; 41%), one study using structural equation models recognized this procedure as promising for the development of a theoretical model of EV for neurocognitive tests (Kieffaber et al., 2007).

What are the factors with a potential impact on the EV of a neurocognitive test?

According to the studies reviewed, there are different factors that add variation in the representativeness and/or generalization (thus, in EV) of a given test in different samples and circumstances, using the same outcome measure (Chaytor and Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007).

The level of representativeness depends on the criteria related to: (a) setting (Barkley, 1991; Bromley et al., 2012; Burgess et al., 2006; Campbell et al., 2009; Dawson et al., 2009; Faith & Rempfer, 2018; Gioia, 2009; Helmstaedter et al., 1998; Holleman et al., 2020; Lewkowicz, 2001; Norris and Tate, 2000; Owen et al., 2004; Parsons, 2015; Phillips et al., 2006; Rumpf et al., 2019; Schmuckler, 2001; van der Ham et al., 2015); (b) task (Bromley et al., 2012; Burgess et al., 2006; Dawson et al., 2009; Faith & Rempfer, 2018; Gioia & Brekke, 2009; Holleman et al., 2020; Owen et al., 2004; Parsons, 2015; Phillips et al., 2006; van der Ham et al., 2015; Weber et al., 2019; Wilson, 1993); (c) stimuli (Dhami et al., 2004; Holleman et al., 2020; Lewkowicz, 2001; Parsons, 2015; Schmuckler, 2001); and (d) assessed behavioural response (Holleman et al., 2020; Lewkowicz, 2001; Schmuckler, 2001). For example, a more natural setting, with complex tasks and stimuli, which requires learning or applying a behaviour that can be used by the individual in real life, favours EV.

The factors with potential impact on the EV of a neurocognitive test were categorized into cognitive and non-cognitive factors (Chaytor and Schmitter-Edgecombe, 2003, Chaytor et al., 2006), either with the potential to increase or decrease EV. The inductive analysis allowed us to identify the following non-cognitive factors that may have an influence on EV: (a) task-related factors (Aubin et al., 2015; Chaytor and Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Cuberos-Urbano et al., 2013; Davies et al., 2011; Farias et al., 2003; Gioia & Brekke, 2009; Horan et al., 2020; Lewkowicz, 2001; Norris and Tate, 2000; Olson et al., 2013; Paiva et al., 2016; Parsons, 2015; Pezzuti, et al. 2013; Phillips et al., 2006; Renison et al., 2012; Rumpf et al., 2019; Thornton et al., 2010; van der Ham et al., 2015; Wilson, 1993); (b) sample-related factors (Bowman, 1996; Bromley et al., 2012; Chaytor and Schmitter-Edgecombe, 2003; Chaytor et al., 2006; Farias et al., 2003; Olson et al., 2013; Pezzuti et al., 2013; Price et al., 2003; Ready et al., 2001; Renison et al., 2012; van der Elst et al., 2008; Wood and Liossi, 2006; Ziemnik & Suchy, 2019; Zgaljardic et al., 2011); (c) patient-related factors (Aubin et al., 2015; Azouvi et al., 2014; Bowman, 1996; Chaytor and Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Dubreuil et al., 2007; Farias et al., 2003; Farley et al., 2011; Kibby et al., 1998; Maeir et al., 2011; Olson et al., 2013; Pezzuti et al., 2013; Phillips et al., 2006; Price et al., 2003; Ready et al., 2001; Renison et al., 2012; van der Elst et al., 2008; Wilson, 1993; Wood and Liossi, 2006; Zgaljardic et al., 2011); (d) informant-related factors (Chaytor and Schmitter-Edgecombe, 2003; Renison et al., 2012); and (e) other factors (Chaytor and Schmitter-Edgecombe, 2003; Chaytor et al., 2007; Lippa et al., 2014) (see Table 1).

Considering the different factors with potential impact on EV, the controversy and disagreement about EV (Kvavilashvili & Ellis, 2004) seem to be related to the need for the individualization of assessment to ensure adequate EV. Specifically, EV refers us to cognitive instrumental activities of daily living that may become more complex as the cognitive demands of the individual's environment increase. Furthermore, similar to neurocognitive training that usually includes activities with increasing complexity, it would be beneficial for EV if neurocognitive tests also included tasks or items with different levels of difficulty, ensuring the correspondence between the cognitive demand of the test and the cognitive demand of the environment in which the person functions, as advocated in several studies reviewed (Aubin et al., 2018; Chaytor et al., 2006; Chaytor & Schmitter-Edgecombe, 2003). Both cognitive and non-cognitive factors have the potential to impact the generalization of test results and, consequently, their EV.

What are the stages and procedures involved in the development of neurocognitive tests with higher EV?

Although none of the studies reviewed mentioned the main stages and procedures involved in the development of ecologically valid tests, they provide several important

Table 1. Factors with potential impact on ecological validity (EV).

	Factors with potential to increase EV	Factors with potential to decrease the EV	Factors that may have the potential to increase or decrease the EV	Role of factors in the interpretation of study results
Cognitive factors	• Use of compensatory strategies (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006; Farley et al., 2011; Goodman & Zarit, 1995; Kibby et al., 1998; Pezzuti et al., 2013; Phillips et al., 2006; Renison et al., 2012; Rumpf et al., 2019; Zgaljardic et al., 2011) • Consideration of the different cognitive functions underlying performance on the task, not just the cognitive function that is intended to be assessed (Burgess et al., 1998; Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006; Drozdick & Cullum, 2011; Dubreuil et al., 2007; Farley et al., 2011; Kieffaber et al., 2007; Olson et al., 2013; Phillips et al., 2006; Possin et al., 2014; Renison et al., 2012; Wood & Liossi, 2006).	• Difficulty in identifying the construct assessed by a performance-based test given the multifaceted nature of real-life tasks, potentially compromising the correspondence between the cognitive function assessed by a neuropsychological assessment test and the outcome measure (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006; Drozdick & Cullum, 2011; Farley et al., 2011)	• Cognitive reserve (Azouvi et al., 2014; Rumpf et al., 2019) • premorbid cognitive level (Bromley et al., 2012; Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Farias et al., 2003; Renison et al., 2012)	
Non-cognitive factors Task-related factors	• Use of dynamic, three-dimensional stimuli (Lewkowicz, 2001; Parsons, 2015a; Rumpf et al., 2019) • Use of realistic stimuli (Aubin et al., 2015; Pezzuti et al., 2013; Rumpf et al., 2019) • Presence of distractor stimuli (Davies et al., 2011; Gioia & Brekke, 2009; Olson et al., 2013; Parsons, 2015a) • Familiarization with the stimuli (Aubin et al., 2015; Pezzuti et al., 2013; Phillips et al., 2006) • Familiarization with the task (Aubin et al., 2015; Thornton et al., 2010) • Presentation of contextual information, making the environment more realistic (Rumpf et al., 2019; van der Ham et al., 2015)	• Small sample behaviour observed during testing (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006) • Disparity between the time available for the task in the test situation and in the real context (Aubin et al., 2015; Wilson, 1993)	• Correspondence between the cognitive demand of the test and the cognitive demand of the environment in which you function in your daily life (Aubin et al., 2015; Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Davies et al., 2011; Farias et al., 2003; Olson et al., 2013; Paiva et al., 2016; Renison et al., 2012; Rumpf et al., 2019) • degree of task structure (Azouvi et al., 2014; Cuberos-Urbano et al., 2013; Davies et al., 2011; Norris & Tate, 2000) • Correspondence between physiological arousal on task and in real-life context (Horan et al., 2020)	
Sample-related factors		• For patients with acquired brain injury, the time between injury and evaluation (Chaytor & Schmitter-Edgecombe, 2003)		• Small sample size (Pezzuti et al., 2013) • Clinical condition (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006; Farias et al., 2003; Ready et al., 2001; Wood & Liossi, 2006; Ziemnik & Suchy, 2019; Zgaljardic et al., 2011) • Heterogeneity of the sample, namely in the results obtained in the assessed cognitive functions (Renison et al., 2012) • Demographic data (Bromley et al., 2012; Chaytor & Schmitter-Edgecombe 2003; Farias et al., 2003; Van der Elst et al., 2008)

(*continued*)

Table 1. Continued.

	Factors with potential to increase EV	Factors with potential to decrease the EV	Factors that may have the potential to increase or decrease the EV	Role of factors in the interpretation of study results
				• chronicity of the injury in the case of samples with Acquired Brain Injury (Zgaljardic et al., 2011) • Injury severity/clinical condition (Bowman, 1996; Chaytor & Schmitter-Edgecombe, 2003; Wood & Liossi, 2006) • Developmental factors in the case of samples with children and adolescents (Olson et al., 2013; Price et al., 2003)
Factors related to the patient		• Emotional distress (Azouvi et al., 2014; Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Farias et al., 2003; Farley et al., 2011; Kibby et al., 1998; Olson et al., 2013; Renison et al., 2012; Van der Elst et al., 2008; Wood & Liossi, 2006; Zgaljardic et al., 2011) • Physical disability/motor function (Bowman, 1996; Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Kibby et al., 1998; Renison et al., 2012) • Poor computer skills of the examinee observed in computerized tests (Phillips et al., 2006) • Health problems, including psychiatric (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006, 2007; Farley et al., 2011; Pezzuti et al., 2013; Renison et al., 2012) • Possible symptoms presented at the time of the evaluation (Zgaljardic et al., 2011) • Perceptual deficits of the evaluated (Aubin et al., 2015)		• Evaluator's awareness of deficits/perception of self-efficacy (Chaytor & Schmitter-Edgecombe, 2003; Maeir et al., 2011; Pezzuti et al., 2013; Wood & Liossi, 2006) • Evaluated personality (Chaytor & Schmitter-Edgecombe, 2003; Ready et al., 2001; Wilson, 1993; Zgaljardic et al., 2011) • Motivation (Olson et al., 2013; Pezzuti et al., 2013; Price et al., 2003; Wilson, 1993) • Fatigue (Pezzuti et al., 2013) • Routinization (Dubreuil et al., 2007) • Family support (Wilson, 1993)
Informant-related factors				• Frequency with which the mentee's deficits impact the informant (Renison et al., 2012) • Nature of informant's relationship with the evaluated, i.e., significant person, clinician, other (Chaytor & Schmitter-Edgecombe, 2003)
Other factors				• Response bias (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2007; Lippa et al., 2014)

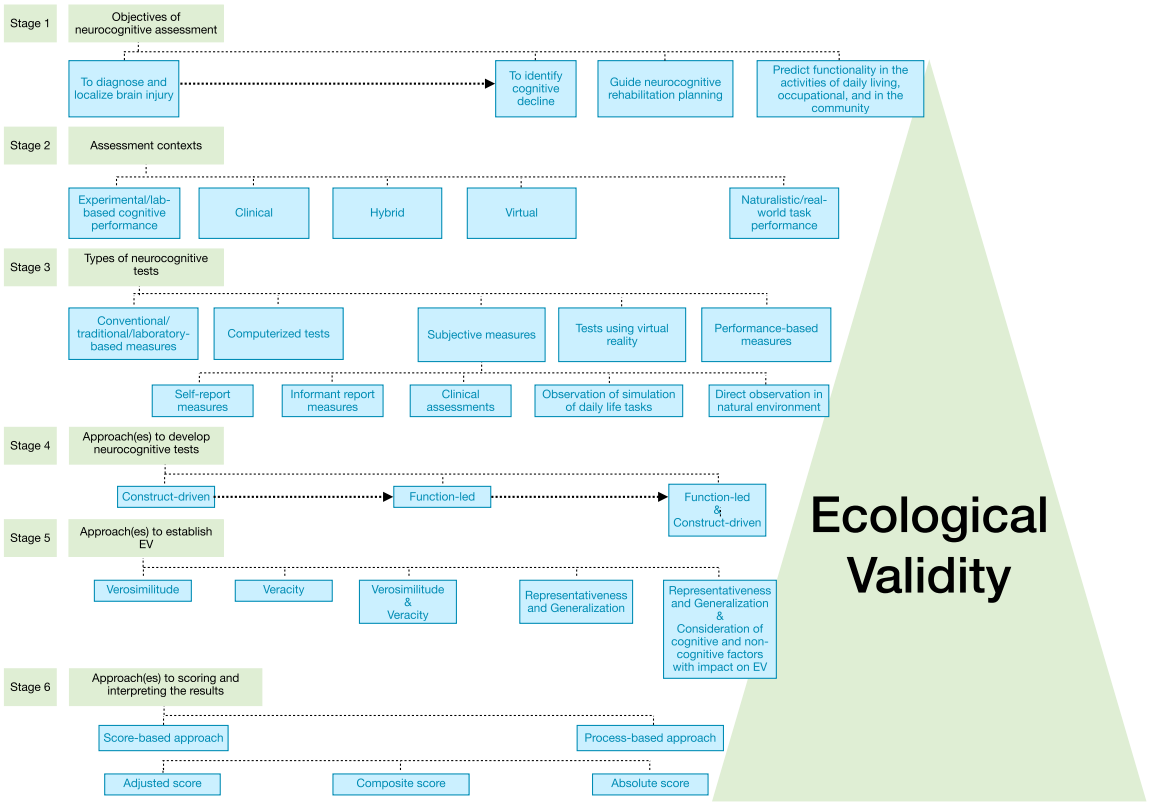

Figure 1. Operationalization stages of the EV.

considerations on this issue, which should be taken into account. These considerations cover relevant aspects to keep in mind when developing tests with higher EV and focus on: (a) objectives of neurocognitive assessment; (b) assessment contexts; (c) types and formats of neurocognitive tests; (d) approaches to develop neurocognitive tests; (e) approaches to establish the EV of neurocognitive measures; and (f) approaches to score and interpret the results of neurocognitive tests. These aspects are discussed in more detail below.

Objectives of neurocognitive assessment. Given that EV implies the generalization of results, it is essential to consider the objectives of the assessment to be performed (Hoc, 2001). According to the content analysis, the main objectives of neurocognitive assessment are to: (a) diagnose and localize brain injury (Adjorlolo, 2016; Chaytor & Schmitter-Edgecombe, 2003, 2007; Kibby et al., 1998; Lippa et al., 2014; Owen et al., 2004; Parsons, 2015; Spooner & Pachana, 2006; van der Elst et al., 2008; Weber et al., 2019; Wen et al., 2006; Yantz et al., 2010; Zgaljardic et al., 2011); (b) identify cognitive decline (Silverberg & Millis, 2009); (c) guide neuropsychological rehabilitation planning (Zgaljardic et al., 2011); and (d) predict functionality, i.e., understand human cognition and behaviour in the "real world", namely in activities of daily living, occupational activities, and active participation in the community (Faith & Rempfer, 2018; Gioia, 2009; Gioia & Brekke, 2009; Holleman et al., 2020; Kibby et al., 1998; Parsons, 2015; Silverberg & Millis, 2009; Tang et al., 2018; Weber et al., 2019). The last objective is the simplest for establishing and obtaining good EV.

Assessment contexts. The assessment contexts identified in the content analysis, ranked from least to most favourable to EV, are: (a) experimental/lab-based (Barkley, 1991; Burgess et al., 2006; Gioia, 2009; Gioia & Brekke, 2009; Helmstaedter et al., 1998; Holleman et al., 2020; Norris & Tate, 2000; Owen et al., 2004; Parsons, 2015; Phillips et al., 2006; Rumpf et al., 2019); (b) hybrid, integrating virtual elements into a natural environment (van der Ham et al., 2015); (c) clinical (Burgess et al., 2006; Owen et al., 2004); (d) virtual reality (Campbell et al., 2009; Holleman et al., 2020; van der Ham et al., 2015); and (e) naturalistic/real-world (Barkley, 1991; Bromley et al., 2012; Dawson et al., 2009; Faith & Rempfer, 2018; Gioia, 2009; Holleman et al., 2020; Norris & Tate, 2000; Parsons, 2015; Phillips et al., 2006; van der Ham et al., 2015).

The EV of a test is assumed to be higher when conducted in a natural environment/real-world setting, compared to experimental settings and natural-like settings (Gioia, 2009; Holleman et al., 2020; Norris & Tate, 2000; Parsons, 2015), but VR can be used to ensure standardization. Individuals' familiarity with assessment settings contributes to increased EV and appears to be high in virtual reality settings (Parsons, 2015). For example, a supermarket may be chosen as an assessment setting to ensure that such environment is similarly familiar or unfamiliar to all participants (Aubin et al., 2015).

Types and formats of neurocognitive tests. Instruments of different types and formats, ranging from least to most favourable to EV, emerged from the content analysis, depending on the objectives of the neurocognitive assessment: (a) conventional/traditional/laboratory-based paper-and-pencil tests (Azouvi et al., 2014; Barkley, 1991;

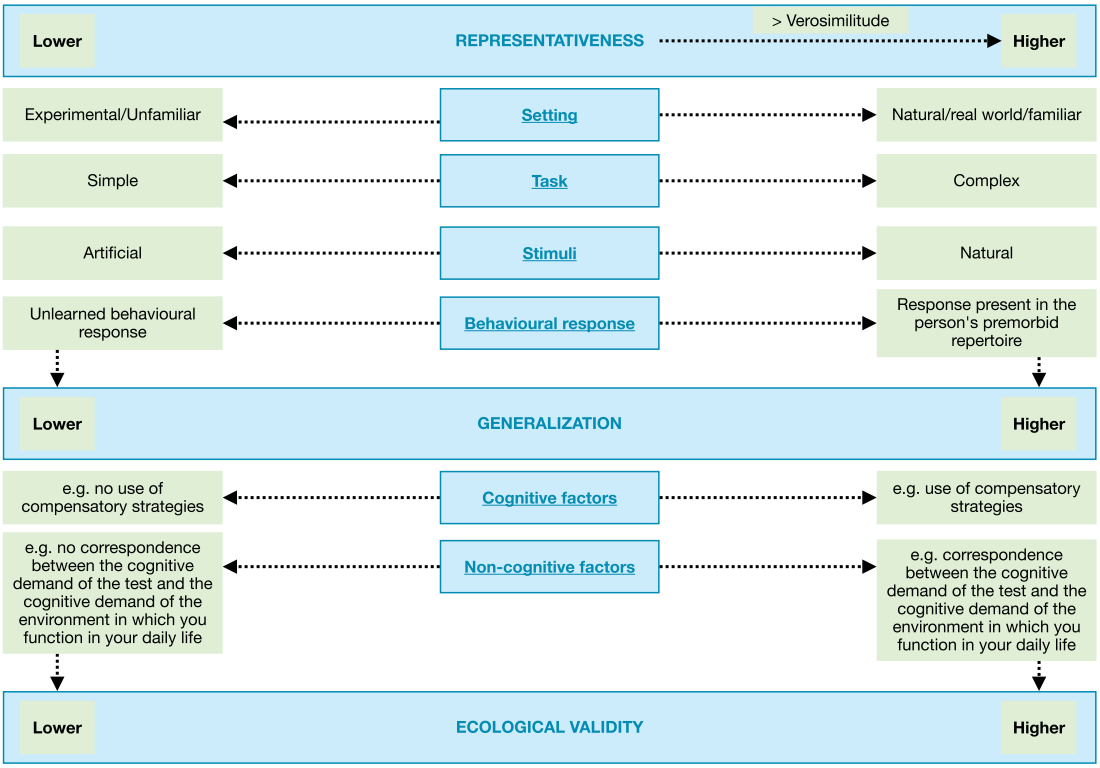

Figure 2. Approaches to establish EV.

Campbell et al., 2009; Chaytor et al., 2006; Cuberos-Urbano et al., 2013; Dawson et al., 2009; Drozdick & Cullum, 2011; Dubreuil et al., 2007; Farley et al., 2011; Gioia, 2009; Goodman & Zarit, 1995; Helmstaedter et al., 1998; Maeir et al., 2011; Parsons, 2015; Spooner & Pachana, 2006; Thornton et al., 2010; Wilson, 1993); (b) computerized tests (Kieffaber et al., 2007; Lippa et al., 2014; Phillips et al., 2006; Rumpf et al., 2019); (c) subjective tests - self-report and informant-report measures, clinical assessments, behaviour observation in simulations of daily life tasks, or direct observation in natural environments (Chaytor et al., 2007; Maeir et al., 2011; Possin et al., 2014); (d) tests using virtual reality (Aubin et al., 2015; Campbell et al., 2009; Horan et al., 2020; Parsons, 2015; Renison et al., 2012; van der Ham et al., 2015); and (e) performance-based naturalistic tasks (Dawson et al., 2009; Faith & Rempfer, 2018; Kenney et al., 2019; Lippa et al., 2014; Yantz et al., 2010).

The evaluation types also influence the EV and were further categorized by Chaytor and Schmitter-Edgecombe (2003) into (a) intentional and (b) incidental. Intentional evaluation involves deliberate effort to encode the stimuli and is considered a more efficient way to retain information; however, it is more attention and executive control demanding (Helmstaedter et al., 1998). Evaluation is said to be incidental when no deliberate effort is required to memorize the primary or content-related information (e.g., asking to copy a picture and unexpectedly asking for its drawing from memory), as well as the secondary or contextual information about the circumstances under which an event took place (i.e., location, date, people present, etc.) and the characteristics of the stimuli presented (i.e., sensory modality, spatial and temporal location, etc.) (Helmstaedter et al., 1998).

Approaches to develop neurocognitive tests. According to the content analysis, two main approaches can be followed for the development of neurocognitive tests: (a) construct-driven (Burgess et al., 2006; Parsons, 2015; Schmuckler, 2001; Wilson, 1993); and (b) function-led (Burgess et al., 2006; Parsons, 2015; Thornton et al., 2010; Wilson, 1993; Ziemnik & Suchy, 2019). While tests developed according to the construct-driven approach aim to assess constructs (e.g., working memory) without considering their ability to predict behavioral functionality (Parsons, 2015), tests developed according to the function-led approach aim to assess functionality (i.e., directly observable everyday behaviors) by analyzing the sequences of actions that trigger a given behavior in normal functioning, as well as possible changes in behavior. Although both approaches are useful for different purposes, function-led approach is considered more ecologically valid, as long as the tests are representative of real-world functions and allow for generalizable results as well as better predictions about functional performance in a variety of situations (Burgess et al., 2006; Parsons, 2015).

Approaches to establish the EV of neurocognitive measures. Verisimilitude and/or veracity (for definition see introduction) were the two approaches identified in the content analysis to establish the EV of neurocognitive measures (Aubin et al., 2015; Bromley et al., 2012; Chaytor & Schmitter-Edgecombe, 2003; Davies et al., 2011; Faith & Rempfer, 2018; Farley et al., 2011; Gioia & Brekke, 2009; Horan et al., 2020; Owen et al., 2004; Parsons, 2015; Poletti, 2010; Price et al., 2003; Renison et al., 2012; Spooner & Pachana, 2006; van der Elst et

al., 2008; Wallisch et al., 2018; Weber et al., 2019; Wood & Liossi, 2006; Yantz et al., 2010; Ziemnik & Suchy, 2019; Zgaljardic et al., 2011). The isolated use of the verisimilitude approach has been criticized for its lack of empirical rigor (Chaytor & Schmitter-Edgecombe, 2003). Regarding veracity, the main criticism is the difficulty in proving the EV of existing tests that were not developed for this purpose (Chaytor & Schmitter-Edgecombe, 2003).

Approaches to scoring and interpreting the results of neurocognitive tests. When it comes to scoring and interpreting test results, two approaches emerged from the content analysis: (a) process-based approach (Cuberos-Urbano et al., 2013; Davies et al., 2011; Dawson et al., 2009); and (b) score-based approach (Silverberg & Millis, 2009; van der Elst et al., 2008).

The process-based approach emphasizes the analysis of errors and strategies used during task performance (Alderman et al., 2003; Cuberos-Urbano et al., 2013; Davies et al., 2011; Owen et al., 2004), as it is assumed that the analysis of qualitative aspects underlying test performance is essential for an ecologically valid interpretation of the results (van der Elst et al., 2008).

Within the score-based approach, we can differentiate three procedures: (a) adjusted score interpretation (Silverberg & Millis, 2009); (b) absolute score interpretation (Silverberg & Millis, 2009); (c) composite score interpretation (Silverberg & Millis, 2009). The adjusted score contrasts the absolute results obtained in the test with the expected results according to the premorbid level (depending on age, education, and other variables that contribute to this prediction), and is assumed by Silverberg and Millis (2009) to be the best option when the objective of the assessment is to identify cognitive decline (Silverberg & Millis, 2009). However, when the objective is to determine whether the person's cognitive abilities are sufficient to cope with the demands of functional tasks (e.g., instrumental activities of daily living), the best option, according to the same authors, is to compare the results obtained with those of a healthy population (Silverberg & Millis, 2009), i.e., a normative sample. Composite results are characterized by a combination of scores obtained in different tests and have been advocated by van der Elst et al. (2008) as a more ecologically valid approach compared to the interpretation of results obtained in single neuropsychological tests (van der Elst et al., 2008). The option that allows for higher EV is to use a process-based approach and a scoring-based approach based on a composite analysis of absolute scores.

Proposal of a theoretical framework to analyze EV in the context of neuropsychological assessment

Based on the reviewed literature and its considerations about the EV of neurocognitive tests, we propose six stages for the operationalization of EV (see Figure 1). Importantly, these stages can be considered either in the development of neurocognitive tests, to improve EV, or in the analysis of the EV of existing tests.

Stage 1: Defining the objectives of neurocognitive assessment

It is difficult to make inferences about domains as diverse as activities of daily living, occupational activities, and community functioning from neurocognitive tests. The objectives of tests to predict functionality in these domains should be directly related to the tasks of the test. For example, if the aim of the assessment is to analyze individual responses to unfamiliar situations, the experimental setting may be appropriate; but if the aim is to understand how individuals behave at home, the experimental setting may not be the most appropriate (Holleman et al., 2020).

Stage 2: Defining the context of neurocognitive assessment

Tests in natural environments have been advocated as more ecologically valid (Bromley et al., 2012; Gioia, 2009), but the decision about the context depends on the assessment objectives.

Stage 3: Defining the type of test

The relevance of studying the EV of incidental evaluation has been suggested in addition to formal tests. This suggestion is based on the idea that in everyday life we use, for example, incidental memory more often than we undertake efforts to retain facts or events (Vingerhoets et al., 2005).

Stage 4: Selecting the underlying approach to test development

Starting with a function-led approach, followed by a construct-driven approach, to the development of the test seems to be the best method to enhance EV (Burgess et al., 2006).

Stage 5: Selecting the approach to establish EV

A combination of both verisimilitude and veracity approaches, favoring representativeness and generalization, seems to allow for a more accurate assessment of the EV of neurocognitive tests than using only one of these approaches, given that none of them is free from criticism (Holleman et al., 2020; Kibby, 1998; Silverberg & Millis, 2009; Tang et al., 2018). However, verisimilitude or veridicality are very simple and superficial approaches that did not provide specific criteria to establish EV, namely for setting and stimuli. Thus, our proposal is to consider the criteria of representativeness and the impact of cognitive and non-cognitive factors on generalization (see Figure 2).

Establishing representativeness entails taking into consideration: (a) the setting of the test—continuum between experimental and natural/real world; (b) the task or tasks—continuum between simplicity and complexity; (c) the stimuli—continuum between artificial and natural; (d) the assessed behavioral responses—continuum between

unlearned behavioral responses and responses present in the person's premorbid repertoire, as well as between the cognitive requirements of the test and the cognitive requirements of the person's daily living environment; the more one resembles the other, the better.

The EV of a test is assumed to be higher when conducted in a natural environment/real world setting, compared to experimental and natural-like settings (Gioia, 2009; Holleman, 2020; Norris & Tate, 2000; Parsons, 2015), but VR can be used to ensure standardization. Individuals' familiarity with assessment settings contributes to increased EV and appears to be high in virtual reality settings (Parsons, 2015). For example, a supermarket may be chosen as an assessment setting to ensure that such environment is similarly familiar or unfamiliar to all participants (Aubin et al., 2018). Although the use of naturalistic settings is a relevant aspect to increase EV, their selection should be guided by theoretical considerations about the goal of the assessment. For example, if the aim of the assessment is to analyze individual responses to unfamiliar situations, the experimental setting may be appropriate; but if the aim is to understand how individuals behave at home, the experimental setting may not be the most appropriate (Holleman et al., 2020).

Regarding the task, on the one hand, we may have simple tasks, which represent single cognitive processes, but do not capture the full complexity of human behavior and, therefore, have a lower EV (Holleman et al., 2020); on the other hand, multitasking, which has different underlying cognitive functions (Romero-Ayuso et al., 2019), making it difficult to evaluate them separately, has higher EV. According to this perspective, we propose to assess task-complexity according to: (a) the activities assessed, on a continuum between general instrumental activities of daily living, on one side, and cognitive instrumental activities of daily living, on the other; and (b) single task versus multitasking.

As far as stimuli are concerned, EV can be analyzed on an artificial-natural continuum. The concept of *artificial* refers us to stimuli specifically designed for research, while natural stimuli are those that are part of the real world (Hoc, 2001). The use of natural, multisensory stimuli is assumed to be more ecologically valid (De Gelder & Bertelson, 2003; Parsons, 2015). Three-dimensional (Lewkowicz, 2001; Rumpf et al., 2019), dynamic (Parsons, 2015), and contextual stimuli (Holleman et al., 2020; Lewkowicz, 2001; Parsons, 2015; Schmuckler, 2001) were found to have greater EV than two-dimensional, static, and non-contextual stimuli. Furthermore, since distractions are present in daily life, introducing distractors also increases EV, while reducing the artificiality of assessment situations (Davies et al., 2011; Farley et al., 2011; Gioia, 2009; Olson et al., 2013; Parsons, 2015).

For a test to allow sound inferences about individuals' functionality in their activities, the targeted behavioral responses should be part of individuals' repertoire (e.g., in a case of a person who never went to a supermarket, a shopping task has low EV, because it is not informative about changes in his or her cognitive performance and has poor generalization). Therefore, a test will have greater EV if the behavioral responses to the tasks used can be generalized to cognitive instrumental activities of daily living, and the cognitive requirements of the tasks should resemble the requirements found in the person's daily living environment (Franzen & Wilhelm, 1996).

Stage 6: Defining scoring approaches and developing an interpretation rationale

The process-based approach has the potential to ensure higher EV, because it allows for the perception that other implicit functions may be influencing task performance. For example, difficulties in reading, comprehension, calculation, or naming may impact performance in memory or executive functioning tests. The process-based approach also considers the familiarity with the task and its complexity, recognizing the role of compensatory strategies that have been reported in different studies as a factor impacting EV (e.g., Farley et al., 2011). Furthermore, using a scoring-based approach based on a composite analysis of absolute scores seems to increase EV.

Summing-up, based on the previous stages, it is assumed that a test has higher EV as more characteristics of the right side of Figures 1 and 2 are reached.

Proposal of a checklist to evaluate the ecological validity of neurocognitive tests

Many criteria and checklists have been developed over the years to analyze the quality of psychological tests (Cizek et al., 2016; Evers, 2001; Evers et al., 2010, 2013; Lindley et al., 2008). Indices for analyzing construct and criterion validity are commonly described in these checklists, but none included indices about EV. Thus, below we propose a checklist that can be used to guide the development of neurocognitive tests with higher EV and/or to assess the EV of existing tests.

The general structure of the checklist was inspired by the checklist for evaluating credentialing testing programmes (Cizek et al., 2016), the Revised Dutch Rating System for Test Quality (Evers, 2001; Evers et al., 2010), the European Federation of Psychologists' Associations review model for the description and evaluation of psychological and educational tests (EFPA; Evers et al., 2013; Lindley et al., 2008), the Standards for educational and psychological testing of the American Educational Research Association (2014), the Guidelines for Test Adaptation (Hambleton, 2005), and the Guidelines on Computer-Based and Internet-Delivered Testing of the International Test Commission (2006).

The checklist has six parts: (a) test development; (b) test administration; (c) evidence of ecological validity; (d) scoring and reporting; (e) specific criteria for technological-based tests; and (f) additional information. The criteria are intentionally simple, allowing for the evaluation of EV using the standard rating system of the EFPA (see checklist in the Supplemental Material). This system comprises the following classifications for test attributes: "n/a" is used when the attribute does not apply to the instrument; "–2" is used

when no rating is possible or not enough information about the attribute is provided; "–1" is used for inadequate attributes; "0" for adequate or reasonable attributes; "1" for good attributes; and "2" for excellent attributes (Bartram, 2011).

Part 1. Test development

Theoretical basis of the test

Among the neurocognitive assessment goals, the one that is most closely related to EV is the prediction of functionality in instrumental cognitive activities in daily life, occupational and/or community activities, ensuring better generalization [see criterion 1.1.1 of the checklist in appendix (Supplemental Material)] (American Educational Research Association, 2014; Cizek et al., 2016; Holleman et al., 2020; Kibby et al., 1998; Silverberg & Millis, 2009; Tang et al., 2018).

Regarding the approach to be followed for developing tests with good generalization and also representativeness potential, the best options are to define the directly observable everyday behaviors to be assessed (function-led) and then identify the constructs to be assessed (construct-driven) [criterion 1.1.2] (Burgess et al., 2006).

It is critical to clearly identify the instrumental cognitive behaviors/activities of daily living to be predicted [criterion 1.1.2.1]. A good strategy for ensuring criterion 1.1.2.1 compliance may be to focus on the different instrumental activities of daily living proposed by the American Association of Occupational Therapy (AOTA, 2014), in adult assessment, and on the activities listed in the International Classification of Functioning, Disability and Health (World Health Organization, 2013), in child assessment. Furthermore, it is important to ensure that the assessed construct is implied in the predicted instrumental cognitive behaviors/activities of daily living [criterion 1.1.2.2], as defined in 1.1.2.3 of the checklist (Supplemental Material).

With respect to the construct-driven approach, the cognitive function(s) targeted for assessment should have practical implications for functionality in the instrumental cognitive activities of daily living to be predicted (Parsons, 2011), and should be described in a detailed and bounded manner [criterion 1.1.2.3] (American Educational Research Association, 2014). For example, in the case of a test designed to assess executive functioning, the executive function(s) to be assessed (e.g., inhibitory control) must be specified. According to Chaytor and Schmitter-Edgecombe (2003), the lack of agreement among researchers on the constructs to be assessed with the different tests compromises their EV. In this regard, sources of variation in performance should be identified, particularly cognitive functions that underlie task performance but are not the focus of the assessment [criterion 1.1.2.4]. Criterion 1.1.2.4 is justified by three reasons: (a) interdependence between cognitive functions (Luria, 1976); (b) impossibility of considering cognitive functions in isolation when it is intended to make predictions about functional cognition (Chaytor & Schmitter-Edgecombe, 2003); (c) as most instrumental cognitive activities of daily living are complex (Romero-Ayuso et al., 2019), a test that intends to predict functional cognition may have various neurocognitive functions involved. On this subject, all tests should have a theoretical framework about neurocognitive functioning that facilitates the identification of the relationship between the involved functions (e.g., Luria's neurocognitive model, 1976). Identifying the neurocognitive functions underlying the tasks may also be useful for inferring about the test difficulty, as more complex tasks are likely to involve more neurocognitive functions (Romero-Ayuso et al., 2019).

Setting

The test should be applied in a naturalistic environment or at least a natural-like environment, for example, simulated through virtual reality [criterion 1.2.1], and it is important to report familiarity conditions [criterion 1.2.2].

Tasks

Tasks should correspond to relevant aspects of real-world activities and environments, i.e., focus on practical problems related to daily functionality, such as the ability to return to work or driving [criterion 1.3.1] (Parsons, 2011). Given that most of our everyday activities involve multitasking (Romero-Ayuso et al., 2019), the test should assess multitasking performance or, if it is a battery of tests, at least one of them should involve multitasking [criterion 1.3.2]. The advantages and limitations of the task should be properly described (Evers, 2001; Evers et al., 2010). With regards to the simplicity-complexity dimension, the test should include tasks with various levels of difficulty [criterion 1.3.3] to ensure correspondence between the cognitive demand of the test and the cognitive demand of the person's naturalistic environment (Aubin et al., 2018; Chaytor et al., 2006, 2007; Chaytor & Schmitter-Edgecombe, 2003).

Ecologically valid tasks must use language that is appropriate to the cultural background of the target population [criterion 1.3.4] (Hambleton, 2005; Parsons, 2011). Moreover, as EV is about the generalization of the results to naturalistic environments, it requires an adequate selection of tasks considering the individual characteristics of patients and their routines, so that appropriate inferences can be made. For example, asking about the *average time to clean the windows of an average-sized house* may be more difficult for certain patients, because cleaning windows was never part of their routines. Thus, it would be useful to have a broad set of questions, and then select those that are best suited to the instrumental activities of each patient. We recommend making the prerequisites for performing each task [criterion 1.3.5] explicit, to ensure that they fit the patient's premorbid functions and cognitive level, allowing to optimize generalization. Motor prerequisites also need to be mentioned (Chaytor et al., 2006, 2007; Kibby et al., 1998; Renison et al., 2012).

Stimuli

The use of multisensory (De Gelder, & Bertelson, 2003; Parsons, 2015), three-dimensional (Lewkowicz, 2001; Rumpf et al., 2019), and dynamic (Parsons, 2015) stimuli is assumed to improve ecological validity [criteria 1.4.1; 1.4.2; 1.4.3]. In addition, naturalistic stimuli are associated with higher EV (Holleman et al., 2020; Lewkowicz, 2001; Parsons, 2015) and, consequently, cultural knowledge of the target population should be considered when selecting stimuli, ensuring its familiarity [criterion, 1.4.4] (Parsons, 2011). The introduction of distracting stimuli is also a key factor in reducing artificiality [criterion 1.4.5] (Davies et al., 2011; Gioia, 2009; Olson et al., 2013; Parsons, 2015).

Sample

The sampling strategy should be reported [criterion 1.5.1], allowing the representativeness of the target population to be gauged. Failure to consider this criterion can compromise generalization (Parsons, 2011). Thus, limitations of the sampling procedure and the sample itself should be reported, particularly the small sample size (Pezzuti et al., 2013) and the heterogeneity of the sample [criterion 1.5.2] (Renison et al., 2012), as should be other characteristics of the sample [criterion 1.5.3], namely demographic data (Bromley et al., 2012; Chaytor & Schmitter-Edgecombe, 2003; Farias et al., 2003; van der Elst et al., 2008), clinical condition and its severity (Chaytor & Schmitter-Edgecombe, 2003; Wood & Liossi, 2006), developmental factors in the case of children and adolescents (Olson et al., 2013; Price et al., 2003), the time between injury and assessment (Chaytor & Schmitter-Edgecombe, 2003), and the chronicity of the injury (Zgaljardic et al., 2011) in the case of acquired brain injury.

Part 2. Test administration

Administration procedures

The development of an administration manual is critical to ensure that the procedures leading to EV are met [criterion 2.1.1] (Chevignard et al., 2012; Cizek et al., 2016). Relevant aspects for the subsequent interpretation of the results should also be clarified: (a) cognitive functions that are critical for the individual's everyday tasks [criterion 2.1.2] (Chaytor & Schmitter-Edgecombe, 2003); (b) degree of familiarity with the tasks [criterion 2.1.3] (Romero-Ayuso et al., 2019); (c) frequency of performing the tasks in a naturalistic environment (i.e., once a week, once a month, etc.) [criterion 2.1.4]; (e) autonomy in performing the tasks in a naturalistic environment (i.e., performs alone, with caregiver support, etc.) [criterion 2.1.5] (Aubin et al., 2018); (c) routinization, i.e., repetition of behaviors and routines [criterion 2.1.6] (Bouisson, 2002; Dubreuil et al., 2007).

Assessment conditions

The characteristics of the testing environment should be described (Cizek et al., 2016), namely its potential to increase EV and its limitations [criterion 2.2.1]. Since EV can be affected by the administration conditions (Chaytor & Schmitter-Edgecombe, 2003), adaptations of the administration procedures [criterion 2.2.2] should be identified (Cizek et al., 2016).

Previous studies have pointed out the disparity between the time available for the task in the test situation and in the real world as a factor with potential impact on reducing EV (Aubin et al., 2018; Wilson, 1993). Thus, the time allotted for the task should approximate the time needed to perform it in everyday life [criterion 2.2.3].

Use of compensatory strategies

The use of compensatory strategies is a cognitive factor that impacts EV (Alderman et al., 2003; Cuberos-Urbano et al., 2013; Davies et al., 2011; Olson et al., 2013). Therefore, it is important to identify and record compensatory strategies that patients eventually use during the assessment [criterion 2.3.1]. This can be achieved either by direct observation of the patient or through open questions such as "Did you use any strategy to perform the task?"; if so "Which one?" If compensatory strategies are used, it is important to ask whether they are used in everyday life [criterion 2.3.2].

Response bias

The American Academy of Clinical Neuropsychology recommends that all neuropsychological assessments should include a measure of response bias (Heilbronner et al., 2009). Response bias refers to the misrepresentation of abilities, namely cognitive, on a test or a self-report measure. Specifically, a negative response bias may occur in the presence of an external gain, impacting the EV of neurocognitive tests (Lippa et al., 2014) and self-reports (Chaytor & Schmitter-Edgecombe, 2003). As response bias may be a reason for considerable disparity between test performance and functional ability in real-world activities (Heilbronner et al., 2009), it should be assessed [criterion 2.4.1].

Part 3. Ecological validity evidence

Verisimilitude and veridicality

Since tests conforming to both verisimilitude and veridicality have been found to be more ecologically valid (Kourtesis et al., 2021), as higher verisimilitude underlies higher representativeness and higher veridicality underlies higher generalization, both approaches must be assessed [criteria 3.1.1 and 3.1.2].

The statistical method used to estimate EV (e.g., correlation with other tests with proven EV, correlation with the same task in a naturalistic environment, regression models to predict patients' functionality in activities of daily living) should be described [criterion 3.1.3].

Since there is no gold-standard measure of everyday functioning that can be used for comparison, as explained in the introduction (Chevignard et al., 2012), and given the factors that can compromise both self-report (e.g., the examinee's awareness of the deficits presented) (Pezzuti et al., 2013; Wood & Liossi, 2006) and other-report measures (e.g., the frequency with which the examinee's deficits impact the

informant, including caregiver burden) (Renison et al., 2012), the use of more than one outcome measure allows obtaining more robust evidence on the EV of the test [criterion 3.1.4]. Measures of the same nature (e.g., self-and other-report measures) should be considered as a single-type of measure. It is worth noting that observation of cognitive functioning in the naturalistic environment, together with self-and other-report measures of cognitive functioning, are the most ecologically sound options (Domensino & van Heugten, 2020).

Part 4. Scoring and reporting

Response sheet

Standardization of scoring procedures is another element that confers clinical reliability to the tests (Dawson et al., 2009) and, therefore, a standardized response sheet and clear scoring procedures should be provided [criterion 4.1.1].

Scoring

Comparing patients' absolute score obtained in the test with the results obtained by healthy individuals may not be the best approach for diagnosis, but it is assumed to be the most ecologically valid option [criterion 4.2.1] (Silverberg & Millis, 2009).

Analysis of response processes

The scoring of tests should be accompanied by the observation and recording of response processes [criterion 4.3.1], as well as their analysis and consideration in the interpretation of results. Thus, the main types of error and the main compensatory strategies observed during the task should be presented in the response sheet, guiding the interpretation of the results [criterion 4.3.2] (Dawson et al., 2009).

Interpretation of results

Regarding the interpretation of results, a theoretical *rationale* should be provided [criterion 4.4.1] (American Educational Research Association, 2014; Cizek et al., 2016). This *rationale* should consider the context in which the results will be used (American Educational Research Association, 2014), e.g., for the characterization of a neurocognitive profile or the analysis of progresses achieved with neuropsychological rehabilitation. In addition, individual factors with potential impact on EV, such as emotional distress (Azouvi et al., 2015; Chaytor et al., 2006, 2007), should be considered. The information necessary for proper interpretation of the test results should be reported [criterion 4.4.2] (Evers, 2001; Evers et al., 2010).

Part 5. Specific criteria for technology-based assessments

In assessing the EV of technology-based tests, specific criteria must be considered in addition to the previous criteria.

Familiarization

Given that patients have different degrees of familiarization with new technologies, familiarization tasks or trials should be given to the patients to get acquainted with the system before computer-based testing or assessment in virtual reality environments (Aubin et al., 2018). Otherwise, poor results can be explained by difficulties in using technological resources [criterion 5.1.1].

Computer-based tests

Regarding computer-based tests, the International Guidelines on Computer-Based and Internet-Delivered Testing (International Test Commission, 2006) should be given due consideration [criterion 5.2.1]. Since the representativeness of the setting, the task, the stimuli, and the evaluated behavioral response are criteria of EV, computerized tasks only make sense when the behavior to be predicted is related to an instrumental cognitive activity of everyday life mediated by computers [criterion 5.2.2]. Furthermore, since difficulties in using computers constitute one of the main criticisms of computer-mediated neurocognitive assessment (Rumpf et al., 2019), the level of computer skills required for the task should be clarified and limitations should be presented [criterion 5.2.3]. Familiarization with the test environment (i.e., how to access the instructions after starting the test, how items are presented, how to respond) should be ensured before starting the assessment (International Test Commission, 2006) [criterion 5.2.4]. The method of test administration should be specified using the categories defined by the International Test Commission: (a) open mode—there is no supervision during the assessment and it does not require registration; (b) controlled mode—administration is remote but requires registration; (c) supervised (proctored) mode—the assessment is done with direct human supervision over the test conditions; (d) managed mode—higher level of human supervision and control over the test environment (International Test Commission, 2006) [criterion 5.2.5]. The potential impact of the method of test administration on EV should be discussed in the interpretation of the results [criterion 5.2.6]. Regarding the method of response, possible accommodations for people with disabilities should be mentioned [criterion 5.2.7] (International Test Commission, 2006).

Virtual reality-based tests

As far as VR-based testing is concerned, the VR-Check Framework criteria suggested by Krohn et al. (2020) is recommended [criterion 5.3].

Part 6. Additional information

Clinical criteria for test selection

In test selection, it is important to ensure that the behavioral response assessed with the test was part of the person's premorbid repertoire [criterion 6.1.1].

Usefulness

Neurocognitive assessment can have two main purposes: (a) support diagnosis; (b) characterize the neurocognitive profile for the purpose of neuropsychological rehabilitation (Kipps & Hodges, 2005). Thus, it is important to understand whether a test with good EV also has diagnostic utility [criterion 6.2.1] and/or utility for planning and assessing the results of neuropsychological rehabilitation programmes [criterion 6.2.2]. As technology-based testing increases, it is also important to assess whether the tests are suitable for clinical use [criterion 6.2.3] (Chevignard, 2012).

Other

Cognitive and non-cognitive factors with potential impact on the results of neurocognitive tests are identified [criterion 6.3.1].

The test name should include a reference to the EV for easy identification of the instruments that were developed with ecological validity in mind [criterion 6.3.2] (Chevignard et al., 2012).

Conclusion

According to the content analysis we have performed, EV in the field of neuropsychological assessment concerns the attributes of neurocognitive tests that make them *representative* of a given cognitive instrumental activity of daily living, in a certain real-world context, allowing the test results to be *generalized* to related activities. Therefore, the main dimensions of EV emerging from this review are representativeness and generalization, with the former depending on the characteristics of setting, task, stimuli, and assessed behavioral response. In turn, generalization is potentially influenced by several cognitive and non-cognitive factors. Regarding EV assessment and operationalization, subjective measures were the most reported outcome measures and none of the reviewed studies provide general guidelines or specify procedures to develop a neurocognitive test with good EV. We propose an EV-checklist based on the six stages for operationalizing EV. The EV-checklist can be used as a quantitative scale, enabling the identification of different levels of EV.

Summing-up, based on this review and content analysis, a test with good EV is an instrument with the following characteristics: (a) requires a familiar setting; (b) involves a complex task and natural stimuli; (c) assesses a behavioral response present in the person's premorbid repertoire; and (d) considers cognitive and non-cognitive factors with potential impact in the generalization of the results.

Given the results of the content analysis, we propose the following assumptions to guide the EV analysis of neurocognitive tests: (a) neurocognitive tests are not universally ecologically valid (Chaytor et al., 2006, 2007; Chaytor & Schmitter-Edgecombe, 2003); (b) cognitive instrumental activities of daily living underlying the development of neurocognitive tests must be adjusted to the developmental stage in which the application is intended (Olson et al., 2013; Price et al., 2003); (c) neurocognitive tests with good EV should allow for some degree of customization of the level of difficulty; and (d) tests with good EV do not replace traditional tests in predicting cognitive decline and should, therefore, be used in a complementary manner (e. g., incidental memory assessment [Helmstaedter et al., 1998]).

Future studies should target the following issues: (a) development of tests with good EV that are brief to apply and that provide information on different cognitive functions, through the inclusion of multitasking measures and the use of a process-based approach for the interpretation of test results, for example, framed in Luria's neurocognitive model that explains the interdependence between cognitive functions; (b) development of batteries with good EV that include a variety of tasks, allowing not only the correspondence between tasks and relevant aspects of performance in the naturalistic environment, but also different levels of cognitive demand; and (c) analyze response bias and practice effects (i.e., changes in performance due to increased familiarity with the test and its items; Goldberg et al., 2015) on tests with good EV, since one of the main dimensions of EV is representativeness, which presumes familiarity with the setting.

Finally, it is important to highlight the specific limitations underlying the use of measures with good EV, for example: (a) increased difficulty in obtaining the instruments; (b) decreased standardization (Diaz-Orueta et al., 2020); (c) increased difficulties in developing tests with higher EV (e.g., based on virtual reality); and (d) feasibility of introducing more ecologically valid measures into a traditional neuropsychological setting.

The main limitation of this work is the high likelihood of bias associated with systematized reviews for not strictly following the recommendations of the Preferred Systematic Review and Meta-analysis (PRISMA; Grant & Booth, 2009; Moher et al., 2009; Page et al., 2021). Furthermore, the restriction of search strategy to titles may have biased the papers included in this content analysis.

Summing-up, this study sheds light on the concept of EV, exploring its conceptualization, operationalization, and estimation. Based on the results, we proposed an EV-checklist and expect it to be useful in helping researchers and clinicians to develop neurocognitive test with good EV or to select tests with higher EV.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Joana Filipa Freire Teixeira de Oliveira Pinto, Bruno Miguel Raposo Távora de Barros Peixoto, Artemisa Agostinha da Rocha Dores, and Fernando Manuel dos Santos Barbosa. The first draft of the manuscript was written by Joana Filipa Freire Teixeira de Oliveira Pinto and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Disclosure statement

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript.

Funding

Artemisa Dores is a researcher of School of Health, Polytechnic Institute of Porto, Porto, Portugal, supported by FCT—Fundação para a Ciência e Tecnologia [Portuguese Foundation for Science and Technology] through R&D Units funding [UIDB/05210/2020].

ORCID

Joana O. Pinto http://orcid.org/0000-0002-0643-8439
Bruno Peixoto http://orcid.org/0000-0002-2427-6330

References

Adjorlolo, S. (2016). Ecological validity of executive function tests in moderate traumatic brain injury in Ghana. *The Clinical Neuropsychologist*, *30*(sup1), 1517–1537. https://doi.org/10.1080/13854046.2016.1172667

Alderman, N., Burgess, P. W., Knight, C., & Henman, C. (2003). Ecological validity of a simplified version of the multiple errands shopping test. *Journal of the International Neuropsychological Society*, *9*(1), 31–44. https://doi.org/10.1017/S1355617703910046

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Occupational Therapy Association (AOTA). (2014). Occupational therapy practice. Framework: Domain & Process (3rd ed.). *American Journal of Occupational Therapy*, *68*(S1), S1–S48.

Aubin, G., Béliveau, M. F., & Klinger, E. (2015). An exploration of the ecological validity of the Virtual Action Planning–Supermarket (VAP-S) with people with schizophrenia. *Neuropsychological Rehabilitation*, *28*(5), 689–708. https://doi.org/10.1080/09602011.2015.1074083

Azouvi, P., Vallat-Azouvi, C., Millox, V., Darnoux, E., Ghout, I., Azerad, S., Ruet, A., Bayen, E., Pradat-Diehl, P., Aegerter, P., Weiss, J., & Jourdan, C. (2014). Ecological validity of the dysexecutive questionnaire: Results from the PariS-TBI study. *Neuropsychological rehabilitation*, *25*(6), 864–878. https://doi.org/10.1080/09602011.2014.990907

Barkley, R. A. (1991). The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. *Journal of Abnormal Child Psychology*, *19*(2), 149–178. https://doi.org/10.1007/BF00909976

Bartram, D. (2011). Contributions of the EFPA standing Committee on Tests and Testing to standards and good practice. *European Psychologist*, *16*(2), 149–159. https://doi.org/10.1027/1016-9040/a000093

Benge, J. F., Artz, J. D., & Kiselica, A. M. (2020). The ecological validity of the Uniform Data Set 3.0 neuropsychological battery in individuals with mild cognitive impairment and dementia. *The Clinical Neuropsychologist*, *36*(6), 1453–1470. https://doi.org/10.1080/13854046.2020.1837246

Bouisson, J. (2002). Routinization preferences, anxiety, and depression in an elderly French sample. *Journal of Aging Studies*, *16*(3), 295–302. https://doi.org/10.1016/S0890-4065(02)00051-8

Bowman, M. L. (1996). Ecological validity of neuropsychological and other predictors following head injury. *The Clinical Neuropsychologist*, *10*(4), 382–396. https://doi.org/10.1080/13854049608406699

Bromley, E., Mikesell, L., Mates, A., Smith, M., & Brekke, J. S. (2012). A video ethnography approach to assessing the ecological validity of neurocognitive and functional measures in severe mental illness: Results from a feasibility study. *Schizophrenia bulletin*, *38*(5), 981–991. https://doi.org/10.1093/schbul/sbr002

Brunswik, E. (1955). Symposium of the probability approach in psychology: Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193–217.

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society*, *4*(6), 547–558. https://doi.org/10.1017/S1355617798466037

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. A., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, *12*(2), 194–209. https://doi.org/10.1017/S1355617706060310

Campbell, Z., Zakzanis, K. K., Jovanovski, D., Joordens, S., Mraz, R., & Graham, S. J. (2009). Utilizing virtual reality to improve the ecological validity of clinical neuropsychology: An FMRI case study elucidating the neural basis of planning by comparing the Tower of London with a three-dimensional navigation task. *Applied Neuropsychology*, *16*(4), 295–306. https://doi.org/10.1080/09084280903297891

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, *13*(4), 181–197. https://doi.org/10.1023/B:NERV.0000009483.91468.fb

Chaytor, N., Schmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology*, *21*(3), 217–227. https://doi.org/10.1016/j.acn.2005.12.002

Chaytor, N., Temkin, N., Machamer, J., & Dikmen, S. (2007). The ecological validity of neuropsychological assessment and the role of depressive symptoms in moderate to severe traumatic brain injury. *Journal of the International Neuropsychological Society*, *13*(3), 377–385. https://doi.org/10.1017/S1355617707070592

Chaytor, N. S., Barbosa-Leiker, C., Germine, L. T., Fonseca, L. M., McPherson, S. M., & Tuttle, K. R. (2020). Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. *The Clinical Neuropsychologist*, *35*(1), 148–164. https://doi.org/10.1080/13854046.2020.1811893

Chen, D., Chen, J., Yang, H., Liang, X., Xie, Y., Li, S., Ding, L., & Li, Q. (2019). Mini-cog to predict postoperative mortality in geriatric elective surgical patients under general anesthesia: A prospective cohort study. *Minerva Anestesiologica*, *85*(11), 1193–1200. https://doi.org/10.23736/S0375-9393.19.13462-1

Chevignard, M. P., Soo, C., Galvin, J., Catroppa, C., & Eren, S. (2012). Ecological assessment of cognitive functions in children with acquired brain injury: A systematic review. *Brain injury*, *26*(9), 1033–1057. https://doi.org/10.3109/02699052.2012.666366

Cizek, G. J., Germuth, A. A., Kosh, A. E., & Schmid, L. A. (2016). *A checklist for evaluating credentialing testing programs.* The Evaluation Center, Western Michigan University. Retrieved from https://wmich.edu/sites/default/files/attachments/u350/2018/credentialing-cizek-etal.pdf

Cuberos-Urbano, G., Caracuel, A., Vilar-López, R., Valls-Serrano, C., Bateman, A., & Verdejo-García, A. (2013). Ecological validity of the Multiple Errands Test using predictive models of dysexecutive problems in everyday life. *Journal of Clinical and Experimental Neuropsychology*, *35*(3), 329–336. https://doi.org/10.1080/13803395.2013.776011

Davies, S. R., Field, A. R., Andersen, T., & Pestell, C. (2011). The ecological validity of the Rey–Osterrieth Complex Figure: Predicting everyday problems in children with neuropsychological disorders. *Journal of Clinical and Experimental Neuropsychology*, *33*(7), 820–831. https://doi.org/10.1080/13803395.2011.574608

Dawson, D. R., Anderson, N. D., Burgess, P., Cooper, E., Krpan, K. M., & Stuss, D. T. (2009). Further development of the Multiple Errands Test: Standardized scoring, reliability, and ecological validity for the Baycrest version. *Archives of Physical Medicine and Rehabilitation*, *90*(11 Suppl), S41–S51. 10.1016/j.apmr.2009.07.012

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*, *7*(10), 460–467. https://doi.org/10.1016/j.tics.2003.08.014

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988. https://doi.org/10.1037/0033-2909.130.6.959

Diaz-Orueta, U., Blanco-Campal, A., Lamar, M., Libon, D. J., & Burke, T. (2020). Marrying past and present neuropsychology: Is the future of the process-based approach technology-based? *Frontiers in Psychology*, *11*, 361. https://doi.org/10.3389/fpsyg.2020.00361

Domensino, A., & van Heugten, C. (2020). Van de 15-woordentest naar 'Heb ik nu wel alle boodschappen gedaan?': Het meten van cognitief functioneren op het continuüm van de kunstmatige testsituatie tot het dagelijks leven. *Tijdschrift voor Neuropsychologie*, *15*(1), 37–49.

Drozdick, L. W., & Cullum, C. M. (2011). Expanding the ecological validity of WAIS-IV and WMS-IV with the Texas Functional Living Scale. *Assessment*, *18*(2), 141–155. https://doi.org/10.1177/1073191110382843

Dubreuil, P., Adam, S., Bier, N., & Gagnon, L. (2007). The ecological validity of traditional memory evaluation in relation with controlled memory processes and routinization. *Archives of Clinical Neuropsychology*, *22*(8), 979–989. https://doi.org/10.1016/j.acn.2007.08.002

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing*, *1*(2), 155–182. https://doi.org/10.1207/S15327574IJT0102_4

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing*, *10*(4), 295–317. https://doi.org/10.1080/15305058.2010.518325

Evers, A., Hagemeister, C., Høstmaelingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). EFPA review model for the description and evaluation of psychological and educational tests. *Test review Form and Notes for Reviewers*, *Version 4.2.6.*

Faith, L. A., & Rempfer, M. V. (2018). Comparison of performance-based assessment and real world skill in people with serious mental illness: Ecological validity of the Test of Grocery Shopping Skills. *Psychiatry Research*, *266*, 11–17. https://doi.org/10.1016/j.psychres.2018.04.060

Farias, S. T., Harrell, E., Neumann, C., & Houtz, A. (2003). The relationship between neuropsychological performance and daily functioning in individuals with Alzheimer's disease: Ecological validity of neuropsychological tests. *Archives of Clinical Neuropsychology*, *18*(6), 655–672. https://doi.org/10.1093/arclin/18.6.655

Farley, K. L., Higginson, C. I., Sherman, M. F., & MacDougall, E. (2011). The ecological validity of clinical tests of visuospatial function in community-dwelling older adults. *Archives of Clinical Neuropsychology*, *26*(8), 728–738. https://doi.org/10.1093/arclin/acr069

Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsychology. In R. J. Sbordone, & C. L. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 91–112). GR Press/St. Lucie Press.

Gioia, D. (2009). Understanding the ecological validity of neuropsychological testing using an ethnographic approach. *Qualitative health Research*, *19*(10), 1495–1503. https://doi.org/10.1177/1049732309348500

Gioia, D., & Brekke, J. S. (2009). Neurocognition, ecological validity, and daily living in the community for individuals with schizophrenia: A mixed methods study. *Psychiatry*, *72*(1), 94–107. https://doi.org/10.1521/psyc.2009.72.1.94

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia*, *1*(1), 103–111. https://doi.org/10.1016/j.dadm.2014.11.003

Goodman, C. R., & Zarit, S. H. (1995). Ecological measures of cognitive functioning: A validation study. *International psychogeriatrics*, *7*(1), 39–50. https://doi.org/10.1017/S1041610295001839

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health information and Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Groth-Marnat, G., & Teal, M. (2000). Block design as a measure of everyday spatial ability: A study of ecological validity. *Perceptual and Motor Skills*, *90*(2), 522–526. https://doi.org/10.2466/pms.2000.90.2.522

Groth-Marnat, G., & Baker, S. (2003). Digit span as a measure of everyday attention: A study of ecological validity. *Perceptual and Motor Skills*, *97*(3 Pt 2), 1209–1218. https://doi.org/10.2466/pms.2003.97.3f.1209

Hambleton, R. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Lawrence Erlbaum Associates.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *23*(7), 1093–1129. https://doi.org/10.1080/13854040903155063

Helmstaedter, C., Hauff, M., & Elger, C. E. (1998). Ecological validity of list-learning tests and self-reported memory in healthy individuals and those with temporal lobe epilepsy. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 365–375. https://doi.org/10.1076/jcen.20.3.365.824

Higginson, C. I., Arnett, P. A., & Voss, W. D. (2000). The ecological validity of clinical tests of memory and attention in multiple sclerosis. *Archives of Clinical Neuropsychology*, *15*(3), 185–204. https://doi.org/10.1093/arclin/15.3.185

Hoc, J. M. (2001). Towards ecological validity of research in cognitive ergonomics. *Theoretical Issues in Ergonomics Science*, *2*(3), 278–288. https://doi.org/10.1080/14639220110104970

Holleman, G. A., Hooge, I. T., Kemner, C., & Hessels, R. S. (2020). The 'real-world approach' and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, *11*, 721. https://doi.org/10.3389/fpsyg.2020.00721

Horan, B., Heckenberg, R., Maruff, P., & Wright, B. (2020). Development of a new virtual reality test of cognition: Assessing the test-retest reliability, convergent and ecological validity of CONVIRT. *BMC Psychology*, *8*(1), 1–10. https://doi.org/10.1186/s40359-020-00429-x

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, *6*(2), 143–171. https://doi.org/10.1207/s15327574ijt0602_4

Kenney, L. E., Margolis, S. A., Davis, J. D., & Tremont, G. (2019). The screening utility and ecological validity of the neuropsychological assessment battery bill payment subtest in older adults with and without Dementia. *Archives of Clinical Neuropsychology*, *34*(7), 1156–1164. https://doi.org/10.1093/arclin/acz033

Kibby, M. Y., Schmitter-Edgecombe, M., & Long, C. J. (1998). Ecological validity of neuropsychological tests: Focus on the California Verbal Learning Test and the Wisconsin Card Sorting Test. *Archives of Clinical Neuropsychology*, *13*(6), 523–534. https://doi.org/10.1093/arclin/13.6.523

Kieffaber, P. D., Marcoulides, G. A., White, M. H., & Harrington, D. E. (2007). Modeling the ecological validity of neurocognitive assessment in adults with acquired brain injury. *Journal of Clinical Psychology in Medical Settings*, *14*(3), 206–218. https://doi.org/10.1007/s10880-007-9075-6

Kipps, C. M., & Hodges, J. R. (2005). Cognitive assessment for clinicians. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(suppl_1), i22–i30. https://doi.org/10.1136/jnnp.2004.059758

Kourtesis, P., Collina, S., Doumas, L. A., & MacPherson, S. E. (2021). Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity. *Journal of the International Neuropsychological Society*, *27*(2), 181–196. https://doi.org/10.1017/S1355617720000764

Krishna, M., Beulah, E., Jones, S., Sundarachari, R., A, S., Kumaran, K., Karat, S. C., Copeland, J. R. M., Prince, M., & Fall, C. (2016). Cognitive function and disability in late life: An ecological validation of the 10/66 battery of cognitive tests among community-dwelling older adults in South India. *International Journal of Geriatric Psychiatry*, *31*(8), 879–891. https://doi.org/10.1002/gps.4404

Krohn, S., Tromp, J., Quinque, E., Belger, J., Klotzsche, F., Rekers, S., Chojecki, P., Mooij, J., Akbal, M., McCal, C., Villringer, A., Gaebler, M., Finke, C., & Thone-Otto, A. (2020). Multidimensional evaluation of Virtual Reality paradigms in clinical neuropsychology: The VR-Check framework. *Journal of Medical Internet Research*, *22*(4), e16724. https://doi.org/10.2196/16724

Kvavilashvili, L., & Ellis, J. (2004). Ecological validity and twenty years of real-life/laboratory controversy in memory research: A critical (and historical) review. *History and Philosophy of Psychology*, *6*(1), 59–80.

Lewkowicz, D. J. (2001). The concept of ecological validity: What are its limitations and is it bad to be invalid? *Infancy*, *2*(4), 437–450. https://doi.org/10.1207/S15327078IN0204_03

Lindley, P. A., Bartram, D., & Kennedy, N. (2008). EFPA Review Model for the description and evaluation of psychological tests: Test review form and notes for reviewers: Version 3.42. *EFPA Standing Committee on Tests and Testing*.

Lippa, S. M., Pastorek, N. J., Romesser, J., Linck, J., Sim, A. H., Wisdom, N. M., & Miller, B. I. (2014). Ecological validity of performance validity testing. *Archives of Clinical Neuropsychology 29*(3), 236–244. https://doi.org/10.1093/arclin/acu002

Luria, A. R. (1976). *The working brain: An introduction to neuropsychology*. Basic Books.

Maeir, A., Krauss, S., & Katz, N. (2011). Ecological validity of the Multiple Errands Test (MET) on discharge from neurorehabilitation hospital. *Occupation, Participation and Health*, *31*(1), S38–S46. https://doi.org/10.3928/15394492-20101108-07

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution (monografia)*. https://www.psychopen.eu/fileadmin/user_upload/books/mayring/ssoar-2014-mayring-Qualitative_content_analysis_theoretical_foundation.pdf

Mitchell, M., & Miller, L. S. (2008). Prediction of functional status in older adults: The ecological validity of four Delis–Kaplan Executive Function System tests. *Journal of Clinical and Experimental Neuropsychology*, *30*(6), 683–690. https://doi.org/10.1080/13803390701679893

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed1000097

Norris, G., & Tate, R. L. (2000). The Behavioural Assessment of the Dysexecutive Syndrome (BADS): Ecological, concurrent and construct validity. *Neuropsychological Rehabilitation*, *10*(1), 33–45. https://doi.org/10.1080/096020100389282

Odhuba, R. A., Van den Broek, M. D., & Johns, L. C. (2005). Ecological validity of measures of executive functioning. *The British Journal of Clinical Psychology*, *44*(Pt 2), 269–278. https://doi.org/10.1348/0144665505X29431

Owen, W. J., Borowsky, R., & Sarty, G. E. (2004). FMRI of two measures of phonological processing in visual word recognition: Ecological validity matters. *Brain and Language*, *90*(1–3), 40–46. https://doi.org/10.1016/S0093-934X(03)00418-8

Olson, K., Jacobson, K. K., & Van Oot, P. (2013). Ecological validity of pediatric neuropsychological measures: Current state and future directions. *Applied neuropsychology*, *2*(1), 17–23. https://doi.org/10.1080/21622965.2012.686330

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Paiva, G. C. D. C., Fialho, M. B., Costa, D. D. S., & Paula, J. J. D. (2016). Ecological validity of the five digit test and the oral trails test. *Arquivos de Neuro-Psiquiatria*, *74*(1), 29–34. https://doi.org/10.1590/0004-282X20150184

Parsons, T. D. (2011). Neuropsychological assessment using virtual environments: Enhanced assessment technology for improved ecological validity. In *Advanced computational intelligence paradigms in healthcare 6. Virtual reality in psychotherapy, rehabilitation, and assessment* (pp. 271–289). Springer.

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective, and social neurosciences. *Frontiers in Human Neuroscience*, *9*, 660. https://doi.org/10.3389/fnhum.2015.00660

Parsons, T. D., Carlew, A. R., Magtoto, J., & Stonecipher, K. (2015). The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical neuropsychology. *Neuropsychological rehabilitation*, *27*(5), 777–807. https://doi.org/10.1080/09602011.2015.1109524

Paula, J. J., Costa, M. V., Bocardi, M. B., Cortezzi, M., De Moraes, E. N., & Malloy-Diniz, L. F. (2013). The Stick Design Test on the assessment of older adults with low formal education: Evidences of construct, criterion-related and ecological validity. *International Psychogeriatrics*, *25*(12), 2057–2065. https://doi.org/10.1017/S1041610213001282

Pezzuti, L., Mastrantonio, E., & Orsini, A. (2013). Construction and validation of an ecological version of the Wisconsin Card Sorting Test applied to an elderly population. *Aging, Neuropsychology, and Cognition*, *20*(5), 567–591. https://doi.org/10.1080/13825585.2012.761668

Phillips, L. H., Kliegel, M., & Martin, M. (2006). Age and planning tasks: The influence of ecological validity. *International journal of Aging & Human Development*, *62*(2), 175–184. https://doi.org/10.2190/EM1W-HAYC-TMLM-WW8X

Poletti, M. (2010). Orbitofrontal cortex-related executive functions in children and adolescents: Their assessment and its ecological validity. *Neuropsychological Trends*, *7*, 7–27. https://doi.org/10.7358/neur-2010-007-pole

Possin, K. L., LaMarre, A. K., Wood, K., Mungas, D. M., & Kramer, J. H. (2014). Ecological validity and neuroantomical correlates of the NIH EXAMINER executive composite score. *Journal of the International Neuropsychological Society*, *20*(1), 20–28. https://doi.org/10.1017/S1355617713000611

Price, K. J., Joschko, M., & Kerns, K. (2003). The ecological validity of pediatric neuropsychological tests of attention. *The Clinical Neuropsychologist*, *17*(2), 170–181. https://doi.org/10.1076/clin.17.2.170.16506

Ready, R. E., Stierman, L., & Paulsen, J. S. (2001). Ecological validity of neuropsychological and personality measures of executive functions. *The Clinical Neuropsychologist*, *15*(3), 314–323. https://doi.org/10.1076/clin.15.3.314.10269

Renison, B., Ponsford, J., Testa, R., Richardson, B., & Brownfield, K. (2012). The ecological and construct validity of a newly developed measure of executive function: The virtual library task. *Journal of the International Neuropsychological Society*, *18*(3), 440–450. https://doi.org/10.1017/S1355617711001883

Romero-Ayuso, D., Castillero-Perea, Á., González, P., Navarro, E., Molina-Massó, J. P., Funes, M. J., Ariza-Veja, P., Toledano-González, A., & Triviño-Juárez, J. M. (2019). Assessment of cognitive instrumental activities of daily living: A systematic review. *Disability and Rehabilitation*, *2019*, 1–17. https://doi.org/10.1080/09638288.2019.1665720

Rumpf, U., Menze, I., Müller, N. G., & Schmicker, M. (2019). Investigating the potential role of ecological validity on change-detection memory tasks and distractor processing in younger and older adults. *Frontiers in Psychology*, *10*, 1046. https://doi.org/10.3389/fpsyg.2019.01046

Sandvoll, A. M., Grov, E. K., & Simonsen, M. (2020). Nursing home residents' ADL status, institution-dwelling and association with outdoor activity: A cross-sectional study. *PeerJ*. *8*, e10202. https://doi.org/10.7717/peerj.10202

Sbordone, R. J. (1996). Ecological validity: Some critical issues for the neuropsychologist. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 15–41). CRC Press.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, *2*(4), 419–436. https://doi.org/10.1207/S15327078IN0204_02

Silver, C. H. (2000). Ecological validity of neuropsychological assessment in childhood traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, *15*(4), 973–988. https://doi.org/10.1097/00001199-200008000-00002

Silverberg, N. D., & Millis, S. R. (2009). Impairment versus deficiency in neuropsychological assessment: Implications for ecological validity. *Journal of the International Neuropsychological Society*, *15*(1), 94–102. https://doi.org/10.1017/S1355617708090139

Smilek, D., Carriere, J. S., & Cheyne, J. A. (2010). Failures of sustained attention in life, lab, and brain: Ecological validity of the SART. *Neuropsychologia*, *48*(9), 2564–2570. https://doi.org/10.1016/j.neuropsychologia.2011.01.037

Solanto, M. V., Abikoff, H., Sonuga-Barke, E., Schachar, R., Logan, G. D., Wigal, T., Hechtman, L., Hinshaw, S., & Turkel, E. (2001). The ecological validity of delay aversion and response inhibition as measures of impulsivity in AD/HD: A supplement to the NIMH multimodal treatment study of AD/HD. *Journal of Abnormal Child Psychology*, *29*(3), 215–228. https://doi.org/10.1023/A:1010329714819

Sousa, F. N., Costa, A. P., & Moreira, A. (2019). *webQDA [programa de computador]*. Microio/Ludomedia.

Spooner, D. M., & Pachana, N. A. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*, *21*(4), 327–337. https://doi.org/10.1016/j.acn.2006.04.004

Tang, S. F., Chen, I. H., Chiang, H. Y., Wu, C. T., Hsueh, I. P., Yu, W. H., & Hsieh, C. L. (2018). A comparison between the original and Tablet-based Symbol Digit Modalities Test in patients with schizophrenia: Test-retest agreement, random measurement error, practice effect, and ecological validity. *Psychiatry research*, *260*, 199–206. https://doi.org/10.1016/j.psychres.2017.11.066

Temple, R. O., Zgaljardic, D. J., Abreu, B. C., Seale, G. S., Ostir, G. V., & Ottenbacher, K. J. (2009). Ecological validity of the neuropsychological assessment battery screening module in post-acute brain injury rehabilitation. *Brain injury*, *23*(1), 45–50. https://doi.org/10.1080/02699050802590361

Thornton, A. E., Kristinsson, H., DeFreitas, V. G., & Thornton, W. L. (2010). The ecological validity of everyday cognition in hospitalized patients with serious mental illness. *Journal of Clinical and Experimental Neuropsychology*, *32*(3), 299–308. https://doi.org/10.1080/13803390903002209

Tupper, D. E., & Cicerone, K. D. (1990). Introduction to the neuropsychology of everyday life. In *The neuropsychology of everyday life: Assessment and basic competencies* (pp. 3–18). Springer.

van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2008). A large-scale cross-sectional and longitudinal study into the ecological validity of neuropsychological test measures in neurologically intact people. *Archives of Clinical Neuropsychology*, *23*(7–8), 787–800. https://doi.org/10.1016/j.acn.2008.09.002

van der Ham, I. J., Faber, A. M., Venselaar, M., van Kreveld, M. J., & Löffler, M. (2015). Ecological validity of virtual environments to assess human navigation ability. *Frontiers in Psychology*, *6*(637), 637–636. https://doi.org/10.3389/fpsyg.2015.00637

Verdejo-Garcia, A., Bechara, A., Recknor, E., & Perez-Garcia, M. (2006). Decision-making and the Iowa Gambling Task: Ecological validity in individuals with substance dependence. *Psychologica Belgica*, *46*(1–2), 55. https://doi.org/10.5334/pb-46-1-2-55

Vingerhoets, G., Vermeule, E., & Santens, P. (2005). Impaired intentional content learning but spared incidental retention of contextual information in non-demented patients with Parkinson's disease. *Neuropsychologia*, *43*(5), 675–681. https://doi.org/10.1016/j.neuropsychologia.2004.09.003

Vordenberg, J. A., Barrett, J. J., Doninger, N. A., Contardo, C. P., & Ozoude, K. A. (2014). Application of the Brixton spatial anticipation test in stroke: Ecological validity and performance characteristics. *The Clinical Neuropsychologist*, *28*(2), 300–316. https://doi.org/10.1080/13854046.2014.881555

Wallisch, A., Little, L. M., Dean, E., & Dunn, W. (2018). Executive function measures for children: A scoping review of ecological validity. *Occupation, Participation and Health*, *38*(1), 6–14. https://doi.org/10.1177/1539449217727118

Weber, E., Goverover, Y., & DeLuca, J. (2019). Beyond cognitive dysfunction: Relevance of ecological validity of neuropsychological tests in multiple sclerosis. *Multiple Sclerosis*, *25*(10), 1412–1419. https://doi.org/10.1177/1352458519860318

Weis, R., & Totten, S. J. (2004). Ecological validity of the Conners' Continuous Performance Test II in a school-based sample. *Journal of Psychoeducational Assessment*, *22*(1), 47–61. https://doi.org/10.1177/073428290402200104

Wen, J. H., Boone, K., & Kim, K. (2006). Ecological validity of neuropsychological assessment and perceived employability. *Journal of Clinical and Experimental Neuropsychology*, *28*(8), 1423–1434. https://doi.org/10.1080/13803390500409609

Wilson, B. A. (1993). Ecological validity of neuropsychological assessment: Do neuropsychological indexes predict performance in everyday activities? *Applied and Preventive Psychology*, *2*(4), 209–215. https://doi.org/10.1016/S0962-1849(05)80091-5

Wilson, B. A., Alderman, N., Burgess, P. W., Esmlie, H., & Evans, J. J. (1996). *Behavioural assessment of the dysexecutive syndrome*. Thames Valley Test Company.

Wood, R. L., & Liossi, C. (2006). The ecological validity of executive tests in a severely brain injured sample. *Archives of Clinical Neuropsychology*, *21*(5), 429–437. https://doi.org/10.1016/j.acn.2005.06.014

Woods, S. P. (2021). Introduction to the special issue on the neuropsychology of daily life. *Neuropsychology*, *35*(1), 1–2. https://doi.org/10.1037/neu0000716

World Health Organization. (2013). International classification of functioning, disability and health (ICF): ICF and ICF-Online. Retrieved from http://apps.who.int/classifications/icfbrowser/.

Yantz, C. L., Johnson-Greene, D., Higginson, C., & Emmerson, L. (2010). Functional cooking skills and neuropsychological functioning in patients with stroke: An ecological validity study. *Neuropsychological rehabilitation*, *20*(5), 725–738. https://doi.org/10.1080/09602011003765690

Zhang, Y., Xiong, Y., Yu, Q., Shen, S., Chen, L., & Lei, X. (2021). The activity of daily living (ADL) subgroups and health impairment among Chinese elderly: A latent profile analysis. *BMC Geriatrics*, *21*(1), 30. https://doi.org/10.1186/s12877-020-01986-x

Zhou, S. Z., Jiang, W., Wang, H., Wei, N., & Yu, Q. T. (2020). Predictive value of pretreatment albumin-to-alkaline phosphatase ratio for overall survival for patients with advanced non-small cell lung cancer. *Cancer Medicine*, *9*(17), 6268–6280. https://doi.org/10.1002/cam4.3244

Ziemnik, R. E., & Suchy, Y. (2019). Ecological validity of performance-based measures of executive functions: Is face validity necessary for prediction of daily functioning? *Psychological assessment*, *31*(11), 1307–1318. https://doi.org/10.1037/pas0000751

Zgaljardic, D. J., Yancy, S., Temple, R. O., Watford, M. F., & Miller, R. (2011). Ecological validity of the screening module and the Daily Living tests of the Neuropsychological Assessment Battery using the Mayo-Portland Adaptability Inventory-4 in postacute brain injury rehabilitation. *Rehabilitation Psychology*, *56*(4), 359–365. https://doi.org/10.1037/a0025466