Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

Validation of an integrated method for determining cognitive ability: Implications for routine assessments and clinical trials

Olivier Godefroy [a,*], Laura Gibbons [b], Momar Diouf [c], David Nyenhuis [d], Martine Roussel [a], Sandra Black [e], Jean Marc Bugnicourt [a] and GREFEX study group[1]

[a] Department of Neurology and Laboratory of Functional Neurosciences University Hospital of Amiens, France
[b] Department of General Internal Medicine, University of Washington, Harborview Medical Center, Seattle, WA, USA
[c] Department of Biostatistics, University Hospital of Amiens, France
[d] Hauenstein Neuroscience Center, Saint Mary's Health Care, Grand Rapids, MI, USA
[e] Brill Chair in Neurology, Dept of Medicine, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 12 July 2013
Reviewed 11 September 2013
Revised 25 September 2013
Accepted 27 January 2014
Action editor Sarah MacPherson
Published online 7 February 2014

Keywords:
Executive functions
Mild cognitive impairment
Dementia
Stroke

ABSTRACT

Introduction: Although accurate diagnosis of deficit of mild intensity is critical, various methods are used to assess, dichotomize and integrate performance, with no validated gold standard. This study described and validated a framework for the analysis of cognitive performance.

Methods: This study was performed by using the *Groupe de Réflexion sur L'Evaluation des Fonctions EXécutives* (GREFEX) database (724 controls and 461 patients) examined by 7 tests assessing executive functions. The first phase determined the criteria for the cutoff scores, the second phase, the effect of test number on diagnostic accuracy and the third phase, the best methods for combining test scores into an overall summary score. Four validation criteria were used: determination of impaired performance as compared to expected one, false-positive rate ≤5%, detection of both single and multiple impairments with optimal sensitivity.

* *Corresponding author*. Service de Neurologie CHU Nord, 80054 Amiens Cedex, France.
E-mail address: godefroy.olivier@chu-amiens.fr (O. Godefroy).

[1] Groupe de Réflexion sur l'Evaluation des Fonctions Exécutives ***(GREFEX) study group***: the following centers and investigators participated in the GREFEX cooperative study (n = number of patients included at each center; investigators): Amiens University Hospital (F) ($n = 183$; O. Godefroy and M. Roussel), Angers University Hospital (F) ($n = 19$; D. Le Gall), Heliomarin Rehabilitation Center Berck (F) ($n = 15$; C. Bertola), Bordeaux University Hospital (F) ($n = 28$; J.M. Giroire and P.A. Joseph), Saint Luc University Hospital Brussels (B) ($n = 6$; X. Seron, F. Coyette), Cholet General Hospital (F) ($n = 8$; E. Bretault and I. Bernard), Ottignies William Lennox Center (B) ($n = 3$; M. Leclercq), Garches University Hospital (F) ($n = 9$; P. Azouvi and C. Vallat-Azouvi), Grenoble University Hospital (F) ($n = 24$; P Pollack, *C Ardouin* and C. Mosca), Lausanne University Hospital (CH) ($n = 9$; C Bindschadler), Lay St Christophe Rehabilitation Center (F) ($n = 3$; M. Krier), Liège Department of Cognitive Sciences (B) ($n = 19$; T. Meulemans and V. Marquet), Lille Stroke Center University Hospital (F) ($n = 26$; D. Leys and M. Roussel), Nantes University Hospital (F) ($n = 8$; P. Renou and M. Vercelletto), Nice University Hospital (F) ($n = 6$; E. Michel and P. Robert), Nîmes University Hospital (F) ($n = 15$; P. Labauge and C. Franconie), Paris-La Salpêtrière University Hospital Neurology Department (F) ($n = 18$; B. Pillon and B. Dubois), Paris-La Salpêtrière University Hospital Geriatrics Department (F) ($n = 13$; B. Dieudonnée and M. Verny), Paris-Broca University Hospital (F) ($n = 5$; H. Lenoir and J. De Rotrou), Rouen University Hospital (F) ($n = 56$; D. Hannequin and S. Bioux), Sion Rehabilitation Clinic (CH) ($n = 12$; J. Fuchs, A. Bellmann and P. Vuadens).

http://dx.doi.org/10.1016/j.cortex.2014.01.016

sensitivity and specificity
Diagnostic accuracy

Results: The procedure based on 5th percentile cutoffs determined from standardized residuals was the most appropriate procedure. Although area under the curve (AUC) increased with the number of scores ($p = .0001$), the false-positive rate also increased ($p = .0001$), resulting in suboptimal sensitivity for detecting selective impairment. Two overall summary scores, the average of the seven process scores and the Item Response Theory (IRT) score, had significantly ($p = .0001$) higher AUCs, even for patients with a selective impairment, and provided higher resulting prevalence of dysexecutive disorders ($p = .0001$).

Conclusions: The present study provides and validates a generative framework for the interpretation of cognitive data. Two overall summary score met all 4 validation criteria. A practical consequence is the need to profoundly modify the analysis and interpretation of cognitive assessments for both routine use and clinical research.

1. Introduction

Given the importance of cognition in contemporary societies and its impact on health, accurate diagnosis of cognitive ability is critical. Cognitive ability is typically assessed in subjects from heterogeneous backgrounds, using a battery of cognitive tests covering language, visuospatial, memory, executive and general cognitive domains. Each test yields between 1 and 15 performance scores, which are dichotomized (normal *vs* impaired) according to norms ideally corrected as appropriate for age and education. Next, the dichotomized scores are integrated to form a clinical diagnosis. Despite major progress in this field, a survey of clinical practice in academic memory clinics and rehabilitation centers (Godefroy et al., 2004) and a review of published studies assessing preclinical (Bateman et al., 2012; Hultsch, MacDonald, Hunter, Levy-Bencheton, & Strauss, 2000; Knopman et al., 2012; Tractenberg and Pietrzak, 2011) and mild cognitive impairment (Clark et al., 2013; Winblad et al., 2004), dementia (Dubois et al., 2010), stroke (Godefroy et al., 2011; Tatemichi et al., 1994), cardiac surgery (Moller et al., 1998; Murkin et al., 1995), multiple sclerosis (Rao, Leo, Bernardin, & Unverzagt, 1991; Sepulcre et al., 2006) and Parkinson's disease (Cooper, Sagar, Tidswell, & Jordan, 1994; Dalrymple-Alford et al., 2011; Litvan et al., 2012) showed that various methods are used to assess, dichotomize and integrate performance, with no reference to a validated gold standard (Brooks and Iverson, 2010; Crawford, Garthwaite, & Gault, 2007; Dalrymple-Alford et al., 2011; Lezak, Howieson, & Loring, 2004; Mungas, Marshall, Weldon, Haan, & Reed, 1996; Sepulcre et al., 2006). Several carefully designed studies have shown that the use of different criteria for impairment dramatically influences the estimated prevalence of cognitive impairment (Clark et al., 2013; Dalrymple-Alford et al., 2011; Sepulcre et al., 2006). Most importantly, a review of these studies failed to provide a rationale for determining the best criterion in the assessment of cognitive impairment. The absence of a standardized method undermines the reliable determination of cognitive status, which in turn has a major impact on both clinical practice and clinical research. This point is especially important because the objective of cognitive assessment has shifted towards the diagnosis of deficit of mild intensity or of selective deficit.

A systematic review of previous studies, of diagnostic criteria of cognitive impairment (e.g., Clark et al., 2013; Knopman et al., 2012; Winblad et al., 2004) and of available normative data of clinical battery shows that methodology differ in three critical respects: the dichotomization of performance, the integration of several dichotomized scores, and the possible use of a global summary score. The first issue concerns the cutoff criteria used to dichotomize performance. Most cutoffs are based on means and standard deviations (SDs) and use varying cutpoints from 1.5 to 1.98 SD. However, the effect of the deviation from normality of most cognitive scores is rarely addressed. Cutoff scores are sometimes based on percentiles, the 10th and 5th percentiles being the most frequently used. Second, cognitive assessment involves multiple tests, thus providing numerous scores. Procedures differ regarding the combination of tests and scores used as criterion of cognitive impairment. Some procedures consider that just one impaired test score is sufficient for classifying a subject as "impaired", whereas others require "impaired" subjects to have at least two (or more) impaired test scores. Other procedures take into account the cognitive domain (each domain being assessed with one to several scores) and classify as impaired subjects with at least one or two or more impaired domain. In clinical practice, the interpretation is usually based on counting the number of impaired scores. Importantly, the use of multiple tests improves sensitivity but it can also artificially increase the false-positive rate (i.e., lowering the specificity) (Brooks and Iverson, 2010), a concern especially important as the scores are often inter-correlated (Crawford et al., 2007). This well-known redundancy artifact is addressed in trials at the stage of interpretation of statistical analyses using correction for multiple analyses, such as Bonferroni correction. However, this artifact has rarely been examined in the field of test battery interpretation, which typically involves 20–50 performance scores (Brooks and Iverson, 2010; Godefroy et al., 2010). Thus, there is currently no rationale for determining the optimal number of tests/scores for diagnostic accuracy (i.e., both sensitivity and specificity). Third, some trials have combined individual test