## VOLUME I: RESEARCH COMPONENT

# Validation of the Neuropsychological Assessment Battery Screening Measure (NAB-S) in participants with Traumatic Brain Injury

By

Thomas Morien Michael

A THESIS SUBMITTED TO THE UNIVERSITY OF BIRMINGHAM FOR THE DEGREE OF DOCTOR OF CLINICAL PSYCHOLOGY

Department of Clinical Psychology
School of Psychology
The University of Birmingham

July 2016

# THESIS OVERVIEW

This thesis is submitted to the University of Birmingham in partial fulfilment of the requirements for the degree of Doctor of Clinical Psychology. This thesis is comprised of two volumes.

Volume I reports a systematic review and an empirical research paper. The systematic review focusses on psychometric properties of rating scales of executive dysfunction due to acquired brain injury. A systematic search of literature databases identified 21 relevant journal articles. The psychometric properties of four rating scales of executive dysfunction were then reviewed according to published quality criteria for psychometric measures. The reviewed rating scales had varying amounts of published data on their psychometric properties, and conclusions were stated where there was sufficient evidence to do so.

The empirical research paper reports the validation of the Neuropsychological Assessment Battery Screening Measure (NAB-S) in participants with traumatic brain injury. The NAB-S was administered to a sample of 44 individuals with TBI, alongside a well validated battery of neuropsychological tests, in order to investigate convergent validity between the two batteries, and to determine to what extent the NAB-S was predictive of impairment as measured by the well-established test battery. Conclusions as to the validity of using the NAB-S with a TBI population are presented.

Volume I also includes a public domain briefing document, which provides a summary of the systematic review and empirical study.

Volume II comprises the clinical component, and contains five Clinical Practice Reports (CPR) completed over the course of the training. These reports are clinical and empirical work conducted during placements in an older adult community mental health team, an adult community mental health team, an outpatient brain injury rehabilitation centre, another older adult mental health team, and a child and adolescent mental health team. CPR1 presents cognitive-behavioural and psychodynamic formulations of an older female client experiencing depression. CPR2 presents an assessment, formulation and intervention of an adult male experiencing depression. CPR3 presents a service evaluation of staff perceptions of client complexity in an outpatient brain injury rehabilitation service. CPR4 presents an assessment, formulation and intervention for an older male experiencing anxiety and depression. CPR5 is an abstract of a presentation of a single-case experimental design of an assessment, formulation and intervention for a teenage girl experiencing obsessive compulsive disorder.

Dedicated to Misty

# ACKNOWLEDGEMENTS

There are many people who have helped me to complete this training course and thesis.

Firstly, I would like to thank my wife, my mother and my friends, for supporting me emotionally through a long and difficult journey.

Secondly, I would like to thank my fellow trainee clinical psychologists, who also helped and supported me throughout.

Finally, I would like to thank all of my supervisors and tutors, both on placements and the course team, for their guidance and encouragement. In particular, I would like to thank Teresa Madurai, Michelle Fisher, Chris Jones, Theresa Madurai and Dave Hacker.

# CONTENTS OF VOLUME I: RESEARCH COMPONENT

## SYSTEMATIC REVIEW: PSYCHOMETRIC PROPERTIES OF RATING SCALES OF EXECUTIVE DYSFUNCTION DUE TO ACQUIRED BRAIN INJURY

# EMPIRICAL PAPER: THE VALIDATION OF THE NEUROPSYCHOLOGICAL ASSESSMENT BATTERY SCREENING MEASURE (NAB-S) FOR PARTICIPANTS WITH TRAUMATIC BRAIN INJURY (TBI)

# PUBLIC DOMAIN BRIEFING PAPER

## SYSTEMATIC REVIEW

## EMPIRICAL PAPER

# CONTENTS OF VOLUME II: CLINICAL COMPONENT

## CPR 1

### A Cognitive and Psychodynamic formulation of Jenny, a 65 year old woman experiencing depression

# CPR 2

## Case Study: Cognitive Behavioural Therapy with Mindfulness for Stephen, a 29 year old man experiencing depression

# CPR 3

## Service Evaluation: Staff beliefs of client and service factors which impact on working with complex clients at a specialist community brain injury rehabilitation unit

# CPR 4

## Case Study: Cognitive Behavioural Therapy for David, a 74 year old man experiencing anxiety and globus sensation

# CPR 5

## Oral Presentation of Single Case Experimental Design: CBT & Mindfulness for Alia, a 14 year old girl experiencing OCD

# LIST OF ILLUSTRATIONS

# LIST OF TABLES

# LIST OF APPENDICES

# SYSTEMATIC REVIEW:

# Psychometric properties of rating scales of executive dysfunction due to acquired brain injury

by

Thomas Morien Michael

## Abstract

### Aims and rationale

Difficulties with executive function are a common sequelae of traumatic brain injury, and are the cause of significant disability and impediment to rehabilitation. Performance based measures of executive dysfunction have experienced difficulty in quantifying the range and diverse of everyday difficulties that can proceed from executive dysfunction. For this reason, self-report and independently rated measures of executive function have been developed. Such measures need to be valid and reliable for both patients and clinicians, as difficulties with executive function are associated with poorer rehabilitation outcomes. This study aims to review the existing evidence on the psychometric properties of such measures and discusses properties which an ideal measure of executive function might have. The review is structured into two parts, the first considers the quality of the journal articles themselves using a quality framework for observational studies. The second part outlines the psychometric properties of individual measures and assesses these against an additional appropriate quality criteria.

### Method

A systematic literature search of the PsychInfo, MEDLINE (R) & EMBASE databases was conducted in two stages. The first combined terms on executive function and psychometric properties to derive a list of self-report and independently rated measures of executive function. The second then combined a list of these measures with terms relating to psychometric properties. Exclusion criteria reduced the list of results to 21 peer reviewed journal articles.

A two part quality review first considers the quality of the journal articles, then examines the evidence for psychometric quality of individual measures, according to two separately published quality criteria.

**Discussion**

The psychometric properties of the rating scales are discussed, focussing on the two most commonly investigated measures, the Dysexecutive Questionnaire (DEX) and the Frontal Systems Behaviour Scale (FrSBe).

**Conclusions**

Currently, the DEX is the most investigated measure of executive function and correspondingly has the largest body of psychometric evidence with this population. More recently developed measures, such as the FrSBe and DEX-R, which consider the heterogeneous nature of executive dysfunction, have better factor structure and construct validity. These rating scales also have convergent validity with neuropsychological tests, and each other in terms of factor structure, as well as their ecological validity at predicting real life impairment. For clinical psychologists and other clinicians, this means that at present, the use of the DEX and FrSBe are most justified with a population of ABI clients.

# Introduction

## Executive function

Executive function (EF) is a psychological construct referring to a system of cognitive processes responsible for goal-based, planned and controlled behaviour; as opposed to impulsive, over-learned or "automatic" behaviour which requires little attention, control or inhibition. One definition of EF is given by Burgess & Alderman (2004):

> "The term 'executive functions' refers to those abilities that enable a person to determine goals, formulate new and useful ways of achieving them, and then follow and adapt this proposed course in the face of competing demands and changing circumstances, often over long periods of time. Crucial aspects of these abilities are thought to be supported by the frontal lobes of the brain, and sometimes the term 'frontal lobe function' is (imprecisely) used as a shorthand to refer to them. Damage to these processes results in a range of symptoms collectively referred to as the *dysexecutive syndrome*.'""

As EF is crucial to a person's ability to achieve goals, the presence of executive deficits can adversely affect rehabilitation efforts following traumatic brain injury (D'Esposito & Gazzaley, 2005; Oddy & Worthington, 2009). For this reason, clinicians need assessment tools which can quickly, validly and reliably assess the presence of executive difficulties.

## Ecological and construct validity of tests and rating scales of executive function

Various psychometric tests of EF have been developed, many of which are reviewed by (Pickens, Ostwald, Murphy-Pace, & Bergstrom, 2010). However, as discussed by Pickens et al., tests of EF have generally been developed according to particular theories of the cognitive processes thought to be critical to EF. For this reason, different tests of EF may measure different aspects of the construct as a whole. This, combined with a greater understanding of the underlying neurophysiology of executive deficits has led some researchers to critique the construct of a unified executive function, or a dysexecutive syndrome (Gilbert & Burgess, 2008; Stuss & Alexander, 2007). This in turn has led some researchers to argue that some tests of EF may lack ecological validity, in that they may not be very predictive of the real-life executive difficulties which can be experienced by a person following traumatic brain injury (Burgess et al., 2006; Burgess, Alderman, Evans, Emslie, & Wilson, 1998). For these reasons, self-report and independently-rated rating scales of EF have been developed, in order to capture the full range of executive difficulties a person can experience. Studies such as that by Burgess et al., (1998) have thus used these rating scales as a measure of ecological validity, with which they can compare a person's results from psychometric tests of EF.

In order for a rating scale of EF to be useful to clinicians, the scale must be able to validly and reliably measure everyday difficulties that proceed from executive dysfunction. To aid in the process of developing valid and reliable rating scales, minimum psychometric criteria for health questionnaires have been proposed (Mokkink et al., 2010; Terwee et al., 2007). These criteria are describe the psychometric properties that are required for reliable and valid clinical interpretation of test performance and are comprised of content validity, internal consistency, criterion validity, construct validity, reproducibility, longitudinal validity, responsiveness, floor and ceiling effects, and interpretability (for definitions of these criteria, see appendix I). In addition to providing a guide by which future rating scales might be developed and validated, the Terwee et al. (2008) criteria enable a comparison of existing rating scales based upon their reported psychometric properties.

Furthermore, in reviewing published literature on psychometric properties of such scales, it is vital to also consider the methodological quality of that literature. For this reason, this review will utilise a list of criteria published by von Elm et al., (2008), developed for the assessment of quality of cross-sectional observational study designs. For definitions of these criteria, see appendix II.

**Objectives of this systematic review**

A previous review of rating scales of executive function was conducted by Malloy & Grace (2005). This review discussed the psychometric properties of these rating scales, but was not performed systematically, had limited criteria to assess the quality of these scales, and included rating scales developed for conditions such as dementia, as well as those developed for traumatic brain injury. In order to expand and improve on this, the present review aims to:

a) Carry out a systematic literature review.

b) Review the psychometric properties of the rating scales identified according to the Terwee et al. and Mokkink et al. criteria on psychometric properties.

c) Review the literature that investigates these measures according to the von Elm et al. criteria for observational study methodology.

d) To focus on self-report or independently-reported rating scales of EF developed for or validated for acquired brain injury.

e) To review additional literature published since the review of Malloy & Grace (2005).

## Method

### Search Terms

The systematic literature search was constructed using search terms listed in tables 1 and 2 below. Initially a search of the PsychInfo database using the terms; "executive function", "dysexecutive" or "executive dysfunction" was performed (1). This was then combined with the terms; "factor analysis" or "validation" or "validity" or "reliability" (2). This combined search produced 1,164 results (3).

**Table 1 – Search terms used to find rating scales of executive function**

| Search | Search Terms | Number of Articles |
|---|---|---|
| 1 | "executive function" or "dysexecutive" or "executive dysfunction" | 13,872 |
| 2 | "factor analysis" or "validation" or "validity" or "reliability" | 194,478 |
| 3 | 1 and 2 | 1,164 |

These 1,164 results were then examined for titles of rating scales of executive function, which found 8 rating scales, listed in table 2 below (4-11). Searches of the PsychInfo, MEDLINE (R) & EMBASE databases were then conducted searching for studies which used these rating scales only, and the results are listed in order of frequency in table 2 below (4-11). These terms were then combined with the term; "factor analysis" or "validation" or "validity" or "reliability" (2). This produced 175 results (12). The abstracts of these 175 results were then examined, according to the inclusion and exclusion criteria listed in table 3 below. A PRISMA diagram (figure 1) describes this process.

**Table 2 – Search terms used to find psychometric properties of rating scales of EF**

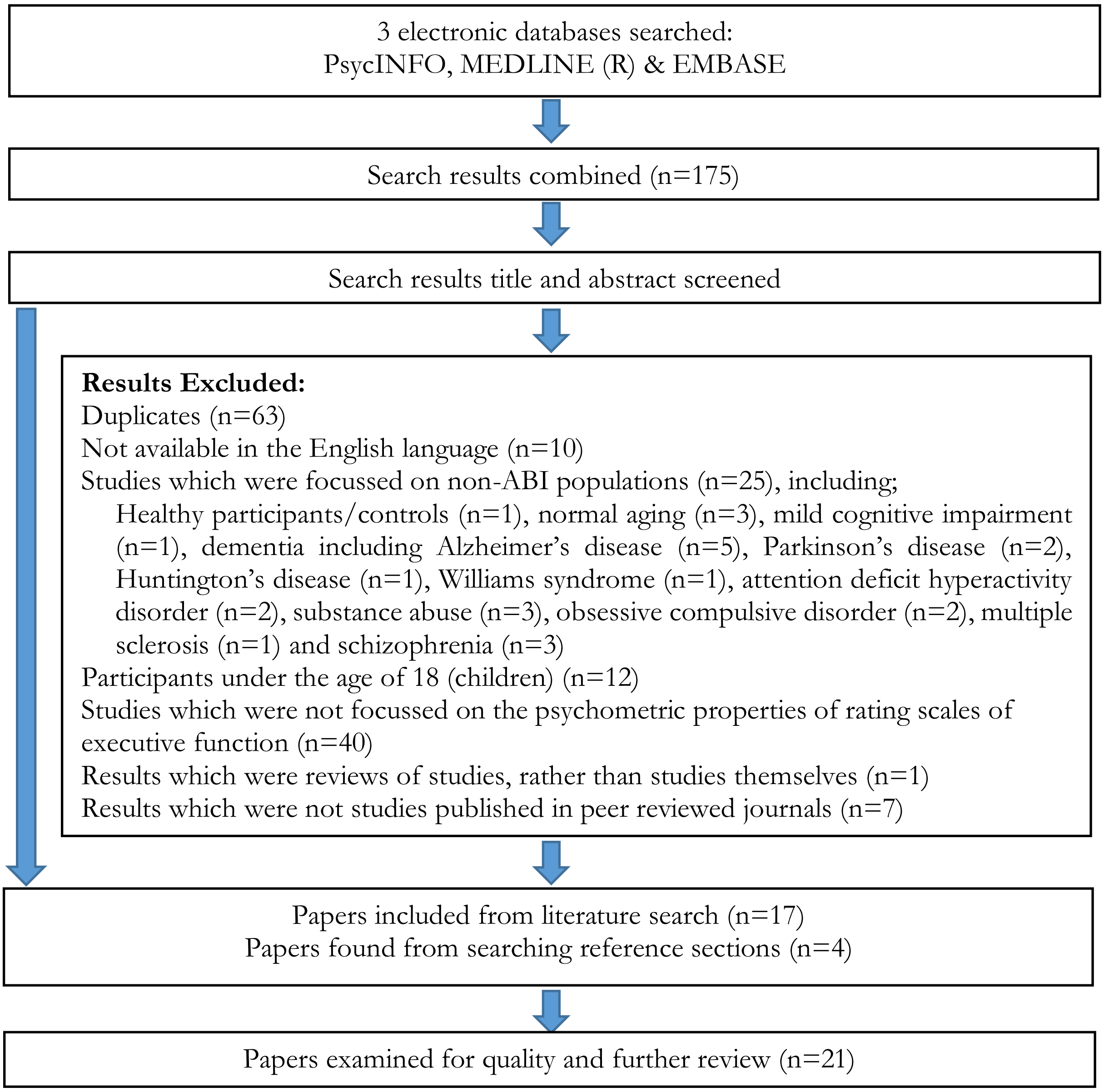
| | | |
|---|---|---|
| 2 | "factor analysis" or "validation" or "validity" or "reliability" | 194,478 |
| 4 | "Dysexecutive Questionnaire" | 211 |
| 5 | "Behaviour Rating Inventory of Executive Function" | 45 |
| 6 | "Frontal Systems Behaviour Scale" | 36 |
| 7 | "Frontal Lobe Personality Scale" | 25 |
| 8 | "Brock Adaptive Functioning Questionnaire" | 12 |
| 9 | "Childhood Executive Functioning Inventory" | 10 |
| 10 | "Iowa Rating Scales of Personality" | 9 |
| 11 | "Frontal Behaviour Inventory" | 2 |
| 12 | Search terms 4–11 combined with OR | 368 |
| 13 | Search terms 12 and 2 combined with AND | 175 |

## Inclusion & exclusion criteria

**Table 3 – Inclusion & Exclusion Criteria**

| **Inclusion Criteria** |
|---|
| 1. Results must be studies in peer reviewed journals. |
| 2. Results must be methodologically focussed on the psychometric properties of rating scales of executive function. |
| 3. Results must be studies focussed on a target population of people with acquired brain injury (ABI). |
| **Exclusion criteria** |
| 1. Results which were duplicates, due to several databases being searched |
| 2. Results that were not available in the English language. |
| 3. Results that referred to or used rating scales of executive function, but for which the methodological focus was not on acquired brain injury. Studies which included non-brain injured participants as a control group were not excluded. |
| 4. Results that used participants under the age of 18 years old (children). |
| 5. Results which did not meet inclusion criteria 2 |

**Figure 1 PRISMA Diagram**



3 electronic databases searched:
PsycINFO, MEDLINE (R) & EMBASE

Search results combined (n=175)

Search results title and abstract screened

**Results Excluded:**
Duplicates (n=63)
Not available in the English language (n=10)
Studies which were focussed on non-ABI populations (n=25), including;
Healthy participants/controls (n=1), normal aging (n=3), mild cognitive impairment (n=1), dementia including Alzheimer's disease (n=5), Parkinson's disease (n=2), Huntington's disease (n=1), Williams syndrome (n=1), attention deficit hyperactivity disorder (n=2), substance abuse (n=3), obsessive compulsive disorder (n=2), multiple sclerosis (n=1) and schizophrenia (n=3)
Participants under the age of 18 (children) (n=12)
Studies which were not focussed on the psychometric properties of rating scales of executive function (n=40)
Results which were reviews of studies, rather than studies themselves (n=1)
Results which were not studies published in peer reviewed journals (n=7)

Papers included from literature search (n=17)
Papers found from searching reference sections (n=4)

Papers examined for quality and further review (n=21)

## Review of methodological quality

Table 4 below provides a summary of methodological quality. Studies were assigned a quality score as a percentage, depending on the proportion of the observational study methodology described by von Elm et al., (2008) that they satisfy. These quality scores are then included in the discussion of the literature below, after the reference, e.g. Burgess et al., (1998, 88%).

## Table 4 – Review of methodological quality

| Reference | Measure(s) | Quality score (%) | Background & Objectives | Study Design | Participant eligibility | Sample Size (n > 50) | Factor analysis Sample size (n > 150 to 300) | Study Setting & Data Source | Quantitative variables | Statistical Methods | Potential Bias | Descriptive data | Main Results | Potential Limitations | Interpretation & |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Azouvi et al., (2015) | **DEX** | **83%** | 🟩 | 🟩 | 🟩 | 🟩 | ⬜ | 🟧 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟧 |
| Barker et al., (2011) | **DEX** | **75%** | 🟩 | 🟧 | 🟩 | 🟩 | ⬜ | 🟧 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟧 | 🟧 |
| Bennett et al., (2005) | **DEX** | **96%** | 🟩 | 🟩 | 🟩 | 🟩 | ⬜ | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 |
| Bodenburg & Dopslaff (2008) | **DEX** | **73%** | 🟧 | 🟩 | 🟩 | 🟩 | 🟧 | 🟧 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟧 |
| Boelen et al., (2009) | **DEX** | **83%** | 🟩 | 🟩 | 🟩 | 🟩 | ⬜ | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟥 | 🟧 |
| Bogod et al., (2003) | **DEX** | **71%** | 🟩 | 🟧 | 🟩 | 🟧 | ⬜ | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟥 | 🟥 |
| Burgess et al., (1998) | **DEX** | **88%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟧 | 🟩 |
| Carvalho et al., (2013) | **FrSBe** | **81%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟧 | 🟧 | 🟩 | 🟧 | 🟧 |
| Chan & Bode (2008) | **DEX** | **71%** | 🟧 | 🟩 | 🟩 | 🟩 | ⬜ | 🟩 | 🟩 | 🟩 | 🟧 | 🟧 | 🟩 | 🟥 | 🟥 |
| Chaytor & Smitter-Edgecombe (2007) | **DEX & BAFQ** | **69%** | 🟩 | 🟩 | 🟧 | 🟧 | 🟥 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟥 | 🟧 |
| Emmanouel et al., (2014) | **DEX** | **92%** | 🟩 | 🟩 | 🟩 | 🟩 | ⬜ | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 |
| Grace et al., (1999) | **FrSBe** | **88%** | 🟩 | 🟩 | 🟩 | 🟧 | ⬜ | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟩 | 🟧 |
| Karzmark et al., (2012) | **FrSBe** | **83%** | 🟧 | 🟩 | 🟩 | 🟩 | ⬜ | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟧 | 🟧 |
| McGuire et al., (2014) | **DEX** | **77%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟥 | 🟧 |
| Shaw et al., (2015) | **DEX** | **88%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟧 | 🟩 | 🟩 | 🟧 |
| Simblett & Bateman (2011) | **DEX** | **92%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟩 | 🟧 |
| Simblett et al., (2012) | **DEX** | **92%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟩 | 🟩 |
| Simpson & Smitter-Edgecombe (2002) | **BAFQ** | **73%** | 🟩 | 🟧 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟧 | 🟧 | 🟧 |
| Stout et al., (2003) | **FrSBe** | **85%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟥 | 🟧 |
| Waid-Ebbs et al., (2012) | **BRIEF** | **85%** | 🟩 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟩 | 🟧 | 🟩 | 🟩 | 🟧 | 🟧 |
| Yamasato et al., (2007) | **Novel & DEX** | **62%** | 🟧 | 🟧 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟥 | 🟥 |

The above journal articles were quality reviewed according to criteria discussed by von Elm et al., (2008) as part of their publication on Strengthening Reporting of Observational Studies in Epidemiology (STROBE). The STROBE statement consist of 22 criteria, not all of which are relevant for the current review. Accordingly, a smaller number of more relevant criteria have been selected for the current review. These criteria are described in appendix II. Studies were rated according to 12 criteria, or 13 in the case of studies that included factor analyses.

**GREEN** – The criterion was adequately discussed by the study. **AMBER** – The criterion was only briefly acknowledged or discussed in the study. **RED** – The criterion was not discussed or acknowledged at all in the study. GREY – The criterion was not applicable for this study (i.e. non-factor analytic studies).

## Discussion of quality scores

In total, twenty-one papers were scored for methodological quality, with scores ranging from 62% to 96%. Rather than discuss the quality of each paper in turn, table 5 below describes the aims, main findings and potential limitations of each paper. In general, studies scored lower if they only briefly discussed a study criteria (amber ratings) or failed to acknowledge or discuss a criteria at all (red ratings). In this regard, briefer papers, with a lower word count and hence less space to describe the study, typically scored lower than lengthier papers.

A number of papers had inadequate sample size for reported factor analyses, and some barely reached the n > 50 threshold recommended by von Elm et al., (2008). These included both studies which investigated the Brock Adaptive Functioning Questionnaire (BAFQ). Of particular note was the fact that very few papers adequately addressed potential sources of bias or limitation in either method or discussion sections, for example, the study by Yamasato et al., (2007, 62%) did not explore whether demographic factors influenced DEX scores, or the possibility of cultural validity of the Japanese version of the DEX in terms of interpretation and generalisability. Studies which explicitly listed inclusion and exclusion criteria, and explored demographic differences between participant-groups (if appropriate) scored better in this regard.

Several papers also used self-rating versions of a rating scale without discussion of limitations due to potential lack of insight, Simblett & Bateman (2011, 92% for example). However, a later study by the same research group (Simblett et al, 2012, 92%) acknowledged this limitation and conducted a study using proxy ratings. In terms of interpretation and generalisability, the many sources of potential limitation (whether discussed by the study authors or not) meant that very few papers scored an adequate (green) score in this area.

# Results

## Overview of study findings

### Table 5 – Overview of study findings

| **Authors, date publication, study location & quality score** | **Executive Measure(s) & Participants** | **Study Aims** | **Main Findings** | **Limitations of the study** |
|---|---|---|---|---|
| Azouvi et al., (2015)<br>*Neuropsychological Rehabilitation*<br>Patients who had experienced acute TBI who were admitted to hospitals in Paris & suburbs at 4 year follow up (2005-2009)<br>Quality score = **83%** | DEX-S<br>504 pts selected following acute TBI<br>257 survived acute stage<br>n= 147<br>At 4 year follow up | Ecological validity of the DEX-S to predict real life impairment.<br>To correlate DEX-S scores with measures of mood disorders (Hospital Anxiety & Depression Rating Scale; HADS), tests of cognitive impairment (Neuro-behavioural rating scale revised; NRS-R), dependency in elementary and extended activities of daily living (ADL), and non-return to work (Glasgow Outcome Scale Extended ; GOS-E).<br>To correlate DEX-S scores with pre-morbid demographic characteristics. | The DEX-S correlated significantly with self-reported mood disorders; (HADS-Anxiety $\rho$ = 0.61, HADS-Depression, $\rho$ = 0.55, HADS-total, $\rho$ = 0.65) and self-rated cognitive impairment (NRS-R, $\rho$ = 0.61).<br>It was also significantly negatively correlated with dependency in elementary ADL ($\rho$ = -0.23) and extended ADL ($\rho$ = -0.49) and non-return to work (GOS-E, $\rho$ = -0.49).<br>The DEX-S was significantly negatively correlated with years of education ($\rho$ = -0.28). | The DEX-S can lead to underestimation of difficulties, as people with TBI can sometimes be unaware of the extent of their difficulties. This is particularly true of people with more severe TBI who were included in the study. Results may have differed for comparisons between subgroups of good and poor self-awareness.<br>There were not enough participants to perform exploratory factor analysis. For this reason, the authors stated that a performed factor analysis was not reported. |
| Barker et al., (2011)<br>*Brain Injury*<br>Head injury charity in the UK.<br>Quality score = **75**% | DEX-S & DEX-I<br>n=60 mixed neurological<br>n=156 relatives or carers of neurological participants | To investigate the inter-rater reliability by comparing DEX-S ratings of a neurological population with DEX-I ratings by two or three carers or relatives of those people. | The DEX showed excellent internal consistency (DEX-S Cronbach's $\alpha$ = 0.92, DEX I $\alpha$ = 0.93-0.95)<br>Inter-rater reliability for 2 and 3 independent raters had an intra-class correlation of 0.47 and 0.52 respectively.<br>There were no significant overall differences between groups of raters (ANOVA). | The authors acknowledge that rater accuracy was not determined according to any criteria in the study. The authors suggest using cognitive tests of executive function in order to test the validity of DEX-I ratings. This could then be used to assess rater accuracy. |

| **Authors, date publication, study location & quality score** | **Executive Measure(s) & Participants** | **Study Aims** | **Main Findings** | **Limitations of the study** |
|---|---|---|---|---|
| Bennett et al., (2005)<br>*Journal of the International Neuropsychological Society*<br>Patient in acute neuropsychological rehabilitation centre in Australia<br>Quality score = **96%** | DEX-I<br>n = 64 participants with TBI (& family members and clinicians working with them) | To investigate sensitivity of the DEX-I to EF difficulties as measured by the Behavioural Assessment of the Dysexecutive Syndrome (BADS) test battery, in comparison with the sensitivity of an expanded 65 item version of the DEX-I (eDEX-I), created from a literature review of the most common EF impairments. | As hypothesized, occupational therapist & neuropsychologist rated DEX-I correlated more highly (and generally significantly) with individual and overall BADS scores than patient or family member rated DEX-I (which were without exception non-significant). The OT ratings were more highly correlated than the neuropsychologist's, suggesting they might be better raters of real life impairment. | The authors suggest that the relatively small sample size of pure TBI participants could in some ways limit the range of EF impairments, which might be quantitatively different in an ABI or other neurological population. Healthy controls could also have been used to increase sample size. |
| Bodenburg & Dopslaff (2008)<br>*Journal of Nervous & Mental Disease*<br>Unselected outpatients at neuropsychological practice in Germany.<br>Quality score = **73%** | DEX-S (German)<br>n=191 acquired brain injury | To investigate internal consistency and discriminative validity of DEX-S (German version).<br>To conduct exploratory varimax rotation principal components factor analysis.<br>To investigate test scores using quartile standards. | No DEX-S items were normally distributed. The distribution of skewed item scores was discussed and item 3 (confabulation) was removed from further analysis on this basis. All items were appropriately discriminating, except for item 11 which was removed from further analysis.<br>Internal consistency (Cronbach's α, r = 0.85).<br>Varimax rotation principle components analysis yielded 4 factors. | The authors discuss how the translation of the DEX-S into German could have reduced the validity of the test, as some items do not directly translate, or have semantic similarity to other items, which was not intended in the original English version. This, in turn, may have affected the extent to which individual items load in the derived factors, and hence any interpretation of those factors should be discussed with caution. |
| Boelen et al., (2009)<br>*Neuropsychological Rehabilitation*<br>ABI clients from seven outpatient clinics in the Netherlands<br>Quality score = **83%** | DEX-I<br>n = 81 ABI including 34 TBI<br>n = 57 healthy controls | To examine the sensitivity of the Behavioural Assessment of the Dysexecutive Syndrome (BADS) test battery, including the DEX-S & DEX-I in terms of sensitivity at correctly classifying ABI participants as having ABI, compared to healthy controls. | Both DEX-S and DEX-I ratings for ABI participants were significantly lower than DEX-S and DEX-I ratings for healthy controls. There were no significant DEX-S to DEX-I ratings within each group, although correlations were modest. DEX-S & DEX-I (relative) ratings correlated with other outcome measures. DEX ratings and neuropsychological tests together were more sensitive than either measure alone. | There was no discussion or acknowledgement of potential limitations of the study. The authors did discuss previous studies which had found proxy DEX ratings to have poor inter-rater reliability, but they did not calculate inter-rater reliability in their study. |

| **Authors, date publication, study location & quality score** | **Executive Measure(s) & Participants** | **Study Aims** | **Main Findings** | **Limitations of the study** |
|---|---|---|---|---|
| Bogod et al., (2003)<br>*Journal of the International Neuropsychological Society*<br>Rehabilitation programmes & local community, British Columbia, Canada.<br>Quality score = **71%** | DEX-I & DEX-S<br>n = 45 TBI<br>(n=40 from rehabilitation programmes & n = 5 from local community) | Criterion validity of DEX-S & DEX-I discrepancy as a measure of self-awareness following acquired brain injury.<br>DEX-S & DEX-I discrepancy was compared to Self-Awareness of Deficits Interview (SADI) and tests of IQ and EF as comparison criteria. | A significant but modest correlation between DEX-S & DEX-I discrepancy and SADI scores was found (r = 0.40).<br>There were no statistically significant correlations between discrepancy scores and tests of EF, except for a modest correlation between DEX-I and go-no go test score (r = 0.27).<br>There was a modest significant negative correlation between discrepancy scores and IQ scores (r = -0.33). | Potential limitations of this study were not specifically discussed by the authors. This is reflected in the quality score.<br>The most obvious limiting factor to the present reviewer is the small sample size, relative to the number of statistical comparisons. This increases the probability of false positive error and limits the interpretability and generalisability of this study. This is also reflected in the quality score. |
| Burgess et al., (1998)<br>*Journal of the International Neuropsychological Society*<br>ABI & Dementia participants from 4 UK neurological centres<br>Quality score = **88%** | DEX-I & DEX-S<br>n = 308<br>Of which mixed aetiology neurological = 92 and healthy controls =216 | Exploratory factor analysis of the DEX into sub-factors, which are then used as criteria to assess the ecological validity of neuropsychological test of EF by correlating these tests and other neuropsychological test with the derived sub-factors. | Control participants rated themselves higher on the DEX than others rated them, whereas neurological participants rated themselves lower than others.<br>Factor analysis of the DEX-I produced a five factor solution, and significant correlations with various neuropsychological tests were reported for several factors, as well as DEX-I but not DEX-S total scores. | The authors discuss the possibility that test and behavioural item sensitivity could be responsible for the correlations. By contrasting the correlations of EF test & DEX-I sub-factors with those of non-EF test and DEX-I sub-factors they suggest the tests have some ecological validity.<br>There was no further discussion of potential limitations, such as the mixed group of participants. |
| Carvalo et al., (2013)<br>*Assessment*<br>Adult relatives of people with mixed neurological conditions from a neuropsychological clinic in the US<br>Quality score = **81%** | FrSBe<br>n = 494 | To conduct a confirmatory factor analysis of the FrSBe to fit the originally proposed factor structure for the rating-scale.<br>To explore internal consistency of the FrSBe. | A number of factor analytic models were produced and reported. The original three factor model of apathy, disinhibition and Executive function had a better fit with the data than a single factor "frontal" model, within which the three factor model was nested.<br>A reduced FrSBe removed 8 items of the original 46, and better fit the original three factor model. Internal consistency of the original model was report as Cronbach's $\alpha$ = 0.95, and for the reduced model as Cronbach's $\alpha$ = 0.93. | The study authors acknowledge that participants rating relatives with Alzheimer's, Parkinson's and mild cognitive impairment may be over represented in the study, whilst participants rating brain injured relatives may be under represented. Demographic data by reported participant condition is not reported.<br>The authors acknowledge that two of them were involved in the original validation and design of the rating scale. |

| **Authors, date publication, study location & quality score** | **Executive Measure(s) & Participants** | **Study Aims** | **Main Findings** | **Limitations of the study** |
|---|---|---|---|---|
| Chan & Bode (2008)<br>*Journal of neurology, neurosurgery & psychiatry*<br>TBI patients and informants at two regional hospitals in Hong Kong<br>Quality score = **71%** | DEX-S & DEX-I<br>n = 92 | To explore inter-rater reliability of the DEX by comparing patient (self) and proxy (other) ratings, using Rasch analysis.<br>To explore the difficulty with lack of insight in a TBI sample. | Inter-rater reliability was in the modest range (intra-class correlation, r = 0.46). Mean ratings of dysexecutive symptoms were similar for the self-ratings (mean = 30.12) and the proxy-ratings (mean = 31.32).<br>However, five of the twenty DEX items showed differential item functioning, i.e. they were consistently reported at different levels of frequency for self and proxy ratings, suggesting that these item ratings cannot be used interchangeably. | There was no discussion of potential limitations to the study by the study authors, although this is perhaps due to the brevity of the study report (three pages in total).<br>A potential limitation in the view of the present reviewer is the interpretability and generalisability of the findings to an English speaking population, given that the version used was a Hong Kong Cantonese version of the DEX. |
| Chaytor & Smitter-Edgecombe (2007)<br>*Brain Injury*<br>Outpatients referred for neuropsychological assessment at US hospital and their relatives or friends<br>Quality score = **69%** | DEX & BAFQ<br>n = 46 mixed neurological participants and their relatives or friends as informants | To determine if EF factors derived from the DEX using exploratory factor analysis can also be derived from the Brock Adaptive Functioning Questionnaire (BAFQ). | There was a strong significant positive correlation between the two measures (r = 0.84)<br>Principal components analysis with varimax rotation found 5 factors for the DEX, with modest to good internal reliability (α = 0.54 to 0.84). Four factors were extracted from the BAFQ with similar internal reliability (α = 0.65 to 0.80). Correlations between DEX and BAFQ factors are reported ranging from r = 0.26 to r = 0.50. | The study utilised a small sample (n = 46) of which only 12 had TBI. This limits the interpretability of the factor analysis in particular. The authors acknowledge these limitations.<br>Furthermore, although individual factors between the two measures correlated with one another, they did so at a level below which individual items would not be considered to be part of a factor using the chosen methodology (r < 0.5). The authors do not discuss this limitation. |
| Emmanouel et al., (2014)<br>*Brain Injury*<br>Clients and therapists at two rehabilitation programme sites in Greece<br>Quality score = **92%** | DEX-S & DEX-I<br>n = 81 ABI<br>(30 anterior lesions vs 22 posterior lesions vs 29 healthy controls) | To investigate the validity of the DEX-S completed by clients themselves and DEX-I completed by therapists at identifying severity of symptoms varying by lesion location (anterior lesions - AL vs. posterior lesions - PL)<br>To examine the strength of association between the DEX-S & DEX-I compared to a range of EF neuropsychological tests. | DEX-S scores were significantly lower for the ABI vs healthy controls, but were not significantly different for AL vs PL groups.<br>DEX-I showed significant differences between all groups, suggesting greater validity at measuring AL impairment than the DEX-S. In addition, EF tests correlated significantly with the DEX-I but not the DEX-S. | Despite having been specifically designed to avoid potential biases and limitations in previous studies, there was no discussion of potential limitations by the authors in this study.<br>Potential limitations in the consideration of the current reviewer include a relatively small (n = 52) mixed aetiology sample, and the fact that healthy controls were rated by family members rather than therapists. |

| Authors, date publication, study location & quality score | Executive Measure(s) & Participants | Study Aims | Main Findings | Limitations of the study |
|---|---|---|---|---|
| Grace et al., (1999)<br>*Assessment*<br>Mixed aetiology ABI participants from rehabilitation and outpatient hospitals in the US<br>Quality score = **83%** | FrSBe<br>n = 39 ABI<br>(24 anterior lesion vs 15 posterior lesion)<br>n = 48 healthy controls<br>n = 87 relatives | To investigate the internal consistency of the FrSBe and to investigate validity at detecting pre and post injury change in anterior lesion (AL) clients, distinguishing AL clients from posterior lesion (PL) clients, and AL clients from healthy controls. | Internal consistency was good (Cronbach's α = 0.94) with most individual items correlating highly. Three items correlated in the 0.18 to 0.20 range.<br>Pre and post ratings were significantly different for AL group but not for the PL group. An optimal cut-off score was derived with 96% sensitivity and 81% specificity | Some effort was made to control for potential bias, noting that age and education did not correlate significantly with FrSBe scores for healthy controls.<br>Amongst ABI participants, education was negatively correlated with pre-injury scores, and males had significantly higher pre-injury scores than females. Sub group were heterogeneous and size was perhaps small. |
| Karzmark et al., (2012)<br>*Applied Neuropsychology*<br>Patients receiving neuropsychological assessment at University medical centre in US<br>Quality score = **83%** | FrSBe<br>n = 100<br>(of which 12 were TBI and a further 50 were various ABI) | To investigate a battery of neuropsychological tests of EF, the Wechsler Adult Intelligence Scale III and the FrSBe in predicting real life impairment in activities of daily living, as measured by the Functional Activities Questionnaire (FAQ). FrSBe and FAQ scores were given by family members of the ABI participants. | Various stepwise multiple regressions were performed, entering an executive index (EI) derived from the tests of EF, and the FrSBe scores, either in total or as sub-scales.<br>One regression explained 44% of FAQ variance (30% EI, then +14% FrSBe total). A further regression found that FrSBe executive dysfunction predicted 33% of the variance, with EI adding +14%. | There was no rationale discussed for the order in which items were entered into multiple regression, and regressions including FSIQ, although discussed, were not reported.<br>The authors did acknowledge that a mixed aetiology sample of participants limits generalisability, but in terms of dementia populations rather than pure TBI or ABI. |
| McGuire et al., (2014)<br>*Frontiers in Behavioural Neuroscience*<br>Brain injury services in the UK and Ireland<br>Quality score = **77%** | DEX-S & DEX-I<br>n = 113<br>(113 ABI, 101 family members and 64 clinician ratings) | To explore the inter-rater reliability of the DEX-S completed by patients, and the DEX-I completed by family members and clinicians, using intra-class correlation coefficients (ICCs)<br>To principal axis factor analysis (PAF) to determine if factor structure varies between the above groups. | Average ICCs for between group agreement was poor (self-clinician = 0.15, self-family = 0.41 and clinician-family = 0.31)<br>PAF found one-factor solutions for each group. | There is no discussion of potential limitations of the study by the authors, other than to suggest that larger sample sizes are needed for future studies.<br>The use of PAF makes direct comparison with studies that utilise other forms of factor analysis difficult. |

| Authors, date publication, study location & quality score | Executive Measure(s) & Participants | Study Aims | Main Findings | Limitations of the study |
|---|---|---|---|---|
| Shaw et al., (2015)<br>*Psychological Assessment*<br>Neurological, psychiatric outpatient and community controls from Australian hospitals<br>Quality score = **88%** | DEX-S<br>n = 997<br>[neurological impairment (n = 120) Psychiatric (n = 212) Community sample (n = 663)] | To explore the validity and internal reliability of the DEX-S in a mixed sample. Exploratory factor analysis of DEX-S, from which a revised questionnaire (DEX-R-S) was derived. Confirmatory factor analysis of the DEX-R-S with new factor structure and previous factor models.<br>Internal consistency, discriminant validity, sensitivity & specificity. | From exploratory factor analysis, a 15 item revised questionnaire (DEX-R-S) was derived.<br>Confirmatory factor analysis suggests a good fit with theoretical models of EF for the DEX-R-S. Confirmatory FA with previous factor models suggested they were less effective than the new factor model.<br>Internal consistency reported as 0.85 (Cronbach's alpha) | DEX-R had insufficient discriminant validity and specificity for neurologically impaired group, incorrectly classifying them as the community group.<br>Participants with high EF deficits also reported high symptoms of depression, anxiety and stress. The direction of causation is not clear and hence these symptoms could be confounding EF measurement in the study.<br>Self-report and medication are other potential limiting factors. |
| Simblett & Bateman (2011)<br>*Neuropsychological Rehabilitation*<br>Patients at a UK neuro rehabilitation centre from 1999-2011.<br>Quality score = **92%** | DEX-S<br>n = 363<br>Of which TBI = 248 And non-TBI = 103 Unknown cause = 12 | Investigation of factor structure of DEX using factor analysis and Rasch analysis.<br>Rasch analysis to ensure that DEX-S meets psychometric assumptions of Rasch model (stricter than classical test theory) and Principal Components factor analysis (PCA) to investigate factor structure. | Internal consistency reliability, measured by person separation index (PSI) = 0.92.<br>The DEX-S did not meet assumptions of uni-dimensionality and several items were not responded to in a consistent manner. The DEX-S did not fit the Rasch model after removal of these. PCA on data corrected to fit Rasch model found 3 factors and removed 3 items. | Despite having the most rigorous psychometric methodology, this study relied on self-report (DEX-S) measures. The authors acknowledge that this may have limited the validity of the study due to insight difficulties in some participants, but suggest that the removal of outliers may have reduced this limitation. |
| Simblett et al., (2012)<br>*Neuropsychological rehabilitation*<br>Carers, relatives, friends or case managers of people with ABI at a UK neuropsychological rehabilitation centre<br>Quality score = **92%** | DEX-I<br>n = 271<br>Independent ratings of people with TBI = 181 or non-TBI = 84 or unknown aetiology = 6 | To measure construct validity and reliability of DEX-I using Rasch analysis (RA).<br>It was hypothesized that the factor structure derived by Simblett & Bateman (2011) would have adequate psychometric properties and internal consistency using RA. | Despite having a good PSI (0.91) the DEX-I did not meet the other RA assumptions for uni-dimensionality.<br>Two of the three factors identified in the Simblett & Bateman (2011) study demonstrated an acceptable level of fit to the RA model. Following item rescoring and removal, all three factors from the previous study satisfied the RA model. | The authors acknowledge that the nature of the relationship of the independent rater and the person experiencing ABI could limit the validity of the study. In addition to the nature of the relationship and how well the independent rater knows the person with ABI, the experience and exposure of the rater to the full range of EF difficulties could be important in terms of rating difficulties. |

| **Authors, date publication, study location & quality score** | **Executive Measure(s) & Participants** | **Study Aims** | **Main Findings** | **Limitations of the study** |
|---|---|---|---|---|
| Simpson & Smitter-Edgecombe (2002)<br>*Brain Injury*<br>Relatives of people with TBI attending brain injury support groups in Washington & Oregon, USA.<br>Quality score = **73%** | BAFQ<br>n = 61 TBI | Using discriminant function analysis to determine the predictive validity of the BAFQ to predict post injury employment status.<br>To use factor analysis to determine if different sub-factors of the BAFQ contribute to the prediction of employment status differentially. | Factor analysis of the BAFQ produced a two factor solution, which the authors named an orbitofrontal factor and a dorsolateral factor, due to items describing difficulties in line with theoretical models of neuropsychological deficits associated with damage to these areas.<br>Discriminant function analysis used these two factors, with background information, to produce a 77.4% accuracy rate at predicting employment. | The authors of the study acknowledge a limited return rate of 22% which could have biased their sample. They also acknowledge the sample size was small as a result of this. The small sample size particularly limits the validity of interpretation of the factor analysis. |
| Stout et al., (2003)<br>*Assessment*<br>Mixed aetiology neurological patients at four research Universities in the US<br>Quality score = **85%** | FrSBe<br>n = 324<br>Mixed aetiology neurological patients (of which 29 were ABI) | To explore whether the factor structure of the FrSBe supports the subscale structure of the measure, using exploratory principle component factor analysis with orthogonal rotation. | A solution with three factors was derived, which broadly agreed with the three subscales of the FrSBe. These three factors each contained a number of items from the original subscales.<br>In addition, each factor correlated significantly with the others, although more weakly than with individual items (r = 0.22 to 0.43) | There was some discussion of using heterogeneous participants, in order to get a wide range of frontal system behaviours, but no discussion of any potential bias in such a varied sample.<br>There was no discussion of possible limitations by the study authors. |
| Waid-Ebbs et al., (2012)<br>*Brain Injury*<br>Inpatients and outpatients and their relatives at hospitals in the US<br>Quality score = **85%** | BRIEF-S & BRIEF-I<br>n = 90 people with TBI.<br>n = 89 relatives of people with TBI. | To investigate whether the two BRIEF indices, Behavioural Regulation Index (BRI) and Metacognitive Index (MI) are unidimensional constructs.<br>To examine item level psychometric properties using Rasch analysis.<br>To determine whether psychometric properties of the BRIEF in a TBI sample warrant its use as a tool with this population. | Informant rating item reliability and person reliability of the BRI were 0.85 & 0.93 and for the MI were 0.86 & 0.94 respectively.<br>Cronbach's alpha of BRI and MI were 0.94 and 0.96.<br>Confirmatory factor analysis generally confirmed the existing factor structure, with 4 items not fitting (load < 0.3)<br>Some items relating to working memory had high item difficulty. | The confirmatory factor analysis, whilst well conducted, used a relatively small sample size, which limits the confidence of the conclusions somewhat.<br>The authors themselves acknowledge this however, and also note that as some severity of TBI data was missing, and there were very few mild TBI cases, that this limits the generalisability of the study to moderate to severe TBI. |

| **Authors, date publication, study location & quality score** | **Executive Measure(s) & Participants** | **Study Aims** | **Main Findings** | **Limitations of the study** |
|---|---|---|---|---|
| Yamasato et al., (2007)<br>*Journal of Psychiatry & Clinical Neurosciences*<br>Relatives of outpatients with TBI visiting a Japanese hospital<br>Quality score = **62%** | Novel measure created and compared to Japanese translation of DEX-I (DEX-I-J)<br>n = 72 | To investigate the validity, reliability and factor structure of a novel questionnaire, designed to measure neurobehavioural disability and personality change following TBI, as reported by relatives of people with TBI. | Reliability was assessed using random split half method for items (r = 0.90) and Spearman's Rho (ρ = 0.95).<br>Validity used the DEX-I-J and Neuropsychiatric Inventory Japanese version (NPI-J). Correlations were r = 0.36 and r = 0.37 respectively. | Exploratory factor analysis was carried out, but the sample size of n = 72 was not sufficiently large for this to have much validity, especially given the number of factors derived (6). Items were YES/NO rather than a Likert scale.<br>There was no discussion by the authors of these or any other potential limitations or potential bias in the study. |

### Demographics of the papers

The twenty-one studies reviewed recruited participants from the US (7), UK (5), Australia (2), Canada (1), France (1), Germany (1), The Netherlands (1), Greece (1), Hong Kong (1), & Japan (1). Total sample sizes of participant populations are displayed in table 5 below:

**Table 6 – Participant population demographics**

| **Study** | **Neurologically Impaired Clients** | **Healthy Controls** | **Independent ratings** |
|---|---|---|---|
| Azouvi et al., (2015) | n = 147 TBI | | |
| Barker et al., (2011) | n = 60 mixed neuro. | | n = 153 relatives |
| Bennett et al., (2005) | n = 64 TBI | | n = 42 relatives |
| Bodenburg & Dopslaff (2008) | n = 191 ABI | | |
| Boelen et al., (2009) | n = 34 TBI<br>n = 47 other ABI | n = 57 | |
| Bogod et al., (2005) | n = 45 TBI | | |
| Burgess et al., (1998) | n = 92 mixed neuro. | n = 216 | |
| Carvalho et al., (2013) | | | n = 494 relatives |
| Chan & Bode (2008) | n = 92 TBI | | n = 92 relatives |
| Chaytor & Smitter-Edgecombe (2007) | n = 46 mixed neuro. | | n = 46 relatives |
| Emmanouel et al., (2014) | n = 81 ABI | | |
| Grace et al., (1999) | n = 39 ABI | n = 48 | n = 87 relatives |
| Karzmark et al., (2012) | n = 100 mixed neuro. | | |
| McGuire et al., (2014) | n = 113 ABI | | n = 101 relatives<br>n = 64 clinicians |
| Shaw et al., (2015) | n = 120 mixed neuro. | n = 663 | |
| Simblett et al., (2012) | | | n = 271 relatives |
| Simblett & Bateman (2011) | n = 248 TBI<br>n = 115 mixed neuro. | | |
| Simpson & Smitter-Edgecombe (2002) | n = 61 TBI | | n = 33 relatives |
| Stout et al., (2003) | | | n = 324 relatives |
| Waid-Ebbs et al., (2012) | n = 90 TBI | | n = 89 relatives |
| Yamasato et al., (2007) | | | n = 72 relatives |
| **Totals** | **Mixed Neurological** | **Healthy Controls** | **Independent ratings** |
| All Mixed Neurological | n = 1,785 | n = 984 | n = 1,804 relatives |
| …of which ABI | n = 1,252 | | n = 64 clinicians |
| …of which TBI | n = 781 | | |

Four existing rating scales of EF were identified in the present review. Fourteen studies focussed on the Dysexecutive Questionnaire (DEX, 20 items), four on the Frontal Systems Behaviour rating scale (FrSBe, 46 items), two on the Brock Adaptive Functioning Questionnaire (BAFQ, 64 items) and one on the Behaviour Rating Inventory of Executive Function (BRIEF, 86 items). In addition, one paper used the DEX as a criterion to measure the validity of a novel measure, and one of the studies above compared the DEX & BAFQ. Clearly, the majority of studies in the present review investigate the DEX, although the proportion of the studies in the present review reflects the relative frequency of DEX-based publications in the literature, as shown in table 2 above (i.e. 211 out of 368 publications).

In terms of psychometrics, six studies investigated content validity, seven investigated criterion validity by correlating the scales with neuropsychological tests or other rating-scales, and three investigated ecological validity, by correlating rating with other real life outcome measures. Eleven studies performed factor analyses to explore and/or confirm construct structural validity, and eleven reported internal consistency statistics, with all four rating scales being represented. Three investigated inter-rater reliability, though only for the DEX. Only one study stated an *a priori* hypothesis, and none investigated longitudinal validity. There was little discussion of cultural validity, even in papers which translated rating-scales into a language other than English. Two studies conducted sensitivity and specificity analysis, and discussed the interpretability of their findings. These psychometric properties, as defined by the Terwee et al., (2007) and Mokking et al., (2010) criteria, are discussed in detail in the next section.

## Discussion

In this section, the measures in the present review are discussed in relation to the rating scale psychometric quality criteria proposed by Mokking et al., (2010) and Terwee et al., (2007), described in detail in Appendix I. The purpose of the discussion will be to determine to what extent existing measures satisfy the quality criteria, and to determine any criteria which are currently not addressed in the literature. This can inform the selection of existing rating scales clinically, and the improvement of existing tests and design of future rating scales from a research standpoint. Rating scales will be discussed in order of their frequency in the reviewed literature, i.e. DEX, FrSBe, BRIEF and BAFQ, with the methodological quality ratings from table 4 cited after the publication year.

## Content Validity

In this context, content validity refers to a clear definition of EF, the reason why its measurement is important, the target population for whom the measurement applies, and justification for inclusion or exclusion of measure items.

The DEX was developed by Wilson et al., (1996) based on a review of common neuropsychological sequelae following frontal lobe brain injury by Stuss & Benson (1984). This emphasises the links between the concepts of frontal lobe neuropsychological function and executive function in defining the concept, and hence the content validity of the DEX. The full DEX consists of 20 behavioural items, and uses a 0-4 Likert Scale for frequency of occurrence, with self and independent-rating versions. In the present review, some studies recommend reducing the number of DEX items. The study by Shaw et al. (2015, 88%) used exploratory factor analysis (FA) and removed 5 items, producing a revised version (DEX-R), and then used confirmatory FA with the second half of their sample, confirming that this model fit the data well. Studies by Simblett & Bateman (2011, 92%) and Simblett et al., (2012, 92%) used Rasch analysis and removed 3 items, with other items being rescored such that the data better fit the model structure. In contrast, the study by Bennett et al. (2005, 96%) used an expanded 65 item e-DEX, developed from a literature review of EF difficulties, conducted by Banich (1997).

The FrSBe was developed by Grace, Stout & Malloy (1999, 83%), based on a observable frontal-behavioural syndromes, itself based on a theoretical model of three frontal-subcortical circuits, later described by Chow & Cummings (2007). Again, this emphasises the link between frontal lobe function and behavioural difficulties that occur following damage or disease to this area of the brain. The FrSBE, originally named the Frontal Lobe Personality Scale (FLOPS), was deliberately designed as a behavioural rating scale for sequelae of frontal lobe injury. The FrSBe consists of 46 behavioural items, and uses a 1-5 Likert scale for frequency of occurrence, with self, family and clinician-rating versions. Unlike the DEX, the FrSBE also asks respondents to rate each item for estimated pre-injury frequency, in order to give a retrospective measure of behavioural change. Also, the FrSBe groups items into three separate subscales, called *Apathy*, *Disinhibition* and *Executive Dysfunction*. In the present review studies by Stout et al. (2003, 85%) and Carvalho et al. (2013, 81%) each used factor analysis and suggest removing items in order that the data better fit the originally proposed structure.

The BAFQ was developed by Dywan & Segalowitz (1996) to assess behavioural functioning in adult brain injured populations following injury to the frontal lobes. The items were developed from qualitative interviews conducted with people who have experienced traumatic brain injury and their family members. Originally intended to measure five separate factors of executive functioning, the Dywan & Segalowitz (1996) study reduced this to two factors, relating to orbitofrontal and dorsolateral frontal lobe function. The scale consists of 68 behavioural items, using a 5 point Likert scale for frequency of occurrence. There are self and independent-rater versions. In the present review, no studies suggest changing or removing any items.

The BRIEF-A (adult version, hereafter BRIEF) was developed based on a child self-report version (the original BRIEF). The original BRIEF was developed to measure EF in children, primarily for use in assessment of attention deficit disorders (Gioia, Isquith, Guy, & Kenworthy, 2000; Isquith, Roth, & Gioia, 2013). The scale consists of 75 items, using a 3 point Likert scale for frequency. In the present review, the study by Waid-Ebbs et al., (2012, 85%) used Rasch analysis to validate the BRIEF-A for use with a TBI population. This analysis recommended removing several items due to low loadings on derived factors, and item difficulty.

In summary, the DEX, FrSBe and BAFQ were designed with for an ABI population, whereas the BRIEF has only recently been validated for this population. The first three scales were also designed with reference to frontal lobe brain injury, whereas the BRIEF was focussed on EF in children with attention deficit disorders. Some studies have proposed removing items from the DEX or FrSBE to increase structural validity. Therefore we might have greatest confidence in the content validity of the revised DEX and FrSBe measures for a TBI population.

**Criterion Validity**

In this context, criterion validity can refer to the extent to which rating scales of EF correlate with other measures of EF, either rating scales or neuropsychological tests, sometimes referred to as convergent validity. Another criterion is the extent to which a rating scale is predictive of real life impairment (e.g. occupational status post ABI), known as ecological validity.

The study by Boelen et al., (2009) discusses the problem of validating neuropsychological measures. The authors suggest that there is really no “gold standard” criterion to validate neuropsychological measures against because any quantitative criterion which we could select to validate a measure by could itself have its validity questioned. It is perhaps for this reason that Burgess et al., (1998, 88%) assume the DEX is a valid criterion for judging ecological validity of tests of EF, whilst other researchers (discussed below) assume the neuropsychological tests are more valid, and use them as criteria to judge the validity of the DEX.

**Convergent validity**

The study by Burgess et al. (1998, 88%) discusses how EF tests and EF rating scales may correlate mainly as a result of both measures being sensitive to multiple aspects of EF. For this reason, their study used factor analysis, and noting distinct correlations between certain EF tests and DEX factors. The study by Bennett et al. (2005, 96%), correlated EF tests from the Behavioural Assessment of the Dysexecutive Syndrome (BADS) with the DEX, finding that occupational therapist (OT) DEX-I ratings had stronger correlations with BADS test scores than family member DEX-I ratings (or clinical psychologist DEX-I ratings), suggesting that OTs might be the most valid independent raters. Emmanouel et al. (2014, 92%) also correlated DEX-S & DEX-I ratings with neuropsychological tests, finding that therapist DEX-I ratings but not DEX-S ratings correlated significantly with EF tests. Finally, the study by Bogod et al. (2005, 71%) investigated the DEX-I to DEX-S discrepancy score, generally considered to indicate a lack of insight on behalf of the ABI person. They found a modest positive correlation with the self-awareness of deficits (SADI) measure (r = 0.40). Taken together, these results suggest greater validity for clinician rated DEX-I scores than for DEX-S scores.

No studies in the present review investigated convergent validity of the FrSBe or the BRIEF, but the study by Chaytor & Smitter Edgecombe (2007, 69%) found a strong positive correlation between the DEX and BAFQ (r = 0.84). This provides some evidence that they measure the same construct, although correlations between derived factors for the two measures were typically lower (r=0.24 to 0.50). Also, the small sample size (n=46) of this study limits the statistical power of this factor structure evidence.

Finally, the study by Yamasato et al., (2007, 62%) used the DEX as a validity criterion to develop novel measure of neurobehavioural disability, finding a weak to moderate correlation (r = 0.36) between the two measures. The small sample size (n = 72) limited the validity of the factor analysis in this study, and arguably the concepts of EF and neurobehavioural disability are related but distinct constructs in some regards.

**Ecological validity**

The study by Azouvi et al. (2015, 83%) investigated the power of the DEX-S to predict real life impairment. This study found positive correlations between the DEX and measures of anxiety (ρ = 0.61), depression (ρ = 0.55), limitations on activities of daily living (ρ = 0.49), and limited return to work (ρ = 0.49). The study by Simpson & Smitter-Edgecombe (2002, 73%), used the BAFQ and other demographic and brain injury factors to predict occupational status post injury. In this study, a derived orbitofrontal factor correlated at r = 0.527 with the second of two discriminant functions. These two functions together correctly predicted employment status post injury with a 77.4% accuracy rate.

The study by Karzmark et al. (2012, 83%) investigated the predictive power of tests of EF and the FrSBe to predict real life functional impairment as measured by the Functional Activities Questionnaire (FAQ). Multiple regressions using an executive index (EI) of neuropsychological tests of EF and the FrSBe total score entered stepwise explained between 44-47% of the variance in FAQ scores. Taken together, these studies provide some evidence that EF difficulties as measured by these tests can indeed be considered predictive of real life impairment, suggesting moderate ecological validity.

Considered all together, the generally moderate strength of these correlational studies suggests that the EF rating scale scores are associated with some, but not all of the variance in the criteria used to assess their validity. This suggests reasonable convergent and ecological validity.

### Inter-rater reliability

Inter-rater reliability (IRR) in this context is the extent to which two or more groups of independent ratings on an EF rating scale correlate with one another. High IRR would indicate that independent raters are capable of a high level of agreement when it comes to the assessment of a person's EF difficulties. This is important, because poor IRR could potentially obscure other areas of validity, by making it hard to distinguish them from issues related to IRR. Relatively few studies reported inter-rater reliability in the present review, with those that did being summarised in table 7 below.

**Table 7 – Studies which report inter-rater reliability statistics**

| **Study, quality rating, rating scale & raters** | **n =** | **Intra-class correlation**[1] |
|---|---|---|
| **Barker et al., (2011, 75%)**<br>DEX-I (between 2 relatives)<br>DEX-I (between 3 relatives) | <br>n = 60<br>n = 36 | Intra-class correlation, r =<br>0.47<br>0.52 |
| **Chan & Bode (2008, 71%)**<br>DEX-S & DEX-I (patient-relative) | <br>n = 92 | Intra-class correlation, r =<br>0.46 |
| **McGuire et al., (2014, 77%)**<br>DEX-S – DEX-I (patient-clinician)<br>DEX-S – DEX-I (patient-family)<br>DEX-S – DEX-I (clinician-family) | n = 113<br>n = 64<br>n = 101<br>n = 64 | Intra-class correlation, r =<br>0.15<br>0.41<br>0.31 |

A complicating factor with regard to IRR is the fact that some people with ABI giving self-rated scores on EF difficulties may also have difficulties with a lack of insight into their difficulties. Indeed, lack of insight can be a consequence of ABI, and particularly frontal lobe injury, which may involve areas thought to be critically involved in metacognitive processes, such as the frontal pole (Stuss, 2007; Stuss, Picton, & Alexander, 2001). It is for this reason that the original authors of the DEX proposed that discrepancy between self-rated and independently rated scores could

[1] The intra-class correlation measures the level of agreement between independent raters or groups of raters, where the measurements are continuous and parametric.

be due to a lack of insight into one's difficulties on the part of the ABI person (Wilson et al., 1996). Indeed, the study by Burgess et al., (1998, 88%) reports a consistent and statistically significant tendency for non-brain injured participants to report more EF difficulties compared to proxy ratings, whereas brain injured participants tend to support significantly less EF difficulties compared to proxy ratings. The discrepancy between DEX-S & DEX-I ratings for ABI persons has been proposed as one method to quantify a lack of insight into one's difficulties (Wilson et al., 1996). For this reason, the study by Bogod et al., (2003) correlated the DEX-S-DEX-I discrepancy with the Self Awareness of Deficits Interview, finding a modest positive correlation (r = 0.40).

Given that a lack of insight can complicate investigation into IRR at least for some ABI participants, studies which investigate multiple independent ratings of EF should be considered. From table 6 above, we can see that the study by McGuire et al. (2014, 77%) gives some tentative support to the idea that ABI people and their relatives have a greater amount of agreement in terms of IRR than clinicians do with either ABI persons, or independent ratings by their relatives. This could be because ABI people and their relatives know each other better than clinicians, or alternatively that family members may have a shared tendency to over or under-report EF difficulties, relative to clinicians, whom we would hope would be more experienced and objective in their understanding of EF. The study by Bennett et al. (2005, 96%) acknowledges this, and uses independent clinician ratings by clinical neuropsychologists and occupational therapists (OTs) compared to those of relatives. In this study, the DEX-I ratings were correlated with tests of EF, which were significant for therapist ratings only, suggesting that therapist DEX-I ratings may have greater validity. Unfortunately, this study does not report clinician-clinician IRR, so we cannot confidently conclude that it will be any better. No studies in the present review report IRR for the other measures, so no conclusions can be drawn as to their IRR.

**Internal Consistency**

Sometimes referred to as internal validity, or internal consistency reliability, this is a measure of the extent to which items within a rating scale or subscale are correlated or homogenous (Cronbach, 1951). Studies reporting an internal consistency statistic, and are listed below:

**Table 8 – Studies which report internal consistency statistics**

| **Study, quality rating, rating scale & raters** | **Sample size** | **Internal consistency statistic** |
|---|---|---|
| **Barker et al., (2011, 75%)** | | Cronbach's α = |
| DEX-S | n = 60 | 0.92 |
| DEX-I (1st family member) | n = 60 | 0.93 |
| DEX-I (2nd family member) | n = 60 | 0.95 |
| DEX-I (3rd family member) | n = 60 | 0.94 |
| **Bennett et al., (2005, 96%)** | | Cronbach's α = |
| DEX-S | n = 55 | 0.92 |
| DEX-I (family members) | n = 42 | 0.93 |
| DEX-I (clinical neuropsychologists) | n = 64 | 0.95 |
| DEX-I (occupational therapists) | n = 45 | 0.94 |
| **Bodenburg & Dopslaff (2008, 73%)** | | Cronbach's α = |
| DEX-S | n = 191 | 0.85 |
| **Carvalho et al., (2013, 81%)** | | Cronbach's α = 0.95 overall |
| FrSBe-S - Apathy, Disinhibition & Executive Dysfunction subscales | n = 494 | 0.88, 0.84, 0.91 |
| **Chaytor & Smitter-Edgecombe (2008, 69%)** | | Cronbach's α = |
| DEX-S (factors 1, 2, 3, 4 & 5) | n = 46 | 0.83, 0.84, 0.75, 0.56 & 0.72 |
| BAFQ-S (factors 1, 2, 3 & 4) | n = 46 | 0.80, 0.65, 0.79 & 0.73 |
| **Grace et al., (1999, 83%)** | | Cronbach's α = |
| FrSBe-I – Pre-injury estimate | n = 87 | 0.94 |
| FrSBe-I – Post-injury ratings | n = 87 | 0.93 |
| **Shaw et al., (2015, 88%)** | | Cronbach's α = |
| DEX-S | n = 997 | 0.85 |
| **Simblett & Bateman (2011, 92%)** | | Person Separation Index (PSI) |
| DEX-S | n = 363 | = 0.92 |
| **Simblett et al., (2012, 92%)** | | Person Separation Index (PSI) |
| DEX-I | n = 271 | = 0.91 |
| **Simpson & Smitter-Edgecombe (2002, 73%)** | | Cronbach's α = |
| BAFQ-I | n = 61 | |
| Dorsolateral factor | | 0.92 |
| Orbitofrontal factor | | 0.86 |
| **Waid-Ebbs et al., (2012, 85%)** | | Cronbach's α = |
| BRIEF-I | n = 89 | |
| Behavioural Regulation Index | | 0.94 |
| Metacognitive Index | | 0.96 |

In the present review, reported internal consistency statistics were in the good ($0.8 < \alpha < 0.9$) to excellent ($0.9 \leq \alpha$) range. Studies typically reported higher internal consistency for DEX-I ratings than for DEX-S ratings, which tended to be in the good range. Good to excellent ratings were also reported for the FrSBe & BAFQ, again showing the same pattern of higher internal consistency for independent ratings, although there are only two studies per measure to compare. There was only one study investigating BRIEF-I internal consistency.

High internal consistency suggests that the rating scale is in fact measuring a single, unidimensional construct. However, the psychometric quality criteria published by Terwee et al., (2007) and Mokkink et al., (2010) state that uni-dimensionality should not be assumed, and should be explored with exploratory factor analysis, unless a theoretical model suggests the construct may be composed of specific factors, in which case confirmatory factor analysis should be conducted. Internal consistency statistics should then be reported for each factor derived. The difficulty of distinguishing separate EF factors from a uni-dimensional construct is discussed by Stuss & Alexander (2007) who suggest that EF may be comprised of a number of relatively independent sub factors, each of which combine to produce observed behaviour, or behavioural difficulties. Thus, EF rating scale can have high internal consistency for the scale as a whole, as well as for sub-factors, which is reflected in the findings in table 8 above. Studies which investigate EF rating scale sub-factors are discussed in the section below.

**Construct & structural validity**

In this context, construct validity refers to the extent to which the EF rating scales correlate with other measures of EF, which has already been discussed in the section on criterion validity above. Structural validity is the extent to which any derived factor structure of EF rating scales agrees with relevant theoretical models of the construct. In the present review, ten studies investigated factor structure, and are summarized in table 9 below:

**Table 9 – Factor structure of reviewed rating scales**

| Study | Rating Scale | Sample Size | Factors Derived (& variance explained if reported) | Total Variance |
|---|---|---|---|---|
| **Bodenburg & Dopslaff (2008, 73%)** | DEX-S | n = 191 | Initiate & sustain<br>Impulse control<br>Excitability<br>Regard for Social Standards | **49.7%** |
| **Burgess et al., (1998, 88%)** | DEX-I | n = 308 | Inhibition (21.3%)<br>Intentionality (15.5%)<br>Executive Memory (11.6%)<br>Positive Affect (10.6%)<br>Negative Affect (8.2%) | **67.2%** |
| **Carvalho et al., (2013, 81%)** | FrSBe-I | n = 494 | Apathy<br>Disinhibition<br>Executive Function | |
| **Chaytor & Smitter-Edgecombe (2007, 69%)** | BAFQ | n = 46 | Behavioural inhibition<br>Intentionality<br>Executive memory<br>Empathy | |
| **McGuire et al., (2014, 77%)** | DEX-I | n = 101<br>n = 64 | DEX-I (relatives) 1 factor (52.3%)<br>DEX-I (clinicians) 1 factor (62.7%) | **62.7%** |
| **Shaw et al., (2015, 88%)** | DEX-R-S | n = 997 | Inhibition<br>Volition<br>Social Regulation | |
| **Simblett & Bateman (2011, 92%)** | DEX-S | n = 363 | Behavioural self-regulation<br>Metacognition<br>Executive cognition | |
| **Simblett et al., (2012, 92%)** | DEX-I | n = 271 | Behavioural self-regulation<br>Metacognition<br>Executive cognition | |
| **Simpson & Smitter-Edgecombe (2002, 73%)** | BAFQ | n = 61 | Dorsolateral factor (55.2%)<br>Orbitofrontal factor (9.2%) | **64.4%** |
| **Stout et al., (2003, 85%)** | FrSBe-I | n = 324 | Apathy (4.4%)<br>Disinhibition (7.2%)<br>Executive function (29.1%) | **41%** |
| **Waid-Ebbs et al., (2012, 85%)** | BRIEF-I | n = 89 | Metacognition<br>Behaviour Regulation | |

When comparing these studies, the critical question is, to what extent do the derived factors agree with theoretical models of EF and frontal lobe function? Of the factors listed above, 2 distinct factors are most common. The first of these is labelled *initiation*, *intentionality* or *volition*, or conversely *apathy*; and the second is labelled *impulse-control*, *inhibition*, *behavioural self-regulation*, an *orbitofrontal factor* or conversely, *disinhibition*. These initiation and inhibition factors correspond well with theoretical model of executive functions (Gilbert & Burgess, 2008; Stuss, 2007; Stuss & Alexander, 2007) which fractionate the concept of EF into energization (critically depending on the anterior cingulate cortex) and behavioural self-regulation (critically dependent on the orbitofrontal cortex). Indeed, some of the studies in table 9 above describe factors relating to *metacognition*, another aspect of EF in the Stuss & Alexander and other models, e.g. Gilbert & Burgess (2008), or *executive dysfunction*, corresponding to executive control of working memory (Baddeley, 1986, 2001). There is less agreement on other aspects of EF however, and some studies describe factors not explicitly described in these models of EF, such as *excitability*, and *positive* and *negative affect*.

In contrast to the multi-factor solutions, the study by McGuire et al., (2014, 77%) found a one factor solution for both clinician's (n=64) and relative's (n=100) independent ratings. The relatively small sample size for factor analysis limits the conclusions that can be drawn from this study however.

The fact that several DEX and FrSBe studies, as well as individual studies for the BRIEF and BAFQ all have some commonalities in the derived factors suggests some convergent validity between the rating scales, at least for these factors. Unfortunately, in the present review, the only study that directly compared the measures was the study by Chaytor & Smitter Edgecombe (2007, 69%) which found a strong positive correlation between the DEX and BAFQ overall (r = 0.84).

Ideally, future studies would conduct factor analyses of two or more of the rating scales, and investigate the correlations between the derived factors. At present, the large sample size Rasch analysis studies by Simblett & Bateman (2011, 92%) and Simblett et al., (2012, 92%) present the most psychometrically robust data exploring the fractionation of EF into distinct factors.

### Hypothesis Testing

In this context, hypothesis testing, sometimes referred to as predictive validity, would refer to an *a priori* prediction, which could be tested using an EF rating scale. Evidence could then either support the hypothesis, or not, in which case a null hypothesis should be accepted.

In the present review, remarkably few studies presented *a priori* hypotheses. This is perhaps because most studies were factor analytic or correlational in design, measuring the extent to which an EF rating scale can be fractionated, or the extent to which the scale or sub-factors correlated with another measure of interest. One notable exception to this was the study by Grace et al., (1999, 83%), which compared 24 participants with anterior lesions (AL), 15 participants with posterior lesions (PL) and 48 normal controls. This study hypothesized that the AL participants would score significantly higher on the FrSBe than the PL group, normal controls, and estimates of their own function pre-injury. These hypotheses were supported by the data. Whilst this is perhaps not a surprising finding for a scale deliberately designed to measure impairment to frontal neuropsychological systems, the importance of testing validity through hypothesis testing is highly emphasised in the quality criteria published by Terwee et al., (2007).

Here, we can only conclude that future studies should also investigate validity by also testing *a priori* hypotheses, as opposed to utilising purely exploratory and correlational designs.

### Cross cultural validity

Does a complex neuropsychological construct such as EF have a conceptually similar construction in a different culture or language? Furthermore, does a different culture or language cause EF to develop in a different way? Finally, does translation of items on an EF rating scale lose meaning or validity, or can they be adequately translated?

The present review includes studies with populations speaking French (Azouvi et al., 2015), German (Bodenburg & Dopslaff, 2008), Dutch (Boelen et al., 2009), Cantonese (Chan & Bode, 2008), Greek (Emmanouel et al., 2014) and Japanese (Yamasato et al., 2007). The remaining studies were all based in the US, UK, Australia & Canada. These English speaking cultures might be considered culturally homogenous enough that findings could be considered generalizable between them. In general, the authors of the non-English speaking participant studies do not discuss whether translation may have affected validity or not. However, Bodenburg & Dopslaff (2008) do discuss this issue for some specific items. Perhaps authors make the assumption that translation of the measures into a different language does not affect the validity, or at least that any effects on validity cannot presently be distinguished from other validity issues.

### Responsiveness or Longitudinal Validity

To what extent can rating scales of EF detect clinically and statistically significant changes over time? And can these changes be distinguished from difficulties of inter-rater reliability at separate times?

In the present review, none of the studies reviewed in the present study investigated longitudinal validity or change over time. This was despite several studies taking place in rehabilitation centres during which time some recovery might be expected to have occurred. This was unfortunately the case for the Azouvi et al., (2015) study also, even though this study was longitudinal in nature, and could have allowed for a comparison of DEX-S scores at times 1 and 2. Perhaps the study authors did not think that meaningful change in EF difficulties would take place during rehabilitation, at least not to the same extent that we might expect changes in anxiety, mood, or activity participation as measured by appropriate measures.

The FrSBe does consider pre-morbid function by asking the raters to estimate how the person being rated functioned before their brain injury or neurological change (e.g. Grace et al., 1999, 83%). This is not a true longitudinal measure however, as the measurement is still cross-sectional, being taken at a single moment in time.

**Interpretability, sensitivity and specificity**

In this context, interpretability refers to the extent to which qualitative meaning can be assigned to quantitative scores, i.e. what does a particular score on an EF rating scale imply about that person's EF? Interpretability can be aided by comparison to the mean average and standard deviation of scores of a group of healthy controls. The sensitivity of an EF rating scale is the extent to which a rating scale can predict membership of a group (i.e. EF impairment due to brain injury). Thus, poor sensitivity produces false-negative error. The specificity of an EF rating scale is the extent to which an EF measure does not mistakenly classify a person as being in the incorrect group (EF dysfunction due to brain injury, when the person is in fact healthy but in the lower normal range of EF). Thus, poor specificity produces false-positive error.

In the present review, the FrSBe has normative data based on 436 men and women at two levels of education (Grace et al., 1999). The BRIEF-A reports specific normative data based on age and gender (see Isquith et al., 2013). The DEX does not have normative data, but some studies have used large control groups. The study by Burgess et al., (1998, 88%) compares mean DEX ratings for a group of 216 normal controls, and 92 neurologically impaired participants. Compared to normal controls, the neurologically impaired group had higher DEX-S, DEX-I and discrepancy scores (DEX-I DEX-S differences). These differences were all statistically significant at the $p < 0.001$ level (for 2 tailed t-tests). However, the magnitude of the differences was greater between DEX-I scores for each group than between DEX-S scores for each group. Again, it is suggested that this is due to the tendency for ABI participants to under-rate their EF difficulties, perhaps due to a lack of insight.

Studies which investigated sensitivity and specificity included the study by Shaw et al., (2015, 88%) and the study by Grace et al., (1999, 83%). Shaw et al., (2015) compared self-ratings of 663 control participants, 214 participants experiencing anxiety or depression, and 120 neurologically impaired clients. They constructed a revised 15 item version of the DEX (the DEX-R) and used a Reciever Operating Characteristic (ROC) analysis to determine the sensitivity and specificity of the DEX-R in categorising neurological and psychiatric participants. They found an optimal cut-off of 37.5 (range 0-60) at which the DEX-R demonstrated a sensitivity of 0.9 and a specificity of 0.7.

Grace et al., (1999) used ROC analysis to investigate the sensitivity and specificity of the FrSBe independent ratings in categorising participants with frontal lobe brain injuries (FL), non-frontal lobe brain injuries (NFL) and healthy controls (HC). An optimal cut-off score of 86 (range X-X) was found which discriminated FL from HC participants with 0.96 sensitivity and 0.81 specificity. An optimal cut-off score of 111 discriminated FL from NFL participants with 0.71 sensitivity and 0.73 specificity.

Although rating-scales of EF are not diagnostic tools for ABI populations, as the presence of ABI has already been established, such analysis could potentially be useful in measuring EF deficits in other client groups. For example, scales such as the DEX and FrSBe can be used to measure EF deficits in clients experiencing dementia, in order to help determine when a meaningful level of EF dysfunction is occurring in a progressive neurological disease. The same could be true for mild non-complicated TBI, in which we might expect any EF dysfunction to be minimal. Optimal cut-offs for sensitivity and specificity, as well as data on group norms and mean ratings for healthy controls are necessary in this regard.

### Floor and Ceiling Effects

In this context, the discussion of floor and ceiling effects refers to the proportion of participants who score the minimum score (floor) or maximum score (ceiling). According to the discussion by Terwee et al. (2007), if a rating scale has > 15% of participants achieving a floor or ceiling effect, this could limit the content validity of the scale, as it fails to capture the full range of EF impairment in either direction.

In the present review, no studies reported any floor effects, and only one study reported a ceiling effect, for one participant who was rated with a maximum score for BRIEF-A-I metacognition sub-scale, (Waid-Ebbs et al., 2012, 85%). Overall, this suggests that the rating-scales in the present review have enough items to adequately measure the full range of EF difficulties

## Conclusions

### Limitations of review

There are several limitations to the present review. Firstly, the focus on studies which focussed on a population of ABI or ABI and other population types means that the present review did not include studies using EF rating scales with other populations. In turn, this also meant that some rating scales are under-represented in the present review. For example, the BRIEF has for the most part been tested and validated on populations with attention deficit disorders, and a review of psychometric properties of the BRIEF with such populations has been recently published (Isquith et al., 2013). For the same reason, the review by Malloy and Grace (2005) includes rating scales not included here, such as the neuro-psychiatric inventory (NPI). This rating scale was developed to measure personality change in dementia, which is a construct that partially overlaps with the concept of EF, due to the involvement of the frontal lobes in both constructs.

A second limitation of this review and the literature that comprise it is that no studies adequately investigated all of the quality criteria within a single study. For this reason, un-investigated psychometric properties may have confounded or limited the validity of interpretation of psychometric properties that were investigated. For example, poor inter-rater reliability, and/or a lack of insight in self-rating studies, could have limited convergent validity with other measures, or factor analysis of structure. Studies which used independent raters attempted to control for this, but the limited inter-rater reliability and no criteria for what constitutes an "objective" rating are still problematic.

### Summary of discussion conclusions for each measure

The DEX has good content validity for measuring EF in an ABI population, but revised versions have better content validity. It correlates reasonably with cognitive tests of EF and with other measures of real life impairment, suggesting reasonable convergent and ecological validity. Its inter-rater reliability is quite modest however, and this cannot be wholly attributed to a lack of insight due to ABI, as multiple independent raters also show modest agreement. Internal consistency is in the good to excellent range, and factor analyses typically agree on initiation and inhibition factors, though with less or no agreement on other sub-factors. The DEX has been translated in to several different languages, and been successfully used in studies with these populations. Data on longitudinal validity is still missing, though some data on normal control participants is available, and cut-off scores for predicting impairment have been suggested.

The FrSBe has also been designed to measure EF in ABI populations. A revised structure (Carvalho et al., 2013) has been proposed, which better fits a theoretical factor structure. No convergent validity with cognitive test data were found in the present review, but one study found that it was predictive of real life impairment (Karzmark et al., 2012). No inter-rater reliability data were available for the FrSBe in the present review, but it is similar to the DEX in terms of internal consistency. The scale is also similar to the DEX in terms of factor structure, and hypothesis testing has supported its ability to measure the consequences of frontal lobe brain injury. Cut-off scores have also been suggested to discriminate frontal ABI, non-frontal ABI and healthy controls. No data on longitudinal or cultural validity were present in the current review.

The BAFQ was only represented by two papers in the present review, but included the only study which correlated two measures in the present review together, finding an r = 0.84 correlation with the DEX (Chaytor & Smitter-Edgecombe, 2007). One study suggested good ecological validity for the BAFQ at predicting real life impairment (Simpson & Schmitter-Edgecombe, 2002). These two studies performed factor analyses of this measure and found similar factor structures to that of the DEX and FrSBe, but the sample sizes in these studies were poor (<50). Data to support other psychometric criteria proposed in this review are lacking for the BAFQ.

The single study on the BRIEF-A in the present review (Waid-Ebbs et al., 2012) limits the conclusions we can draw. Its factor structure includes factors found on other measures, and excellent internal consistency is reported. No other data are available for a TBI population, although the BRIEF has been extensively used and validated for children with attention deficit disorders, and is the only measure in the present review with normative data for age and gender.

**Overall conclusions**

Overall, this review aimed to focus on the psychometric properties of rating scales of EF in an ABI population, and to build on a previous review by Grace & Malloy (2005). Adequate literature has been published on the DEX and FrSBe measures to answer some of these questions, but there is not enough published research on the BAFQ and BRIEF-A in an ABI population to adequately review their psychometric properties at present.

For clinical psychologists and other clinicians, this means that at present, the use of the DEX and FrSBe are most justified with a population of ABI clients. Their convergent validity with neuropsychological tests, and each other in terms of factor structure, as well as their ecological validity at predicting real life impairment (e.g. future unemployment) makes them useful adjuncts to traditional neuropsychological assessment. The poor inter-rater reliability for the DEX needs to be acknowledged however, and it would be reasonable to hypothesize that this limitation would be equally the case for the FrSBe.

### Recommendations for future EF rating scale design or study

Future research should focus on investigating the following psychometric properties in order to improve existing measures, or create novel ones.

Firstly, collect multiple independent ratings of the same ABI participant from trained clinicians (preferably clinical neuropsychologists or occupational therapists) who have had the opportunity to observe the participant's EF related behaviour in a range of environments over a period of time, e.g. a neuropsychological rehabilitation centre. Inter-rater reliability can then be investigated. Repeated rating at a suitably later time can be used to assess longitudinal validity (responsiveness) of the measure, which is currently missing from the literature, and could then be used to investigate the extent to which functional recovery of EF is possible over time.

Secondly, these independent ratings should be correlated with existing cognitive tests and rating scales of EF as criteria for convergent validity. This data could be used to validate rater objectivity, to further investigate inter-rater reliability.

Exploratory and confirmatory factor analysis should be conducted on a large enough sample size (n > 300) in order to both a) test whether the derived factor structure fits a theoretical model of subscales and b) to eliminate items which do not fit into derived factors. Use of Rasch analysis will improve the psychometric validity of such analysis. Once ill-fitting items have been removed, further factor analysis can confirm whether the factors fit the proposed subscale structure. Internal consistency of overall rating scale scores, and sub-scale scores should then be conducted

A new measure should have reasonable sensitivity and specificity to the presence or absence of EF impairment, due to a variety of neurological condition types and severity. Analysis using Area Under Receiver Operator Characteristic Curve is most appropriate for this. A rating scale designed according to the above criteria should then be translated into different languages and cross-validated for use in different cultures.

## References

Azouvi, P., Vallat-Azouvi, C., Millox, V., Darnoux, E., Ghout, I., Azerad, S., … Jourdan, C. (2015). Ecological validity of the Dysexecutive Questionnaire: Results from the PariS-TBI study. *Neuropsychological Rehabilitation*, *25*(6), 864–878. http://doi.org/10.1080/09602011.2014.990907

Baddeley, A. D. (1986). The Central Executive and its Malfunctions. In A. D. Baddeley (Ed.), *Working Memory*. Oxford: Oxford University Press. Retrieved from http://books.google.com/books?id=ZKWbdv__vRMC&printsec=frontcover

Baddeley, A. D. (2001). Is working memory still working? *American Psychologist*, *56*, 851–864.

Banich, M. T. (1997). *Neuropsychology: The neural basis of mental function*. Boston, Massachusetts: Houghton Mifflin.

Barker, L. A., Morton, N., Morrison, T. G., & McGuire, B. E. (2011). Inter-rater reliability of the Dysexecutive Questionnaire (DEX): Comparative data from non-clinician respondents-all raters are not equal. *Brain Injury*, *25*(10), 997–1004.

Bennett, P. C., Ong, B. E. N., & Ponsford, J. (2005). Measuring executive dysfunction in an acute rehabilitation setting: Using the dysexecutive questionnaire (DEX). *Journal of the International Neuropsychological Society*, *11*(4), 376–385.

Bodenburg, S., & Dopslaff, N. (2008). The Dysexecutive Questionnaire Advanced: Item and rest score characteristics, 4-factor solution, and severity classification. *Journal of Nervous and Mental Disease*, *196*(1), 75–78.

Boelen, D. H. E., Spikman, J. M., Rietveld, A. C. M., & Fasotti, L. (2009). Executive dysfunction in chronic brain-injured patients: Assessment in outpatient rehabilitation. *Neuropsychological Rehabilitation*, *19*(5), 625–644. http://doi.org/10.1080/09602010802613853

Bogod, N. M., Mateer, C. A., & Macdonald, S. W. (2003). Self-awareness after traumatic brain injury: A comparison of measures and their relationship to executive functions. *Journal of the International Neuropsychological Society*, *9*(3), 450–458.

Burgess, P. W., & Alderman, N. (2004). Executive Dysfunction. In L. H. Goldstein & J. E. McNeil, *Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians*. John Wiley & Sons Ltd.

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

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L., Dawson, D. R., … 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.

Carvalho, J. O., Ready, R. E., Malloy, P., & Grace, J. (2013). Confirmatory factor analysis of the frontal systems behavior scale (FrSBe). *Assessment*, 1073191113492845.

Chan, R. C. K., & Bode, R. K. (2008). Analysis of patient and proxy ratings on the Dysexecutive Questionnaire: an application of Rasch analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*(1), 86–88. http://doi.org/10.1136/jnnp.2007.117184

Chaytor, N., & Smitter-Edgecombe, N. (2007). Fractionation of the dysexecutive syndrome in a heterogeneous neurological sample: comparing the Dysexecutive Questionnaire and the Brock Adaptive Functioning Questionnaire. *Brain Injury*, *21*(6), 615–621.

Chow, T. W., & Cummings, J. L. (2007). Frontal-Subcortical Circuits. In B. L. Miller & J. L. Cummings (Eds.), *The Human Frontal Lobes* (2nd ed.). London: Guildford Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

D'Esposito, M., & Gazzaley, A. (2005). Neurorehabilitation of executive function. In M. Selzer, S. Clarke, L. Cohen, G. Kwakkel and R. Miller (Eds.), *Textbook of Neural Repair and Rehabilitation.* Cambridge: Cambridge University Press.

Dywan, J., & Segalowitz, S. (1996). Self and family ratings of adaptive behavior after traumatic brain injury: Psychometric scores and frontally generated ERPs. *Journal of Head Trauma Rehabilitation*, *11*, 79–95.

Emmanouel, A., Mouza, E., Kessels, R. P. C., & Fasotti, L. (2014). Validity of the Dysexecutive Questionnaire (DEX). Ratings by patients with brain injury and their therapists. *Brain Injury*, *28*(12), 1581–1589. http://doi.org/10.3109/02699052.2014.942371

Gilbert, S. J., & Burgess, P. W. (2008). Executive Function. *Current Biology*, *18*(3), 110–114.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Behavior rating inventory of executive function. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *6*(3), 235–238. http://doi.org/10.1076/chin.6.3.235.3152

Grace, J., Stout, J. C., & Malloy, P. (1999). Assessing frontal behavior syndromes with the Frontal Lobe Personality Scale. *Assessment*, *6*(3), 269–284.

Henson, R. K. (2006). Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, *66*(3), 393–416. http://doi.org/10.1177/0013164405282485

Isquith, P. K., Roth, R. M., & Gioia, G. (2013). Contribution of Rating Scales to the Assessment of Executive Functions. *Applied Neuropsychology: Child*, *2*(2), 125–132. http://doi.org/10.1080/21622965.2013.748389

Karzmark, P., Llanes, S., Tan, S., Deutsch, G., & Zeifert, P. (2012). Comparison of the Frontal Systems Behavior Scale and Neuropsychological Tests of Executive Functioning in Predicting Instrumental Activities of Daily Living. *Applied Neuropsychology*, *19*(2), 81–85. http://doi.org/10.1080/09084282.2011.643942

Malloy, P., & Grace, J. (2005). A Review of Rating Scales for Measuring Behavior Change Due to Frontal Systems Damage. *Journal of Cognitive & Behavioral Neurology*, *18*(1), 18–27.

McGuire, B. E., Morrison, T. G., Barker, L. A., Morton, N., McBrinn, J., Caldwell, S., … Walsh, J. (2014). Impaired self-awareness after traumatic brain injury: inter-rater reliability and factor structure of the Dysexecutive Questionnaire (DEX) in patients, significant others and clinicians. *Frontiers in Behavioral Neuroscience*, *8*. http://doi.org/10.3389/fnbeh.2014.00352

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research*, *19*(4), 539–549. http://doi.org/10.1007/s11136-010-9606-8

Oddy, M., & Worthington, A. (2009). *The Rehabilitation of Executive Disorders - A Guide to Theory and Practice*. Oxford: Oxford University Press.

Pickens, S., Ostwald, S. K., Murphy-Pace, K., & Bergstrom, N. (2010). Systematic review of current executive function measures in adults with and without cognitive impairments: *International Journal of Evidence-Based Healthcare*, *8*(3), 110–125. http://doi.org/10.1111/j.1744-1609.2010.00170.x

Shaw, S., Oei, T. P. S., & Sawang, S. (2015). Psychometric validation of the Dysexecutive Questionnaire (DEX). *Psychological Assessment*, *27*(1), 138–147. http://doi.org/10.1037/a0038195

Simblett, S. K., Badham, R., Greening, K., Adlam, A., Ring, H., & Bateman, A. (2012). Validating independent ratings of executive functioning following acquired brain injury using Rasch analysis. *Neuropsychological Rehabilitation*, *22*(6), 874–889. http://doi.org/10.1080/09602011.2012.703956

Simblett, S. K., & Bateman, A. (2011). Dimensions of the Dysexecutive Questionnaire (DEX) examined using Rasch analysis. *Neuropsychological Rehabilitation*, *21*(1), 1–25. http://doi.org/10.1080/09602011.2010.531216

Simpson, A., & Schmitter-Edgecombe, M. (2002). Prediction of employment status following traumatic brain injury using a behavioural measure of frontal lobe functioning. *Brain Injury*, *16*(12), 1075–1091. http://doi.org/10.1080/02699050210155249

Stout, J. C., Ready, R. E., Grace, J., & Paulsen, J. S. (2003). Factor Analysis of the Frontal Systems Behaviour Scale. *Assessment*, *10*(1), 79–85.

Stuss, D. T. (2007). New Approaches to Prefrontal Lobe testing. In B.L. Miller & J.M. Cummings (Eds.), *The Human Frontal Lobes* (Vol. 2). New York: Guildford Press.

Stuss, D. T., & Alexander, M. P. (2007). Is there a Dysexecutive Syndrome? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 901–915.

Stuss, D. T., & Benson, D. F. (1984). Neuropsychological Studies of the Frontal Lobes. *Psychological Bulletin*, *95*(1), 3–28.

Stuss, D. T., Picton, T. W., & Alexander, M. P. (2001). Consciousness, self-awareness and the frontal lobes. In S. Salloway, P. Malloy, & J. Duffy (Eds.), *The Frontal Lobes and Neuropsychiatric Illness* (pp. 101–109). Washington, DC: American Psychiatric Press.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42. http://doi.org/10.1016/j.jclinepi.2006.03.012

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, *61*(4), 344–349. http://doi.org/10.1016/j.jclinepi.2007.11.008

Waid-Ebbs, J. K., Wen, P.-S., Heaton, S. C., Donovan, N. J., & Velozo, C. (2012). The item level psychometrics of the behaviour rating inventory of executive function-adult (BRIEF-A) in a TBI sample. *Brain Injury*, *26*(13–14), 1646–1657. http://doi.org/10.3109/02699052.2012.700087

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., & Evans, J. J. (1996). *The Behavioural Assessment of the Dysexecutive Syndrome*. London: Harcourt Assessment.

Yamasato, M., Satoh, S., Ikejima, C., Kotani, I., Senzaki, A., & Asada, T. (2007). Reliability and validity of Questionnaire for Neurobehavioral Disability following traumatic brain injury: Questionnaire for Disability with TBI. *Psychiatry and Clinical Neurosciences*, *61*(6), 658–664. http://doi.org/10.1111/j.1440-1819.2007.01720.x

## EMPIRICAL PAPER:

# Validation of the Neuropsychological Assessment Battery Screening Measure (NAB-S) in participants with Traumatic Brain Injury

By

Thomas Morien Michael

## Abstract

**Background:** This study validates the Neuropsychological Assessment Battery Screening Tool (NAB-S) in comparison with a battery of well validated neuropsychological tests, used as a convergent validity test battery (CVTB).

**Method:** Forty-four participants with mild-complicated to severe traumatic brain injury (TBI) were recruited from a cohort of patients attending an outpatient clinic at a major UK trauma centre, and a residential rehabilitation centre in the same city. The NAB-S and CVTB were administered to the sample as part of their routine clinical assessment.

**Results:** Highly significant and strong positive correlations were observed between NAB-S overall indices, subtest indices and a NAB-S TBI index in comparison with indices for distinct cognitive domains from the CVTB, as well as the convergent validity test battery mean (CVTBM). There was a high degree of collinearity between NAB-S subtest indices, and poor internal consistency for some of these indices. Semi-partial correlations reveal the unique variance between NAB-S indices and CVTB indices, which were highly significant for the NAB-S attention and memory indices. An area under the receiver operator characteristic curve (AUROC) analysis revealed that the NAB-S index and NAB-S TBI index are highly predictive of impairment as measured by the CVTB.

**Conclusions:** The NAB-S has good predictive validity of overall impairment as measured by the CVTB. Overall, the measure is an adequate screen of cognitive impairment following TBI, particularly for the attention and memory subtest indices. Other indices had poor internal consistency and high collinearity, suggesting that further assessment with more sophisticated tests would be warranted for people with TBI.

## Introduction and rationale

Head injuries, and accompanying traumatic brain injury (TBI) is a reasonably common cause of emergency department hospital admission in the UK, responsible for 3.4% of admissions according to a study by Yates (2006). Whilst most of these TBI were mild, the Yates study (2006) classified 10.9% of TBI cases as moderate to severe, having a Glasgow Coma Scale (GCS) rating of < 12. Moderate to severe TBI is associated with a range of long term sequelae, including cognitive, emotional and behavioural difficulties (Rao & Lyketsos, 2000) and poor occupational outcomes and return to work (Azouvi et al., 2015).

Given that the sequelae of TBI are associated with poor psychosocial outcomes, it is vital for clinicians to have valid and reliable tests of cognitive ability for TBI patients. However, the acute stage following TBI is normally a period of rapid restitution of function, which can continue up to approximately 2 years post injury, with increasingly slowing rate of improvement after that (Sbordone, Liter, & Pettler-Jennings, 1995). Due to this, lengthy neuropsychological assessment at an early post-acute stage could be considered an expensive use of time and resources, given that the TBI person is likely to experience some natural recovery of function over the following months. For this reason, brief neuropsychological batteries such as the Repeatable Battery for the Assessment of Neuropsychological Syndromes (RBANS; Randolph, Tierney, Mohr, & Chase, 1998) and the Neuropsychological Assessment Battery Screening Tool (NAB-S; Stern & White, 2003) have been developed. These brief test batteries take about 30-40 minutes to administer, which for pragmatic purposes save clinician time, a valuable resource in both public and private practice.

Due to the relative brevity of these shorter test batteries (compared to more comprehensive neuropsychological tests) and the fact that they have not generally been developed for or normed using TBI populations, validation studies are necessary to determine to what extent these test batteries can be considered valid and reliable measures of cognitive deficits post TBI. A study by McKay et al., (Mckay, Casey, Wertheimer, & Fichtenberg, 2007) investigated convergent validity of the RBANS, i.e. the extent to which sub-tests of the RBANS correlated with their equivalent sub-tests in a range of more comprehensive neuropsychological tests, for a group of 57 participants with TBI. This study found good convergent validity for a range of sub-tests, particularly for tests measuring attention, immediate and delayed memory. Other sub-tests were also significantly correlated, albeit at a more modest level of correlation. Unfortunately, the RBANS does not include sub-tests for executive function (EF) which is one limitation of the battery, given the prevalence of EF difficulties following TBI (Burgess & Alderman, 2004).

The NAB-S includes sub-tests designed to measure EF, and the convergent validity of the NAB-S compared to other neuropsychological tests has been explored by Zgaljardic & Temple (2010b) although with a group of 42 mixed acquired brain injury (ABI) participants, rather than a TBI group. This study found that NAB-S sub-tests had significant correlations with equivalent subtests from other neuropsychological test batteries, but not for EF subtests or shape learning subtest. In addition, it found that NAB-S domain indices had weak internal consistency, and there were significant correlation between individual subtests, suggesting that individual subtests might rely on shared cognitive resources, not specific to a single sub-test or domain index.

In terms of sensitivity and specificity, studies have investigated the ability of the NAB-S to predict injury status and discriminate group membership in comparison with healthy controls. A study by Zgaljardic & Temple (2010a) used the full Neuropsychological Assessment Battery (NAB) which includes the NAB-S in a sample of 20 participants with moderate to severe TBI. This study investigated which subtests were most sensitive to impairment following TBI, defining impairment as less than 10th percentile in any given domain. Impairment was most frequently observed in Numbers and Letters, List Learning, Story Learning, Daily Living Memory and Categories tests.

There were no significant impairments on the spatial or language cognitive domain indices. A recent study by Hacker et al., (2016, submitted) investigates the ability of the NAB-S to discriminate between people with TBI and demographically matched healthy controls, as well as its ability to discriminate between people with mild-complicated, and severe TBI. By constructing a novel NAB-S TBI index, the authors reported excellent sensitivity and good specificity at distinguishing these client groups, using an area under the receiver operator characteristic curve (AUROC) method. The NAB-S authors report that the NAB-S has good sensitivity to the full NAB, suggesting that studies which use the full NAB can be interpreted alongside studies which use the NAB-S only. A study by Donders & Levitt (2012) investigated the full NAB EF, Attention and Memory modules in 54 TBI participants, compared to 54 healthy controls. These indices all showed significant differences between these groups, and had strong negative correlations with duration of coma for the TBI group. In this study, the Numbers & Letters, and Mazes subtests showed greatest sensitivity to TBI, which the authors suggest could be due to psychomotor speed impairment affecting performance on these subtests.

In non-TBI groups, studies have investigated the sensitivity of NAB-S to measure impairment in mixed neurological groups (Iverson, Williamson, Ropacki, & Reilly, 2007), mild cognitive impairment (Brooks, Iverson, & White, 2009), and groups of substance abusing patients (Cannizzaro, Elliott, Stohl, Hasin, & Aharonovich, 2014; Grohman & Fals-Stewart, 2004).

In terms of ecological validity, studies have compared NAB-S test scores with measures of functional independence following TBI, generally finding significant relationships, which the study authors argue suggests good validity in terms of predicting real life impairment following TBI (Temple et al., 2009; Zgaljardic, Yancy, Temple, Watford, & Miller, 2011).

### Aim and hypotheses of present study

In the present study, I aim to investigate convergent validity of the NAB-S in a TBI population using other neuropsychological tests as criteria for convergent validity. As noted above, the Zgaljardic & Temple (2010b) study utilized 42 mixed ABI participants, and found poor internal consistency of NAB-S domain indices. For this reason, the present study will further investigate NAB-S domain internal consistency, and will correlate individual both subtest indices and subtests themselves with their equivalent indices from the criterion neuropsychological tests, as such correlations might be considered to be more valid than index-index correlations.

The present study will utilize a range of well-established and validated neuropsychological tests as convergent validity criteria, as well as utilising tests to estimate pre-morbid functioning, and tests of effort, so as to ensure a valid interpretation of results. A complete list of tests utilised in the present study is described in the method section.

I hypothesize that the strongest correlations will be between indices comprised of subtests which are most similar to each other across the batteries, i.e. forwards & backwards digit span. In addition, we might expect weakest correlations between tests in domains that themselves have low internal consistency, such as EF. This could be due to such domains being comprised of a number of different cognitive processes, such that different tests of EF could be measuring distinct cognitive difficulties. However, the tests of EF in both the NAB-S and CVTB involve processing speed to a certain extent, and the relationship between EF, attention and working memory is also well established in the literature and theoretical constructs of EF (Sohlberg & Mateer, 2001). For this reason, we might expect a relationship between NAB-S EF and attention indices, and the CVTB indices for working memory, processing speed and executive function.

## Method

### Ethical approval

Ethical approval was initially granted by the South Birmingham NHS Research Ethics Committee (REC) on 23rd July 2014 (Reference 14/WM/1006). Ethical amendments were then submitted, and were approved by the same NHS REC on 1st July 2015 and 8th December 2015. These amendments allowed the inclusion of the Test of Memory Malingering (TOMM) which had been intended to be included in the original ethics application, and allowed the retrospective recruitment of potential participants for whom researchers already had data as part of their routine clinical assessment, but for whom prior informed consent had not been obtained. The amendments also allowed the author to gather data by taking part in the study as an assessing clinician. In addition to ethical approval from the NHS REC, approval was granted from the Research and Development (R&D) department at the University of Birmingham, which submitted the ethical application to the REC, and from R&D at University Hospitals Birmingham (UHB), one of the participating sites.

### Recruitment and Informed Consent

Potential participants were approached by members of the clinical teams at each participating site. Potential participants consisted of patients at each site who were undergoing routine clinical assessment at that site. Each person was given an information sheet and had the aims of the study explained to them, as well as being given the opportunity to ask questions about the study. If the person was happy to give informed consent to take part in the study, this was gained either prior to the first, or prior to the second testing session, in order to give the potential participant time to consider the information and to decide if they wished to take part. In any case, potential participants were tested as part of their routine clinical assessment, but their data was not included in the study if they either declined to participate, failed to meet inclusion criteria, or met exclusion criteria.

Inclusion criteria for the study were applied, and potential participants for the study were only approached to be recruited if:

a) They have experienced a mild-complicated, moderate or severe TBI (as determined by a combination of Glasgow Coma Scale score, length of unconsciousness, and imaging data, if available);

b) They are between 18 and 69 years of age;

c) The have experienced their brain injury within the last 3 years;

d) They have mental capacity to be able to give informed consent to participate, as judged by recruiting clinicians at their site.

Exclusion criteria for the study were applied, and potential participants for the study were not recruited if:

a) English was their second language

b) They had experienced previous head injury

c) They were still experiencing post-traumatic amnesia, as judged by consultant clinical neuropsychologist in clinical interview, and clinical discharge notes where available.

d) They have apparent sensory or motor deficits

e) They have enduring mental health conditions

f) They have a diagnosed learning disability, or developmental or acquired dyslexia

g) They have a diagnosed organic brain disease such as dementia

h) They fail one or more of the effort tests used in the study, according to standardised cut-off scores detailed in the manuals used to score these effort tests.

In addition to the above method of recruitment, there were a number of potential participants (17) who had already taken part in testing as part of their routine clinical assessment, but who had not been approached for recruitment prior to testing. An ethical amendment allowed researchers to retrospectively approach these potential participants, in order to gain their informed consent to take part in the study. This recruitment process was identical to the above process, except that it took place after testing had been completed, and used an appropriately modified participant information sheet and informed consent form to reflect this. The same inclusion and exclusion criteria applied.

### Participants

Participants consisted of 44 people who had experienced TBI and were attending an outpatient neuropsychology clinic in a UK hospital. Participant demographics are detailed in table 10 below, with severity of brain injury determined according to World Health Organisation criteria (2001).

**Table 10 – Participant demographics and TBI characteristics**

| **Participant Demographics** | | | | | |
|---|---|---|---|---|---|
| **Age** (years) | | | **Gender** | | |
| Mean | 35.41 | | | | |
| SD | 15.62 | | Male | 30 | 68.2% |
| Range | 18-66 | | Female | 14 | 31.8% |
| **Initial Glasgow Coma Scale (GCS)** | | | **Overall severity Classification** | | |
| Unavailable | 10 | 22.7% | Mild-complicated | 5 | 11.4% |
| 13-15 | 16 | 36.4% | Moderate | 17 | 38.6% |
| 9-12 | 3 | 6.8% | Severe | 22 | 50% |
| 3-8 | 15 | 34.1% | | | |
| **Post Traumatic Amnesia duration (PTA)** | | | **Imaging data** | | |
| | | | Unavailable | 2 | 4.5% |
| Unavailable | 6 | 13.6% | Frontal contusions | 17 | 38.6% |
| <24 hour | 6 | 13.6% | Temporal contusions | 10 | 22.7% |
| >24 hours <1 week | 16 | 36.4% | Parietal contusions | 8 | 18.2% |
| >1 week | 16 | 36.4% | Occipital contusions | 2 | 4.5% |
| | | | Any skull fracture | 22 | 50% |
| **Months since injury** | **Mean** | **SD** | Subarachnoid haemorrhage | 22 | 50% |
| Range 1.5 to 27.5 | 11.7 | 7.3 | Other Intracranial haemorrhage | 9 | 20.5% |

### Measures

A complete list of the measures used in the study is as follows:

### NAB-S

The NAB-S was co-normed with the full NAB on a sample of 1448 individuals. Test score distributions are reported by age (18-97), gender and level of education (Stern & White, 2003). The NAB-S is similar to the RBANS in that it has a number of sub-tests and indices, has two alternate forms, and takes 35-40 minutes to administer. It assesses domains of Attention, Executive Function, Language, Memory and Visuospatial reasoning. It is co-normed with the Neuropsychological Assessment Battery (NAB) which assesses the same cognitive domains, but in more detail, using a dedicated module for each domain. The NAB-S is therefore designed to quickly screen for cognitive difficulties, which can then be further assessed by the cognitive module if necessary. A list of NAB-S subtests is described in appendix III.

### Pre-morbid estimate measure

In order to obtain an estimate of a participant's general level of cognitive functioning prior to their traumatic brain injury, the UK version of the Test of Premorbid Functioning (TOPF; Wechsler, 2011) was used. This is a revised and updated version of the Wechsler Test of Adult Reading.

### Effort tests

In order to be confident that the test battery provides a valid assessment of symptom validity, there is a clinical and scientific need to use tests which measure participant effort. In the United States, the National Academy of Neuropsychology considers use of effort tests to be a vital component of neuropsychological test batteries (Bush et al., 2005). For this reason, the current study utilises the Test of Memory Malingering (TOMM; Tombaugh, 1997) and Green's Word Memory Test (WMT; Green, Allen, & Astner, 1995). The TOMM is a picture memory test, while the WMT is a word memory test. Both tests claim high sensitivity to participant effort, whilst simultaneously being very insensitive to the effects of traumatic brain injury on memory, i.e. being relatively easy to complete even in the presence of severe cognitive impairment.

A recent study investigated the validity of both the TOMM and WMT alongside several tests of memory, using factor analysis to separate performance into effort and memory factors (Heyanka et al., 2015). This study found a two-factor solution, in which TOMM and WMT performance loaded onto an effort factor, whilst the memory tests loaded onto a distinct memory factor. The authors argue that this provides good evidence for the validity of such tests in a traumatic brain injured population. Thus, poor participant performance on these tests should reduce our confidence in the validity of other test results in the battery.

### Convergent validity test battery

A battery of existing and well validated neuropsychological tests was chosen to correspond to the cognitive domains assessed by the NAB-S. This battery consisted of specific subtests from the following neuropsychological tests; the Wechsler Adult Intelligence Scale IV (WAIS-IV; Wechsler, 2008), the Wechsler Memory Scale IV (WMS-IV; Wechsler, 2009) and the Delis-Kaplan Executive Function System (D-KEFS; Delis, Kaplan, & Kramer, 2001). The specific subtests used from the WAIS-IV were: Block Design, Coding, Symbol Search, Digit Span, Arithmetic, Matrix Reasoning, Vocabulary and Information. Specific subtests from the WMS-IV were: Logical Memory I & II, Visual Memory I & II. Specific subtests from the D-KEFS were: The Towers Test, Colour Word Interference, Verbal Fluency and Trail Making. WAIS-IV indices were calculated using the eight subtests as described in the manual. WMS-IV immediate and delayed memory indices were calculated according to a method described by Miller et al., (2012). D-KEFS scores were marked according to the manual and an executive functioning index (EFI) was constructed using a software programme, according to a method described by Crawford et al. (2011).

### Test Procedure

Sub-tests were administered in approximately the order shown in table 11 below. Occasionally, tests were administered in a slightly different order, so as to ensure a consistent period of time had passed before administering delayed memory tasks, present in the WMT, WMS & NAB-S test batteries. As fatigue is a common issue, particularly in the relatively post-acute stage of TBI, participants were encouraged to take breaks if necessary, so as not to bias their results (Mollayeva et al., 2014).

**Table 11 – Neuropsychological test administration**

**Test Session 1**

| **Subtest** | **Battery** |
|---|---|
| Block Design | WAIS-IV |
| Test of Premorbid Functioning | TOPF |
| WMT Immediate Recognition (IR) | WMT (computerised) |
| Coding | WAIS-IV |
| Symbol Search | WAIS-IV |
| Trail Making | D-KEFS |
| Digit Span | WAIS-IV |
| Colour Word Interference | D-KEFS |
| WMT Delayed Recognition (DR) | WMT (computerised, 30 minutes after IR) |
| WMT - MC, PA and FR (if required) | WMT (computerised) |

**Test Session 2**

| **Subtest** | **Battery** |
|---|---|
| Test of Memory Malingering | TOMM |
| Visual Reproduction I | WMS-IV |
| Logical Memory I | WMS-IV |
| Verbal Fluency | D-KEFS |
| Arithmetic | WAIS-IV |
| Visual Reproduction II | WMS-IV |
| Logical Memory II | WMS-IV |
| Matrix reasoning | WAIS-IV |
| The Towers Test | D-KEFS |
| Vocabulary | WAIS-IV |
| Information | WAIS-IV |
| Neuropsychological Assessment Battery Screening Module | NAB-S |

**Results**

Table 12 presents descriptive statistics for the NAB-S test data. These statistics are presented both including effort test fails (n=44) and excluding effort test fails (n=40). Additionally, an internal consistency statistic (Cronbach's α) was calculated for the overall NAB-S index and individual NAB-S subtest indices. Finally, a NAB-S TBI index was calculated, consisting of the mean average of 6 NAB-S subtests, transformed to a T-score, as described in Hacker et al., (2016, submitted). The 6 subtests were; Screening Numbers & Letters Speed (S-N&L-A), Screening Numbers and Letters Efficiency (S-N&L-B), Screening Story Learning Delayed Recall (S-STL-Drc), Screening Digit Span Backwards (S-DGB), Screening Mazes (S-MAZE) and Screening Designs (S-DES).

**Table 12 – NAB-S total index & sub-test index statistics**

| Index & sub-tests | n = | Mean | SD | Range | Cronbach's α |
|---|---|---|---|---|---|
| **NAB-S Total Index** | 44 | 95.6 | 20.9 | 51 – 141 | α = 0.811 |
| - Excluding effort test fails | 40 | 97.9 | 20.1 | 51 – 141 | α = 0.791 |
| **Attention Index (S-ATT)** | 44 | 88.4 | 18.3 | 54 – 118 | α = 0.829 |
| - Excluding effort test fails | 40 | 90.2 | 17.9 | 54 – 118 | α = 0.819 |
| **Executive Function Index (S-EXE)** | 44 | 91.2 | 17.4 | 59 – 127 | α = 0.382 |
| - Excluding effort test fails | 40 | 92.8 | 17.2 | 59 – 127 | α = 0.437 |
| **Language (S-LAN)** | 44 | 102.1 | 23.6 | 45 – 134 | α = 0.489 |
| - Excluding effort test fails | 40 | 103.4 | 22.3 | 45 – 134 | α = 0.470 |
| **Memory (S-MEM)** | 44 | 95.2 | 17.2 | 61 – 142 | α = 0.668 |
| - Excluding effort test failures | 40 | 97.1 | 16.5 | 61 – 142 | α = 0.703 |
| **Visuospatial Reasoning (S-SPT)** | 44 | 106.1 | 16.9 | 74 – 135 | α = 0.238 |
| - Excluding effort test failures | 40 | 107.4 | 16.4 | 74 – 135 | α = 0.212 |
| **NAB-S TBI Index (T-Score)** | 44 | 42.6 | 10 | 21.9 – 59.6 | α = 0.777 |
| - Excluding effort test failures | 40 | 43.3 | 10 | 22.3 – 60.1 | α = 0.774 |

As can be seen in table 12 above, when effort test fail data were excluded from the sample, the mean average scores were slightly increased, but the ranges were unchanged, and the internal consistency statistics remained similar also. The National Academy of Neuropsychology in the United States argues that failure to control for effort test failure significantly biases data analysis for empirical studies (Bush et al., 2005). For this reason, effort test fail data (from 4 participants) were excluded from all further analyses.

Table 13 presents descriptive statistics for the convergent validity test battery (CVTB). Internal consistency statistics were not calculated for the CVTB.

**Table 13 – Convergent Validity Test Battery (CVTB) indices**

| Index & sub-tests | n = | Mean | SD | Range |
|---|---|---|---|---|
| **TOPF Full Scale IQ (FSIQ) Predicted** | 40 | 96.6 | 8.8 | 82.5 to 117.0 |
| **WAIS-IV FSIQ** | 37 | 93.9 | 13.7 | 69 to 126 |
| Verbal Comprehension Index (VCI) | 40 | 96.0 | 17.7 | 63 to 134 |
| Perceptual Reasoning Index (PRI) | 34 | 97.0 | 15.1 | 71 to 133 |
| Working Memory Index (WMI) | 40 | 96.8 | 13.2 | 69 to 122 |
| Processing Speed Index (PSI) | 40 | 89.8 | 11.1 | 62 to 117 |
| **WMS-IV** | | | | |
| Immediate Memory Index (IMI) | 40 | 95.8 | 16.0 | 55 to 130 |
| Delayed Memory Index (DMI) | 40 | 95.1 | 18.3 | 49 to 141 |
| **D-KEFS** | | | | |
| Executive Function Index (EFI) | 40 | 95.4 | 17.6 | 55 to 130 |
| **CVTB Mean (CVTBM)** | 40 | 94.6 | 12.3 | 71.6 to 118.6 |

In addition to the indices derived directly from the CVTB, I constructed a CVTB mean (CVTBM) according to the following formula: CVTBM = (WMI+PSI+IMI+DMI+EFI)/5. This used the most sensitive to change indices from the WAIS-IV (the WMI & PSI). Our data are consistent with findings from the WAIS-IV TBI clinical data group, published by Iverson et al., (2013) which focussed on a moderate to severe TBI group. Similarly to that group, PSI was most impaired.

Table 14 presents Pearson's correlations, comparing the NAB-S overall index and NAB-S TBI index with CVTB indices and the CVTBM. A discrepancy score representing the magnitude of loss of function were also calculated. This was the discrepancy between CVTBM and the TOPF FSIQ (predicted) score and was also correlated with the NAB-S index and NAB-S TBI index.

**Table 14 – Correlation of NAB-S and CVTB indices**

| **Index** | **n =** | NAB-S<br>Overall Index | NAB-S<br>TBI Index |
|---|---|---|---|
| IMI | 40 | 0.533** | 0.373* |
| DMI | 40 | 0.499** | 0.444** |
| EFI | 40 | 0.605** | 0.5266** |
| WMI | 40 | 0.602** | 0.609** |
| PSI | 40 | 0.365* | 0.445** |
| CVTBM | 40 | 0.673** | 0.646** |
| TOPF FSIQ – CVTBM | 40 | -0.472** | -0.481** |

Legend: * p < 0.05, ** p < 0.01

### Comparisons of subtest indices

Table 15 presents a zero order correlation matrix for NAB-S subtest indices and CVTB indices. The shaded correlations represent cognitive domains which we might expect to be reasonably independent from one another, and have weaker or non-significant correlations. The non-shaded correlations represent hypothesized relationships.

**Table 15 – Direct comparisons of subtest indices**

| Indices | IMI | DMI | EFI | WMI | PSI |
|---|---|---|---|---|---|
| NAB-S S-ATT | 0.255 | 0.317* | 0.508** | 0.660** | 0.366* |
| NAB-S S-EXE | 0.123 | 0.176 | 0.339* | 0.395* | 0.233 |
| NAB-S S-MEM | 0.559** | 0.497** | 0.491** | 0.457** | 0.184 |
| NAB-S S-SPT | 0.621** | 0.496** | 0.455** | 0.311 | 0.290 |
| NAB-S S-LAN | 0.431** | 0.369* | 0.453** | 0.405** | 0.251 |

Legend * p < 0.05, ** p < 0.01 (all correlations two-tailed)

Amongst the correlations of indices which have hypothesized relationships, the NAB-S Attention index (S-ATT) correlated most highly with the WMI (r = 0.660, p < 0.01). Indeed this was the strongest correlation in the matrix. The weakest correlations of these hypothesized indices were the NAB-S executive function index with PSI (non-significant) and EFI (r = 0.339, p < 0.05). Several of the shaded comparisons also showed highly significant correlations. Given the high degree of collinearity between NAB-S and CVTB indices, we repeated these comparisons using semi-partial correlations, which are reported in Table 16.

**Table 16 – Semi Partial Correlations of Subtest Indices**

| Indices | IMI | DMI | EFI | WMI | PSI |
|---|---|---|---|---|---|
| NAB-S S-ATT | 0.0237 | 0.1057 | 0.2435* | 0.4333*** | 0.2230 |
| NAB-S S-EXE | -0.1176 | -0.0734 | 0.0070 | -0.0116 | 0.0254 |
| NAB-S S-MEM | 0.2659* | 0.2611* | 0.2460* | 0.2247* | 0.0448 |
| NAB-S S-SPT | 0.3804** | 0.2404 | 0.0864 | -0.1272 | 0.1263 |
| NAB-S S-LAN | 0.0313 | 0.0325 | 0.0679 | 0.0922 | 0.0293 |

Legend: *p < 0.05, ** p < 0.01 *** p < 0.001 (all semi-partial correlations two tailed)

Semi-partial correlations remove the variance due to other NAB-S indices, leaving only the variance due to the NAB-S index in question. Therefore, this correlation matrix describes the relationship between the CVTB indices and the unique variance of each NAB-S subtest index score.

Again, shaded correlations represent indices for cognitive domains which we might expect to be reasonably independent from one another. The non-shaded correlations represent hypothesized relationships. In this correlation matrix, the IMI and DMI memory indices are most strongly associated with the memory and spatial processing indices of the NAB-S, whereas the EFI and WMI indices are most strongly associated with the NAB-S attention and memory indices. In this analysis, most of the shaded comparisons were not significant. The NAB-S executive function index correlations were also non-significant however.

Given the high degree of collinearity within NAB-S subtest indices, it is difficult to interpret the relationship between individual NAB-S subtest indices and the CVTB indices. For this reason, I conducted a series of forward stepwise regressions to identify significant covariates. Covariates with less than 0.10 level significance were removed from the stepwise equation, however, having been removed these covariates were allowed to re-enter the stepwise model if they subsequently evidenced significance (at 0.05) when the number of covariates in the stepwise model was reduced. The standardized Beta coefficients for the NAB-S subtests are reported in table 17 below.

**Table 17 – Forward stepwise regressions of NAB-S subtests and CVTB indices**

| Variable | IMI | DMI | EFI | WMI | PSI |
|---|---|---|---|---|---|
| SSTL Delayed | 0.294** | | 0.266* | | |
| SVIS | 0.443*** | 0.361** | | | |
| SDES | 0.414*** | 0.355** | | | |
| SN&L A Speed | -0.310** | -0.321** | | -0.248* | |
| SDGB | | 0.413*** | | 0.620*** | |
| SN&L B Efficiency | | | 0.378** | 0.351** | 0.485*** |
| SSHL Immediate | | | 0.305* | | |
| | | | | | |
| $R^2$ | 0.638 | 0.564 | 0.466 | 0.611 | 0.236 |
| Adjusted $R^2$ | 0.601 | 0.520 | 0.462 | 0.582 | 0.217 |

Legend: * p<.05; ** p<.01; *** p<.001

Our sample size of n=40 is perhaps too small to provide enough statistical power to confidently interpret the results of this regression analysis. However, I note that other researchers have also published regression analyses with NAB-S data for a similar number of participants (Zgaljardic et al., 2001, n=47). In this analysis, the Screening Numbers and Letters A Speed test had a negative beta coefficient, as its score is the time to complete in seconds.

The IMI was significantly associated (F = 17.17, p < 0.001) with delayed free recall of a two-sentence story (Screening Story Learning: Delayed Recall), a visual match to target test (Screening Visual Discrimination), a visuospatial construction task (Screening Design Construction) and speed to complete a letter cancellation task (Screening Numbers & Letters). The combination of these four subtests explained approximately 64% of the variance of the immediate memory index.

The DMI was significantly associated (F = 12.62, p < 0.001) with a visual match to target test (Screening Visual Discrimination), a visuospatial construction task (Screening Design Construction), speed to complete a letter cancellation task (Screening Numbers & Letters) and reverse digit span (SDGB). The combination of these four subtests explained approximately 52% of the variance of the delayed memory index.

The EFI was significantly associated (F = 11.62, p < 0.001) with delayed free recall of a two-sentence story (Screening Story Learning: Delayed Recall), efficiency on a divided attention task (SN&L B Efficiency) and immediate recognition on a single trial learning task (SSHL Immediate). The combination of these three subtests explained approximately 46% of the variance of the Executive Functioning Index.

The WMI was significantly correlated (F = 20.93, p < 0.001) with speed to complete a letter cancellation task (Screening Numbers & Letters), reverse digit span (SDGB) and efficiency on a divided attention task (SN&L B Efficiency). The combination of these three subtests explained approximately 58% of the variance of the Working Memory Index.

Finally, the Processing Speed Index what significantly associated (F = 12.95, p < 0.001) with the efficiency on a divided attention task (SN&L B Efficiency) and this single subtest explained approximately 24% of the variance of the Processing Speed Index.

Overall, the subtests of the NAB-S reliably predicted the CVTB indices. Large effect sizes were observed for the WMS-IV IMI and DMI, and the WAIS-IV WMI, a moderate effect size was observed for the D-KEFS EFI and a small effect size was observed for the WAIS-IV PSI. It should also be noted that the subtests selected in the stepwise regression models show a good correspondence with the subtests used to calculate the NAB-S TBI index (N&L (A) Speed, N&L (B) Efficiency, STLDrc, DGB, and MAZ).

Table 18 provides base rate frequencies for the number of CVTB indices (the IMI, DMI, EFI, WMI and PSI) that are below different cut off points. I chose the 25th, 16th, 10th and 5th percentiles, to represent differing severities of impairment. I then calculated the area under the receiver operator characteristic curve (AUROC) for the NAB-S index and the NAB-S TBI index for the number of CVTB indices that are below different cut off points.

**Table 18 – Base rate frequencies for CVTB indices below specified percentile cut-offs and Area under receiver operator characteristic curve (AUROC) values**

| | | At least<br>1 index | At least<br>2 indices | At least<br>3 indices | At least<br>4 indices | At least<br>5 indices |
|---|---|---|---|---|---|---|
| < 25th Percentile | % | 82.5% | 55.0% | 45.0% | 32.5% | 22.5% |
| AUROC | NAB-S | 0.805 | 0.823 | 0.765 | 0.823 | 0.714 |
| | NAB TBI | 0.779 | 0.765 | 0.707 | 0.765 | 0.792 |
| < 16th Percentile | % | 55.0% | 32.5% | 22.5% | 12.5% | 10.0% |
| AUROC | NAB-S | 0.714 | 0.851 | 0.824 | 0.805 | 0.782 |
| | NAB-S TBI | 0.792 | 0.829 | 0.849 | 0.777 | 0.794 |
| < 10th Percentile | % | 40.0% | 30.0% | 17.5% | 10.0% | 5.0% |
| AUROC | NAB-S | 0.697 | 0.866 | 0.861 | 0.854 | 0.960 |
| | NAB-S TBI | 0.783 | 0.869 | 0.852 | 0.875 | 0.934 |
| < 5th Percentile | % | 30.0% | 17.5% | 7.5% | 0.0% | 0.0% |
| AUROC | NAB-S | 0.866 | 0.878 | 0.864 | N/A | N/A |
| | NAB-S TBI | 0.869 | 0.861 | 0.901 | N/A | N/A |

Legend: AUROC = Area under the receiver operator characteristic

The area under the receiver operator characteristic curve can be conceived as the ratio of the number of true positive classifications (i.e., sensitivity) divided by the number of false positive predictions (i.e., 1-specificity). An AUROC value of 0.5 is associated with random classification, so the greater the AUROC value is above 0.5, the greater the ratio of true positive classifications.

The NAB-S index produced AUROC values ranging from 0.697 to 0.960. The NAB-S TBI index produced similar AUROC values, ranging from 0.707 to 0.934. Both indices evidenced good to excellent classification accuracy and neither index evidenced a statistically significant advantage in the prediction of any of the full battery indices at any of the cut off points.

## Discussion

### Convergent validity overall

There were multiple, highly statistically significant correlations between the NAB-S overall index and NAB-S TBI index, and the CVTB indices and CVTB mean (Table 5). This suggests that the overall level of performance on one test battery is associated with the level of performance on the other test battery. There was also a statistically significant correlation between the NAB-S indices and the TOPF FSIQ – CVTBM discrepancy, suggesting that the NAB-S overall index and NAB-S TBI index could be predictive of the magnitude of a loss of function. We would expect this correlation to be negative, i.e. a larger loss of function would be associated with a lower NAB-S index score.

### Convergent validity between sub-test indices

I then investigated the strength of correlation of individual indices in both the NAB-S and CVTB. When Pearson's correlations were derived, these evidenced a pattern of strong positive correlation, presented in the correlation matrix in table 15. In interpreting this correlation matrix, we should expect strong positive correlations between sub-tests which purport to measure the same construct, or which measure constructs which are theoretically related to one another in the literature. For this reason, correlations of specific interest were the relationship between memory indices (the NAB-S S-MEM vs the IMI and DMI), and the relationship between Attention, Working Memory, Processing Speed and Executive Function indices (NAB-S S-ATT & NAB-S S-EXE vs the WMI, EFI and PSI).

Equally, we should also expect weaker, or non-significant correlations between indices purport to measure neuro-psychologically distinct indices. For example, NAB-S S-LAN (language) and WAIS-IV PSI (processing speed). Comparisons between indices which we might expect to be neuro-psychologically distinct have a shaded background, whereas comparisons between similar constructs have no shading.

In some cases, these correlations presented a pattern which we might expect, given the content of the tests that comprised them. For example, the NAB-S S-ATT index correlated strongly and significantly with the WAIS-IV WMI ($r = 0.660$, $p < 0.01$). Given that both of these indices contain a digit span test as one of the tests that comprise them, this is perhaps not surprising, as the digit span tests in both batteries were the most similar tests to each other across both batteries. For this reason, it could also be argued that each digit span test is measuring the same neuropsychological construct and perhaps the neurological processes which are its substrate. This finding supports our hypothesis that the indices containing digit span subtests would be amongst the strongest correlations in the battery.

Equally, the NAB-S S-EXE and EFI indices should be expected to correlate with one another, as both claim to be measuring the same neuropsychological construct, executive function (EF). However, given that the tests of EF in each test battery are in some ways quite different to one another, I hypothesised that the indices comprised of these tests would be amongst the weakest of the correlations of a shared construct. This correlation was $r = 0.339$, $p < 0.01$. This was the weakest of the correlations between indices purporting to measure the same construct, perhaps reflecting that tests of EF each measure different aspect of executive function.

However, there were also statistically significant correlations between indices purporting to measure constructs which we might expect to be neuro-psychologically distinct. For example, NAB-S S-SPT (spatial) and NAB-S S-LAN (language) correlated highly with several of the other indices. Other indices in table 6 did correlate more weakly or non-significantly, as we might expect.

How can we make sense of the high degree of collinearity shared between NAB-S indices and all indices of the CVTB? One interpretation of this would be to suggest that every test, and therefore every index, partially relies upon common cognitive processes, which, when impaired, would lead to a general loss of performance across the test battery. Such common cognitive processes could be considered analogous to Spearman's g, a general intelligence factor (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002).

A more cautious interpretation would be to consider the fact that some of the NAB-S indices showed poor internal consistency. Notably, the executive function index (S-EXE, $\alpha = 0.437$), the language index (S-LAN, $\alpha = 0.470$) and the spatial index (S-SPT, $\alpha = 0.212$). This suggests that the sub-tests which comprise these indices may in fact be measuring different constructs, rather than a single unidimensional construct. This may suggest that the NAB-S indices are poor operationalisations of the cognitive domains in question. For this reason, whilst NAB-S indices may show good sensitivity in predicting variance of a CVTB index we might expect to be similar, the common variance with other indices is perhaps evidence of poor specificity to the index domain. However, it is worth noting that the internal consistency statistics were considerably higher in the present study than in the study of Zgaljardic & Temple (2010b).

Given the high degree of collinearity between the indices, I then used semi-partial correlations to determine the unique variance shared between each NAB-S index and the CVTB indices, whilst removing any variance shared with other NAB-S factors (Table 16). These correlations maintained the strength of association between the NAB-S attention index and WMI ($r = 0.433$, $p < 0.001$) which was the strongest correlation in the matrix. Many of the comparisons between indices purporting to measure constructs which we might expect to be neuro-psychologically distinct were non-significant using this method. The EF comparisons also became non-significant though, and significant correlations were also found between NAB-S memory and EFI and WMI. Finally, the NAB-S S-SPT (spatial index) showed a significant correlation with the IMI.

At the NAB-S subtest level, I used forward stepwise regression to measure the extent to which different NAB-S subtests predicted variance in CVTB indices. Using this method, I again observed expected relationships, such as the NAB-S digit span backwards having a high Beta coefficient with the CVTB working memory index (Beta = 0.620, $p < 0.001$). However, I also observed relationships between NAB-S subtests and CVTB indices which we might expect to be neuro-psychologically distinct, such as the highly significant associations between NAB-S visuospatial tasks and CVTB immediate and delayed memory indices. These results are difficult to interpret, but the high degree of significance could suggest that a common cognitive factor is shared between tests on either battery that we might expect to measure distinct neuropsychological domains.

Overall, our findings suggest that the NAB-S indices vary with an adequate degree of sensitivity with indices of the CVTB. In the cases of the NAB-S attention and NAB-S memory indices, the strength of association is good. In other cases, indices or subtests correlate with indices for which there is not a sound neuropsychological theory to explain the association. This could represent measurement of a common cognitive factor, or that the index has poor specificity to the cognitive domain.

**Prediction of impairment**

The AUROC analysis suggests that the NAB-S overall index and the NAB-S TBI index can be used to predict overall impairment as measured by the CVTB. For example, if we choose the 5th percentile as a cut off to signify impairment, then 30% of our sample (12 of 40 participants that passed effort testing) would have at least one CVTB index below the 5th percentile. At this level of cut-off, the area under the ROC curve for the NAB-S Index was 0.866 and the area under the ROC curve for the NAB-S TBI index was 0.869. Overall, these AUROC values ranged from good to excellent, suggesting that the NAB-S can be used to predict impairment on the full CVTB. Our data suggest that the impairment present in the TBI group is highly variable, suggesting that reliance on the overall mean may be misleading. The data in table 18 suggests that, consistent with Iverson et al (2013) the interpretation of neuropsychological test scores at higher cut-offs but across multiple tests simultaneously well may be a more efficient and sensitive measure of TBI impairment. Given that this is a group which are at the less severe range of TBI impairment, the sensitivity of the NAB-S index and NAB-S TBI index at predicting impairment is encouraging.

### Clinical implications

Given that the NAB-S overall index and NAB-S TBI index correlate strongly with the CVTBM, and these indices can also be used to predict impairment on the CVTB, the NAB-S could be argued to be a useful measure of cognitive impairment following TBI, particularly for the attention and memory indices. Given the relatively short administration time, and age and education normative data, the NAB-S could be very useful in a clinical setting.

However, given the poor internal consistency of the executive, spatial and language indices, and the high degree of collinearity shared between these indices, caution should be taken in over interpreting the meaning of these index scores as a measure of specific impairment in a single domain. For this reason, supplementary testing, with either the full NAB, or equivalent tests from other batteries, would be desirable in order to better measure executive function, language and visuospatial ability following TBI.

### Limitations of the current study

The present study had a sample size of 44, reduced to 40 after excluding effort test fail data. Whilst this compares similarly to the Zgaljardic & Temple (2010b) study, which had 42 participants, it is less than the 57 participants in the McKay et al., (2007) RBANS validation study, and is below the 50 participant minimum recommended by Crawford & Garthwaite (2008).

Our study also lacked control participants, meaning that proper normative data for the NAB-S was unavailable for this study. However, Hacker, et al. (2016, submitted) have recently investigated this for the NAB-S with 104 TBI participants and 98 demographically matched orthopaedic participants as controls. Furthermore, I have controlled for effort test failure in this study, which is consistent with the  US National Academy of Neuropsychology statement that failure to control for effort test failure significantly biases data analysis for empirical studies (Bush et al., 2005).

Another limitation is that the NAB-S index scores are normed by both age and years of education, whereas WAIS-IV, WMS-IV and DKEFS test batteries are normed only by age.  This limits the precision of direct comparison of index scores from each battery.

**Conclusions**

Overall, our results suggest that the NAB-S and NAB-S TBI indices are good measures for the detection of cognitive impairment following TBI in the mild-complicated to severe range.  From a pragmatic clinical standpoint, the predictive validity of the NAB-S TBI is equivalent to the total NAB-S is a useful finding as it has half the administration time.  This finding from an outpatient TBI sample supports the use of the NAB-S TBI index developed for an acute inpatient TBI population by Hacker Jones et al (2016 in press).  It is also suggest that the index is sensitive enough to predict impairment in a group further along in their recovery.  However, the domain specific measures of the NAB-S do not all show sufficient convergent validity to the CVTB to afford interpretation on their own merit.  This is consistent with previous findings of low internal consistency for some of the NAB-S indices.  These findings suggest that if more extensive analysis of strengths and weaknesses post-injury is required, a more substantive test battery should be used.  The present study also builds on earlier NAB-S studies by controlling for participant effort.

## References

Azouvi, P., Vallat-Azouvi, C., Millox, V., Darnoux, E., Ghout, I., Azerad, S., … Jourdan, C. (2015). Ecological validity of the Dysexecutive Questionnaire: Results from the PariS-TBI study. *Neuropsychological Rehabilitation*, *25*(6), 864–878. http://doi.org/10.1080/09602011.2014.990907

Brooks, B. L., Iverson, G. L., & White, T. (2009). Advanced Interpretation of the Neuropsychological Assessment Battery with Older Adults: Base Rate Analyses, Discrepancy Scores, and Interpreting Change. *Archives of Clinical Neuropsychology*, *24*(7), 647–657. http://doi.org/10.1093/arclin/acp061

Burgess, P. W., & Alderman, N. (2004). Executive Dysfunction. In L. H. Goldstein & J. E. McNeil, *Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians*. John Wiley & Sons Ltd.

Bush, S., Ruff, R., Troster, A., Barth, J., Koffler, S., Pliskin, N., … Silver, C. (2005). Symptom validity assessment: Practice issues and medical necessity NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology*, *20*(4), 419–426. http://doi.org/10.1016/j.acn.2005.02.002

Cannizzaro, D. L., Elliott, J. C., Stohl, M., Hasin, D. S., & Aharonovich, E. (2014). Neuropsychological Assessment Battery-Screening Module (S-NAB): Performance in treatment-seeking cocaine users. *The American Journal of Drug and Alcohol Abuse*, *40*(6), 476–483. http://doi.org/10.3109/00952990.2014.916718

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. http://doi.org/10.1016/S0160-2896(01)00096-4

Crawford, J. R., & Garthwaite, P. H. (2008). On the “Optimal” Size for Normative Samples in Neuropsychology: Capturing the Uncertainty When Normative Data Are Used to Quantify the Standing of a Neuropsychological Test Score. *Child Neuropsychology*, *14*(2), 99–117. http://doi.org/10.1080/09297040801894709

Crawford, J. R., Garthwaite, P. H., Sunderland, D., & Borland, N. (2011). Some supplementary methods for the analysis of the Delis-Kaplan Executive Function System. *Psychological Assessment*, *23*, 888–898.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system (D-KEFS).* Psychological Corporation.

Donders, J., & Levitt, T. (2012). Criterion Validity of the Neuropsychological Assessment Battery after Traumatic Brain Injury. *Archives of Clinical Neuropsychology*, *27*(4), 440–445. http://doi.org/10.1093/arclin/acs043

Green, P., Allen, L., & Astner, K. (1995). *The Word Memory Test: A manual for the oral and computerized forms.* Edmonton, Canada: Green's Publishing Inc.

Grohman, K., & Fals-Stewart, W. (2004). The detection of cognitive impairment among substance-abusing patients: the accuracy of the neuropsychological assessment battery-screening module. *Experimental and Clinical Psychopharmacology*, *12*(3), 200–207. http://doi.org/10.1037/1064-1297.12.3.200

Hacker, D., Jones, C. A., Clowes, Z., Belli, A., Su, Z., Sitaraman, M., … Pettigrew, Y. (2016). The Development and Psychometric Evaluation of a Supplementary Index Score of the Neuropsychological Assessment Battery Screening Module that is Sensitive to Traumatic Brain Injury.

Heyanka, D. J., Thaler, N. S., Linck, J. F., Pastorek, N. J., Miller, B., Romesser, J., & Sim, A. H. (2015). A Factor Analytic Approach to the Validation of the Word Memory Test and Test of Memory Malingering as Measures of Effort and Not Memory. *Archives of Clinical Neuropsychology*, *30*(5), 369–376. http://doi.org/10.1093/arclin/acv025

Iverson, G. L., Holdnack, J. A., & Lange, R. T. (2013). Using the WAIS–IV/WMS–IV/ACS Following Moderate-Severe Traumatic Brain Injury. In *WAIS-IV, WMS-IV, and ACS* (pp. 485–544). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780123869340000109

Iverson, G. L., Williamson, D. J., Ropacki, M., & Reilly, K. J. (2007). Frequency of Abnormal Scores on the Neuropsychological Assessment Battery Screening Module (S-NAB) in a Mixed Neurological Sample. *Applied Neuropsychology*, *14*(3), 178–182. http://doi.org/10.1080/09084280701508952

Mckay, C., Casey, J., Wertheimer, J., & Fichtenberg, N. (2007). Reliability and validity of the RBANS in a traumatic brain injured sample. *Archives of Clinical Neuropsychology*, *22*(1), 91–98. http://doi.org/10.1016/j.acn.2006.11.003

Miller, J. B., Axelrod, B. N., Rapport, L. J., Millis, S. R., VanDyke, S., Schutte, C., & Hanks, R. A. (2012). Parsimonious prediction of Wechsler Memory Scale, Fourth Edition scores: Immediate and delayed memory indexes. *Journal of Clinical and Experimental Neuropsychology*, *34*(5), 531–542. http://doi.org/10.1080/13803395.2012.665437

Mollayeva, T., Kendzerska, T., Mollayeva, S., Shapiro, C. M., Colantonio, A., & Cassidy, J. D. (2014). A systematic review of fatigue in patients with traumatic brain injury: The course, predictors and consequences. *Neuroscience & Biobehavioral Reviews*, *47*, 684–716. http://doi.org/10.1016/j.neubiorev.2014.10.024

Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary clinical validity. *The Journal of Clinical and Experimental Neuropsychology*, *20*, 310–319.

Rao, V., & Lyketsos, C. (2000). Neuropsychiatric Sequelae of Traumatic Brain Injury. *Psychosomatics*, *41*, 95–103.

Sbordone, R. J., Liter, J. C., & Pettler-Jennings, P. (1995). Recovery of function following severe traumatic brain injury: A retrospective 10-year follow-up. *Brain Injury*, *9*(3), 285–299. http://doi.org/10.3109/02699059509008199

Sohlberg, M. M., & Mateer, C. A. (2001). Management of Dysexecutive Symptoms. In *Cognitive Rehabilitation: An Integrative Neuropsychological Approach*. New York: Guildford Press.

Stern, R. A., & White, T. (2003). *NAB, Neuropsychological Assessment Battery: psychometric and technical manual*. Psychological Assessment Resources Inc.

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 : [BI]*, *23*(1), 45–50. http://doi.org/10.1080/02699050802590361

Tombaugh, T., N. (1997). The Test of Memory Malingering (TOMM): Normative Data from Cognitively Intact and Cognitively Impaired Individuals. *Psychological Assessment*, *9*(3), 260–268. http://doi.org/http://dx.doi.org/10.1037/1040-3590.9.3.260

Wechsler, D. (2008). *Wechsler adult intelligence scale–Fourth Edition (WAIS–IV)*. San Antonio, TX: NCS Pearson.

Wechsler, D. (2009). *WMS-IV: Wechsler Memory Scale- Administration and Scoring Manual*. Psychological Corporation.

Wechsler, D. (2011). *Test of Premorbid Functioning UK Verson Manual*. Pearson Assessment.

World Health Organisation. (2001). *WHO | International Classification of Functioning, Disability and Health (ICF)*. World Health Organisation. Retrieved from http://www.who.int/classifications/icf/en/

Yates, P. J. (2006). An epidemiological study of head injuries in a UK population attending an emergency department. *Journal of Neurology, Neurosurgery & Psychiatry*, *77*(5), 699–701. http://doi.org/10.1136/jnnp.2005.081901

Zgaljardic, D. J., & Temple, R. O. (2010a). Neuropsychological Assessment Battery (NAB): Performance in a Sample of Patients with Moderate-to-Severe Traumatic Brain Injury. *Applied Neuropsychology*, *17*(4), 283–288. http://doi.org/10.1080/09084282.2010.525118

Zgaljardic, D. J., & Temple, R. O. (2010b). Reliability and Validity of the Neuropsychological Assessment Battery-Screening Module (NAB-SM) in a Sample of Patients with Moderate-to-Severe Acquired Brain Injury. *Applied Neuropsychology*, *17*(1), 27–36. http://doi.org/10.1080/09084280903297909

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. http://doi.org/10.1037/a0025466

# PUBLIC DOMAIN BRIEFING PAPER

# Validation of the Neuropsychological Assessment Battery Screening Measure (NAB-S) in participants with Traumatic Brain Injury

By

Thomas Morien Michael

Department of Clinical Psychology

School of Psychology

The University of Birmingham

July 2016

# SYSTEMATIC REVIEW:

## Psychometric properties of rating scales of executive dysfunction due to acquired brain injury

### Introduction

Executive function (EF) is a term psychologists use to talk about planning, organising, and self-control. EF can become impaired if we suffer a brain injury. EF is crucial to our ability to achieve goals, and this can make rehabilitation harder after suffering a brain injury. For this reason, psychologists need ways of measuring EF quickly, validly and reliably. Some psychologists say that tests of EF don't measure all of the difficulties that people can have with EF after brain injury. For this reason, questionnaires that people answer about themselves (if they have a brain injury) or their relatives (if the relative has a brain injury) have been developed.

### Aim

To systematically review all the published the literature on self-report and independently rated questionnaires about EF, and discuss them according to quality criteria, which suggest what a good questionnaire measure of EF needs to consider and measure well.

### Method

We searched three databases of psychological research, for research articles about four commonly used questionnaire scales of EF, combined with terms relating to validity and reliability. The resulting papers were then reviewed, and rated for quality, according to another quality criteria to make sure the studies were of good quality. Four other journal articles were found from the reference sections of these articles.

### Results

The search strategy resulted in twenty-one journal articles being included in the review. These were reviewed, rated for quality and summarised, before being discussed according to the important qualities that these questionnaires should measure. There were fourteen studies focussed on the Dysexecutive Questionnaire (DEX, 20 items), four on the Frontal Systems Behaviour rating scale (FrSBe, 46 items), two on the Brock Adaptive Functioning Questionnaire (BAFQ, 64 items) and one on the Behaviour Rating Inventory of Executive Function (BRIEF, 86 items).

## Conclusions

At present, most studies have been published on the DEX and FrSBe, but there is not enough published research on the BAFQ and BRIEF-A in a brain injured population to properly review these measures. The DEX and FrSBe questionnaires were good at measuring difficulties that people also had when they did psychological tests of EF. They also measure different types of EF difficulty that people can have quite consistently, and could predict real life difficulties quite well. This means that at present, the DEX and FrSBe are probably the best questionnaires to use with people after brain injury.

## EMPIRICAL PAPER:

# Validation of the Neuropsychological Assessment Battery Screening Measure (NAB-S) in participants with Traumatic Brain Injury

### Introduction

Traumatic brain injury (TBI) can be a common consequence of head injury. Head injury is a common cause of emergency department hospital admission in the UK, responsible for 3.4% of hospital admissions according to one study. Measuring the cognitive difficulties, such as memory difficulties, caused by TBI can be expensive and take a long time. For this reason, briefer screening tests have been developed, such as the Neuropsychological Assessment Battery Screening Measure (NAB-S).

### Aim

To compare the NAB-S longer and more widely used neuropsychological tests, to work out how well the NAB-S measures difficulties following TBI, both overall and in specific areas like memory.

### Method

Forty-four people with TBI were recruited from an outpatient clinic at a major UK hospital, and from a residential rehabilitation centre in the same city. They were then assessed using the NAB-S and the longer, more widely used tests, as well as being assessed using tests to estimate how well they were functioning before their TBI, and the effort they were putting into the tests.

### Results

Four people failed effort tests, so we didn't consider their results further. The NAB-S overall scores were similar to those on the longer and more widely used tests. This was particularly true for attention/concentration and memory scores. Some areas of the NAB-S scores were not as similar to those on the longer and more widely used tests, such as executive function tests. In addition, some tests were similar, when we might not expect them to be so. This might be because our thinking abilities in general help us with a lot of different areas that can be tested. The NAB-S was also very good at predicting overall difficulties as measured by the established battery.

## Conclusions

As it is quick to test people with, the NAB-S (30 minute administration time) could potentially be very useful to screen people for difficulties following TBI. However, some specific tests on the NAB-S were less good at predicting difficulties in comparison with their equivalent tests on the established test battery. This means psychologists sometimes need to use the longer and more detailed tests to measure specific difficulties that people can have.

# VOLUME I: APPENDICES

## Appendix I – Definitions of rating scale psychometric quality criteria

| **Criteria** | **Concise definition of rating scale psychometric quality criteria, based on papers by Mokking et al., (2010) and Terwee et al., (2007)** |
|---|---|
| Content Validity | A clear definition of the concept being measured, the measurement aim, and target population, as well as justification for selection and reduction of scale items, and interpretability of those items. |
| Criterion Validity | The extent to which the rating scale correlates with a "gold standard" criterion. A strong positive correlation of > 0.7 with that criterion measure would therefore suggest good criterion validity. |
| Internal Consistency | A measure of the extent to which individual items within a measurement scale or subscale are correlated or homogenous. Subscales within a measure can be explored or confirmed using appropriate factor analysis, and these subscales can then be rated for internal consistency using Cronbach's alpha; with a score of 0.7-0.95 being considered good |
| Construct validity<br>Structural validity<br>Hypotheses testing<br>Cross-cultural validity | The extent to which scores on a measure relate to other measures in a manner consistent with the theories by which each measure is constructed.<br>Exploration of structural validity, using exploratory and confirmatory factor analysis should be conducted referring to relevant theoretical models<br>This should be determined by specific hypothesis testing with group sizes of n > 50 and at least 75% agreement required for good construct validity. The extent to which the construct remains valid in cross-comparison with other cultures, including when translated into other languages, if appropriate. |
| Responsiveness | The extent to which a measure can detect clinically important changes over time. Terwee et al., (2007) consider this an aspect of longitudinal validity. Actual (true) change over time needs to be distinguished from measurement error (false change). On way to measure responsiveness is to use Area Under Receiver Operator Characteristics (ROC) Curve (AUC) to compare respondents which have changed vs. those which have not, according to an external criterion (another measure). |
| Interpretability | The extent to which qualitative meaning can be interpreted from quantitative scores. Interpretation of scores can be aided by a comparison of means and SDs of; a) a comparison group (norms), b) a contrast group, differing in terms of e.g. gender, severity, condition, c) a group having undergone a treatment or intervention. A sample size of n > 50 is required for such comparisons. |
| Reproducibility<br>Reliability<br>Measurement Error | Includes both reliability and measurement error, which need to be distinguished from one another.<br>The extent to which repeated measurement in a stable person can reproduce the same score. This is test-retest reliability, which needs to be at a time after which individual items cannot be remembered, but not so late that meaningful change might have occurred.<br>Measurement of error is critical to being able to determine reliability. Small measurement error is required for accurate and precise reliability. |
| Floor and ceiling effects | If > 15% of respondents achieve the lowest or highest score achievable on the measure, then the measure may have floor or ceiling effects which limit its content validity. More items, or more extreme ratings for items are thus required. A good scale will have no floor or ceiling effects for a sample of n > 50 participants. |

## Appendix II – Definitions of observational study methodology criteria

| Criteria | Concise definition of von Elm et al., (2008) STROBE criteria |
|---|---|
| **Background & Objectives** | Inclusion and exclusion criteria need to be clearly defined, and the sources and methods of recruitment need to be clearly described. |
| **Study Design** | The study design needs to be well described, early on in the paper, and should be appropriate to the aims and objectives of the study. |
| **Participant eligibility** | Participant eligibility should be appropriate to the objectives and design of the study. Recruitment criteria should be described. |
| **Sample Size (n > 50)**<br>**Factor Analysis Sample Size (n > 150 to 300)** | The sample size must be sufficient to enable sufficient statistical power to determine significance or non-significance of results. Accordingly, Terwee et al., (2007) suggest a sample size of n > 50 for most of their quantitative criteria.<br>For factor analytic studies von Elm et al., (2008) state that power calculations should be used. A sample of n < 150 is considered poor, n > 150-299 is considered acceptable, and n > 300 is considered good, according to a review of factor analytic sample size literature by Henson (2006). |
| **Study Setting & Data Source** | The setting for the study and the source(s) of data must be adequately described, including dates and methods of recruitment and data collection. The data source must be valid for the study objectives. |
| **Quantitative variables** | Quantitative variables used should be described and the rationale for their analysis detailed. |
| **Statistical Methods** | All statistical methods should be described, including analyses of subgroups, methods used to control for confounding, how any missing data were addressed, and statistical methods used in sampling strategy, if applicable. |
| **Potential Bias** | Potential sources of bias should be identified, as well as strategies used to reduce the possibility of bias, if present. |
| **Descriptive data** | Characteristics of study participants (e.g. demographic & clinical) should be described, and any missing data acknowledged. |
| **Main Results** | Statistical analyses should be reported correctly, including confidence intervals, if appropriate. |
| **Potential Limitations** | Potential limitations of the study should be discussed, as well as the magnitude and direction of any limitations, if possible to determine. |
| **Interpretation & Generalizability** | A cautious overall interpretation of the study should be given, considering limitations, analysis, and other relevant evidence. Generalizability (external validity) should be discussed. |

## Appendix III – Sub-tests of the Neuropsychological Assessment Battery Screening Tool

| NAB-S Indices | Sub-test comprising each index |
|---|---|
| Attention (S-ATT) | Digit Span Forwards (S-DGF) |
| | Digit Span Backwards (S-DGB) |
| | Sequencing Numbers & Letters Part A Speed (S-N&L(A)Spd) |
| | Sequencing Numbers & Letters Part A Errors (S-N&L(A)Err) |
| | Sequencing Numbers & Letters Part A Efficiency (S-N&L(A)Eff) |
| | Sequencing Numbers and Letters Part B Efficiency (S-N&L(B)Eff) |
| Language (S-LAN) | Auditory comprehension (S-AUD) |
| | Naming (S-NAM) |
| Executive function | Mazes (S-MAZ) |
| (S-EXE) | Word Generation (S-WGN) |
| Memory (S-MEM) | Shape Learning Immediate Recognition (S-SHL-Irg) |
| | Shape Learning Immediate Recognition (S-SHL-Drg) |
| | Story Learning Immediate Recall (S-STL-irc) |
| | Story Learning Delayed Recall (S-STL-drc) |
| Spatial (S-SPT) | Design Construction (S-DES) |
| | Visual Discrimination (S-VIS) |