Cognitive & Behavioral Assessment

Improving the quality of cognitive screening assessments: ACEmobile, an iPad-based version of the Addenbrooke's Cognitive Examination-III

Craig G. J. Newman[a,*], Adam D. Bevins[b], John P. Zajicek[c], John R. Hodges[d], Emil Vuillermoz[e], Jennifer M. Dickenson[b], Denise S. Kelly[b], Simona Brown[f], Rupert F. Noad[g]

[a]Plymouth University Peninsula Schools of Medicine and Dentistry (PU PSMD), Plymouth, United Kingdom
[b]Older People's Psychology and Psychological Therapies Department, Devon Partnership NHS Foundation Trust, Exeter, United Kingdom
[c]School of Medicine, Medical & Biological Sciences, St Andrews, United Kingdom
[d]The University of Sydney, Brain & Mind Centre, Sydney, Australia
[e]School of Psychology, Plymouth University, Plymouth, United Kingdom
[f]Devon Partnership NHS Trust, OPMH Teignbridge Team, Exeter, United Kingdom
[g]Plymouth Hospitals NHS Trust, Department of Neuropsychology, Level 7, Derriford Hospital, Plymouth, United Kingdom

Abstract

Introduction: Ensuring reliable administration and reporting of cognitive screening tests are fundamental in establishing good clinical practice and research. This study captured the rate and type of errors in clinical practice, using the Addenbrooke's Cognitive Examination-III (ACE-III), and then the reduction in error rate using a computerized alternative, the ACEmobile app.

Methods: In study 1, we evaluated ACE-III assessments completed in National Health Service (NHS) clinics (n = 87) for administrator error. In study 2, ACEmobile and ACE-III were then evaluated for their ability to capture accurate measurement.

Results: In study 1, 78% of clinically administered ACE-IIIs were either scored incorrectly or had arithmetical errors. In study 2, error rates seen in the ACE-III were reduced by 85%–93% using ACEmobile.

Discussion: Error rates are ubiquitous in routine clinical use of cognitive screening tests and the ACE-III. ACEmobile provides a framework for supporting reduced administration, scoring, and arithmetical error during cognitive screening.

© 2017 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer's Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Screening assessment; Cognitive assessment; Alzheimer's; Dementia; Computerized; App; Usability; Validity; Administrator error

1. Introduction

The psychometric properties of cognitive screening tools for dementia are routinely reported, yet far less is known about the clinician's ability to administer and score these tests accurately. Evidence suggests that users make many more errors than expected [1–3]. There is surprisingly little detail in the literature on how well the cognitive screening tests perform in the hands of the clinicians for whom they are designed.

Despite the brevity and perceived simplicity of two of the most commonly used cognitive assessment instruments in the United Kingdom—the Mini–Mental State Examination [4] and the Addenbrooke's Cognitive Examination–Revised (ACE-R) [5]—test scoring simulation studies have revealed high rates of errors on both measures [6,7]. Both the Mini–Mental State Examination and ACE-R use cutoffs for determining caseness, and this influences subsequent diagnostic/treatment pathways, highlighting the importance of accurate assessment.

Computerized approaches to cognitive assessment have the potential to improve the standards of administration, scoring, and reporting, as by automating processes, the possibility of human error is reduced. For example, it is possible to more closely control administration instructions, thus reducing the chance of intrarater and interrater variation. The scoring of a test can also be supported or automated, reducing the chance of arithmetical errors.

The use of a product in real-world settings cannot be presumed and must be tested. Such *usability* research is a worldwide standard in medical device evaluation [8–10] but until recently has been largely neglected in the validation of cognitive screening assessments. Furthermore, there are few studies that have evaluated the use of these tools in routine clinical practice.

The present study set out to explore the nature of scoring errors using the ACE-III in routine National Health Service (NHS) practice. This was followed by a comparison of the ACE-III and ACEmobile, a new iPad-based version developed by the authors. The aim was to assess the ability of each tool to support a clinician to capture accurate measurement, that is, the hypothetical score that the patient would receive with zero measurement error. ACEmobile was designed to support users of the ACE-III by guiding and automating the administration, rule adherence, scoring, and reporting.

1.1. Aims and hypotheses

Aim 1: To establish the presence, nature, and extent of scoring errors on the ACE-III in standard clinical practice, via the analysis of completed ACE-III assessments from NHS memory assessments (study 1).
Hypothesis 1: High rates of administration and arithmetical errors will be observed in ACE-III assessments from NHS memory clinics.
Aim 2: To compare the measurement accuracy of ACE-III and ACEmobile (study 2).
Hypothesis 2: Administration and reporting errors will be significantly less for ACEmobile than for ACE-III.

2. Study 1: Identification of scoring errors on the ACE-III in standard NHS clinical practice

2.1. Sample

ACE-III scoresheets ($N = 132$) were extracted from the electronic patient records of four Older People's Community Mental Health teams in Devon, UK (with NHS ethical approval). Of these, 45 (34%) were subsequently excluded from the analysis because they were not suitable for further analysis (i.e., incomplete assessments, scores omitted, illegible, and older version of ACE used [i.e., ACE-R]). A total of 87 ACE-III scoresheets were subsequently analyzed. The ACE-IIIs were administered by community psychiatric nurses (63%, $n = 55$), psychiatrists (26%, $n = 23$), and occupational therapists (10%, $n = 9$). Details of specific training undertaken

by each administrator were not collected but were assumed to be the standard required for that clinical service. This was deemed to be representative of standard NHS clinical practice.

2.2. Measures

The ACE-III is a cognitive screening tool to detect mild dementia and distinguish between Alzheimer's disease and frontotemporal dementia [11]. It contains 24 individual test items contributing to five subdomains—attention (18 points), memory (26 points), fluency (14 points), language (26 points), and visuospatial functioning (16 points), with a total score of 100. The ACE-III shows high sensitivity and specificity for dementia using a cutoff of 88 or 82, respectively [12].

2.3. Procedure

Two anonymized copies of each ACE-III were produced. Rescoring was conducted by two raters, strictly following the published scoring guidelines. The two data sets were compared for consistency using an Excel formula. There were 121 discrepancies between raters, equating to an error rate of 2.58%. The second author adjudicated on the discrepancies to reconcile the differences and produce a single data set with an accurate score at the individual item level, subdomain level, and ACE-III total score.

Data were double entered, and any discrepancies were adjudicated by the lead researcher. Finalized ACE-IIIs were then compared back to the clinician-scored ACE-IIIs. *Scoring errors* (points deducted or added in error by each clinician, for each subtest), *arithmetic errors* (mental arithmetic errors made in adding the scores together), and *total error* (scoring and arithmetic errors combined) were calculated.

2.4. Results

The range of clinician ACE-III total scores in the sample was from 30 to 88 points ($\mu_x = 64.80$, SD $= 13.24$).

Scoring errors were observed in 68% of the ACE-IIIs. Arithmetic errors were observed in 24% of ACE-IIIs, with a range of −10 to 10. Only 22% of ACE-IIIs had no errors at all. The total error rate ranged from 0 to 22, with a mean of 3.3 (SD $= 4.2$). In 22% of the sample, the total error rate was 5 or more points (Fig. 1).

At the subdomain level, 46% and 44.8% of clinicians made at least one error on the visuospatial and language domains, respectively. Errors were present but observed less frequently for the memory (20%), fluency (15%), and attention and orientation (12%) domains. At the individual item level, 39% of clinicians made at least one error on sentences, 34% on clock drawing, and 11% on animal fluency.

2.5. Summary

In NHS settings, clinician errors in scoring, mental arithmetic, and reporting the ACE-III were commonplace. However, this is likely to be an underestimate of the error rate because

it did not include administration accuracy (e.g., using the wrong verbal instructions, incorrectly prompting, not following the rules in the manual). This is addressed in study 2.

3. Study 2: Usability study—Evaluation of ACEmobile versus ACE-III, on a user's ability to accurately measure cognitive functioning

3.1. Sample

According to usability research methodology [13], to capture 94% of user errors, a sample size of 10 per group was required (assuming 25% of users demonstrate errors). This number was considered more than adequate, given the high rates of scoring errors observed in study 1.

The sample included trainee clinical psychologists (trainees) and PhD/postgraduate health science students. Participants were recruited during training sessions and randomly allocated to an ACE-III or ACEmobile group. The ACE-III group ($N = 10$ mean age 25) included five trainees and five postgraduates, and the ACEmobile group ($N = 11$, mean age 27) included six trainees and five postgraduates. Of these, one ACEmobile participant and three ACE-III participants reported previous experience using ACE-III in a clinic setting.

3.2. Procedure

The ACE-III group was asked to read the ACE-III administration manual alongside the test sheets. This was considered the minimal training requirement that could be expected for a new user of the ACE-III. The ACEmobile group was asked to complete the in-App automated training module.

Using data from study 1, a mock patient script was generated that would test the rater's accuracy of administration, scoring, and reporting of the test. This script was used by the experimenter in the role of a mock patient. The development of the script was informed by patient responses and scoring errors from study 1. The mock patient's total score was 1 point below the mean ACE-III score in study 1 (i.e., 64/100).
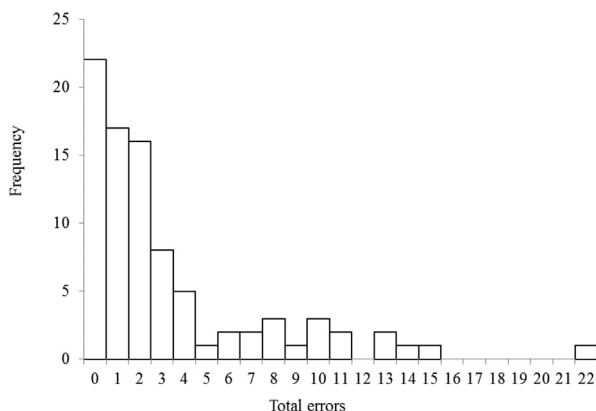

Fig. 1. Histogram showing the frequency of total errors in 87 ACE-IIIs from NHS Older People's Mental Health teams. Abbreviations: ACE-III, Addenbrooke's Cognitive Examination-III; NHS, National Health Service.

Participants were instructed to complete their assessment as if they were "working in a memory clinic." Participants were instructed to prioritize accuracy over time. Participants in the ACE-III group were allowed access to the administration manual throughout.

3.3. Data preparation

Inspection of the data revealed that 90% of ACE-III and 75% of ACEmobile participants scored the cube subtest incorrectly, making the same error (awarding 1 point rather than 0). It was later considered that the cube drawn by the mock patient was not representative of a likely dementia performance and was too difficult to match with the scoring criteria. These data were considered an artifact of the experimental design, and so in the following analysis, all participants were credited with a correct score.

3.4. Results

Table 1 summarizes the difference between ACEmobile and ACE-III scoring accuracy. The table provides summaries of ACE-III "total error" comprising administration, scoring, and arithmetic errors. The ACEmobile "total error" is the sum of administration and scoring errors only as the arithmetic component of the measure is automated and computerized.

3.4.1. Administration/scoring errors

The total error rate between the ACEmobile group ($\mu_x = 1.5$ points) and the ACE-III group ($\mu_x = 7.4$ points) revealed significantly more errors in the ACE-III group [$U(19) = 7$, $Z = -3.3$, $P < .01$]. The range of total errors was large for the ACE-III group, ranging from 2 to 16 points.

3.4.2. Arithmetic scoring errors

For the total reported scores, none of the participants in the ACE-III group summed up correctly. The mean arithmetic error rate was -4.1 points (SD = 5.2) with a range of 1 to -12 points.

Comparison of errors made within each subdomain is shown in Table 2.

For the ACE-III group, errors were observed across all domains. Errors for ACEmobile were mostly in the language domain, in particular, the naming task (44%, $N = 7$) and reading task (19%, $N = 3$). A usability issue was observed for the naming task, which is discussed below.

3.4.3. Time to complete assessments

Assessment time, from initiation to final reported score, was on average 26 minutes in the ACE-III group (SD = 6.2) and 21 minutes in the ACEmobile group (SD = 2.5). This was a significant difference favoring the ACEmobile group (t value 2.20; $P < .05$). This includes arithmetic calculation time for the ACE-III, which was not required by ACEmobile.

Table 1
Scoring accuracy for ACEmobile and ACE-III, with and without arithmetic errors (excluding cube scoring)

	ACEmobile (n = 11) Administration/scoring errors (mean [SD], range)	ACE-III (n = 10) Administration/scoring errors (mean [SD], range)	ACE-III (n = 10) Administration plus arithmetic errors (mean [SD], range)
Total scoring errors per rater	1.5 (1.6), 0–4	7.7 (3.9)*, 2–16	10.3 (4.8)*, 4–20

Abbreviations: ACE-III, Addenbrooke's Cognitive Examination-III; SD, standard deviation.
NOTE. Arithmetic errors for ACEmobile are zero as this step is automated.
*Significant difference compared to ACEmobile performance, $P < .01$.

3.4.4. ACEmobile usability error—Results adjustment

The naming task accounted for 44% of ACEmobile administration/scoring errors. A significant proportion (35%) of the group showed similar errors on the first few items of this subtest, seemingly reflecting a usability error with the App that was self-corrected by all participants (it is not clear which side of the screen to tap for "correct" and "incorrect," and they were not given a prompt until after tapping). Adjusting for the usability errors by removing the incorrect naming items reduced the distance from the correct score from a mean of 2.5 to 0.6. The revised score was significantly different from the ACE-III group [U(19) = 10.5,

Z = −3.1, $P < .01$]. The total rate of scoring errors was reduced from a mean of 1.5 to 0.8 following this correction.

4. Discussion

Study 1 demonstrated that scoring and arithmetic errors on cognitive screening tests are made in routine clinical practice by specialist staff. Error rates are high with 22% of ACE-IIIs containing error rates of 5 points or more. Sentences, clock drawing, and animal fluency subtests showed the highest error rates. Study 1 was limited in its ability to capture administration error because it was not possible to observe clinicians conducting the test. Study 2 assessed administration of ACE-III and ACEmobile. Error rates were significantly higher in the ACE-III group.

Both studies reveal that ACE-III, when used by appropriately skilled individuals, is still prone to rater error. In study 2, the participants of the ACE-III group were provided with time to familiarize themselves with the manual and were able to refer to it during the assessment. In the context of a challenging cognitive profile, with an examinee script that tested areas known to be susceptible to error, administration using ACE-III revealed large margins of error. In this context, ACEmobile increased the accuracy of measurement by a minimum of 85%. This increased to 93% accuracy when a usability issue with the naming subtest was resolved.

The impact of these findings is relevant in numerous contexts. In clinical and research settings, the value of a single score in supporting diagnosis, disease staging, and prescribing is essential. In situations requiring serial assessment (e.g., follow-up appointments), the ability to measure disease progression and evaluate the effectiveness of medication is dependent upon the accuracy of a screening tool to capture change over time. The existence of an error rate of 10.3 points (ranges 4–20) would potentially undermine a clinician's ability to be confident in the reliability of the obtained score, where a score of 82 suggests clinical impairment and error impacts on clinical decision making. ACEmobile demonstrates that a computer-supported approach, which is underpinned by a clear understanding of how the tool performs in the hands of the user, can improve the accuracy of measurement, even with a challenging score profile.

The method in this article is arguably both a model for future development of existing or new tools and an invitation for other cognitive assessments to be similarly evaluated.

Table 2
Summary of domain and sub-domain errors on ACE-III and ACEmobile

Domain	Subtest	ACE-III percentage of total errors (number of errors)	ACEmobile percentage of total errors (number of errors)
Orientation/attention	Orient 1	6 (5)	0 (0)
	Orient 2	1 (1)	0 (0)
	Registration	1 (1)	13 (2)
	Attention	14 (11)	0 (0)
	Domain total	23 (18)	13 (2)
Visuospatial	Infinity	0 (0)	0 (0)
	Clock	4 (3)	0 (0)
	Dots	6 (5)	6 (1)
	Frag letters	0 (0)	0 (0)
	Domain total	10 (8)	6 (1)
Memory	LKB recall	0 (0)	0 (0)
	Anterograde	5 (4)	0 (0)
	Retrograde	1 (1)	0 (0)
	Delayed	5 (4)	0 (0)
	Recog	17 (13)	0 (0)
	Domain total	29 (22)	0 (0)
Language	Comprehension	5 (4)	6 (1)
	Writing	4 (3)	6 (1)
	Word repetition	5 (4)	6 (1)
	Phrase repetition	1 (1)	0 (0)
	Naming	5 (4)	44 (7)
	Comprehension pics	5 (4)	0 (0)
	Reading	1 (1)	19 (3)
	Domain total	27 (21)	81 (13)
Fluency	Phonetic	4 (3)	0 (0)
	Semantic	6 (5)	0 (0)
	Domain total	10 (8)	0 (0)
Total		100 (77)	100 (16)

Cognitive screening has historically been perceived as intuitively simple, with paper tools freely available to download. It is feasible to consider that tools have been developed that are intuitive to their authors yet complex and challenging to those who try to use them in the real world. This is particularly pertinent in the context of increasing pressure on services to see more patients with fewer resources and greater reliance on staff with less training and experience in the use of such tests. There is a need for human factors research to be an integral part of the evaluation of current cognitive assessment and design of future measures.

Undertaking this study highlighted some issues with both the ACE-III and ACEmobile. Problems with sentences and clock drawing highlighted in study 1 were communicated to the developers of the ACE-III. The scoring criteria were subsequently updated, and these changes were incorporated into ACEmobile. It was also clear that ACEmobile did have some inherent administration error and, after evaluation, it was attributed to training failures. To address these, a training video was made and has now been embedded into the App.

A limitation of the article is the use of skilled but not specialist participants in study 2 insofar as the observed error rate could have been an artifact of experience. Although this is possible, it was felt this method was not outside of ecological validity. The amount of training available to participants was felt to reflect routine clinical practice. In addition, participants in study 2 had ready and continued access to the manual and replicated the observed error rates seen in study 1.

The delivery of this study highlighted the challenges of scoring where subjective judgments are required, as with visuospatial tests, particularly the cube. This emphasizes the need for further research to either better support scoring in ACEmobile or to introduce less subjective assessment paradigms or more automated scoring technologies (machine learning, etc.).

It is worth acknowledging that in study 2, ACEmobile administration was supported by a training video, but an equivalent was not provided for ACE-III. This adds an additional confound. However, this was intentional as the study sought to create a context that simulated standard clinical practice as closely as possible.

This study also has an implication for research where the challenge of achieving high levels of interrater reliability has been well documented. Inconsistent administration and scoring between raters offer a major threat to the success of clinical trials. Using automation, ACEmobile significantly reduces the chance of interrater and intrarater error and offers a potentially useful tool to support outcome research in the field of dementia.

There are potential drawbacks of computerized assessments. These include forgetting passwords, battery failures, software failures, and loss of internet connection, which can all threaten the utility of computerized tests. Any additional value added by using a computerized test needs to be weighed against the ease of the use of a paper-and-pencil test.

Finally, there is a need for further evaluation and validation of ACEmobile to ensure that the changes in data collection described do not have impact on the ACE-III's ability to discriminate dementia and/or change the normative data and established cutoffs.

5. Summary

Administration, scoring, and reporting errors are common in routine use of cognitive screening tools. ACEmobile is very effective at reducing errors when compared with the standard paper-and-pen test. This is likely to offer considerable value in clinical and research settings where the ability to accurately measure cognition is a key component of the diagnostic process or the success of a clinical trial in demonstrating a treatment effect.

ACEmobile is currently provided as a free tool, with no restrictions for clinical use, available on iTunes and is already registered for use in over 1100 clinical settings worldwide.

Acknowledgments

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature using traditional sources (e.g., PubMed), conference proceedings and meeting abstracts. A small number of citations highlight the importance of usability research in the evaluation of medical devices (i.e., how a person interacts with the systems and technologies they use). Nonetheless, this promising approach has hitherto been neglected in the validation of cognitive tests.

2. Interpretation: Our findings reveal the nature and extent of errors on a commonly used cognitive screening tool in NHS memory clinics (ACE-III). These same errors were largely overcome by a computerized assessment (ACEmobile), designed and evaluated using a human factors approach.

3. Future directions: The manuscript calls for the use of human factors approaches in the comprehensive evaluation of new and existing cognitive screening tests (both traditional pen-and-paper and digital assessment tools). In clinical and research contexts, computerized cognitive tests have the potential to dramatically reduce scoring errors.

References

[1] Sherrets S, Gard G, Langner H. Frequency of clerical errors on WISC protocols. Psychol Sch 1979;16:495–6.

[2] Sullivan K. Examiners' errors on the Wechsler Memory Scale–Revised. Psychol Rep 2000;87:234–40.

[3] Kozora E, Kongs S, Hampton M, Zhang L. Effects of examiner error on neuropsychological test results in a multi-site study. Clin Neuropsychol 2008;22:977–88.

[4] Folstein MF, Folstein SE, McHugh PR, Roth M, Shapiro MB, Post F, et al. Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.

[5] Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR. The Addenbrooke's Cognitive Examination revised (ACE-R): a brief cognitive test battery for dementia screening. Int J Geriatr Psychiatry 2006; 21:1078–85.

[6] Crawford S, Whitnall L, Robertson J, Evans JJ. A systematic review of the accuracy and clinical utility of the Addenbrooke's Cognitive Examination and the Addenbrooke's Cognitive Examination-Revised in the diagnosis of dementia. Int J Geriatr Psychiatry 2012;27:659–69.

[7] Queally VR, Evans JJ, McMillan TM. Accuracy in scoring vignettes using the mini mental state examination and the short orientation memory concentration test. J Geriatr Psychiatry Neurol 2010; 23:160–4.

[8] Hegde V. Role of human factors/usability engineering in medical device design. Reliab Maintainab Symp (RAMS) 2013:1–5.

[9] Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. Annales francaises d'anesthesie et de reanimation 2000;21:453–4.

[10] Medicines and Healthcare Products Regulatory Agency. Human Factors and Usability Engineering – Guidance for Medical Devices Including Drug-device Combination Products, version 1.0. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/645862/HumanFactors_Medical-Devices_v1.0.pdf. Published September 2017. Accessed February 6, 2018.

[11] Hsieh S, Schubert S, Hoon C, Mioshi E, Hodges JR. Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. Dement Geriatr Cogn Disord 2013; 36:242–50.

[12] Larner AJ, Mitchell AJ. A meta-analysis of the accuracy of the Addenbrooke's Cognitive Examination (ACE) and the Addenbrooke's Cognitive Examination-Revised (ACE-R) in the detection of dementia. Int Psychogeriatr 2014;26:555–63.

[13] Sauro J, Lewis JR. Quantifying the user experience. Practical statistics for user research. Waltham, MA: Morgan Kaufmann; 2012. p.312.