Title:

Inter-rater variability in scoring of Addenbrooke's Cognitive Examination – third edition (ACE-III) protocols

Authors

Miranda Julia Say[1,2]; Ciarán O'Driscoll[3]

Institutional affiliations:

1. Barts Health NHS Trust, London, UK
2. Concord Repatriation General Hospital, Sydney, Australia
3. Centre for Outcomes Research and Effectiveness (CORE), Research Department of Clinical, Educational, and Health Psychology, University College London, London, UK

Corresponding author's email: Miranda.Say@health.nsw.gov.au

This is a pre-copyedited, author-produced version of an article accepted for publication in Applied Neuropsychology: Adult following peer review. The version of record Say, M. J., & O'Driscoll, C. (2022). Inter-rater variability in scoring of Addenbrooke's Cognitive Examination-(ACE-III) protocols. Applied Neuropsychology: Adult, 1-5 is available online at: https://doi.org/10.1080/23279095.2022.2083964

Abstract

Background: Despite its wide use in dementia diagnosis on the basis of cut-off points, the inter-rater variability of the ACE-III has been poorly studied.

Methods: 31 healthcare professionals from an older adults' mental health team scored two ACE-III protocols based on mock patients in a computerised form. Scoring accuracy, as well as total and domain-specific scoring variability, were calculated; factors relevant to participants were obtained, including their level of experience and self-rated confidence administering the ACE-III.

Results: There was considerable inter-rater variability (up to 18 points for one of the cases), and one case's mean score was significantly higher (by nearly four points) than the true score. The Fluency, Visuospatial and Attention domains had greater levels of variability than Language and Memory. Higher scoring accuracy was not associated with either greater levels of experience or higher self-confidence in administering the ACE-III.

Conclusions: The results suggest that the ACE-III is susceptible to scoring error and considerable inter-rater variability, which highlights the critical importance of initial, and continued, administration and scoring training.

Key words:

Inter-rater variability ACE-III scoring dementia

Introduction

The Addenbrooke's Cognitive Examination – third edition (ACE-III) is a widely-used, free cognitive screening tool. Along with its two predecessors (the ACE and the ACE-R), the ACE-III has been frequently relied upon to distinguish dementia from healthy ageing, and to aid in differential diagnosis of dementia subtypes and other cognitive disorders (Matias-Guiu et al., 2017; Mioshi et al., 2006; Dudas et al., 2005). The ACE-III is administered by a range of healthcare professionals including nurses, psychologists, doctors and occupational therapists. However, use of the ACE-III is not limited to those with training in its administration or scoring. Understanding variations in scoring of the test is important for appreciating its reliability in diagnosing dementia.

Surprisingly, the inter-rater reliability of the ACE-III has been scantly investigated. This is highly relevant as the briefer Mini-Mental Status Examination (MMSE) has been demonstrated to have poor inter-rater reliability amongst clinicians who have not received standardised training (Bowie et al., 1999; Molloy & Standish, 1997). In fact, even small sections of the MMSE (e.g. spelling WORLD backwards; serial 7s backwards) had poor inter-rater reliability amongst neurologists (Davey & Jamieson, 2004). To our awareness, no studies have been published assessing previous or current ACE versions for inter-rater reliability in English. Where established in translated versions, inter-rater reliability has been shown to be high (e.g., an intra-class correlation [ICC] = 0.996 for the Japanese version [Takenoshita *et al*., 2019] and ICC = 0.92 for a Greek version [Kourtesis et al., 2020]). However, these studies used only a few highly trained clinicians. Examining the reliability between the scores of two clinicians who work in the same clinic or trial may over-represent accuracy across multiple raters and, hence, be unreflective of routine scoring in a more diverse clinical setting.

A review of ACE-III scoring within clinical practice highlighted substantial errors (Newman et al., 2017), and it has been argued that there is likely to be a lack of standardised implementation in general clinical practice (Jones et al., 2020). In contrast to inter-rater reliability, inter-rater variability, which measures the amount of variability in the scores of numerous raters scoring the same case, provides a more clinically-meaningful insight into how widely the scores differ when clinicians are scoring the ACE-III. Score variability is particularly relevant given the clinical implications of using specific cut-off points to distinguish between dementia versus mild cognitive impairment (MCI) versus no cognitive impairment (Potts et al., 2021; Hsieh et al., 2013).

The primary aim of the current study was to establish the accuracy and variability in raters' scoring of the ACE-III, and the variability within each domain being assessed (e.g. memory, attention). Secondary to this, we aimed to establish whether the level of rater accuracy is influenced by experience (time and frequency), and by self-rated confidence, administering the ACE-III.

Materials and Methods

Participants

The study used purposive sampling to obtain a diverse sample of mental health professionals. Participants were invited to take part if they were a health professional who routinely used the ACE-III with patients in the older adults' mental health service of a London Trust. Thirty-three people agreed and consented to participate, with thirty-one completing in the study.

Materials

Participation was online and anonymous. The first section contained a question asking the participant to identify their profession; two Likert-scale questions related to (i) how frequently they administered the ACE-III ('Frequency': 1= less than monthly; 2 = once a month; 3 = a few times a month; 4 = a few times a week), and (ii) the duration of their ACE-III administration experience ('Duration': 1 = less than six months; 2 = six months to a year; 3 = more than a year); and a subjective rating of their scoring Confidence (from 0 [no confidence] to 10 [most confident]). The second section contained two mock (completed) ACE-III (UK version A) forms that were to be scored online by the participant. Participants would have been able to access scoring guidelines online. This was not mentioned in the participant instructions. The overall experience of practitioners ('Experience') was calculated by multiplying each participant's Frequency of using the ACE-III by Duration of use.

The two cases' true scores were 74 and 84 respectively (hereafter referred to as Case 74 and Case 84). True scores were agreed by both authors (a clinical neuropsychologist and a clinical psychologist specialising in neuropsychological assessment), according to the scoring criteria. The two ACE-III protocols were presented in randomised order. Item responses that were presented in still visual form (i.e. the participant viewed a transcript of the mock patient's verbal response or viewed the mock patient's writing or drawing) were Attention: orientation and subtraction; Memory: recognition; Language: writing, naming, and semantic knowledge; and all Visuospatial items. Simulated auditory verbal responses were presented for: Attention: repetition; Memory: repetition, learning, semantic memory, and delayed recall; both Fluency items; and Language: single-word repetition, sentence repetition, and irregular word reading. Finally, the praxis item from Language was presented in the form of

a video recording of simulated responses. It should be noted that responses were made challenging to score in the following ways: for Fluency subtests, some responses were repeated and variations of a root word were generated. Participants scored each response and then added these to form a total score for each item. Domain scores and the total score were generated automatically by the computer programme. Scoring accuracy (Accuracy) was calculated for each case by subtracting a participant's total score from the case's true score.

The study was approved by the ethics review committee of the Joint Research Management Office for Barts Health NHS Trust and Queen Mary University of London (ref. QMREC1249).

Results

The participants' range of health professions is representative of a diagnostic dementia service (see Table 1). The majority (74.2%) had administered the ACE-III for greater than six months; and 61.3% administered it at least once a month. Administrator confidence ranged considerably (from 1-9/10).

Table 1: Participants' profession, duration of experience administering the ACE-III, frequency of administering the ACE-III, administration confidence; and case total scores.

Scoring accuracy (see Figure 1)

For Case 74, the mean participant score (77.8) was significantly higher than the true score [$t(30) = 6.23$, $p<0.0001$, $d = 1.12$]. There was notable variability in scores for Case 74,

ranging from 71 to 86. For Case 84, there was no significant difference between the mean participant score (83.4) and the true score [$t(30) = -0.905$, $p<0.38$, $d = 0.15$], though the score range was very wide (75 to 93).

In order to assess the relative variability within cognitive domains, the coefficient of variation (the degree of variability relative to the mean) for subtests was calculated for each Case protocol separately (Figure 1). Subtest variability in each Case (74 and 84) was (respectively) higher for Fluency (9.13%, 13.31%); Visuospatial (12.92%, 9.90%); and Attention (9.26%, 11.75%), and lower for Language (6.5%, 5.32%); and Memory (4.79%, 5.89%) subtests.

There was a moderate to high positive correlation between Experience and Confidence ($r=0.63$, $p<0.001$). The association between Experience and Accuracy was insignificant for both Case 74 ($r=0.30$, $p=0.10$) and Case 84 ($r=0.34$, $p=0.07$). There was a significant, but negative correlation between Confidence and Accuracy for Case 84 ($r= -0.38$, $p< 0.04$), where higher confidence was related to lower accuracy. This association was insignificant for Case 74 ($r = 0.08$, $p=0.67$).

Figure 1: (top row) Coefficient of variation for subscales in each test, higher values represent a greater degree of relative variability; (bottom row) accuracy plots for each test, dotted lines are 95% limits of agreement, solid line represents true score.

Discussion

The present results suggest that the ACE-III can be susceptible to considerable inter-rater scoring variability, with the two cases' scores varying by 15 and 18 points. The mean participant score for Case 74 was significantly higher than the true score by nearly four points. These findings highlight potential limitations in terms of relying on the total score (in particular in relation to specific cut-off scores) for dementia and MCI diagnosis. Test-retest accuracy is also a possible issue with such scoring variation affecting the ability to accurately monitor progression over time.

For both cases, there was a similar pattern regarding which subtests were more susceptible to scoring error. High levels of variability were not surprising for Fluency, given that the scoring can be prone to misinterpretation and requires close review of the scoring instructions. The Visuospatial subtest scoring was also more variable, largely owing to disagreement on scoring of the drawings (as opposed to the dot counting or fragmented letter identification).

Unsurprisingly, the amount of experience (combination of duration and frequency) of ACE-III administration correlated with participants' confidence in administering the test. Interestingly, neither greater experience nor greater confidence was associated with higher levels of scoring accuracy. Indeed, greater confidence was associated with more errors for Case 84. Experienced clinicians employ heuristics, and it is possible that overconfidence leads to cognitive errors (Croskerry, 2003).

There are limitations of the study. Studying inter-rater variability of a cognitive task is complicated by the difficulty replicating the exact conditions for different raters. We attempted to address this by developing artificial, computerised representations of simulated

patient protocols. Concernedly, we note that additional variability in scores is likely to be introduced as a result of inconsistencies in administration, and also errors made when adding up item scores to calculate domain and total scores, which were not captured in this study. The range of professions reflects those routinely undertaking cognitive assessments in clinical practice. Administration of the ACE-III requires limited training; however, background training in psychometrics and cognitive assessment could contribute to variance. The small sample does not allow us to examine this robustly. The study focuses on scores relevant to clinical decision making where the ACE-III cut-off may inform diagnosis. It may also be of interest to explore the inter-rater variability in cases with a very high or very low score.

The results highlight the critical importance of initial, and continued, training in ACE-III administration and scoring. Scoring variability is likely to be even greater in real-life settings in view of possible deviation from uniform administration, as well as other distractions that occur in clinical practice. The use of ACEmobile may help to reduce variability and errors (Newman et al., 2017). The data also highlight the importance of taking a more tentative interpretation than one based solely on cut-offs for dementia and MCI diagnosis. Larger scale studies focusing on inter-rater reliability and variability of scoring and administration of the ACE-III are certainly warranted.

Acknowledgements

We thank the clinician participants for their time, and A/Prof Laurie Miller for her comments on an earlier version of the manuscript.

Declaration of Interest Statement

The authors report there are no competing interests to declare.

References

Bowie, P., Branton, T. & Holmes, J. (1999). Should the Mini Mental State Examination be used to monitor dementia treatments? *Lancet, 354.* 1527–1528.

Davey, R.J. & Jamieson, S. (2004). The validity of using the mini mental state examination in NICE dementia guidelines. *Journal of Neurology, Neurosurgery and Psychiatry, 75,* 343-344.

Dudas, R.B., Berrios, G.E. & Hodges, J.R. (2005). The Addenbrooke's Cognitive Examination (ACE) in the differential diagnosis of early dementias versus affective disorder. *American Journal of Geriatric Psychiatry, 13(3),* 218-226.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures,1(1),* 77-89.

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E. & Hodges, J.R. (2013). Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *36(3-4)*, 242-50.

Jones, D., Wilkinson, R., Jackson, C. & Drew, P. (2020). Variation and interactional non-standardisation in neuropsychological tests: The case of the Addenbrooke's Cognitive Examination. *Qualitative Health Research*, *30(3),* 458-470.

Kourtesis, P., Margioti, E., Demenega, C., Christidi, F. & Abrahams, S. (2020). A Comparison of the Greek ACE-III, M-ACE, ACE-R, MMSE, and ECAS in the assessment and identification of Alzheimer's disease. *Journal of the International Neuropsychological Society*, *26(8),* 825-834.

Matias-Guiu, J.A., Cortés-Martínez, A., Valles-Salgado, M., Rognoni, T., Fernández-Matarrubia, M., Moreno-Ramos, T. & Matías-Guiu, J. (2016). Addenbrooke's cognitive examination III: Diagnostic utility for mild cognitive impairment and dementia and correlation with standardised neuropsychological tests. *International Psychogeriatrics, 29(1),* 105-113.

Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., *et al.* (2006). The Addenbrooke's Cognitive Examination Revised (ACE-R): A brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry, 21(11),* 1078-1085.

Molloy, W. & Standish, T. (1997) A Guide to the standardised Mini-Mental State Examination. *International Psychogeriatrics, 9(1),* 87-94.

Newman, C.G.J., Bevins, A.D., Zajicek, J.P., Hodges, J.R., Vuillermoz, E., Dickenson, J.M., Kelly, D.S., Brown, S. & Noad, R.F. (2017). Improving the quality of cognitive screening assessments: ACEmobile, an iPad-based version of the Addenbrooke's Cognitive Examination-III. *Alzheimer's Dementia (Amsterdam), 10,* 182-187.

Takenoshita, S., Terada, S., Yoshida, H., Yamaguchi, M., Yabe, M., Imai, N., Horiuchi, M., Miki, T., Yokota, O. & Yamada, N. (2019). Validation of Addenbrooke's

Cognitive Examination III for detecting mild cognitive impairment and dementia in Japan. *BMC Geriatrics, 19(1)*, 123.

Data availability statement

The data that support the findings of this study are available from the corresponding author, [M.J.S.], upon reasonable request.

Table 1: Participants' profession, duration of experience administering the ACE-III, frequency of administering the ACE-III, administration confidence; and case total scores.

Participants	n (%)
Social Worker	1 (3.2)
Psychiatrist/Medical Professional	6 (19.4)
Nurse	8 (25.4)
Psychologist	10 (32.3)
Occupational Therapist	5 (16.1)
Other	1 (3.2)
Duration of using the ACE-III	
less than 6 months	8 (25.8)
6 months to a year	9 (29)
more than a year	14 (45.2)
Frequency of use of ACE-III	
less than monthly	12 (38.7)
once a month	2 (6.5)
a few times a month	12 (38.7)
a few times a week	5 (16.1)
Administration confidence	
Mean (SD)	6.58 (2.36)
Median [Min, Max]	8 [1, 9]
ACE-III (Case 74) rater scores	
Mean (SD)	77.8 (3.40)
Median [Min, Max]	77 [71, 86]
ACE-III (Case 84) rater scores	
Mean (SD)	83.4 (3.97)
Median [Min, Max]	83 [75, 93]

Figure 1: (top row) Coefficient of variation for subscales in each test, higher values represent a greater degree of relative variability; (bottom row) accuracy plots for each test, dotted lines are 95% limits of agreement, solid line represents true score.