Assessments of ADAS-Cog administration quality are comparable across expert and non-expert reviewers

Rachel Kindellan[1], Rachel Newsome[1], Celia Fidalgo[1], Jessica Robin[1]

(1) Winterlight Labs, Toronto, ON, Canada

Background

- Ecological validity and inter-/intra-rater reliability of clinical assessments depends on consistent rater adherence to the clinical administration guidelines unique to each assessment.
- For this reason, quality assurance (QA) review of clinical assessments is necessary to identify issues in administration. These reviews are used to identify and remediate inconsistencies in subsequent administration.
- Clinical experts typically provide QA reviews based on their familiarity with assessments and clinical judgment, which is a costly process.
- Here, we explore whether non-expert reviewers may provide similar ratings of ADAS-Cog administration QA to an expert clinician reviewer.
- We hypothesized that non-expert reviewers can identify QA issues with assessment administration comparable to that of expert reviewers.
- High agreement between reviewer types would indicate a viable, scalable model for QA review of clinical assessments by non-experts, thereby reducing cost and turnaround time of QA of clinical assessments.

Methods

- Fifteen audio recordings of ADAS-Cog assessments were reviewed by an expert reviewer and by three non-expert reviewers.
- The expert reviewer had more than 20 years of clinical experience administering and reviewing CNS clinical assessments. Non-expert reviewers had no experience neither administering nor reviewing CNS clinical assessments, nor any other relevant clinical experience (a BA in psycholinguistics, a specialist in Linguistics, and an MA in psychology).
- For each assessment, all reviewers filled out a QA rubric developed based on the ADAS-Cog administration manual. The rubric consisted of 8 potential administration inconsistencies, called QA issues.
- The QA issues were as follows: presence of instruction, severe script deviation (e.g."complete as quickly as possible" versus "take your time"), hints ("you forgot the word that sounds like…"), overly positive encouragement ("you've got them all right so far"), correct instruction order, rater clarity (i.e. rater's speech interferes with participant's understanding of the task), rater interruption of task, and rater preparedness (rater interrupts assessment preparing for next task).
- Non-expert reviewers received a 30 minute training on the rubric.
- Each instruction in the ADAS-Cog assessment was given a binary score on each QA issue based on the presence or absence of that issue (see Table 1). For example, if a rater skipped an instruction, they would receive 0 as opposed to a 1 for a present instruction. Similarly, if a rater deviated severely from the script, they would receive a score of 1 as opposed to a 0 for verbatim adherence to the assessment script.
- Intraclass correlation (ICC) was used to calculate agreement between the expert reviewer and each non-expert reviewer across all QA issues.
- Additionally, simple agreement proportions for individual QA issues averaged across all non-experts were calculated to further examine the the impact of individual QA issue on agreement.

Table 1: QA Rubric for ADAS-Cog Assessment*

ID	Task	Instruction	Did the rater give this instruction?	Did the rater adhere to the script verbatim?	Did the rater provide hints?
1	Word recall	"I'm going to show you a list of words…"	1	1	0
2	Word recall	"Read it out loud and try to remember it"	0	NA	NA
3	Word recall	"Good, now tell me all the words you can remember."	1	1	0
4	Commands	"Now I am going to ask you to do a few things. Ready?"	1	1	0
5	Commands	"Make a fist"	1	0	1
6	Commands	"Point to the ceiling then the floor"	1	1	1
7	Constructional praxis	"Please draw a figure like this one…"	1	1	0
8	Constructional praxis	*"Take your time"*	1	0	1

Table 1: This table shows a subset of the QA rubric used to review a rater's administration of ADAS-Cog assessments. This was developed using instructions outlined in the ADAS-Cog manual. On the left, task denotes the sub-item in the assessment, and the instruction, which defines the individual questions. The remaining three columns are examples of QA issues for which the reviewers provided a score.

* Table includes only a subset of ADAS-Cog tasks and QA issues for demonstration purposes. Instructions in italics are optional. Raters were not penalized for not including optional instructions.

Figure 1: Pairwise ICC

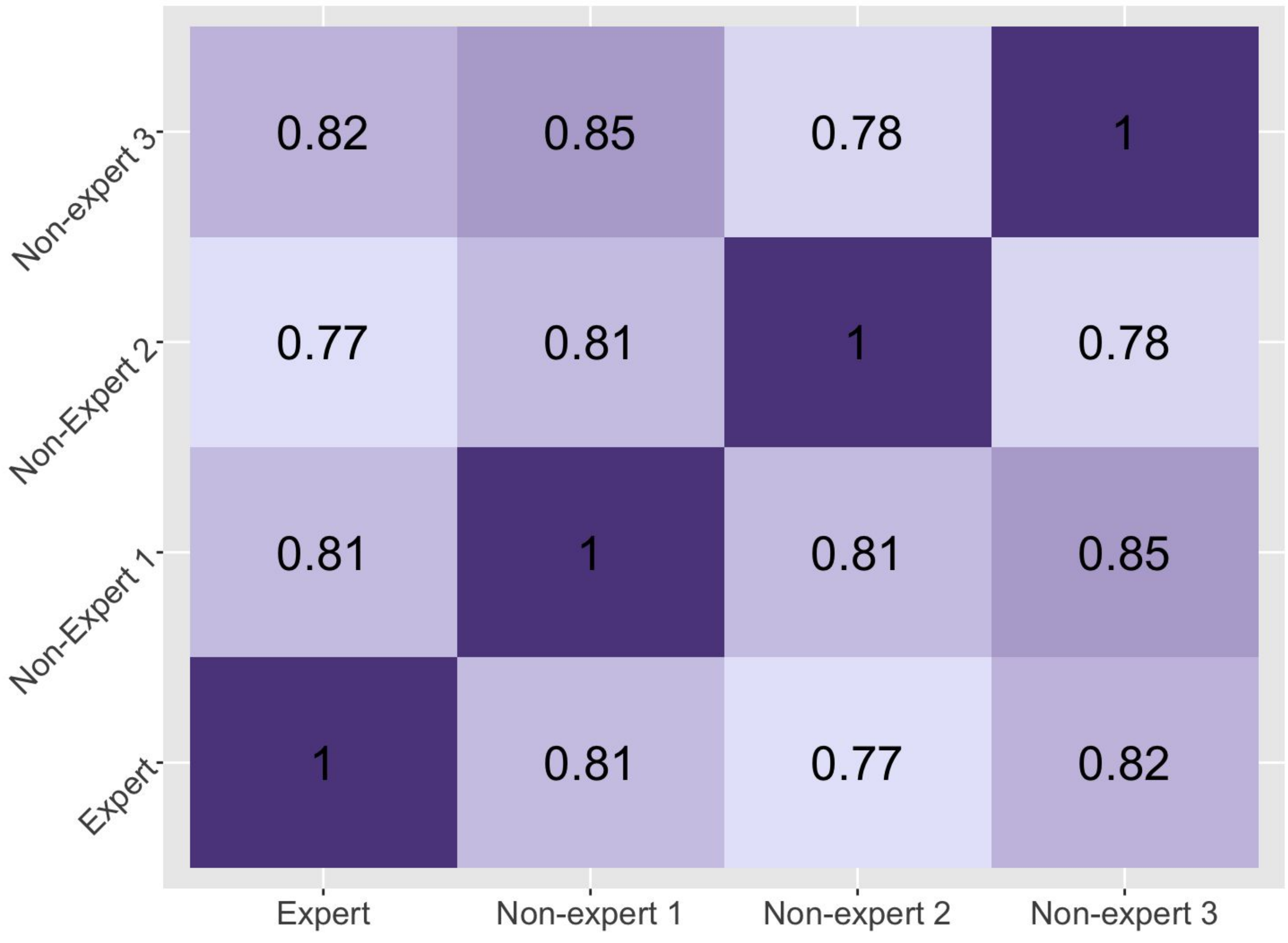

Figure 1: This heat map shows the ICC values between each reviewer. The ICC values of interest are between the expert and all other reviewers (mean = 0.80, SD = 0.03) . The ICC between the non-experts are included for comparison (mean = 0.82, SD = 0.03).

Figure 2: Mean Agreement by QA issue

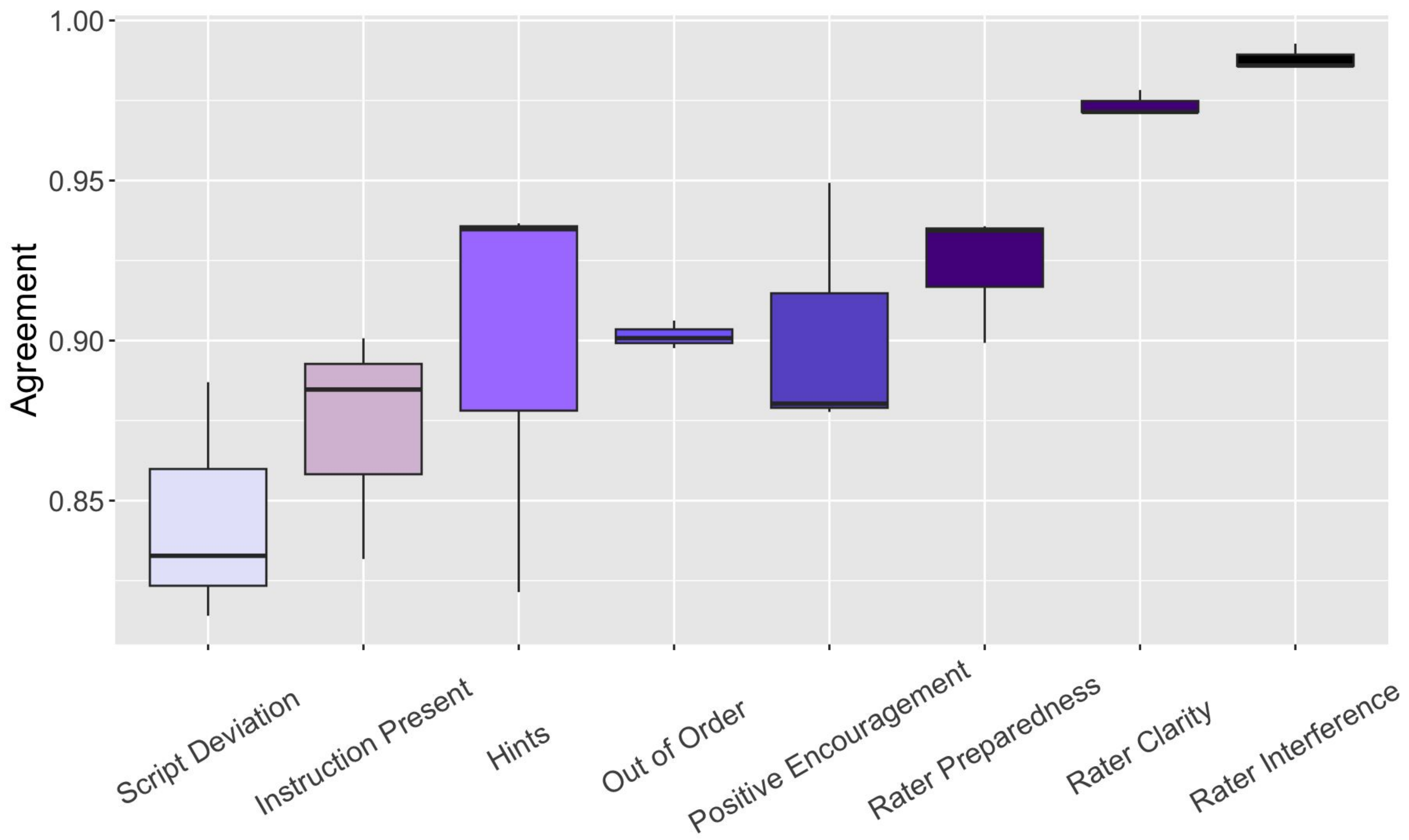

Figure 2: This boxplot shows agreement scores with the expert reviewer, averaged across non-expert reviewers per QA issue. In order of increasing agreement: script deviation mean = 0.85, SD = 0.04; instruction present mean = 0.87; SD = 0.04; hints mean = 0.90, SD = 0.07; item order mean = 0.90, 0.001; positive encouragement mean = 0.90; SD = 0.04; rater preparedness = 0.92, SD = 0.04; rater clarity mean = 0.97 SD = 0.003; rater interference mean = 0.99, SD = 0.004.

Discussion

- The ICC values shown in Figure 1 (mean = 0.80, SD = 0.03) between expert and non-expert reviewers represent good agreement, supporting the feasibility of seeking QA review of cognitive assessments beyond experienced clinicians, thereby reducing cost, reducing turnaround time, and remediating administration issues. The non-expert reviewers had a higher average ICC score (mean = 0.82, SD = 0.03), indicating that disagreements were not due to differences in clinical experience only.
- Looking closer at agreement scores by QA issue (Figure 2), the lowest and most variable agreement was seen in script deviation (mean = 0.85, SD = 0.04), hints (mean = 0.90, SD = 0.07), instruction presence (mean = 0.87, SD = 0.04), and positive encouragement (mean = 0.90, SD = 0.04). These QA issues required the most subjective evaluation, presenting the greatest opportunity for disagreement.
- Two major limitations of this study were a small number of reviewers and differential rates of different QA issues. For example, script deviation is a common issue thus had more balanced scores (both 1s and 0s) compared to rater clarity, which was seldom an issue and heavily skewed toward correct administration. This inherent inequality between QA issues inflated agreement among those with lower variance.
- Future directions for this work include enhancing data collection to both increase reviewer number and include other assessments. These data would allow further investigation into the lowest agreement QA issues and inform more detailed training of non-experts in order to better define the more subjective evaluations to optimize agreement scores.

References

(1) Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. The American Journal of Psychiatry.

Presented at ISCTM, February 16, 2023