Scoring the Quick Aphasia Battery: Training, definitions, fidelity

Katarina L. Haley¹, Adam Jacks¹, Marcia Rodriguez¹, Lorelei Johnson²
¹University of North Carolina, Chapel Hill NC, ²Atrium Health, Charlotte NC

Acknowledgements: The project was supported by grant R01DC018569 from the National Institutes of Health. Reliability samples were obtained from AphasiaBank, which is supported by NIH R01 DC008524.

Background
- The Quick Aphasia Battery (QAB) is a multidimensional and reliable assessment tool with the desirable quality of an approximate administration time of 15 minutes.
- Scores inform both severity estimation and qualitative language profiling. Subtests use a graded scoring system.
- We recently adopted the QAB for a large study on speech production after left hemisphere stroke. To maximize assessment fidelity, we standardized administration procedures and trained our team on what we thought was the most subjective component of the test—the Connected Speech rating.

Connected Speech Rating (trained):
- 10 features: length/complexity, speech rate, agrammatism, connected speech rating.

Methods
- Two experienced clinicians were trained to score the QAB alongside the rest of our research team. One had comprehensive clinical research experience; the other had significant research experience.
- Training progressed from reviewing and clarifying manual instructions to administering the test to healthy adults and subsequently to volunteers portraying different types of aphasia.
- Scoring definitions were reviewed and expanded. The research team constructed supplementary manual notes and checklists and devised a slightly refined Connected Speech rating rubric (partially illustrated below). This work was followed by practice ratings on video recorded assessment sessions.

Results
- Inter-rater agreement was strong for the trained subtest (“Connected Speech” and satisfactory, albeit somewhat lower, for most of the subtests that were not trained.

Point-to-point inter-rater agreement (%):

<table>
<thead>
<tr>
<th>QAB subtest</th>
<th>Perfect agreement</th>
<th>Agreement within one scale level</th>
<th>Agreement within two scale levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected Speech</td>
<td>69</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Word Comprehension</td>
<td>95</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>Sentence Comprehension</td>
<td>87</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>Picture Naming</td>
<td>94</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>Repetition</td>
<td>76</td>
<td>93</td>
<td>100</td>
</tr>
<tr>
<td>Reading aloud</td>
<td>69</td>
<td>90</td>
<td>90</td>
</tr>
</tbody>
</table>

Discussion
- The trained connected speech ratings showed satisfactory rating fidelity (somewhat stronger than the reliability data provided with the original validation and reliability study, where agreement was 94% within one scale level and 52% for a perfect match).
- Because scoring for the other subtests appeared straightforward, we assumed that observer agreement would be strong. This was not the case. Some issues were resolved by consulting the manual. Other reasons that were consequential for scoring reliability:
- a. Unclear how to time 3 and 6 seconds
- b. Unclear what is “a complete attempt” (e.g. pe-pale-pe-pale-pal-pal)
- c. Difficult to differentiate fragment-complete response vs self-correction
- d. Unclear what constitutes an “apraxic error”

We refined operational definitions, added observation checklists, and are currently implementing video vignettes for rater training and calibration purposes.

Our teams will calibrate monthly and monitor interobserver agreement throughout the study.

There is no reason to believe that the QAB is unique in generating disagreement among scorers. It is our experience that researchers and clinicians report that scoring manuals do not answer many of their questions, leaving them to develop their own definitions.

Less than 7% of treatment studies provide information about scoring or assessment training. It seems wise to radically increase that percentage and consider sharing training resources across clinics and laboratories.

15 minutes administration time is accompanied by significant time for scoring reliability. This work was followed by practice ratings on video recorded assessment sessions.

References