This is the accepted version of the paper.

This version of the publication may differ from the final published version.

1

Feasibility and reliability of the Adapted Kagan Scales for rating conversations for people with acquired brain injury: A multi-phase iterative mixed methods design

Nicholas Behn[a] Emma Power[b] Penny Prodger[a] Leanne Togher[c] Madeline Cruice[a] Jane Marshall[a] and Rachael Rietdijk[c]

[a]Department of Language and Communication Science, City St Georges, University of London, London, UK.

[b]Graduate School of Health, University of Technology, Sydney, Australia

[c]Faculty of Medicine and Health, The University of Sydney, Australia.

PURPOSE: Rating the quality of conversations can assess communication skills in both people with acquired brain injury (ABI) and their communication partners. This study explored the clinical feasibility and reliability of two conversation rating scales: The Adapted Measure of Participation in Conversation (MPC) and Adapted Measure of Support in Conversation (MSC)

METHOD: Raters were final-year speech and language therapy students (n = 14) and qualified clinicians (n = 2). Raters attended training on the Adapted MPC and MSC, watched 5 or 10 minutes of videotaped conversations (n = 23) and then scored them on the MPC and MSC scales. Data was collected over four phases which varied according to the length of the training, sample length, number of samples rated and level of clinical expertise. Feasibility data (time taken to score conversations and ease of use) was collected. Inter-rater reliability was assessed using intra-class correlations (ICCs: absolute agreement, single measures).

27 RESULTS: Raters took 30 - 45 minutes to score a 10-minute sample; and 20 - 30 minutes to

28 score a 5-minute sample. Ease of use was rated highly across all phases. Overall reliability

29 for rating 5-minutes of conversation (ICC = 0.52-0.73) was better than for 10-minutes of

30 conversation (ICC = 0.33 - 0.68). Reliability for the MPC was moderate for both students

31 (ICC = 0.69) and clinicians (ICC = 0.55); and for the MSC, moderate for both students (ICC

32 = 0.73) and clinicians (ICC= 0.58). Reliability was better for students compared with

33 clinicians.

34 CONCLUSION: Rating a 5-minute conversation in under 30 minutes was feasible, with more

35 reliable results for 5-minute compared with 10-minute conversations. Implications for

36 assessing conversation in the future are discussed.

37

38 **Corresponding author**

39 Dr Nicholas Behn, Nicholas.behn@city.ac.uk

40

43

46

47
48 **INTRODUCTION**

49   Impaired communication is common for people with acquired brain injury (ABI).

50 People may talk too much or too little; perseverate on a topic or go off on a tangent; lack

51 initiation or frequently interrupt; have difficulty with taking turns and talk over people; not

52   listen to others and be disruptive; or be socially inappropriate in their interactions with others

53   (Coelho et al., 1991; Hartley & Jensen, 1992; Sim et al., 2013; Spence et al., 1993). These

54   impairments have often been described as lying on a spectrum from impoverished (lack

55   initiation, sparse, reduced content) to excessive (over talkative, tangential, repetitive)

56   (MacDonald, 2017; Sim et al., 2013). These impairments are often referred to as a *cognitive-*

57   *communication disorder* (CALSPO, 2015) to highlight the impact of impaired cognitive

58   processes on a person's ability to communicate. Over two-thirds of people who sustain an

59   ABI present with some form of cognitive communication impairment (Hewetson et al., 2017;

60   Kelly et al., 2017; Shorland et al., 2022). This disorder is heterogeneous (Hartley & Jensen,

61   1992; Snow et al., 1997) with impairments known to be both long-term and pervasive (Knox

62   & Douglas, 2009; Olver et al., 1996; Ponsford et al., 2014; Snow et al., 1998). The impacts of

63   this disorder are far-reaching, negatively affecting a person's ability to return to work

64   (Meulenbroek & Turkstra, 2016; Rietdijk et al., 2013), integrate socially (Dahlberg et al.,

65   2006; Knox & Douglas, 2009; Struchen et al., 2008) and achieve a better quality of life

66   (Dahlberg et al., 2006; Galski et al., 1998).

67   Communication is a dynamic, two-way process involving both the person with ABI

68   and their communication partner, whether a family member, friend, or carer. The skills of the

69   partner can either hinder or facilitate a conversation (Togher et al., 1997). Just as a person

70   with ABI may struggle in conversation, so may the communication partner. Partners may

71   frequently ask questions that test a person's knowledge, limit opportunities for the person

72   with ABI to participate and/or not give the person with ABI a turn to respond (Mann et al.,

73   2015; Sim et al., 2013). Conversely, an increased use of a supportive questioning style and

74   use of positive communication strategies by partners (e.g., use of short, simple direct

75   sentences and questions) may improve interactions (Mann et al., 2015; Shelton and Shryock,

76   2007).

77   Given the important role that communication partners play in conversational

78   interactions, training partners is recommended within international guidelines (Togher et al.,

79   2023) and recent systematic reviews (Behn et al., 2020; Wiseman-Hakes et al., 2020). As part

80   of the training process, assessment of conversation is integral to establishing an

81   understanding of the skills of the person with ABI, the ability of the communication partner

82   to support interactions, and to subsequently guide planning of relevant interventions.

83   Conversation is also considered a key outcome for any cognitive-communication intervention

84   (Lê et al., 2022; Tobar-Fredes & Salas, 2022), particularly for determining whether training

85   communication partners has been beneficial to the dyad (Togher et al., 2023).

86   Assessing conversation provides insights into real-life communication with relevant

87   partners and may illuminate communication skills that have been impaired by the brain injury

88   (Keegan et al., 2023; MacDonald, 2017). However, assessing conversation can be difficult

89   due to its dynamic and interactive nature; and may vary according to the type of conversation

90   (e.g., casual, purposeful, task-specific) and the communication partner involved (e.g., family

91   member, sibling, friend, carer). Furthermore, there is a lack of tools that objectively and

92   reliably evaluate conversation in ecologically valid ways (Sohlberg et al., 2019). Pragmatic or

93   observational scales are common (Keegan et al., 2023; Sohlberg et al., 2019; Steel & Togher,

94   2019) though these measures are limited by reduced reliability and consistency (Coelho et al.,

95   2005).

96   Detailed assessment of the quality of conversation is not routinely assessed in clinical

97   practice. An international survey of 265 speech and language therapists from a range of

98   clinical settings found under 10% of clinicians directly assess functional performance,

99   pragmatics, and discourse (Frith et al., 2014). Findings are similar for therapists (n=182) in

100   acute settings, with fewer than 20% assessing conversation (Morrow et al., 2020). A recent

101   international survey of speech and language therapists from mainly rehabilitation and

102 community settings (n = 70) found that 80% of clinicians assessed conversation (Steel et al.,

103 2022). However, the most common type of analysis (> 90%) focussed only on pragmatic

104 features (e.g., eye contact, topic maintenance) of the conversation. Common barriers to both

105 assessment and detailed analysis include the lack of resources and time, and limited

106 availability of tools (Frith et al., 2014; Kelly et al., 2017; Maddy et al., 2015; Morrow et al.,

107 2020; Steel et al., 2022). More detailed analyses beyond pragmatic features alone are needed

108 to guide intervention that enables people with ABI and their communication partners to

109 participate effectively in conversation and in their social lives.

110   Therefore, access to assessments that can feasibly and reliably measure conversation

111 in clinical practice is needed. Sohlberg and colleagues (2019) described feasibility of a

112 measure in terms of time and complexity of administration. A measure that took no longer

113 than 60 minutes to administer and did not require a complex analysis procedure such as

114 transcription and hand coding was considered feasible. In that study, only one (of six)

115 measures the Profile of Pragmatic Impairments in Communication (Linscott et al., 1996) was

116 not considered to be feasible. Iwashita and Sohlberg (2019) described a feasible measure for

117 clinicians as one that was acceptable to clinicians and administrated in 30 minutes or less.

118 The Modified Pragmatic Rating Scale was compared to the Profile of Pragmatic Impairments

119 in Communication. The former was found to be quicker to rate (in under 5 minutes) and

120 described by raters as easier to use. However, a limitation of these conversational scales is

121 that they focus on the skills of the person with ABI, and do not score or rate the skills of the

122 communication partner within a conversation.

123   One commonly reported measure of conversation that focuses on both the person with

124 ABI and their communication partner is the Adapted Kagan Scales (Togher et al., 2010).

125 These scales are clinician-rated, do not require transcription or detailed linguistic analyses,

126 and have demonstrated sensitivity to change from communication partner training (Behn et

127   al., 2012; Rietdijk et al., 2020a; Togher et al., 2013). Originally designed to rate

128   conversations involving people with aphasia (Kagan et al., 2004), these scales were adapted

129   for people with brain injury and their communication partners (Togher et al., 2010). The

130   Adapted Kagan Scales comprise two scales, each with several sub-scales. The first, the

131   Adapted Measure of Participation in Conversation (MPC) rates the interaction and

132   transactional skills of the person with brain injury. The second, the Adapted Measure of

133   Support in Conversation (MSC) rates the ability of the communication partner to both

134   acknowledge and reveal the competence of the person with brain injury within the

135   conversation. These tools are the only available scales that rate the skills of both people in the

136   dyad. The scales have excellent inter-rater and intra-rater reliability when rated by

137   experienced clinicians (Togher et al., 2010), good ecological validity, and based on the

138   parameters described by Sohlberg et al. (2019), would be considered feasible in terms of time

139   to rate and ease of use.

140   Although the Adapted Kagan Scales have been found to have acceptable reliability in

141   research contexts, the clinical feasibility of the Adapted Kagan Scales is likely to be affected

142   by a range of factors. Empirical studies have reported varying degrees of inter-rater reliability

143   (Behn et al., 2019a; Behn et al., 2012; Chia et al., 2019; Rietdijk et al., 2020b; Togher et al.,

144   2013) with the time taken to train raters ranging from 2.5 hours to 35 hours with better

145   reliability reported for longer training times of at least 14 hours (Behn et al., 2019a; Behn et

146   al., 2012; Chia et al., 2019; Rietdijk et al., 2020b). Raters have ranged from students studying

147   speech and language therapy with limited experience of people with brain injury to clinicians

148   with little to extensive clinical experience. The length of conversation has ranged from 5- to

149   10-minutes and the type of conversation has included casual and purposeful (or structured)

150   conversation. Casual conversations have involved a dyad talking about a topic of interest,

151   while purposeful conversations require the dyad to complete a task (e.g., plan a holiday) or

152 ask structured questions. Reliability results across different lengths and types of conversation

153 have been comparable in some studies (Behn et al., 2012; Togher et al., 2010) and better for

154 purposeful than casual conversations in other studies (Rietdijk et al., 2020a; Rietdijk et al.,

155 2020b). All these factors may impact the extent and ease of implementation of the Adapted

156 Kagan Scales in clinical practice.

157 The purpose of this study was to determine whether the Adapted Kagan Scales could

158 be established as a clinically feasible method (i.e., completed in 30 minutes or less) for

159 assessing a single type of conversation; and could achieve acceptable levels of inter-rater

160 reliability with limited training. The face, ecological and construct validity of the measures

161 has already been established (Kagan et al., 2004; Sohlberg et al., 2019; Togher et al., 2010).

162 The same conversations were used across multiple phases to allow direct comparison; with

163 consideration of training length; scales rated; and rater experience. This study aims to address

164 the following research questions:

165   1. Can videotaped conversations involving people with ABI and their communication

166   partners be feasibly rated in terms of time taken (30 minutes or less) using the

167   Adapted Kagan Scales?

168   2. Can acceptable (i.e., moderate) reliability be achieved by students and experienced

169   clinicians?

170   3. Can raters achieve acceptable (i.e., moderate) levels of reliability with limited training

171   (<8 hours of training) in the use of the scales?

172   4. Can similar levels of reliability be achieved from rating 5-minute compared with 10-

173   minute videotaped conversations?

174   5. What is raters feedback on their experience of using the Adapted Kagan Scales?

175 **METHODS**

176

Design

A four-phase iterative mixed-methods design was conducted using data collected from a previous feasibility trial examining communication skills in people with ABI (Behn et al., 2019a). The four phases were conducted over the period from 2019-2023. Ethical approval was initially granted as part of the trial by City, University of London School of Health Ethics Committee (PhD/12-13/14), and the Brain Injury Rehabilitation Trust Ethics Committee (dated 21st May 2013). Further approval for this study was granted by the City, University of London Language and Communication Science Proportionate Review Committee (ETH1920-0181/ETH2021-0421/ETH2122-0209).

Participants

Video samples from a total of 21 participants with acquired brain injury and their communication partners from the United Kingdom were included. The participants had previously given informed consent to participate as part of a published feasibility trial on a social communication skills group treatment (Behn et al., 2019a). Table 1 presents the demographic variables for participants with ABI and their communication partners. All participants were at least 12 months post-injury, determined to have a moderate-to-severe brain injury based on the period of post-traumatic amnesia, the Glasgow Coma Scale score, or the participants' clinical presentation. All participants were reported to have a diagnosis of a cognitive communication disorder, as determined by a practicing speech and language therapist. All participants had significant cognitive impairment based on the Repeatable Battery of the Assessment of Neuropsychological Status (RBANS) (Randolph, 1998) and Wisconsin Card Sorting Test (WCST) (Heaton et al., 1993). Communication partners were identified by people with ABI as someone who they interact with regularly on a weekly basis and who would be able to attend assessment sessions and contribute to goal setting. For the

202   21 participants, there were 17 female communication partners and four male communication

203   partners.

204

205   [insert Table 1 about here]

206

207   *Measures*

208   The Adapted Kagan Scales (Togher et al., 2010) comprise two main scales. The first,

209   the Adapted Measure of Participation in Conversation (MPC) is used to rate the

210   conversational participation of the person with ABI, specifically evaluating how they socially

211   connect, engage, and share the conversation with their communication partner. The scale is

212   further divided into two subscales: Interaction (social connection) and Transaction

213   (exchanging content).

214   The second scale is the Adapted Measure of Skill in Supported Conversation (MSC),

215   which rates the skills of the communication partner during the conversation. This scale is

216   divided into two subscales: Acknowledging competence (AC) and Revealing competence

217   (RC). The Revealing Competence subscale involves three elements: (RC1) Ensure the adult

218   understands; (RC2) Ensure the adult has a means of responding; and (RC3) Verification.

219   Each scale is rated on a 9-point Likert scale presented as a range of 0 – 4 with 0.5

220   intervals. There are behavioural descriptors and five anchor points to help guide the rater's

221   judgement. For the MPC the anchor points range from 0 (no participation) to 4 (full

222   participation in conversation) while the MSC anchor points range from 0 (not supportive) to 4

223   (highly skilled support). In total, six ratings are obtained: one for each subscale of the MPC

224   (interaction and transaction), one for the Acknowledge Competence subscale of the MSC,

225   and one for each of the three elements from the Revealing Competence subscale, which can

226   later be averaged to give a total subscale score.

227

Raters

229   Fourteen final-year speech and language therapy students were recruited from City,

230   University of London. All students had limited to no knowledge of working with people with

231   brain injury; though had received six hours of lectures on the topic by the first author. In

232   addition, two experienced speech and language therapists were recruited, who had 12 and 20

233   years clinical experience working with people with brain injury.

234

Procedure

236   Raters scored the Adapted Kagan Scales to evaluate casual conversations involving

237   people with ABI and their communication partners. There were 73 conversations recorded in

238   the original feasibility trial (Behn et al., 2019a). These recordings were either taken pre-

239   treatment, post-treatment, or at follow-up. Conversations were recorded using a Flip Video

240   Camera HD mounted on a tripod. Dyads were instructed to discuss a topic of interest for 10

241   minutes, while the researcher (NB) left the room. A proportion (n = 23, 32%) of these

242   conversations were randomly selected to check inter-rater reliability in the original study.

243   Several conversations involved the same dyad, but at different time points. These same 23

244   conversations were used in the current study to directly compare the results of the current

245   study with that study.

246   The procedure for this study is divided across four phases, where the results of the

247   previous phase influenced the procedure for the successive phase. Detailed information that

248   informed the decisions made for each phase including, the statistical results (both feasibility

249   and reliability) and discussions among the research team are reported in the results section.

250   The phases are as follows:

251

Phase I (Student raters, half-day versus full-day training, 10 min samples, six scales): The

aim of this phase was to examine different lengths of training. Six final-year speech and

language therapy students were recruited as raters (two males, four females). Three raters

received four hours of direct training (half-day) on the scales, while the other three raters

received eight hours (full-day) of direct training. All raters were required to rate the full 10-

minutes of the conversations, using all six scales (two for the MPC, four for the MSC).

Phase II (Student raters, half-day training, 5 versus 10 min samples, four scales versus three

scales): The aim of this phase was to examine different lengths of conversation and a reduced

number of scales to rate. Six different final-year speech and language therapy students were

recruited as raters (all female). All raters received four hours of direct training on the scales

and a further four hours of self-directed training using the TBI Bank Grand Rounds training

package (Elbourn et al., 2023). Two raters rated the full 10-minutes of conversation using

four scales, while two raters rated only the first 5-minutes of the conversation using the same

four scales (two for the MPC; and two elements of the Revealing competence scale – ensure

the adult understands and ensuring the adult has a means of responding). Two raters rated the

full 10-minutes of conversation using three scales (MPC Interaction; and two elements of the

Revealing Competence scale – ensure the adult understands and ensure the adult has a means

of responding).

Phase III (Student raters, half-day training, 5 min samples, two scales): The aim of this phase

was to examine a further reduced number of scales to rate. Two different final-year speech

and language therapy students were recruited as raters (both female). Both raters received

four hours of direct training on the scales and a further four hours of self-directed training

276 using the TBI Bank Grand Rounds training package (Elbourn et al., 2023). The raters rated

277 only the first 5-minutes of conversation using two scales (MPC Interaction; and one element

278 of the Revealing Competence scale – ensure the adult understands).

279

280 *Phase IV (Experienced raters, half-day training, 5 min samples, two scales)*: The aim of this

281 phase was to examine the ratings of experienced raters compared with students from the

282 previous phase. Two qualified speech and language therapists were recruited as raters (both

283 female). Both raters received four hours of direct training on the scales. The raters followed

284 the exact procedure from the previous phase. They rated only the first 5-minutes of

285 conversation using two scales (MPC Interaction; and one element of the Revealing

286 Competence scale – ensure the adult understands).

287

288   Raters in phase one were required to watch the conversation at least once, with the

289 option of repeat viewing. They were asked to record the number of times they watched the

290 conversation. In phases 2 - 4, raters were required to watch each conversation twice; in

291 clusters of three samples at a time to reduce rater fatigue (Eriksson et al., 2014). Raters in all

292 phases were required to record the following feasibility information for each rating: (1) the

293 time taken to watch (and re-watch, where appropriate), consider behaviours observed and

294 decide on a final rating of the conversation; (2) the ease of rating the conversation on a 10-

295 point scale (1 = not easy; 10 = very easy); (3) qualitative feedback about the ease of rating;

296 and (4) qualitative feedback about the descriptors in the scale, and the rating process. In

297 phases three and four, a think-aloud protocol was included to gather more detailed

298 information about the rating process. Think Aloud is a technique in which participants

299 verbalise their thoughts while simultaneously carrying out a task to gain in-depth qualitative

300 data, in this case on anything which affected the rating of the conversations (Durning et al.,

301   2013). Raters in these phases were observed online via *Zoom* (Version 5.12.8) scoring the

302   same three conversations while recorded by a researcher (PP).

303

304   *Training*

305   All direct training on the scales was led by the first author who had more than 20

306   years experience in working with people with ABI, and who had more than 10 years'

307   experience in training and using the Adapted Kagan Scales for research purposes (Behn et al.,

308   2019a; Behn et al., 2012). The first author collaborated with other co-authors (EP, LT, and

309   RR) to refine training in the use of the scales. Training was delivered in groups of two-to-four

310   raters; in-person in phase one then online for phases 2-4 due to the COVID-19 pandemic.

311   While the decision to do training online was influenced by external factors, studies have

312   shown online training to be as effective as face-to-face methods (Cook et al., 2008; Soffer &

313   Nachmias, 2018).

314   Training began with a general familiarisation of the scales and the rating process.

315   Raters then watched several sample conversations. These conversations were accessed

316   through TBI Bank, which is an online repository of materials and resources including

317   videotaped conversations of people with traumatic brain injury and their communication

318   partners (Elbourn et al., 2023). The full-day training (8-hours) involved eight sample

319   conversations while the half-day training (4-hours) involved four-to-five sample

320   conversations. Raters independently scored each sample conversation individually then

321   discussed their scoring and any discrepancies with the group with reference to the descriptors

322   and anchor points. Common issues that could influence the rating process were discussed and

323   examined in relation to the sample conversations (e.g.,, relationship of partner, weighing up

324   different descriptors, imbalanced conversations). Final ratings for the sample training

325   conversations were agreed on via consensus, which became anchors for different points on

326 the 9-point scale, to provide a reference point when rating. All raters across all phases were

327 permitted to review anchor videos as often as needed.

328   After phase one, rater feedback suggested a need for student raters to gain further

329 knowledge of the communication problems that can occur from an ABI. Therefore, raters in

330 phases two and three were required to do the online TBI Bank Grand Rounds training

331 (Modules 1-3, 5 and 7) (Elbourn et al., 2023). This training was self-directed and took raters

332 approximately four hours to complete in addition to the direct rater training already received.

333 Topics covered were cognitive-communication disorders, discourse, and variability of

334 discourse across contexts.

335

336 *Data analysis*

337   Feasibility information (time taken to rate; ease of rating scores) was compiled in a

338 Microsoft Excel spreadsheet and analysed descriptively. Qualitative feedback on the

339 experience of using the rating scales was initially compiled and analysed by the first and/or

340 third author (NB or PP) using conventional content analyses (Hsieh & Shannon, 2005). The

341 data was coded, identifying similarities and differences in the feedback of raters, with

342 categories of information identified. Qualitative data from the think-aloud sessions were

343 transcribed verbatim and analysed by the second author using conventional content analysis

344 and checked for accuracy by the first author. As recommended in mixed methods research

345 (Fetters, Curry & Creswell, 2013), we collected the quantitative and qualitative data in

346 parallel and analysed the data for integration prior to the commencement of successive

347 phases. Therefore, the preliminary integration of data was completed at the end of phase one,

348 two and three. At the end of the study all qualitative data was synthesised with the

349 quantitative data from all four phases to explain the findings of the study.

350   Reliability data from each phase were examined separately including, the effect of

351   relevant variables such as the length of training, length of conversation viewed, and the

352   experience of the rater. Interrater reliability was assessed using intra-class correlations (ICC)

353   2,1 procedure with absolute agreement, single measures (Shrout & Fleiss, 1979). The 95%

354   confidence intervals were reported as percent agreement in line with reporting guidelines for

355   reporting reliability results (Kottner et al., 2011). Excellent reliability was defined as ICCs

356   greater than 0.90, good reliability as between 0.75 and 0.90, moderate reliability between

357   0.50 and 0.75 and poor reliability as less than 0.5 (Koo & Li, 2016). Acceptable reliability in

358   this study was determined as moderate. As both the MPC and MSC include several subscales,

359   Spearman's rank-order correlations (r_s) were calculated for each scale to evaluate the strength

360   of the association between subscales. Strong correlations have values between 0.7 and 0.9,

361   moderate correlations between 0.4 and 0.6, and weak correlations between 0.1 and 0.3

362   (Dancey & Reidy, 2007). Throughout all phases the strength of the ICCs and correlations and

363   feasibility information were examined by the research team to inform each phase. All

364   statistical analyses were computed using IBM SPSS Statistics (Version 28).

365

366   **RESULTS**

367   Quantitative results are presented for each phase, followed by overall qualitative results.

368   *Phase I*. The mean time taken to watch and rate each 10-minute length conversation by raters

369   was 29 mins (range 14 – 56 mins)(Figure 1). Mean ease-of-use ratings on a scale of 1 - 10

370   was 6.8 with a range of scores from 5 - 10. Raters who completed the 4-hour training rated

371   the conversation from a single viewing 81% of the time. Raters who completed the 8-hour

372   training rated the conversation from a single viewing 43% of the time.

373

374   [insert Figure 1 about here]

375

376   Reliability for both half-day and full-day of training was poor-to-moderate (ICCs =

377   0.43 - 0.62) with confidence intervals poor-through-good (Table 2). Percent agreement within

378   0.5 ranged from 17% to 43% across both conditions.

379

380   [insert Table 2 about here]

381

382   There were strong positive correlations, $r_s(21) > 0.73 - 0.92$, $p < .001$ between the

383   MPC Interaction and Transaction subscales for all six raters. Strong positive correlations

384   were found between the MSC Acknowledging Competence subscale and each of the three

385   elements of the Revealing Competence scales for most raters, $r_s(21) > 0.74 - 0.96$, $p < .001$.

386   For one rater, the correlation between the RC1 and RC2 elements was moderate $r_s(21) =$

387   0.68, $p < .001$.

388

389   *Phase II.* As there was minimal difference in ICCs between half-day and full-day training,

390   Phase II used half-day training only; and the strong correlations between specific scales led to

391   the number of scales rated being reduced to either four scales (MPC Interaction and

392   Transaction, MSC RC1 and RC2); or three scales (MPC Interaction, MSC RC1 and RC2)

393   given stronger ICCs for MPC Interaction over MPC Transaction. Additionally, phase II

394   compared long (10 minute) and short (5 minute) conversation samples.

395   For raters who viewed 10-minute samples, the mean time taken to rate each sample

396   was 34 minutes (range 25-53 mins) when four scales were rated and 38 minutes (range 23 -

397   60 mins) when three scales were rated (Figure 1). For raters who viewed 5-minute samples,

398   the mean time taken to rate each sample was 23 mins (range 18 - 26 mins). Mean ease-of-use

399   ratings on a scale of 1 - 10 was 6.6 with a range of scores from 2 to 10.

400   Reliability for rating four scales with 10-minutes of conversation was moderate for

401   MPC Interaction, RC1 and RC2 (ICCs = 0.56 - 0.59) and poor for MPC Transaction (ICC =

402   0.33) (Table 3). Reliability for rating four scales with 5-minutes of conversation was

403   moderate for all four scales (ICCs = 0.56 - 0.68). Reliability was moderate for rating three

404   scales with 5-minutes of conversation (ICCs = 0.52 – 0.68). Confidence intervals were poor-

405   through-good for all ICCs across all conditions. Percent agreement within 0.5 ranged from 39

406   – 83% across the three conditions.

407

408   [insert Table 3 about here]

409

410

411   There were strong positive correlations, $r_s (21) > 0.86 - 0.91$, $p < .001$ between the

412   MPC Interaction and Transaction subscales for all raters in this phase. There were also strong

413   positive correlations, $r_s (21) > 0.85 - 0.97$, $p < .001$ between the RC1 and RC2 for all raters.

414

415   *Phase III*. As there were more favourable ICCs for 5-minutes compared to 10-minutes of

416   conversation, the next two phases used 5-minute conversations only. As the correlations were

417   strong and ICCs higher for MPC Interaction and MSC RC1, only these two scales were used

418   in the next two phases.

419   The mean time taken to rate conversations in this phase was 20 mins (range 15 - 26

420   mins) (Figure 1). Mean ease-of-use ratings on a scale of 1 - 10 was 7.6 with a range of scores

421   from 2-10.

422   Reliability for rating the two scales with 5-minutes of conversation was moderate for

423   MPC Interaction (ICC = 0.69) and MSC RC1 (ICCs = 0.73) (Table 4). Confidence intervals

424   were poor-through-good. Percent agreement within 0.5 ranged from 57 – 78%.

425

426   [insert Table 4 about here]

427

428

429   *Phase IV*. The mean time taken by experienced clinicians to rate conversations in this phase

430   was 22 mins (range 19 - 35 mins) (Figure 1). Mean ease-of-use ratings on a scale of 1 - 10

431   was 7.3 with a range of scores from 4 to 9.

432   Reliability for rating the two scales with 5-minutes of conversation was moderate for

433   MPC Interaction (ICC = 0.55) and MSC RC1 (ICCs = 0.58) (Table 3). Confidence intervals

434   were poor-through-good. Percent agreement within 0.5 ranged from 48 – 70%.

435

436   *Overall summary*

437   Table 5 presents a summary of each of the four phases. Overall, the time taken to rate

438   conversations decreased across the phases, particularly as shorter conversations were rated;

439   ease of use for rating conversations on the scales improved slightly across phases; and

440   measures of reliability (ICCs) generally improved across each of the four phases, most

441   notably when fewer scales were used.

442

443   [insert Table 5 around here]

444

445   *Qualitative Data*

446

447 Qualitative data revealed two broad categories across all four phases around: (i) scale use;

448 and (ii) conversation ratings. *Scale use* referred to the raters' actual use of the scales to inform

449 their final rating. Raters from all phases found it difficult to know how weigh-up one

450 behaviour or descriptor over another.

451

452 *"Finding it difficult to finalise scores between 2 and 3 and decide what gives enough*

453 *weight to lower or increase a score" (Student rater, Phase I)*

454

455 There were issues with the clarity of the descriptors where some raters reported lack

456 of detail, ambiguous, or imprecise descriptors (e.g., "share responsibility for feel/flow").

457 Some raters reported descriptors were not helpful, that some partially met or absent

458 descriptors were difficult to rate and that overall, there were simply too many descriptors to

459 rate and/or consider at once (particularly in phase I). Raters did report it easier to rate more

460 concrete and overt behaviours (e.g., "listening attitude, supportive questioning"). Many of

461 these reports were reduced in frequency in later phases (when fewer scales were rated) and

462 the clinician raters reported fewer concerns than students regarding the usefulness of the

463 descriptors.

464

465 *"Found the descriptors helpful to go through for CP [communication partner] as*

466 *although she had a warm manner and was interested in her son, very few of the*

467 *criteria for supporting understanding were explicitly met" (SLT, Phase IV).*

468

469 Finally, raters reported that the most and least successful conversations were easier to

470 judge; and conversations that fell in the middle of the scale harder to rate. Clinicians reported

finding it hard to use the half-point ratings due to familiarity with scales in clinical practice

with full points only.

The second category *conversation ratings* referred to how a rater reflected on the

conversation viewed to make a rating. Raters reported challenges with rating a conversation

without personal knowledge of the dyad and context of the conversation (e.g., their sense of

humour and usual dynamics). Some raters wanted additional information about how the dyad

were at baseline and/or prior to injury to judge the conversation. Some raters were aware of

their own biases and emotional response to the interactions and how they may affect ratings

(either positively or negatively)

"[I was] worried that my emotional response to the video would affect my scoring"

(student rater, Phase I)

Raters reported difficulty when the behaviours of the dyad changed throughout the

conversation and struggled with resolving how the behaviours of an individual affect the

other and in turn, the ratings given to each person in the dyad. These challenges were raised

mainly by the student raters.

"pragmatics again can be mixed throughout with some examples of flat affect/ blank

expression and others of good pragmatics" (student rater, Phase I)

"Do I score the person with brain injury lower on interaction because they didn't

initiate, or the CP [communication partner] lower on RC2 for not giving enough time

and silence to allow the person with brain injury to initiate?" (student rater, Phase II)

496   In later phases, rating a conversation with fewer scales and making a judgement of the

497   impact of individual behaviours relative to the whole conversation was a challenge,

498   particularly for clinicians. Student raters also provided insightful comments describing this

499   challenge.

500

501   *"I'm just gonna stick to what I'm rating. There's so many things that play a part in*

502   *making the conversation great and I'm only focusing on do they ensure that the other*

503   *person understands" (student rater, phase III)*

504

505   Clinicians sometimes reported using clinical intuition to make a judgement of the

506   conversation as the rating score was not felt to reflect their observations.

507

508   *"found myself judging the score on gut feeling once all descriptors considered, rather*

509   *than any one descriptor carrying more weight" (SLT, Phase IV).*

510

511   **DISCUSSION**

512   The aim of this study was to explore the feasibility of the Adapted Kagan Scales for

513   clinical practice, and reliability under different training and rating conditions. Overall, the

514   training required to achieve proficiency, and the time to view and rate conversations would

515   be considered feasible. Across all phases of the study, raters were able to view and rate a 5-

516   or 10-minute conversation in under 60 minutes. Rating time was reduced to 30 minutes for a

517   5-minute conversation. This result is consistent with the findings of Iwashita and Sohlberg

518   (2019) where raters could rate a 10-minute conversation using two scales of social

519   communication ability in less than 30 minutes. Training was also feasible to deliver in either

520   a half-day or full-day training program. While it was not the intention to explore the delivery

521   mode, training was able to be successfully delivered both face-to-face and online. While

522   longer training (i.e., full-day) offered increased opportunities for practice and discussion,

523   when compared to shorter training duration (i.e., half-day) there was no discernible difference

524   in the reliability results. The time taken to train the scales was significantly less than the 14 to

525   35 hours reported elsewhere (Behn et al., 2019a; Behn et al., 2012; Chia et al., 2019; Rietdijk

526   et al., 2020b), with these other studies involving a procedure in which raters demonstrated

527   reliability on training samples to be considered competent in rating. The potential for reduced

528   training time and quicker scoring demonstrated in the present study is important, as it enables

529   the scales to be more clinically accessible to speech and language therapists, who have

530   restrictions on their time (Frith et al., 2014; Kelly et al., 2017; Maddy et al., 2015).

531   The reliability results are encouraging and considered acceptable, with moderate

532   reliability for most scales and improved reliability for the student raters when fewer scales

533   were rated. The results were not as strong as for the original study that used the same

534   conversations (Behn et al., 2019a) however, that study involved 18 hours of training over

535   multiple days. Given previous studies have reported moderate-to-excellent reliability for

536   rating casual conversations with longer training, the finding is optimistic (Behn et al., 2019a;

537   Behn et al., 2012; Rietdijk et al., 2020a; Rietdijk et al., 2020b; Togher et al., 2010).

538   Moreover, reducing the number of scales yielded positive reliability results and addressed

539   rater burden raised by some student raters who made comments about too many scales to rate

540   and descriptors to consider. However, the same positive results cannot be said of the scales

541   rated by experienced clinicians. In interpreting the significance of these results, researchers

542   have suggested that intra-class correlations need to be at least 0.80 for high-risk clinical

543   decisions, such as making clinical diagnoses (Slagle et al., 2002), 0.70 for research purposes

544   (Nunnally, 1978) and 0.60 to be clinically useful (Chinn, 1991). While this would seem to

545   suggest that the Adapted Kagan Scales have potential as a clinical measure, closer inspection,

546   and interpretation of the 95% confidence intervals, suggest the picture to be less clear, with

547   most confidence intervals showing great variability between poor-to-good. These results

548   require additional thought about the complexity of conversations, how conversations are rated

549   using the scales, the influence of the individual raters, and how they are trained to use the

550   scales.

551   Several of the study's findings raise an important issue about the complexity of

552   conversation and its variable nature; and whether a set of scales can reliably capture the

553   subtle behaviours and nuances that may in turn, be difficult to objectively define (Eriksson et

554   al., 2014). The conversations that occur for people with brain injury are highly heterogeneous

555   (Hartley & Jensen, 1992; Snow et al., 1997). The environment, social context, goals and

556   demands of the conversation, the communication partner, and social and cultural roles they

557   assume, may all impact the nature of conversation and support provided to someone with a

558   brain injury (Keegan & Müller, 2022; MacDonald, 2017). A rater is then required to observe

559   and rate subtle communicative behaviours that occur in a fast-moving, dynamic interaction.

560   Several raters in this study highlighted the need for additional personal information of the

561   dyad and how they communicated prior to the injury, and the context of the conversation to

562   make accurate judgements. Therefore, raters were required to make their own judgements

563   about the relationship between the dyad and the amount of shared knowledge and experience

564   for the conversation they rated.

565   The process of rating conversation is potentially therefore, susceptible to rater bias

566   (Eriksson et al., 2014; Sohlberg et al., 2019). Certainly, in this study several raters were

567   aware of personal bias and how this may have positively or negatively affected their own

568   ratings. This bias has been found in previous studies where a raters' judgement of the

569   significance of behaviours in performance varied widely (Yeates et al., 2013a). A clinician

570   may identify impaired communication when those involved in the interaction including the

571 communication partner may not identify any impairment at all. A clinician may not share the

572 person's culture or social background or have experience of situations or contexts being

573 discussed in the conversation, which may affect their judgement. Further, a clinician may

574 have an unconscious bias on factors such as gender, culture, race, and ethnicity, that may

575 influence their judgement of the interaction (Badon et al., 2005; Harrison et al., 2017).

576 Eriksson et al (2014) identified the effect of raters' personal biases as one of the key factors

577 undermining the reliability and validity of clinical rating scales. Longer training that

578 explicitly addresses many of these issues may need to be considered and evaluated in the

579 future to determine whether they can be mitigated (Behn et al., 2012; Eriksson et al., 2014).

580 There are several types of rater error and bias that may influence the rater's ability to

581 make a judgement about a conversation. These have been described by Eriksson et al (2014)

582 (2014) including, primacy/recency effects when ratings are based on observations made early

583 or late in the conversation, or contrast effects where ratings are higher or lower relative to

584 previously assessed samples (Feldman et al., 2012; Yeates et al., 2013b). One reflection by

585 several raters was a difficulty in deciding how to weigh one-off behaviours when scoring. For

586 example, a communication partner may demonstrate good listening skills throughout most of

587 the conversation, but then dismiss contributions from the person with ABI at one point in the

588 conversation. The relative weight (and thus rating) given to one behaviour over another may

589 differ between raters (Yeates et al., 2013a). This effect was particularly noticeable when a

590 behaviour was brief but had significant impact on the other person. Raters rarely agreed on

591 which conversations were the most challenging for weighing up behaviours. This finding

592 may suggest the presence of "halo errors" whereby ratings are based on one positive or

593 negative observation (Jacobs & Kozlowski, 1985). Rating a conversation is a complex

594 process, and there is likely to be variability (and bias) in how individual raters place emphasis

595 or perceive value on different aspects of an interaction.

596   Experienced clinicians reported difficulty with a scale that contained half-points,

597   which suggests a reduced scale may be more favourable. Eriksson et al (2014) attempted to

598   address this issue (and that of bias) by shortening the rating periods (e.g., to one minute each)

599   and using a reduced scale from 9-points to 4-points (e.g., 1 to 4, predominantly poor support,

600   consistently satisfactory support). However, 10-30 hours of training was required, and

601   reliability was poor to moderate. In another study, a more reduced scale (of 1-3 points:

602   predominantly poor support, OK but not satisfactory, predominantly satisfactory) was found

603   to achieve better reliability (Saldert et al., 2013) although a reduced scale may potentially

604   limit the validity and the scales' sensitivity to change.

605   The impact of several factors on reliability was considered in this study, including the

606   length of the conversation (i.e., 5 and 10 minutes) and experience of the rater (i.e., student

607   and experienced clinician). Other studies have used either five minutes (Rietdijk et al., 2020b;

608   Togher et al., 2010) or ten minutes of conversation (Behn et al., 2019a; Behn et al., 2012;

609   Iwashita & Sohlberg, 2019); and a study for people with post-stroke aphasia reported that 3-5

610   minutes of conversation was sufficient for analysis (Correll et al., 2010). The reliability

611   findings from this study were generally more favourable for conversations of 5-minutes in

612   length when raters were rating the same scales, which suggests that clinicians could adopt the

613   same length of conversation in clinical practice. Reliability was less favourable for

614   experienced clinicians compared to student raters. Togher et al (2010) reported good-to-

615   excellent reliability when raters were experienced clinicians. However, in that study raters

616   rated all six scales and in the current study (phase IV), clinicians rated only two scales (MPC

617   Interaction and MSC RC1). Qualitative reports suggest that the clinicians tended to use their

618   wider clinical experience and intuition when rating. Certainly, for one clinician, they found it

619   challenging to focus on the two scales and gave ratings that reflected the overall

620   conversation. While the earlier study of the original Kagan scales found a significant positive

621 correlation between clinical intuition and ratings (Kagan et al., 2004), the raters rated all six

622 scales rather than the two in this study.

623 An additional factor to raise relates to the training process itself. The training

624 familiarised the raters with the scales, provided sample conversations to rate, and discussed

625 common issues. While the training process used was like other studies using the same scales

626 (Behn et al., 2012; Behn et al., 2019a; Rietdijk et al., 2020b), greater consideration of some

627 of the issues raised by raters in this study may be needed (e.g., managing personal bias,

628 weighing up behaviours, changing behaviours, influence of a person's behaviour on another).

629 Longer training and/or greater use of challenging sample videos may help. There may also be

630 an issue with how raters listen and engage during training and apply what they have learnt.

631 Future research may need to closely examine the training process using think aloud

632 techniques to more robustly identify how raters observe and interpret what they are seeing

633 and where the specific differences may lie when they rate the same video. This research will

634 contribute to our understanding of how best to train the use of the scales thus, standardising

635 training for the future.

636 Finally, there may be a tension between the concise clarity of the rating scales and the

637 subtle insights from a rater who has either greater clinical experience, more training, or who

638 is rating a longer conversation sample. Individual raters' reported issues with the clarity and

639 weighting of the descriptors, including unhelpful and/or an excessive number of descriptors

640 to consider. Visible communicative behaviours (e.g., eye contact, questions asked, turn-

641 taking) were certainly considered easier to rate than more abstract, ambiguous behaviours

642 (e.g., appropriate amount of information, organisation of information). However, qualitative

643 comments from the experienced clinicians suggest that clinical intuition may lead raters to

644 identify or describe more subtle, abstract and difficult to describe behaviours that may not be

645   listed, highlighting the inherent conflict for raters during the process. Striking the right

646   balance between these factors may prove challenging.

647   Reliability and feasibility of the measures may be improved through modifications to

648   the scale including, reducing the number of descriptors, and linking them to more concrete

649   behaviours. However, these changes may negatively influence the validity of the scales and

650   their ability to adequately explain differences in ratings. Measures like the Modified

651   Pragmatic Rating Scale (Iwashita and Sohlberg, 2019) have simple scales and few descriptors

652   (e.g., eye contact, gesture, and initiation of new topics), however, the reliability results are

653   comparable to the Adapted Kagan Scales (Iwashita & Sohlberg, 2019).

654   Inclusion of the Adapted Kagan Scales is important as they are the only known scales

655   to measure support provided by communication partners during conversation. Therefore,

656   future recommendations may include the use of larger participant numbers and the potential

657   integration of automated analysis of some conversational skill behaviours that are

658   quantifiable such as percentage of speaking time and facial expressions (Liu et al 2016).

659   Additionally, there may be consideration of other scales focused on measuring the skills of

660   the person with ABI and the degree of communicative effectiveness (e.g., Conversational

661   discourse scale of the Montreal Evaluation of Communication, Joanette et al., 2015) or

662   inclusion of patient-reported outcome measures of perceived communicative ability and

663   participant experiences that are psychometrically robust such as the Communication

664   Participation Item Bank (Baylor et al., 2013); La Trobe Communication Questionnaire

665   (Douglas et al., 2000) and Social Skills Questionnaire-Traumatic Brain Injury (Francis et al.,

666   2017). Such changes will ensure ongoing data may be collected for feasibility and reliability

667   of the scales with consideration of their validity.

668

669   *Limitations*

670   Overall, this study was limited by its small sample size of 23 conversations, which as

671   a convenience sample may not represent the full range of scores from these scales. In

672   addition, a specific measure of cognitive-communication disorder was not used to recruit

673   participants. Researchers suggest for reliability studies, there should be at least 30 samples

674   with three raters (Koo & Li, 2016). Therefore, low ICCs may be attributable to fewer raters in

675   each phase and potentially a lack of variation among the sampled people with brain injury

676   and their communication partners (Eriksson et al., 2014) given the use of a convenience

677   sample. In addition, the conversation samples for this study were drawn from people who had

678   sustained both traumatic and non-traumatic injuries where previous studies have used

679   samples from only people with TBI (Behn et al., 2012; Rietdijk et al., 2018; Togher et al.,

680   2010). There was a dependence on a high proportion of student raters who were

681   predominantly female, however this is consistent with the speech and language therapy

682   profession. The student raters had limited knowledge and experience of brain injury and

683   associated communication problems, which may reduce the generalisability of the study

684   findings to how experienced brain injury clinicians may feasibly use these scales in clinical

685   practice.

686   While the think-aloud protocol was intended to provide rich qualitative data, this was

687   not consistently the case. A concurrent think-aloud interview while raters were viewing the

688   conversation was initially attempted but proved to be cognitively challenging, so a

689   retrospective think-aloud interview was used. Further training and consideration of rater

690   prompts through the think aloud process may be required in the future (Hu & Gao, 2017).

691   Despite this, the rater logs provided clear information during all four phases and was helpful

692   to developing a clear understanding of the challenges faced by raters. Another limitation

693   could relate to the statistics used in this study. To be transparent; correlations, significance

694   value, confidence intervals and percent agreement was reported. Bland Altman plots (1986)

695   may also be used to visualise disagreements in rating, degree of differences and assessor bias

696   (Eriksson et al., 2014) and may have generated further insights into the nature of the data.

697   This study did not further examine the content validity of the scales nor consider intra-rater or

698   test-retest reliability, with the latter relevant to the use of the scales as an outcome measure. A

699   reduction in the number of scales had little to no effect on reliability however, it may affect

700   sensitivity to change so future research would need to consider whether there is a trade-off

701   between reliability and sensitivity for the measures.

702

703   *Clinical implications and future directions*

704   Rating scales of conversation offer a useful starting point for clinicians who are

705   conducting assessments with the goal of making clinical decisions and setting goals for

706   treatment. For example, they could help guide the clinician and the dyad as to the aims of

707   intervention (e.g., improving the communication partner's ability to reveal the competence of

708   the person with ABI by ensuring they can respond) and thus identify relevant target

709   behaviours for treatment (e.g., asking questions, take turns, give time to respond). In doing

710   so, the clinician can select targets that focus on person-centred and contextually relevant

711   conversations and topics that align with the values and needs of the person with brain injury

712   and their communication partner (Keegan et al., 2023; Sohlberg et al., 2019). Those target

713   behaviours could be translated to a goal setting framework such as Goal Attainment Scaling,

714   in collaboration with the dyad. The outcome of treatment would therefore be a positive

715   change to a discrete communicative behaviour or use of a specific strategy by the person with

716   brain injury and/or communication partner, to achieve a social activity or participation goal

717   (Behn et al., 2019b; Keegan et al., 2020), with the potential to evaluate progress using the

718   Adapted Kagan Scales on conversation samples collected across different timepoints. The

719   Adapted Kagan Scales have been found to be a sensitive outcome measure for demonstrating

720   positive change in conversations after communication partner training in multiple studies

721   (Behn et al., 2012; Rietdijk et al., 2020; Togher et al., 2013), which indicates they may be

722   clinically useful for this purpose. Future research that strengthens the psychometric properties

723   of the scales including for example, test-retest reliability, will be important to progressing the

724   use of these measures in research and clinical practice.

725   One additional solution to the problem of reliably evaluating conversation could be in

726   the form of emerging technologies and artificial intelligence. Computerised discourse

727   analysis programs and software programs for rating conversational discourse may be a future

728   innovation (Steel & Togher, 2019). Artificial intelligence has already been used for rating

729   conversational discourse to evaluate communication partner training for discrete conversation

730   behaviours that are identified by human review of videotaped conversations (e.g., open and

731   closed questions, long pauses, and yes/no questions) (Croteau et al., 2018). Artificial

732   intelligence has also been used to conduct a conversational assessment to help predict

733   depression (Weisenburger et al., 2024). Such technologies may be able to be adapted and

734   repurposed for rating conversations of people with brain injury and their communication

735   partners.

736

737   Conclusion

738   There is a need for reliable and valid measures of conversation that can be easily used

739   to assess social communication impairments, and which are time efficient. In this study, the

740   Adapted Kagan Scales were used to rate conversations involving people with brain injury and

741   their communication partners. A short training period (of four hours) enabled students and

742   clinicians to view and rate 5-minute conversations using two subscales in under 30 minutes:

743   with acceptable/moderate reliability. Conversation is dynamic, interactive, and complex; and

744   requires a clinician to make many judgements about the communicative behaviours of

745 participants. Use of several Adapted Kagan Scales (MPC-Interaction; and MSC-RC1) was

746 feasible and future research could evaluate how these scales may influence the goal setting

747 process and outcome measurement in communication partner training interventions. This

748 paper is intended to raise the importance of measuring social communication in dyads and

749 present a clinically feasible method for assessing these skills.

750

755

756 **Data Availability Statement**

757 Data available upon request

758

759 **References**

760

761 Badon, L. C., Oller Jr, J. W., & Oller, S. D. (2005). Ratings within and across ethnic

762   boundaries of methods of one on one reading instruction. *Journal of Communication*

763   *Disorders*, *38*(6), 445-457.

764 Baylor, C., Yorkston, K., Eadie, T., Kim, J., Chung, H., & Amtmann, D. (2013). The

765   Communicative Participation Item Bank (CPIB): Item bank calibration and

766   development of a disorder-generic short form. *Journal of Speech, Language, and*

767   *Hearing Research*, 56, 1190-1208.

768   Behn, N., Francis, J. J., Power, E., Hatch, E., & Hilari, K. (2020). Communication partner

769   training in traumatic brain injury: a UK survey of Speech and Language Therapists'

770   clinical practice. *Brain Injury*, *34*(7), 934-944.

771   Behn, N., Marshall, J., Togher, L., & Cruice, M. (2019a). Feasibility and initial efficacy of

772   project-based treatment for people with ABI. *International Journal of Language &*

773   *Communication Disorders*, *54*(3), 465-478.

774   Behn, N., Marshall, J., Togher, L., & Cruice, M. (2019b). Setting and achieving

775   individualized social communication goals for people with acquired brain injury

776   (ABI) within a group treatment. *International Journal of Language & Communication*

777   *Disorders*, *54*(5), 828-840.

778   Behn, N., Togher, L., Power, E., & Heard, R. (2012). Evaluating communication training for

779   paid carers of people with traumatic brain injury. *Brain Injury*, *26*(13-14), 1702-1715.

780   Bond, F., & Godfrey, H. (1997). Conversation with traumatically brain-injured individuals: A

781   controlled study of behavioural changes and their impact. *Brain Injury*, *11*(5), 319-

782   329.

783   CALSPO. (2015). *Practice standards and guidelines for acquired cognitive communication*

784   *disorders*.

785   Chia, A. A., Power, E., Kenny, B., Elbourn, E., McDonald, S., Tate, R., MacWhinney, B.,

786   Turkstra, L., Holland, A., & Togher, L. (2019). Patterns of early conversational

787   recovery for people with traumatic brain injury and their communication partners.

788   *Brain Injury*, *33*(5), 690-698.

789   Chinn, S. (1991). Statistics in respiratory medicine. 2. Repeatability and method comparison.

790   *Thorax*, *46*(6), 454.

791   Coelho, C., Ylvisaker, M., & Turkstra, L. S. (2005). Nonstandardized assessment approaches

792   for individuals with traumatic brain injuries. *Seminars in Speech and Language*,

793   *26*(4), 223-241.

794   Coelho, C. A., Liles, B. Z., & Duffy, R. J. (1991). Analysis of conversational discourse in

795   head-injured adults. *Journal of Head Trauma Rehabilitation*, *6*(2), 92-99.

796   Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M.

797   (2008). Internet-based learning in the health professions: a meta-analysis. *JAMA*,

798   *300*(10), 1181-1196.

799   Correll, A., Steenbrugge, W., & Scholten, I. (2010). Judging conversation: How much is

800   enough? *Aphasiology*, *24*(5), 612-622.

801   Croteau, C., McMahon-Morin, P., Le Dorze, G., Power, E., Fortier-Blanc, J., & Davis, G. A.

802   (2018). Exploration of a quantitative method for measuring behaviors in conversation.

803   *Aphasiology*, *32*(3), 247-263.

804   Dahlberg, C., Hawley, L., Morey, C., Newman, J., Cusick, C. P., & Harrison-Felix, C. (2006).

805   Social communication skills in persons with post-acute traumatic brain injury: Three

806   perspectives. *Brain Injury*, *20*(4), 425-435.

807   Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

808   Douglas, J. M., O'Flaherty, C. A., & Snow, P. C. (2000). Measuring perception of

809   communicative ability: The development and evaluation of the La Trobe

810   communication questionnaire. *Aphasiology*, *14*(3), 251-268.

811   Durning, S. J., Artino Jr, A. R., Beckman, T. J., Graner, J., Van Der Vleuten, C., Holmboe, E.,

812   & Schuwirth, L. (2013). Does the think-aloud protocol reflect thinking? Exploring

813   functional neuroimaging differences with thinking (answering multiple choice

814   questions) versus thinking aloud. *Medical teacher*, *35*(9), 720-726.

815   Elbourn, E., MacWhinney, B., Fromm, D., Power, E., Steel, J., & Togher, L. (2023).

816   TBIBank: An international shared database to enhance research, teaching and

817   automated language analysis for traumatic brain injury populations. *Archives of*

818   *Physical Medicine and Rehabilitation*, *104*(5), 824-829.

819   Eriksson, K., Bergstrom, S., Carlsson, E., Hartelius, L., Johansson, C., Schwarz, A., &

820   Saldert, S. (2014). Aspects of rating communicative interaction: Effects on reliability

821   and agreement. *Journal of Interactional Research in Communication Disorders*, *5*(2),

822   245-267.

823   Farrell, A. D., Rabinowitz, J. A., Wallander, J. L., & Curran, J. P. (1985). An evaluation of

824   two formats for the intermediate-level assessment of social skills. *Behavioral*

825   *Assessment*.

826   Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to

827   support high-stakes simulation-based assessments. *Journal of Continuing Education*

828   *in the Health Professions*, *32*(4), 279-286.

829   Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed

830   methods designs - principles and practices. *Health Services Research*, *48*(6pt2), 2134-

831   2156.

832   Finch, E., Copley, A., Cornwell, P., & Kelly, C. (2016). Systematic Review of Behavioral

833   Interventions Targeting Social Communication Difficulties After Traumatic Brain

834   Injury. *Archives of Physical Medicine and Rehabilitation*, *97*(8), 1352-1365.

835   Francis, H. M., Osborne-Crowley, K., & McDonald, S. (2017). Validity and reliability of a

836   questionnaire to assess social skills in traumatic brain injury: a preliminary study.

837   *Brain Injury*, *31*(3), 336-343.

838   Frith, M., Togher, L., Ferguson, A., Levick, W., & Docking, K. (2014). Assessment practices

839   of speech-language pathologists for cognitive communication disorders following

840   traumatic brain injury in adults: an international survey. *Brain Injury*, *28*(13-14),

841   1657-1666.

842   Galski, T., Tompkins, C., & Johnston, M. (1998). Competence in discourse as a measure of

843   social integration and quality of life in persons with traumatic brain injury. *Brain*

844   *Injury*, *12*(9), 769-782.

845   Harrison, A. J., Long, K. A., Tommet, D. C., & Jones, R. N. (2017). Examining the role of

846   race, ethnicity, and gender on social and behavioral ratings within the Autism

847   Diagnostic Observation Schedule. *Journal of autism and developmental disorders*, *47*,

848   2770-2782.

849   Hartley, L. L., & Jensen, P. J. (1992). Three discourse profiles of closed-head-injury speakers:

850   Theoretical and clinical implications. *Brain Injury*, *6*(3), 271-282.

851   Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card*

852   *Sorting Test*. Psychological Assessment Resources, Inc.

853   Hewetson, R., Cornwell, P., & Shum, D. (2017). Cognitive-communication disorder

854   following right hemisphere stroke: exploring rehabilitation access and outcomes.

855   *Topics in Stroke Rehabilitation*, *24*(5), 330-336.

856   Howell, S., Varley, R., Sinnott, E. L., Pring, T., & Beeke, S. (2021). Measuring group social

857   interactions following acquired brain injury: an inter-rater reliability evaluation.

858   *Aphasiology*, *35*(11), 1505-1517.

859   Hsieh, H.-F., & Shannon, S. (2005). Three approaches to qualitative content analysis.

860   *Qualitative Health Research*, *15*(9), 1277-1288.

861   Hu, J., & Gao, X. A. (2017). Using think-aloud protocol in self-regulated reading research.

862   *Educational Research Review*, *22*, 181-193.

863 Iwashita, H., & Sohlberg, M. M. (2019). Measuring conversations after acquired brain injury

864 in 30 minutes or less: a comparison of two pragmatic rating scales. *Brain Injury*,

865 *33*(9), 1219-1233.

866 Jacobs, R., & Kozlowski, S. W. J. (1985). A closer look at halo error in performance ratings.

867 *Academy of Management Journal*, *28*(1), 201-212.

868 Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgment decomposition, and

869 their impact on interrater agreement and rating accuracy. *Journal of applied*

870 *psychology*, *75*(5), 500.

871 Joanette, Y., Ska, B., Cote, H., Ferre, P., LaPointe, L., Coppens, P., & Small, S. (2015).

872 *Montreal Protocol for the Evaluation of Communication*. ASSBI Resources.

873 Kagan, A., Winckel, J., Black, S. E., Duchan, J. F., Simmons-Mackie, N., & Square, P.

874 (2004). A set of observational measures for rating support and participation in

875 conversation between adults with aphasia and their conversation partners. *Topics in*

876 *Stroke Rehabilitation*, *11*(1), 67-83.

877 Keegan, L. C., Hoepner, J. K., Togher, L., & Kennedy, M. (2023). Clinically applicable

878 sociolinguistic assessment for cognitive-communication disorders. *American Journal*

879 *of Speech-Language Pathology*, *32*(2S), 966-976.

880 Keegan, L. C., & Müller, N. (2022). The influence of context on identity construction after

881 traumatic brain injury. *Journal of Interactional Research in Communication*

882 *Disorders*, *13*(2), 171-195.

883 Keegan, L. C., Murdock, M., Suger, C., & Togher, L. (2020). Improving natural social

884 interaction: Group rehabilitation after Traumatic Brain Injury. *Neuropsychological*

885 *Rehabilitation*, *30*(8), 1497-1522.

886   Kelly, M., McDonald, S., & Frith, M. H. J. (2017). A survey of clinicians working in brain

887   injury rehabilitation: Are social cognition impairments on the radar? *Journal of Head*

888   *Trauma Rehabilitation*, *32*(4), E55-E65.

889   Knox, L., & Douglas, J. (2009). Long-term ability to interpret facial expression after

890   traumatic brain injury and its relation to social integration. *Brain and Cognition*,

891   *69*(2), 442-449.

892   Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation

893   coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

894   Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts,

895   C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and

896   agreement studies (GRRAS) were proposed. *International journal of nursing studies*,

897   *48*(6), 661-671.

898   Lê, K., Coelho, C., & Fiszdon, J. (2022). Systematic Review of Discourse and Social

899   Communication Interventions in Traumatic Brain Injury. *American Journal of Speech-*

900   *Language Pathology*, 1-32.

901   Linscott, R. J., Knight, R. G., & Godfrey, H. P. D. (1996). The Profile of Functional

902   Impairment in Communication (PFIC): a measure of communication impairment for

903   clinical use. *Brain Injury*, *10*.

904   Liu C, Lim RL, McCabe KL, Taylor S, Calvo RA (2016). A Web-Based Telehealth Training

905   Platform Incorporating Automated Nonverbal Behavior Feedback for Teaching

906   Communication Skills to Medical Students: A Randomized Crossover Study. *Journal*

907   *of Medical Internet Research*. *18(9),* e246.

908   MacDonald, S. (2017). Introducing the model of cognitive-communication competence: A

909   model to guide evidence-based communication interventions after brain injury. *Brain*

910   *Injury*, *31*(13-14), 1760-1780.

911   Maddy, K., Howell, D., & Capilouto, G. (2015). Current practices regarding discourse

912   analysis and treatment following non-aphasic brain injury: A qualitative study.

913   *Journal of Interactional Research in Communication Disorders*, *6*(2), 211-236.

914   Mann, K., Power, E., Barnes, S., & Togher, L. (2015). Questioning in conversations before

915   and after communication partner training for individuals with traumatic brain injury.

916   *Aphasiology*, *29*(9), 1082-1109.

917   McDonald, S., Tate, R., Togher, L., Bornhofen, C., Long, E., Gertler, P., & Bowen, R. (2008).

918   Social skills treatment for people with severe, chronic acquired brain injuries: A

919   multicenter trial. *Archives of Physical Medicine and Rehabilitation*, *89*(9), 1648-1659.

920   Meulenbroek, P., & Turkstra, L. S. (2016). Job stability in skilled work and communication

921   ability after moderate-severe traumatic brain injury. *Disability and Rehabilitation*,

922   *38*(5), 452-461.

923   Morrow, E. L., Hereford, A. P., Covington, N. V., & Duff, M. C. (2020). Traumatic brain

924   injury in the acute care setting: assessment and management practices of speech-

925   language pathologists. *Brain Injury*, *34*(12), 1590-1609.

926   Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical diagnosis of

927   mental disorders: A handbook*, 97-146.

928   Olver, J., Ponsford, J., & Curran, C. (1996). Outcome following traumatic brain injury: A

929   comparison between 2 and 5 years after injury. *Brain Injury*, *10*(11), 841-848.

930   Ponsford, J. L., Downing, M. G., Olver, J., Ponsford, M., Acher, R., Carty, M., & Spitz, G.

931   (2014). Longitudinal follow-up of patients with traumatic brain injury: outcome at

932   two, five, and ten years post-injury. *Journal of Neurotrauma*, *31*(1), 64-77.

933   Portney, L. G., & Watkins, M. P. (2014). *Foundations of Clinical Research: Applications to

934   Practice* (3rd ed.). Pearson Education Limited.

935   Randolph, C. (1998). *Repeatable Battery for the Assessment of Neuropsychological Status*.

936   The Psychological Corporation.

937   Rietdijk, R., Power, E., Attard, M., Heard, R., & Togher, L. (2020a). Improved conversation

938   outcomes after social communication skills training for people with traumatic brain

939   injury and their communication partners: a clinical trial investigating in-person and

940   telehealth delivery. *Journal of Speech, Language, and Hearing Research*, *63*(2), 615-

941   632.

942   Rietdijk, R., Power, E., Brunner, M., & Togher, L. (2020b). The reliability of evaluating

943   conversations between people with traumatic brain injury and their communication

944   partners via videoconferencing. *Neuropsychological rehabilitation*, *30*(6), 1074-1091.

945   Rietdijk, R., Simpson, G., Togher, L., Power, E., & Gillett, L. (2013). An exploratory

946   prospective study of the association between communication skills and employment

947   outcomes after severe traumatic brain injury. *Brain Injury*, *27*(7-8), 812-818.

948   Saldert, C., Backman, E., & Hartelius, L. (2013). Conversation partner training with spouses

949   of persons with aphasia: A pilot study using a protocol to trace relevant

950   characteristics. *Aphasiology*, *27*(3), 271-292.

951   Shelton, C., & Shryock, M. (2007). Effectiveness of communication/interaction strategies

952   with patients who have neurological injuries in a rehabilitation setting. *Brain Injury*,

953   *21*(12), 1259-1266.

954   Shorland, J., Douglas, J., & O'Halloran, R. (2022). Age-based trends in cognitive-

955   communication management for adults in subacute rehabilitation following new onset

956   traumatic brain injury. *American Journal of Speech-Language Pathology*, *31*(6),

957   2557-2568.

958   Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability.

959   *Psychological Bulletin*, *86*(2), 420-428.

960 Sim, P., Power, E., & Togher, L. (2013). Describing conversations between individuals with

961   traumatic brain injury (TBI) and communication partners following communication

962   partner training: Using exchange structure analysis. *Brain Injury*, *27*(6), 717-742.

963 Slagle, J., Weinger, M. B., Dinh, M.-T. T., Brumer, V. V., & Williams, K. (2002). Assessment

964   of the intrarater and interrater reliability of an established clinical task analysis

965   methodology. *The Journal of the American Society of Anesthesiologists*, *96*(5), 1129-

966   1139.

967 Snow, P., Douglas, J., & Ponsford, J. (1997). Conversational assessment following traumatic

968   brain injury: A comparison across two control groups. *Brain Injury*, *11*(6), 409-429.

969 Snow, P., Douglas, J., & Ponsford, J. (1998). Conversational discourse abilities following

970   severe traumatic brain injury: A follow-up study. *Brain Injury*, *12*(11), 911-935.

971 Soffer, T., & Nachmias, R. (2018). Effectiveness of learning in online academic courses

972   compared with face-to-face courses in higher education. *Journal of Computer assisted*

973   *learning*, *34*(5), 534-543.

974 Sohlberg, M. M., MacDonald, S., Byom, L., Iwashita, H., Lemoncello, R., Meulenbroek, P.,

975   Ness, B., & O'Neil-Pirozzi, T. M. (2019). Social communication following traumatic

976   brain injury part I: State-of-the-art review of assessment tools. *International Journal*

977   *of Speech Language Pathology*, *21*(2), 115-127.

978 Spence, S. E., Godfrey, H. P. D., Knight, R. G., & Bishara, S. N. (1993). First impressions

979   count: A controlled investigation of social skill following closed head injury. *British*

980   *Journal of Clinical Psychology*, *32*, 309-318.

981 Steel, J., & Togher, L. (2019). Social communication assessment after TBI: a narrative review

982   of innovations in pragmatic and discourse assessment methods. *Brain Injury*, *33*(1),

983   48-61.

984   Struchen, M., Clark, A., Sander, A., Mills, M., Evans, G., & Kurtz, D. (2008). Relation of

985   executive functioning and social communication measures to functional outcomes

986   following traumatic brain injury. *NeuroRehabilitation*, *23*(2), 185-198.

987   Tobar-Fredes, R., & Salas, C. (2022). Rehabilitation of communication in people with

988   traumatic brain injury: a systematic review of types of intervention and therapeutic

989   ingredients (Rehabilitación de la comunicación en personas con traumatismo

990   encefalocraneal: una revisión sistemática de tipos de intervención e ingredientes

991   terapéuticos). *Studies in Psychology*, *43*(1), 88-131.

992   Togher, L., Douglas, J., Turkstra, L. S., Welch-West, P., Janzen, S., Harnett, A., Kennedy, M.,

993   Kua, A., Patsakos, E., & Ponsford, J. (2023). INCOG 2.0 guidelines for cognitive

994   rehabilitation following traumatic brain injury, part IV: cognitive-communication and

995   social cognition disorders. *Journal of Head Trauma Rehabilitation*, *38*(1), 65-82.

996   Togher, L., Hand, L., & Code, C. (1997). Analysing discourse in the traumatic brain injury

997   population: Telephone interactions with different communication partners. *Brain*

998   *Injury*, *11*(3), 169-189.

999   Togher, L., McDonald, S., Tate, R., Power, E., & Rietdijk, R. (2013). Training

1000   communication partners of people with severe traumatic brain injury improves

1001   everyday conversations: A multicenter single blinded clinical trial. *Journal of*

1002   *Rehabilitation Medicine*, *45*, 637-645.

1003   Togher, L., Power, E., Tate, R., McDonald, S., & Rietdijk, R. (2010). Measuring the social

1004   interactions of people with traumatic brain injury and their communication partners:

1005   The adapted Kagan scales. *Aphasiology*, *24*(6-8), 914-927.

1006   Weisenburger, R. L., Mullarkey, M. C., Labrada, J., Labrousse, D., Yang, M. Y., MacPherson,

1007   A. H., Hsu, K. J., Ugail, H., Shumake, J., & Beevers, C. G. (2024). Conversational

1008   assessment using artificial intelligence is as clinically useful as depression scales and

1009   preferred by users. *Journal of affective disorders*.

1010   Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013a). Seeing the same thing differently:

1011   Mechanisms that contribute to assessor differences in directly-observed performance

1012   assessments. *Advances in Health Sciences Education*, *18*, 325-341.

1013   Yeates, P., O'Neill, P., Mann, K., & W Eva, K. (2013b). 'You're certainly relatively

1014   competent': assessor bias due to recent experiences. *Medical education*, *47*(9), 910-

1015   922.

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034   Table 1. *Demographic variables*

	ALL people with ABI (n = 21)
Age	45.80 ± 14.47
Gender	
Male	12
Female	9
Years post-injury	11.95 ± 12.69
Injury type	
Trauma	13
Non-trauma	8
Injury severity (n=13)[a]	
Severe	12
Moderate	1
Living arrangements	
Alone	5
With others	15
Care home	1
Employment status	
Full-time	1
Part-time	2
Unemployed	18
Communication partner	
Family member	11
Spouse	4
Friend	3
Paid carer	3
RBANS	
Total score	70.85 ± 15.27
WCST	
Categories	3.62 ± 1.78
Perseverative errors	25.24 ± 15.47

1035   [a]Injury severity can only be determined for traumatic injuries
1036   *Note.* Values are mean ± SD. RBANS = Repeatable Battery of
1037   Assessment of Neuropsychological Status (average score = 90 - 109);
1038   WCST = Wisconsin Card Sorting Test (average categories = 5.07;
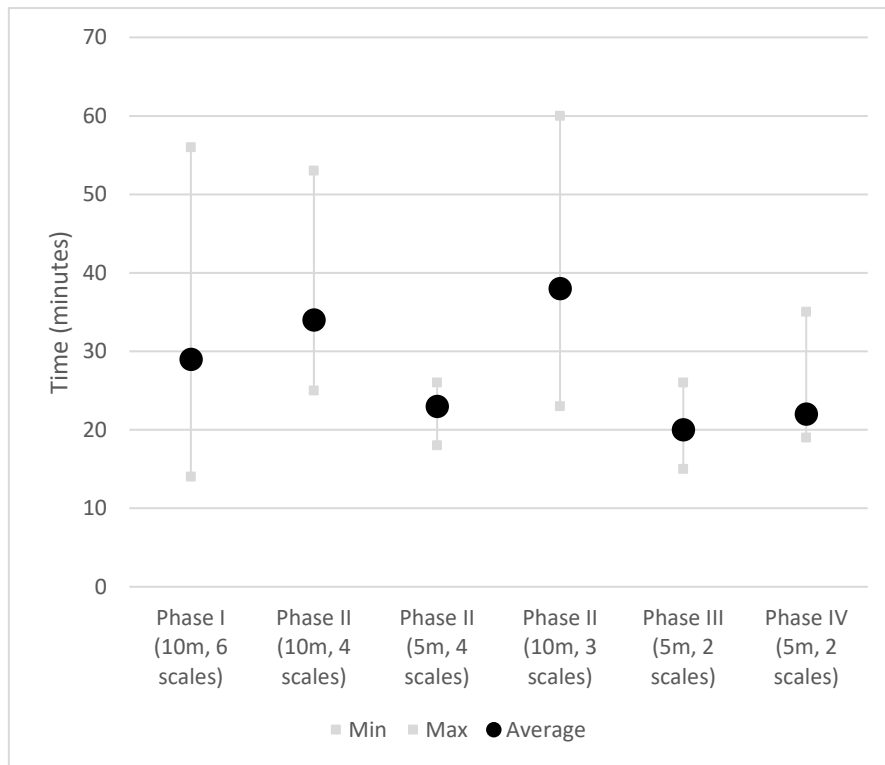1039   average perseverative errors = 15.78).

1040

1041

1042

1043

1044

Figure 1. Time (in minutes) to watch and rate conversations

1045

1046

1047
1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060   Table 2. ICCs for Phase I conditions

	Half-day training (rate 10 mins)			Full-day training (rate 10 mins)		
	ICC[a]	95% CI	% agreement within 0.5	ICC[a]	95% CI	% agreement within 0.5
MPC						
Interaction	0.60	[0.36, 0.78]	39%	0.58	[0.34, 0.78]	43%
Transaction	0.49	[0.50, 0.88]	35%	0.47	[0.23, 0.70]	39%
MSC						
AC	0.51	[0.27, 0.73]	26%	0.54	[0.30, 0.75]	39%
RC1	0.52	[0.26, 0.73]	22%	0.53	[0.28, 0.74]	22%
RC2	0.62	[0.40, 0.80]	26%	0.43	[0.17, 0.67]	30%
RC3	0.49	[0.22, 0.72]	17%	0.62	[0.39, 0.80]	30%

1061   ICC, Intraclass Correlation; CI, Confidence Intervals; MPC, Measure of Participation in Conversation; MSC, Measure of
1062   Support In Conversation; AC, Acknowledging Competence; RC, Revealing Competence; RC1, Ensure the adults
1063   understands; RC2, Ensure the adult has a means of responding; and RC3, Verification.
1064   [a]$p < .001$

1065
1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081   Table 3. ICCs for Phase II conditions

	Rate 10 mins, 4 scales			Rate 5 mins, 4 scales			Rate 10 mins, 3 scales		
	ICC	95% CI	% agreement within 0.5	ICC	95% CI	% agreement within 0.5	ICC	95% CI	% agreement within 0.5
MPC									
Interaction	0.58[a]	[0.17, 0.80]	52%	0.63[a]	[0.30, 0.82]	83%	0.66[a]	[0.19, 0.86]	74%
Transaction	0.33[c]	[-0.05, 0.64]	39%	0.59[b]	[0.25, 0.80]	78%	-	-	-
MSC									
RC1	0.56[a]	[0.17, 0.79]	61%	0.68[a]	[0.38, 0.85]	70%	0.68[a]	[0.26, 0.87]	65%
RC2	0.59[b]	[0.24, 0.80]	52%	0.56[b]	[0.20, 0.79]	70%	0.52[a]	[0.11, 0.77]	48%

1082   ICC, Intraclass Correlation; CI, Confidence Intervals; MPC, Measure of Participation in Conversation; MSC, Measure of Support In Conversation; RC, Revealing Competence; RC1, Ensure the
1083   adults understands; RC2, Ensure the adult has a means of responding; and RC3, Verification.
1084   [a]$p < .001$
1085   [b]$p < .01$
1086   [c]$p < .05$
1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100   Table 4. ICCs for Phase III (student) and IV (experienced clinicians) conditions

	Student raters			Experienced clinicians		
	ICC	95% CI	% agreement within 0.5	ICC	95% CI	% agreement within 0.5
MPC						
Interaction	0.69[a]	[0.40, 0.86]	78%	0.55[b]	[0.19, 0.78]	70%
MSC						
RC1	0.73[a]	[0.47, 0.88]	57%	0.58[b]	[0.22, 0.80]	48%

1101   ICC, Intraclass Correlation; CI, Confidence Intervals; MPC, Measure of Participation in Conversation; MSC, Measure of
1102   Support In Conversation; RC, Revealing Competence; RC1, Ensure the adults understands.
1103   [a]p < .001
1104   [b]p < .01
1105
1106
1107
1108

1109   *Table 5*. Summary of time taken to rate, ease of use and reliability measures across all four

1110   phases

Phase Conversation length, scales rated	Average time to rate	Average ease of use score (range)	Reliability (ICCs)
I			
10 mins, 6 scales	29 mins	6.8 (5-10)	Poor-to-moderate (.43 - .62)
II			
10 mins, 4 scales	34 mins	6.8 (2-10)	Poor-to-moderate (.33 - .59)
5 mins, 4 scales	23 mins	6.4 (4-9)	Moderate (.56 - .68)
10 mins, 3 scales	38 mins	6.5 (4-9)	Moderate (.52 - .68)
III			
5 mins, 2 scales	20 mins	7.6 (2-10)	Moderate (.69 - .73)
IV			
5 mins, 2 scales	22 mins	7.3 (4-9)	Moderate (.55 - .58)

1111   ICC, Intraclass Correlation

1112