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).

RESULTS: Raters took 30 - 45 minutes to score a 10-minute sample; and 20 - 30 minutes to score a 5-minute sample. Ease of use was rated highly across all phases. Overall reliability for rating 5-minutes of conversation (ICC = 0.52-0.73) was better than for 10-minutes of conversation (ICC = 0.33 - 0.68). Reliability for the MPC was moderate for both students (ICC = 0.69) and clinicians (ICC = 0.55); and for the MSC, moderate for both students (ICC = 0.73) and clinicians (ICC= 0.58). Reliability was better for students compared with clinicians.

CONCLUSION: Rating a 5-minute conversation in under 30 minutes was feasible, with more reliable results for 5-minute compared with 10-minute conversations. Implications for assessing conversation in the future are discussed.

Corresponding author

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

Conflicts of interest

No financial or other nonprofessional benefits to declare

Keywords

Social communication; brain injury; assessment

INTRODUCTION

Impaired communication is common for people with acquired brain injury (ABI). People may talk too much or too little; perseverate on a topic or go off on a tangent; lack initiation or frequently interrupt; have difficulty with taking turns and talk over people; not

listen to others and be disruptive; or be socially inappropriate in their interactions with others (Coelho et al., 1991; Hartley & Jensen, 1992; Sim et al., 2013; Spence et al., 1993). These impairments have often been described as lying on a spectrum from impoverished (lack initiation, sparse, reduced content) to excessive (over talkative, tangential, repetitive) (MacDonald, 2017; Sim et al., 2013). These impairments are often referred to as a *cognitive-communication disorder* (CALSPO, 2015) to highlight the impact of impaired cognitive processes on a person's ability to communicate. Over two-thirds of people who sustain an ABI present with some form of cognitive communication impairment (Hewetson et al., 2017; Kelly et al., 2017; Shorland et al., 2022). This disorder is heterogeneous (Hartley & Jensen, 1992; Snow et al., 1997) with impairments known to be both long-term and pervasive (Knox & Douglas, 2009; Olver et al., 1996; Ponsford et al., 2014; Snow et al., 1998). The impacts of this disorder are far-reaching, negatively affecting a person's ability to return to work (Meulenbroek & Turkstra, 2016; Rietdijk et al., 2013), integrate socially (Dahlberg et al., 2006; Knox & Douglas, 2009; Struchen et al., 2008) and achieve a better quality of life (Dahlberg et al., 2006; Galski et al., 1998).

Communication is a dynamic, two-way process involving both the person with ABI and their communication partner, whether a family member, friend, or carer. The skills of the partner can either hinder or facilitate a conversation (Togher et al., 1997). Just as a person with ABI may struggle in conversation, so may the communication partner. Partners may frequently ask questions that test a person's knowledge, limit opportunities for the person with ABI to participate and/or not give the person with ABI a turn to respond (Mann et al., 2015; Sim et al., 2013). Conversely, an increased use of a supportive questioning style and use of positive communication strategies by partners (e.g., use of short, simple direct sentences and questions) may improve interactions (Mann et al., 2015; Shelton and Shryock, 2007).

Given the important role that communication partners play in conversational interactions, training partners is recommended within international guidelines (Togher et al., 2023) and recent systematic reviews (Behn et al., 2020; Wiseman-Hakes et al., 2020). As part of the training process, assessment of conversation is integral to establishing an understanding of the skills of the person with ABI, the ability of the communication partner to support interactions, and to subsequently guide planning of relevant interventions. Conversation is also considered a key outcome for any cognitive-communication intervention (Lê et al., 2022; Tobar-Fredes & Salas, 2022), particularly for determining whether training communication partners has been beneficial to the dyad (Togher et al., 2023).

Assessing conversation provides insights into real-life communication with relevant partners and may illuminate communication skills that have been impaired by the brain injury (Keegan et al., 2023; MacDonald, 2017). However, assessing conversation can be difficult due to its dynamic and interactive nature; and may vary according to the type of conversation (e.g., casual, purposeful, task-specific) and the communication partner involved (e.g., family member, sibling, friend, carer). Furthermore, there is a lack of tools that objectively and reliably evaluate conversation in ecologically valid ways (Sohlberg et al., 2019). Pragmatic or observational scales are common (Keegan et al., 2023; Sohlberg et al., 2019; Steel & Togher, 2019) though these measures are limited by reduced reliability and consistency (Coelho et al., 2005).

Detailed assessment of the quality of conversation is not routinely assessed in clinical practice. An international survey of 265 speech and language therapists from a range of clinical settings found under 10% of clinicians directly assess functional performance, pragmatics, and discourse (Frith et al., 2014). Findings are similar for therapists (n=182) in acute settings, with fewer than 20% assessing conversation (Morrow et al., 2020). A recent international survey of speech and language therapists from mainly rehabilitation and

community settings (n = 70) found that 80% of clinicians assessed conversation (Steel et al., 2022). However, the most common type of analysis (> 90%) focussed only on pragmatic features (e.g., eye contact, topic maintenance) of the conversation. Common barriers to both assessment and detailed analysis include the lack of resources and time, and limited availability of tools (Frith et al., 2014; Kelly et al., 2017; Maddy et al., 2015; Morrow et al., 2020; Steel et al., 2022). More detailed analyses beyond pragmatic features alone are needed to guide intervention that enables people with ABI and their communication partners to participate effectively in conversation and in their social lives.

Therefore, access to assessments that can feasibly and reliably measure conversation in clinical practice is needed. Sohlberg and colleagues (2019) described feasibility of a measure in terms of time and complexity of administration. A measure that took no longer than 60 minutes to administer and did not require a complex analysis procedure such as transcription and hand coding was considered feasible. In that study, only one (of six) measures the Profile of Pragmatic Impairments in Communication (Linscott et al., 1996) was not considered to be feasible. Iwashita and Sohlberg (2019) described a feasible measure for clinicians as one that was acceptable to clinicians and administrated in 30 minutes or less. The Modified Pragmatic Rating Scale was compared to the Profile of Pragmatic Impairments in Communication. The former was found to be quicker to rate (in under 5 minutes) and described by raters as easier to use. However, a limitation of these conversational scales is that they focus on the skills of the person with ABI, and do not score or rate the skills of the communication partner within a conversation.

One commonly reported measure of conversation that focuses on both the person with ABI and their communication partner is the Adapted Kagan Scales (Togher et al., 2010). These scales are clinician-rated, do not require transcription or detailed linguistic analyses, and have demonstrated sensitivity to change from communication partner training (Behn et

al., 2012; Rietdijk et al., 2020a; Togher et al., 2013). Originally designed to rate conversations involving people with aphasia (Kagan et al., 2004), these scales were adapted for people with brain injury and their communication partners (Togher et al., 2010). The Adapted Kagan Scales comprise two scales, each with several sub-scales. The first, the Adapted Measure of Participation in Conversation (MPC) rates the interaction and transactional skills of the person with brain injury. The second, the Adapted Measure of Support in Conversation (MSC) rates the ability of the communication partner to both acknowledge and reveal the competence of the person with brain injury within the conversation. These tools are the only available scales that rate the skills of both people in the dyad. The scales have excellent inter-rater and intra-rater reliability when rated by experienced clinicians (Togher et al., 2010), good ecological validity, and based on the parameters described by Sohlberg et al. (2019), would be considered feasible in terms of time to rate and ease of use.

Although the Adapted Kagan Scales have been found to have acceptable reliability in research contexts, the clinical feasibility of the Adapted Kagan Scales is likely to be affected by a range of factors. Empirical studies have reported varying degrees of inter-rater reliability (Behn et al., 2019a; Behn et al., 2012; Chia et al., 2019; Rietdijk et al., 2020b; Togher et al., 2013) with the time taken to train raters ranging from 2.5 hours to 35 hours with better reliability reported for longer training times of at least 14 hours (Behn et al., 2019a; Behn et al., 2012; Chia et al., 2019; Rietdijk et al., 2020b). Raters have ranged from students studying speech and language therapy with limited experience of people with brain injury to clinicians with little to extensive clinical experience. The length of conversation has ranged from 5- to 10-minutes and the type of conversation has included casual and purposeful (or structured) conversation. Casual conversations have involved a dyad talking about a topic of interest, while purposeful conversations require the dyad to complete a task (e.g., plan a holiday) or

ask structured questions. Reliability results across different lengths and types of conversation have been comparable in some studies (Behn et al., 2012; Togher et al., 2010) and better for purposeful than casual conversations in other studies (Rietdijk et al., 2020a; Rietdijk et al., 2020b). All these factors may impact the extent and ease of implementation of the Adapted Kagan Scales in clinical practice.

The purpose of this study was to determine whether the Adapted Kagan Scales could be established as a clinically feasible method (i.e., completed in 30 minutes or less) for assessing a single type of conversation; and could achieve acceptable levels of inter-rater reliability with limited training. The face, ecological and construct validity of the measures has already been established (Kagan et al., 2004; Sohlberg et al., 2019; Togher et al., 2010). The same conversations were used across multiple phases to allow direct comparison; with consideration of training length; scales rated; and rater experience. This study aims to address the following research questions:

1. Can videotaped conversations involving people with ABI and their communication partners be feasibly rated in terms of time taken (30 minutes or less) using the Adapted Kagan Scales?
2. Can acceptable (i.e., moderate) reliability be achieved by students and experienced clinicians?
3. Can raters achieve acceptable (i.e., moderate) levels of reliability with limited training (<8 hours of training) in the use of the scales?
4. Can similar levels of reliability be achieved from rating 5-minute compared with 10-minute videotaped conversations?
5. What is raters feedback on their experience of using the Adapted Kagan Scales?

METHODS

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

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

[insert Table 1 about here]

Measures

The Adapted Kagan Scales (Togher et al., 2010) comprise two main scales. The first, the Adapted Measure of Participation in Conversation (MPC) is used to rate the conversational participation of the person with ABI, specifically evaluating how they socially connect, engage, and share the conversation with their communication partner. The scale is further divided into two subscales: Interaction (social connection) and Transaction (exchanging content).

The second scale is the Adapted Measure of Skill in Supported Conversation (MSC), which rates the skills of the communication partner during the conversation. This scale is divided into two subscales: Acknowledging competence (AC) and Revealing competence (RC). The Revealing Competence subscale involves three elements: (RC1) Ensure the adult understands; (RC2) Ensure the adult has a means of responding; and (RC3) Verification.

Each scale is rated on a 9-point Likert scale presented as a range of 0 – 4 with 0.5 intervals. There are behavioural descriptors and five anchor points to help guide the rater's judgement. For the MPC the anchor points range from 0 (no participation) to 4 (full participation in conversation) while the MSC anchor points range from 0 (not supportive) to 4 (highly skilled support). In total, six ratings are obtained: one for each subscale of the MPC (interaction and transaction), one for the Acknowledge Competence subscale of the MSC, and one for each of the three elements from the Revealing Competence subscale, which can later be averaged to give a total subscale score.

227

228 *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

235 *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:

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

using the TBI Bank Grand Rounds training package (Elbourn et al., 2023). The raters rated only the first 5-minutes of conversation using two scales (MPC Interaction; and one element of the Revealing Competence scale – ensure the adult understands).

Phase IV (Experienced raters, half-day training, 5 min samples, two scales): The aim of this phase was to examine the ratings of experienced raters compared with students from the previous phase. Two qualified speech and language therapists were recruited as raters (both female). Both raters received four hours of direct training on the scales. The raters followed the exact procedure from the previous phase. They rated only the first 5-minutes of conversation using two scales (MPC Interaction; and one element of the Revealing Competence scale – ensure the adult understands).

Raters in phase one were required to watch the conversation at least once, with the option of repeat viewing. They were asked to record the number of times they watched the conversation. In phases 2 - 4, raters were required to watch each conversation twice; in clusters of three samples at a time to reduce rater fatigue (Eriksson et al., 2014). Raters in all phases were required to record the following feasibility information for each rating: (1) the time taken to watch (and re-watch, where appropriate), consider behaviours observed and decide on a final rating of the conversation; (2) the ease of rating the conversation on a 10-point scale (1 = not easy; 10 = very easy); (3) qualitative feedback about the ease of rating; and (4) qualitative feedback about the descriptors in the scale, and the rating process. In phases three and four, a think-aloud protocol was included to gather more detailed information about the rating process. Think Aloud is a technique in which participants verbalise their thoughts while simultaneously carrying out a task to gain in-depth qualitative data, in this case on anything which affected the rating of the conversations (Durning et al.,

2013). Raters in these phases were observed online via *Zoom* (Version 5.12.8) scoring the same three conversations while recorded by a researcher (PP).

Training

All direct training on the scales was led by the first author who had more than 20 years experience in working with people with ABI, and who had more than 10 years' experience in training and using the Adapted Kagan Scales for research purposes (Behn et al., 2019a; Behn et al., 2012). The first author collaborated with other co-authors (EP, LT, and RR) to refine training in the use of the scales. Training was delivered in groups of two-to-four raters; in-person in phase one then online for phases 2-4 due to the COVID-19 pandemic. While the decision to do training online was influenced by external factors, studies have shown online training to be as effective as face-to-face methods (Cook et al., 2008; Soffer & Nachmias, 2018).

Training began with a general familiarisation of the scales and the rating process. Raters then watched several sample conversations. These conversations were accessed through TBI Bank, which is an online repository of materials and resources including videotaped conversations of people with traumatic brain injury and their communication partners (Elbourn et al., 2023). The full-day training (8-hours) involved eight sample conversations while the half-day training (4-hours) involved four-to-five sample conversations. Raters independently scored each sample conversation individually then discussed their scoring and any discrepancies with the group with reference to the descriptors and anchor points. Common issues that could influence the rating process were discussed and examined in relation to the sample conversations (e.g.,, relationship of partner, weighing up different descriptors, imbalanced conversations). Final ratings for the sample training conversations were agreed on via consensus, which became anchors for different points on

the 9-point scale, to provide a reference point when rating. All raters across all phases were permitted to review anchor videos as often as needed.

After phase one, rater feedback suggested a need for student raters to gain further knowledge of the communication problems that can occur from an ABI. Therefore, raters in phases two and three were required to do the online TBI Bank Grand Rounds training (Modules 1-3, 5 and 7) (Elbourn et al., 2023). This training was self-directed and took raters approximately four hours to complete in addition to the direct rater training already received. Topics covered were cognitive-communication disorders, discourse, and variability of discourse across contexts.

Data analysis

Feasibility information (time taken to rate; ease of rating scores) was compiled in a Microsoft Excel spreadsheet and analysed descriptively. Qualitative feedback on the experience of using the rating scales was initially compiled and analysed by the first and/or third author (NB or PP) using conventional content analyses (Hsieh & Shannon, 2005). The data was coded, identifying similarities and differences in the feedback of raters, with categories of information identified. Qualitative data from the think-aloud sessions were transcribed verbatim and analysed by the second author using conventional content analysis and checked for accuracy by the first author. As recommended in mixed methods research (Fetters, Curry & Creswell, 2013), we collected the quantitative and qualitative data in parallel and analysed the data for integration prior to the commencement of successive phases. Therefore, the preliminary integration of data was completed at the end of phase one, two and three. At the end of the study all qualitative data was synthesised with the quantitative data from all four phases to explain the findings of the study.

Reliability data from each phase were examined separately including, the effect of relevant variables such as the length of training, length of conversation viewed, and the experience of the rater. Interrater reliability was assessed using intra-class correlations (ICC) 2,1 procedure with absolute agreement, single measures (Shrout & Fleiss, 1979). The 95% confidence intervals were reported as percent agreement in line with reporting guidelines for reporting reliability results (Kottner et al., 2011). Excellent reliability was defined as ICCs greater than 0.90, good reliability as between 0.75 and 0.90, moderate reliability between 0.50 and 0.75 and poor reliability as less than 0.5 (Koo & Li, 2016). Acceptable reliability in this study was determined as moderate. As both the MPC and MSC include several subscales, Spearman's rank-order correlations (r_s) were calculated for each scale to evaluate the strength of the association between subscales. Strong correlations have values between 0.7 and 0.9, moderate correlations between 0.4 and 0.6, and weak correlations between 0.1 and 0.3 (Dancey & Reidy, 2007). Throughout all phases the strength of the ICCs and correlations and feasibility information were examined by the research team to inform each phase. All statistical analyses were computed using IBM SPSS Statistics (Version 28).

RESULTS

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

Phase I. The mean time taken to watch and rate each 10-minute length conversation by raters was 29 mins (range 14 – 56 mins)(Figure 1). Mean ease-of-use ratings on a scale of 1 - 10 was 6.8 with a range of scores from 5 - 10. Raters who completed the 4-hour training rated the conversation from a single viewing 81% of the time. Raters who completed the 8-hour training rated the conversation from a single viewing 43% of the time.

[insert Figure 1 about here]

Reliability for both half-day and full-day of training was poor-to-moderate (ICCs = 0.43 - 0.62) with confidence intervals poor-through-good (Table 2). Percent agreement within 0.5 ranged from 17% to 43% across both conditions.

[insert Table 2 about here]

There were strong positive correlations, $r_s(21) > 0.73 - 0.92$, $p < .001$ between the MPC Interaction and Transaction subscales for all six raters. Strong positive correlations were found between the MSC Acknowledging Competence subscale and each of the three elements of the Revealing Competence scales for most raters, $r_s(21) > 0.74 - 0.96$, $p < .001$. For one rater, the correlation between the RC1 and RC2 elements was moderate $r_s(21) = 0.68$, $p < .001$.

Phase II. As there was minimal difference in ICCs between half-day and full-day training, Phase II used half-day training only; and the strong correlations between specific scales led to the number of scales rated being reduced to either four scales (MPC Interaction and Transaction, MSC RC1 and RC2); or three scales (MPC Interaction, MSC RC1 and RC2) given stronger ICCs for MPC Interaction over MPC Transaction. Additionally, phase II compared long (10 minute) and short (5 minute) conversation samples.

For raters who viewed 10-minute samples, the mean time taken to rate each sample was 34 minutes (range 25-53 mins) when four scales were rated and 38 minutes (range 23 - 60 mins) when three scales were rated (Figure 1). For raters who viewed 5-minute samples, the mean time taken to rate each sample was 23 mins (range 18 - 26 mins). Mean ease-of-use ratings on a scale of 1 - 10 was 6.6 with a range of scores from 2 to 10.

Reliability for rating four scales with 10-minutes of conversation was moderate for MPC Interaction, RC1 and RC2 (ICCs = 0.56 - 0.59) and poor for MPC Transaction (ICC = 0.33) (Table 3). Reliability for rating four scales with 5-minutes of conversation was moderate for all four scales (ICCs = 0.56 - 0.68). Reliability was moderate for rating three scales with 5-minutes of conversation (ICCs = 0.52 – 0.68). Confidence intervals were poor-through-good for all ICCs across all conditions. Percent agreement within 0.5 ranged from 39 – 83% across the three conditions.

[insert Table 3 about here]

There were strong positive correlations, $r_s(21) > 0.86 - 0.91$, $p < .001$ between the MPC Interaction and Transaction subscales for all raters in this phase. There were also strong positive correlations, $r_s(21) > 0.85 - 0.97$, $p < .001$ between the RC1 and RC2 for all raters.

Phase III. As there were more favourable ICCs for 5-minutes compared to 10-minutes of conversation, the next two phases used 5-minute conversations only. As the correlations were strong and ICCs higher for MPC Interaction and MSC RC1, only these two scales were used in the next two phases.

The mean time taken to rate conversations in this phase was 20 mins (range 15 - 26 mins) (Figure 1). Mean ease-of-use ratings on a scale of 1 - 10 was 7.6 with a range of scores from 2-10.

Reliability for rating the two scales with 5-minutes of conversation was moderate for MPC Interaction (ICC = 0.69) and MSC RC1 (ICCs = 0.73) (Table 4). Confidence intervals were poor-through-good. Percent agreement within 0.5 ranged from 57 – 78%.

[insert Table 4 about here]

Phase IV. The mean time taken by experienced clinicians to rate conversations in this phase was 22 mins (range 19 - 35 mins) (Figure 1). Mean ease-of-use ratings on a scale of 1 - 10 was 7.3 with a range of scores from 4 to 9.

Reliability for rating the two scales with 5-minutes of conversation was moderate for MPC Interaction (ICC = 0.55) and MSC RC1 (ICCs = 0.58) (Table 3). Confidence intervals were poor-through-good. Percent agreement within 0.5 ranged from 48 – 70%.

Overall summary

Table 5 presents a summary of each of the four phases. Overall, the time taken to rate conversations decreased across the phases, particularly as shorter conversations were rated; ease of use for rating conversations on the scales improved slightly across phases; and measures of reliability (ICCs) generally improved across each of the four phases, most notably when fewer scales were used.

[insert Table 5 around here]

Qualitative Data

Qualitative data revealed two broad categories across all four phases around: (i) scale use; and (ii) conversation ratings. *Scale use* referred to the raters' actual use of the scales to inform their final rating. Raters from all phases found it difficult to know how weigh-up one behaviour or descriptor over another.

> *"Finding it difficult to finalise scores between 2 and 3 and decide what gives enough weight to lower or increase a score" (Student rater, Phase I)*

There were issues with the clarity of the descriptors where some raters reported lack of detail, ambiguous, or imprecise descriptors (e.g., "share responsibility for feel/flow"). Some raters reported descriptors were not helpful, that some partially met or absent descriptors were difficult to rate and that overall, there were simply too many descriptors to rate and/or consider at once (particularly in phase I). Raters did report it easier to rate more concrete and overt behaviours (e.g., "listening attitude, supportive questioning"). Many of these reports were reduced in frequency in later phases (when fewer scales were rated) and the clinician raters reported fewer concerns than students regarding the usefulness of the descriptors.

> *"Found the descriptors helpful to go through for CP [communication partner] as although she had a warm manner and was interested in her son, very few of the criteria for supporting understanding were explicitly met" (SLT, Phase IV).*

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

In later phases, rating a conversation with fewer scales and making a judgement of the impact of individual behaviours relative to the whole conversation was a challenge, particularly for clinicians. Student raters also provided insightful comments describing this challenge.

> *"I'm just gonna stick to what I'm rating. There's so many things that play a part in making the conversation great and I'm only focusing on do they ensure that the other person understands" (student rater, phase III)*

Clinicians sometimes reported using clinical intuition to make a judgement of the conversation as the rating score was not felt to reflect their observations.

> *"found myself judging the score on gut feeling once all descriptors considered, rather than any one descriptor carrying more weight" (SLT, Phase IV).*

DISCUSSION

The aim of this study was to explore the feasibility of the Adapted Kagan Scales for clinical practice, and reliability under different training and rating conditions. Overall, the training required to achieve proficiency, and the time to view and rate conversations would be considered feasible. Across all phases of the study, raters were able to view and rate a 5- or 10-minute conversation in under 60 minutes. Rating time was reduced to 30 minutes for a 5-minute conversation. This result is consistent with the findings of Iwashita and Sohlberg (2019) where raters could rate a 10-minute conversation using two scales of social communication ability in less than 30 minutes. Training was also feasible to deliver in either a half-day or full-day training program. While it was not the intention to explore the delivery

mode, training was able to be successfully delivered both face-to-face and online. While longer training (i.e., full-day) offered increased opportunities for practice and discussion, when compared to shorter training duration (i.e., half-day) there was no discernible difference in the reliability results. The time taken to train the scales was significantly less than the 14 to 35 hours reported elsewhere (Behn et al., 2019a; Behn et al., 2012; Chia et al., 2019; Rietdijk et al., 2020b), with these other studies involving a procedure in which raters demonstrated reliability on training samples to be considered competent in rating. The potential for reduced training time and quicker scoring demonstrated in the present study is important, as it enables the scales to be more clinically accessible to speech and language therapists, who have restrictions on their time (Frith et al., 2014; Kelly et al., 2017; Maddy et al., 2015).

The reliability results are encouraging and considered acceptable, with moderate reliability for most scales and improved reliability for the student raters when fewer scales were rated. The results were not as strong as for the original study that used the same conversations (Behn et al., 2019a) however, that study involved 18 hours of training over multiple days. Given previous studies have reported moderate-to-excellent reliability for rating casual conversations with longer training, the finding is optimistic (Behn et al., 2019a; Behn et al., 2012; Rietdijk et al., 2020a; Rietdijk et al., 2020b; Togher et al., 2010). Moreover, reducing the number of scales yielded positive reliability results and addressed rater burden raised by some student raters who made comments about too many scales to rate and descriptors to consider. However, the same positive results cannot be said of the scales rated by experienced clinicians. In interpreting the significance of these results, researchers have suggested that intra-class correlations need to be at least 0.80 for high-risk clinical decisions, such as making clinical diagnoses (Slagle et al., 2002), 0.70 for research purposes (Nunnally, 1978) and 0.60 to be clinically useful (Chinn, 1991). While this would seem to 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

communication partner may not identify any impairment at all. A clinician may not share the person's culture or social background or have experience of situations or contexts being discussed in the conversation, which may affect their judgement. Further, a clinician may have an unconscious bias on factors such as gender, culture, race, and ethnicity, that may influence their judgement of the interaction (Badon et al., 2005; Harrison et al., 2017). Eriksson et al (2014) identified the effect of raters' personal biases as one of the key factors undermining the reliability and validity of clinical rating scales. Longer training that explicitly addresses many of these issues may need to be considered and evaluated in the future to determine whether they can be mitigated (Behn et al., 2012; Eriksson et al., 2014).

There are several types of rater error and bias that may influence the rater's ability to make a judgement about a conversation. These have been described by Eriksson et al (2014) (2014) including, primacy/recency effects when ratings are based on observations made early or late in the conversation, or contrast effects where ratings are higher or lower relative to previously assessed samples (Feldman et al., 2012; Yeates et al., 2013b). One reflection by several raters was a difficulty in deciding how to weigh one-off behaviours when scoring. For example, a communication partner may demonstrate good listening skills throughout most of the conversation, but then dismiss contributions from the person with ABI at one point in the conversation. The relative weight (and thus rating) given to one behaviour over another may differ between raters (Yeates et al., 2013a). This effect was particularly noticeable when a behaviour was brief but had significant impact on the other person. Raters rarely agreed on which conversations were the most challenging for weighing up behaviours. This finding may suggest the presence of "halo errors" whereby ratings are based on one positive or negative observation (Jacobs & Kozlowski, 1985). Rating a conversation is a complex process, and there is likely to be variability (and bias) in how individual raters place emphasis or perceive value on different aspects of an interaction.

Experienced clinicians reported difficulty with a scale that contained half-points, which suggests a reduced scale may be more favourable. Eriksson et al (2014) attempted to address this issue (and that of bias) by shortening the rating periods (e.g., to one minute each) and using a reduced scale from 9-points to 4-points (e.g., 1 to 4, predominantly poor support, consistently satisfactory support). However, 10-30 hours of training was required, and reliability was poor to moderate. In another study, a more reduced scale (of 1-3 points: predominantly poor support, OK but not satisfactory, predominantly satisfactory) was found to achieve better reliability (Saldert et al., 2013) although a reduced scale may potentially limit the validity and the scales' sensitivity to change.

The impact of several factors on reliability was considered in this study, including the length of the conversation (i.e., 5 and 10 minutes) and experience of the rater (i.e., student and experienced clinician). Other studies have used either five minutes (Rietdijk et al., 2020b; Togher et al., 2010) or ten minutes of conversation (Behn et al., 2019a; Behn et al., 2012; Iwashita & Sohlberg, 2019); and a study for people with post-stroke aphasia reported that 3-5 minutes of conversation was sufficient for analysis (Correll et al., 2010). The reliability findings from this study were generally more favourable for conversations of 5-minutes in length when raters were rating the same scales, which suggests that clinicians could adopt the same length of conversation in clinical practice. Reliability was less favourable for experienced clinicians compared to student raters. Togher et al (2010) reported good-to-excellent reliability when raters were experienced clinicians. However, in that study raters rated all six scales and in the current study (phase IV), clinicians rated only two scales (MPC Interaction and MSC RC1). Qualitative reports suggest that the clinicians tended to use their wider clinical experience and intuition when rating. Certainly, for one clinician, they found it challenging to focus on the two scales and gave ratings that reflected the overall conversation. While the earlier study of the original Kagan scales found a significant positive

correlation between clinical intuition and ratings (Kagan et al., 2004), the raters rated all six scales rather than the two in this study.

An additional factor to raise relates to the training process itself. The training familiarised the raters with the scales, provided sample conversations to rate, and discussed common issues. While the training process used was like other studies using the same scales (Behn et al., 2012; Behn et al., 2019a; Rietdijk et al., 2020b), greater consideration of some of the issues raised by raters in this study may be needed (e.g., managing personal bias, weighing up behaviours, changing behaviours, influence of a person's behaviour on another). Longer training and/or greater use of challenging sample videos may help. There may also be an issue with how raters listen and engage during training and apply what they have learnt. Future research may need to closely examine the training process using think aloud techniques to more robustly identify how raters observe and interpret what they are seeing and where the specific differences may lie when they rate the same video. This research will contribute to our understanding of how best to train the use of the scales thus, standardising training for the future.

Finally, there may be a tension between the concise clarity of the rating scales and the subtle insights from a rater who has either greater clinical experience, more training, or who is rating a longer conversation sample. Individual raters' reported issues with the clarity and weighting of the descriptors, including unhelpful and/or an excessive number of descriptors to consider. Visible communicative behaviours (e.g., eye contact, questions asked, turn-taking) were certainly considered easier to rate than more abstract, ambiguous behaviours (e.g., appropriate amount of information, organisation of information). However, qualitative comments from the experienced clinicians suggest that clinical intuition may lead raters to identify or describe more subtle, abstract and difficult to describe behaviours that may not be

listed, highlighting the inherent conflict for raters during the process. Striking the right balance between these factors may prove challenging.

Reliability and feasibility of the measures may be improved through modifications to the scale including, reducing the number of descriptors, and linking them to more concrete behaviours. However, these changes may negatively influence the validity of the scales and their ability to adequately explain differences in ratings. Measures like the Modified Pragmatic Rating Scale (Iwashita and Sohlberg, 2019) have simple scales and few descriptors (e.g., eye contact, gesture, and initiation of new topics), however, the reliability results are comparable to the Adapted Kagan Scales (Iwashita & Sohlberg, 2019).

Inclusion of the Adapted Kagan Scales is important as they are the only known scales to measure support provided by communication partners during conversation. Therefore, future recommendations may include the use of larger participant numbers and the potential integration of automated analysis of some conversational skill behaviours that are quantifiable such as percentage of speaking time and facial expressions (Liu et al 2016). Additionally, there may be consideration of other scales focused on measuring the skills of the person with ABI and the degree of communicative effectiveness (e.g., Conversational discourse scale of the Montreal Evaluation of Communication, Joanette et al., 2015) or inclusion of patient-reported outcome measures of perceived communicative ability and participant experiences that are psychometrically robust such as the Communication Participation Item Bank (Baylor et al., 2013); La Trobe Communication Questionnaire (Douglas et al., 2000) and Social Skills Questionnaire-Traumatic Brain Injury (Francis et al., 2017). Such changes will ensure ongoing data may be collected for feasibility and reliability of the scales with consideration of their validity.

Limitations

Overall, this study was limited by its small sample size of 23 conversations, which as a convenience sample may not represent the full range of scores from these scales. In addition, a specific measure of cognitive-communication disorder was not used to recruit participants. Researchers suggest for reliability studies, there should be at least 30 samples with three raters (Koo & Li, 2016). Therefore, low ICCs may be attributable to fewer raters in each phase and potentially a lack of variation among the sampled people with brain injury and their communication partners (Eriksson et al., 2014) given the use of a convenience sample. In addition, the conversation samples for this study were drawn from people who had sustained both traumatic and non-traumatic injuries where previous studies have used samples from only people with TBI (Behn et al., 2012; Rietdijk et al., 2018; Togher et al., 2010). There was a dependence on a high proportion of student raters who were predominantly female, however this is consistent with the speech and language therapy profession. The student raters had limited knowledge and experience of brain injury and associated communication problems, which may reduce the generalisability of the study findings to how experienced brain injury clinicians may feasibly use these scales in clinical practice.

While the think-aloud protocol was intended to provide rich qualitative data, this was not consistently the case. A concurrent think-aloud interview while raters were viewing the conversation was initially attempted but proved to be cognitively challenging, so a retrospective think-aloud interview was used. Further training and consideration of rater prompts through the think aloud process may be required in the future (Hu & Gao, 2017). Despite this, the rater logs provided clear information during all four phases and was helpful to developing a clear understanding of the challenges faced by raters. Another limitation could relate to the statistics used in this study. To be transparent; correlations, significance value, confidence intervals and percent agreement was reported. Bland Altman plots (1986)

may also be used to visualise disagreements in rating, degree of differences and assessor bias (Eriksson et al., 2014) and may have generated further insights into the nature of the data. This study did not further examine the content validity of the scales nor consider intra-rater or test-retest reliability, with the latter relevant to the use of the scales as an outcome measure. A reduction in the number of scales had little to no effect on reliability however, it may affect sensitivity to change so future research would need to consider whether there is a trade-off between reliability and sensitivity for the measures.

Clinical implications and future directions

Rating scales of conversation offer a useful starting point for clinicians who are conducting assessments with the goal of making clinical decisions and setting goals for treatment. For example, they could help guide the clinician and the dyad as to the aims of intervention (e.g., improving the communication partner's ability to reveal the competence of the person with ABI by ensuring they can respond) and thus identify relevant target behaviours for treatment (e.g., asking questions, take turns, give time to respond). In doing so, the clinician can select targets that focus on person-centred and contextually relevant conversations and topics that align with the values and needs of the person with brain injury and their communication partner (Keegan et al., 2023; Sohlberg et al., 2019). Those target behaviours could be translated to a goal setting framework such as Goal Attainment Scaling, in collaboration with the dyad. The outcome of treatment would therefore be a positive change to a discrete communicative behaviour or use of a specific strategy by the person with brain injury and/or communication partner, to achieve a social activity or participation goal (Behn et al., 2019b; Keegan et al., 2020), with the potential to evaluate progress using the Adapted Kagan Scales on conversation samples collected across different timepoints. The Adapted Kagan Scales have been found to be a sensitive outcome measure for demonstrating

positive change in conversations after communication partner training in multiple studies (Behn et al., 2012; Rietdijk et al., 2020; Togher et al., 2013), which indicates they may be clinically useful for this purpose. Future research that strengthens the psychometric properties of the scales including for example, test-retest reliability, will be important to progressing the use of these measures in research and clinical practice.

One additional solution to the problem of reliably evaluating conversation could be in the form of emerging technologies and artificial intelligence. Computerised discourse analysis programs and software programs for rating conversational discourse may be a future innovation (Steel & Togher, 2019). Artificial intelligence has already been used for rating conversational discourse to evaluate communication partner training for discrete conversation behaviours that are identified by human review of videotaped conversations (e.g., open and closed questions, long pauses, and yes/no questions) (Croteau et al., 2018). Artificial intelligence has also been used to conduct a conversational assessment to help predict depression (Weisenburger et al., 2024). Such technologies may be able to be adapted and repurposed for rating conversations of people with brain injury and their communication partners.

Conclusion

There is a need for reliable and valid measures of conversation that can be easily used to assess social communication impairments, and which are time efficient. In this study, the Adapted Kagan Scales were used to rate conversations involving people with brain injury and their communication partners. A short training period (of four hours) enabled students and clinicians to view and rate 5-minute conversations using two subscales in under 30 minutes: with acceptable/moderate reliability. Conversation is dynamic, interactive, and complex; and requires a clinician to make many judgements about the communicative behaviours of

participants. Use of several Adapted Kagan Scales (MPC-Interaction; and MSC-RC1) was feasible and future research could evaluate how these scales may influence the goal setting process and outcome measurement in communication partner training interventions. This paper is intended to raise the importance of measuring social communication in dyads and present a clinically feasible method for assessing these skills.

Acknowledgements

We wish to acknowledge the 14 final-year speech and language therapy students from City, University of London who rated the conversations of this study; and the two speech and language therapists for also taking the time and effort to also rate conversations.

Data Availability Statement

Data available upon request

References

Badon, L. C., Oller Jr, J. W., & Oller, S. D. (2005). Ratings within and across ethnic boundaries of methods of one on one reading instruction. *Journal of Communication Disorders*, *38*(6), 445-457.

Baylor, C., Yorkston, K., Eadie, T., Kim, J., Chung, H., & Amtmann, D. (2013). The Communicative Participation Item Bank (CPIB): Item bank calibration and development of a disorder-generic short form. *Journal of Speech, Language, and 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.

Coelho, C., Ylvisaker, M., & Turkstra, L. S. (2005). Nonstandardized assessment approaches for individuals with traumatic brain injuries. *Seminars in Speech and Language*, *26*(4), 223-241.

Coelho, C. A., Liles, B. Z., & Duffy, R. J. (1991). Analysis of conversational discourse in head-injured adults. *Journal of Head Trauma Rehabilitation*, *6*(2), 92-99.

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008). Internet-based learning in the health professions: a meta-analysis. *JAMA*, *300*(10), 1181-1196.

Correll, A., Steenbrugge, W., & Scholten, I. (2010). Judging conversation: How much is enough? *Aphasiology*, *24*(5), 612-622.

Croteau, C., McMahon-Morin, P., Le Dorze, G., Power, E., Fortier-Blanc, J., & Davis, G. A. (2018). Exploration of a quantitative method for measuring behaviors in conversation. *Aphasiology*, *32*(3), 247-263.

Dahlberg, C., Hawley, L., Morey, C., Newman, J., Cusick, C. P., & Harrison-Felix, C. (2006). Social communication skills in persons with post-acute traumatic brain injury: Three perspectives. *Brain Injury*, *20*(4), 425-435.

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

Douglas, J. M., O'Flaherty, C. A., & Snow, P. C. (2000). Measuring perception of communicative ability: The development and evaluation of the La Trobe communication questionnaire. *Aphasiology*, *14*(3), 251-268.

Durning, S. J., Artino Jr, A. R., Beckman, T. J., Graner, J., Van Der Vleuten, C., Holmboe, E., & Schuwirth, L. (2013). Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Medical teacher*, *35*(9), 720-726.

Elbourn, E., MacWhinney, B., Fromm, D., Power, E., Steel, J., & Togher, L. (2023). TBIBank: An international shared database to enhance research, teaching and automated language analysis for traumatic brain injury populations. *Archives of Physical Medicine and Rehabilitation*, *104*(5), 824-829.

Eriksson, K., Bergstrom, S., Carlsson, E., Hartelius, L., Johansson, C., Schwarz, A., & Saldert, S. (2014). Aspects of rating communicative interaction: Effects on reliability and agreement. *Journal of Interactional Research in Communication Disorders*, *5*(2), 245-267.

Farrell, A. D., Rabinowitz, J. A., Wallander, J. L., & Curran, J. P. (1985). An evaluation of two formats for the intermediate-level assessment of social skills. *Behavioral Assessment*.

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279-286.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs - principles and practices. *Health Services Research*, *48*(6pt2), 2134-2156.

Finch, E., Copley, A., Cornwell, P., & Kelly, C. (2016). Systematic Review of Behavioral Interventions Targeting Social Communication Difficulties After Traumatic Brain Injury. *Archives of Physical Medicine and Rehabilitation*, *97*(8), 1352-1365.

Francis, H. M., Osborne-Crowley, K., & McDonald, S. (2017). Validity and reliability of a questionnaire to assess social skills in traumatic brain injury: a preliminary study. *Brain Injury*, *31*(3), 336-343.

Frith, M., Togher, L., Ferguson, A., Levick, W., & Docking, K. (2014). Assessment practices of speech-language pathologists for cognitive communication disorders following

traumatic brain injury in adults: an international survey. *Brain Injury*, *28*(13-14), 1657-1666.

Galski, T., Tompkins, C., & Johnston, M. (1998). Competence in discourse as a measure of social integration and quality of life in persons with traumatic brain injury. *Brain Injury*, *12*(9), 769-782.

Harrison, A. J., Long, K. A., Tommet, D. C., & Jones, R. N. (2017). Examining the role of race, ethnicity, and gender on social and behavioral ratings within the Autism Diagnostic Observation Schedule. *Journal of autism and developmental disorders*, *47*, 2770-2782.

Hartley, L. L., & Jensen, P. J. (1992). Three discourse profiles of closed-head-injury speakers: Theoretical and clinical implications. *Brain Injury*, *6*(3), 271-282.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test*. Psychological Assessment Resources, Inc.

Hewetson, R., Cornwell, P., & Shum, D. (2017). Cognitive-communication disorder following right hemisphere stroke: exploring rehabilitation access and outcomes. *Topics in Stroke Rehabilitation*, *24*(5), 330-336.

Howell, S., Varley, R., Sinnott, E. L., Pring, T., & Beeke, S. (2021). Measuring group social interactions following acquired brain injury: an inter-rater reliability evaluation. *Aphasiology*, *35*(11), 1505-1517.

Hsieh, H.-F., & Shannon, S. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277-1288.

Hu, J., & Gao, X. A. (2017). Using think-aloud protocol in self-regulated reading research. *Educational Research Review*, *22*, 181-193.

Iwashita, H., & Sohlberg, M. M. (2019). Measuring conversations after acquired brain injury in 30 minutes or less: a comparison of two pragmatic rating scales. *Brain Injury*, *33*(9), 1219-1233.

Jacobs, R., & Kozlowski, S. W. J. (1985). A closer look at halo error in performance ratings. *Academy of Management Journal*, *28*(1), 201-212.

Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of applied psychology*, *75*(5), 500.

Joanette, Y., Ska, B., Cote, H., Ferre, P., LaPointe, L., Coppens, P., & Small, S. (2015). *Montreal Protocol for the Evaluation of Communication*. ASSBI Resources.

Kagan, A., Winckel, J., Black, S. E., Duchan, J. F., Simmons-Mackie, N., & Square, P. (2004). A set of observational measures for rating support and participation in conversation between adults with aphasia and their conversation partners. *Topics in Stroke Rehabilitation*, *11*(1), 67-83.

Keegan, L. C., Hoepner, J. K., Togher, L., & Kennedy, M. (2023). Clinically applicable sociolinguistic assessment for cognitive-communication disorders. *American Journal of Speech-Language Pathology*, *32*(2S), 966-976.

Keegan, L. C., & Müller, N. (2022). The influence of context on identity construction after traumatic brain injury. *Journal of Interactional Research in Communication Disorders*, *13*(2), 171-195.

Keegan, L. C., Murdock, M., Suger, C., & Togher, L. (2020). Improving natural social interaction: Group rehabilitation after Traumatic Brain Injury. *Neuropsychological 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.

Maddy, K., Howell, D., & Capilouto, G. (2015). Current practices regarding discourse analysis and treatment following non-aphasic brain injury: A qualitative study. *Journal of Interactional Research in Communication Disorders*, *6*(2), 211-236.

Mann, K., Power, E., Barnes, S., & Togher, L. (2015). Questioning in conversations before and after communication partner training for individuals with traumatic brain injury. *Aphasiology*, *29*(9), 1082-1109.

McDonald, S., Tate, R., Togher, L., Bornhofen, C., Long, E., Gertler, P., & Bowen, R. (2008). Social skills treatment for people with severe, chronic acquired brain injuries: A multicenter trial. *Archives of Physical Medicine and Rehabilitation*, *89*(9), 1648-1659.

Meulenbroek, P., & Turkstra, L. S. (2016). Job stability in skilled work and communication ability after moderate-severe traumatic brain injury. *Disability and Rehabilitation*, *38*(5), 452-461.

Morrow, E. L., Hereford, A. P., Covington, N. V., & Duff, M. C. (2020). Traumatic brain injury in the acute care setting: assessment and management practices of speech-language pathologists. *Brain Injury*, *34*(12), 1590-1609.

Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical diagnosis of mental disorders: A handbook*, 97-146.

Olver, J., Ponsford, J., & Curran, C. (1996). Outcome following traumatic brain injury: A comparison between 2 and 5 years after injury. *Brain Injury*, *10*(11), 841-848.

Ponsford, J. L., Downing, M. G., Olver, J., Ponsford, M., Acher, R., Carty, M., & Spitz, G. (2014). Longitudinal follow-up of patients with traumatic brain injury: outcome at two, five, and ten years post-injury. *Journal of Neurotrauma*, *31*(1), 64-77.

Portney, L. G., & Watkins, M. P. (2014). *Foundations of Clinical Research: Applications to Practice* (3rd ed.). Pearson Education Limited.

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

Rietdijk, R., Power, E., Attard, M., Heard, R., & Togher, L. (2020a). Improved conversation outcomes after social communication skills training for people with traumatic brain injury and their communication partners: a clinical trial investigating in-person and telehealth delivery. *Journal of Speech, Language, and Hearing Research*, *63*(2), 615-632.

Rietdijk, R., Power, E., Brunner, M., & Togher, L. (2020b). The reliability of evaluating conversations between people with traumatic brain injury and their communication partners via videoconferencing. *Neuropsychological rehabilitation*, *30*(6), 1074-1091.

Rietdijk, R., Simpson, G., Togher, L., Power, E., & Gillett, L. (2013). An exploratory prospective study of the association between communication skills and employment outcomes after severe traumatic brain injury. *Brain Injury*, *27*(7-8), 812-818.

Saldert, C., Backman, E., & Hartelius, L. (2013). Conversation partner training with spouses of persons with aphasia: A pilot study using a protocol to trace relevant characteristics. *Aphasiology*, *27*(3), 271-292.

Shelton, C., & Shryock, M. (2007). Effectiveness of communication/interaction strategies with patients who have neurological injuries in a rehabilitation setting. *Brain Injury*, *21*(12), 1259-1266.

Shorland, J., Douglas, J., & O'Halloran, R. (2022). Age-based trends in cognitive-communication management for adults in subacute rehabilitation following new onset traumatic brain injury. *American Journal of Speech-Language Pathology*, *31*(6), 2557-2568.

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

Struchen, M., Clark, A., Sander, A., Mills, M., Evans, G., & Kurtz, D. (2008). Relation of executive functioning and social communication measures to functional outcomes following traumatic brain injury. *NeuroRehabilitation*, *23*(2), 185-198.

Tobar-Fredes, R., & Salas, C. (2022). Rehabilitation of communication in people with traumatic brain injury: a systematic review of types of intervention and therapeutic ingredients (Rehabilitación de la comunicación en personas con traumatismo encefalocraneal: una revisión sistemática de tipos de intervención e ingredientes terapéuticos). *Studies in Psychology*, *43*(1), 88-131.

Togher, L., Douglas, J., Turkstra, L. S., Welch-West, P., Janzen, S., Harnett, A., Kennedy, M., Kua, A., Patsakos, E., & Ponsford, J. (2023). INCOG 2.0 guidelines for cognitive rehabilitation following traumatic brain injury, part IV: cognitive-communication and social cognition disorders. *Journal of Head Trauma Rehabilitation*, *38*(1), 65-82.

Togher, L., Hand, L., & Code, C. (1997). Analysing discourse in the traumatic brain injury population: Telephone interactions with different communication partners. *Brain Injury*, *11*(3), 169-189.

Togher, L., McDonald, S., Tate, R., Power, E., & Rietdijk, R. (2013). Training communication partners of people with severe traumatic brain injury improves everyday conversations: A multicenter single blinded clinical trial. *Journal of Rehabilitation Medicine*, *45*, 637-645.

Togher, L., Power, E., Tate, R., McDonald, S., & Rietdijk, R. (2010). Measuring the social interactions of people with traumatic brain injury and their communication partners: The adapted Kagan scales. *Aphasiology*, *24*(6-8), 914-927.

Weisenburger, R. L., Mullarkey, M. C., Labrada, J., Labrousse, D., Yang, M. Y., MacPherson, A. H., Hsu, K. J., Ugail, H., Shumake, J., & Beevers, C. G. (2024). Conversational

assessment using artificial intelligence is as clinically useful as depression scales and preferred by users. *Journal of affective disorders*.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013a). Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education*, *18*, 325-341.

Yeates, P., O'Neill, P., Mann, K., & W Eva, K. (2013b). 'You're certainly relatively competent': assessor bias due to recent experiences. *Medical education*, *47*(9), 910-922.

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

[a]Injury severity can only be determined for traumatic injuries

Note. Values are mean ± SD. RBANS = Repeatable Battery of Assessment of Neuropsychological Status (average score = 90 - 109); WCST = Wisconsin Card Sorting Test (average categories = 5.07; average perseverative errors = 15.78).

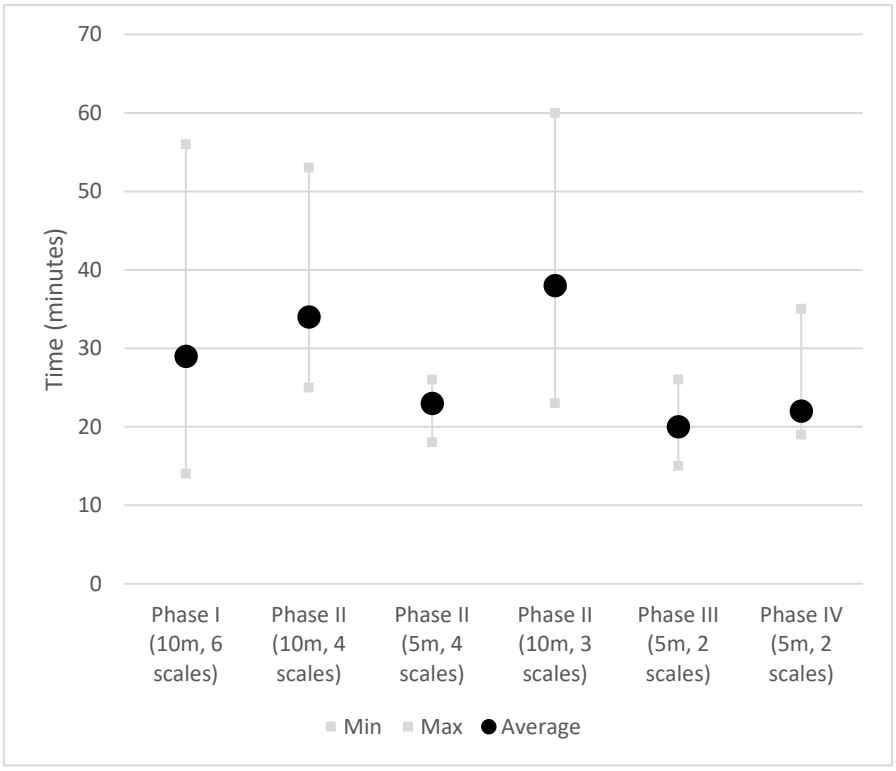

Figure 1. Time (in minutes) to watch and rate conversations

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%

ICC, Intraclass Correlation; CI, Confidence Intervals; MPC, Measure of Participation in Conversation; MSC, Measure of Support In Conversation; AC, Acknowledging Competence; RC, Revealing Competence; RC1, Ensure the adults understands; RC2, Ensure the adult has a means of responding; and RC3, Verification.

[a]$p < .001$

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%

ICC, Intraclass Correlation; CI, Confidence Intervals; MPC, Measure of Participation in Conversation; MSC, Measure of Support In Conversation; RC, Revealing Competence; RC1, Ensure the adults understands; RC2, Ensure the adult has a means of responding; and RC3, Verification.

[a] $p < .001$

[b] $p < .01$

[c] $p < .05$

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%

ICC, Intraclass Correlation; CI, Confidence Intervals; MPC, Measure of Participation in Conversation; MSC, Measure of Support In Conversation; RC, Revealing Competence; RC1, Ensure the adults understands.

[a] $p < .001$

[b] $p < .01$

Table 5. Summary of time taken to rate, ease of use and reliability measures across all four 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)

ICC, Intraclass Correlation