

An examination of sample length and reliability of the Interactional Network Tool, a new measure of group interactions in acquired brain injury

Susan Howell, Suzanne Beeke, Emma Louise Sinnott, Rosemary Varley & Tim Pring

To cite this article: Susan Howell, Suzanne Beeke, Emma Louise Sinnott, Rosemary Varley & Tim Pring (2022): An examination of sample length and reliability of the Interactional Network Tool, a new measure of group interactions in acquired brain injury, Aphasiology, DOI: [10.1080/02687038.2022.2118517](https://doi.org/10.1080/02687038.2022.2118517)

To link to this article: <https://doi.org/10.1080/02687038.2022.2118517>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 04 Sep 2022.



Submit your article to this journal [↗](#)



Article views: 61








View related articles [↗](#)



View Crossmark data [↗](#)

An examination of sample length and reliability of the Interactional Network Tool, a new measure of group interactions in acquired brain injury

Susan Howell PhD ^a, Suzanne Beeke PhD ^a, Emma Louise Sinnott ^b,
Rosemary Varley PhD ^a and Tim Pring PhD ^c

^aDivision of Psychology and Language Sciences, University College London; ^bAneurin Bevan University Health Board, Wales; ^cDivision of Language and Communication Science, City, University of London

ABSTRACT

Background: Conversation is challenging to measure. Quantitative and qualitative measures need to be sensitive to the conversation context, the purpose and the variable contributions of participants in order to capture meaningful change. Measurements also need to be consistent across independent raters. The reliability of global observational rating scales across differing sample lengths has previously been investigated. An investigation into the effects of sample length on inter-rater reliability using a behavioural frequency measure is a new field of research.

Aims: This study reports on the inter-rater reliability of the Interactional Network Tool (INT), a behavioural coding system for use with group interaction data. It examines the effects of sample length on reliability using a refined coding system designed to improve the speed and efficiency of use in clinical settings.

Methods: Fourteen video samples of group interactions for people with acquired brain injury were prepared for analysis. Two raters independently coded the films using the INT coding system. Individual code reliability was calculated using intra-class correlations (ICCs). Codes were combined to form a new coding structure. Reliability of the new codes was calculated using intra-class correlations across four sample lengths (5,10,15 and 20 minutes). A one-way analysis of variance was used to compare the means of the four sample lengths.

Outcomes and Results: Acceptable inter-rater reliability was achieved using the refined INT coding system. There was no difference between the four sample lengths.

Conclusions: These findings indicate that trained clinicians using the INT in clinical practice can achieve a reliable measure of participation in a group interaction from short samples. Validation with other clinical groups is now indicated.

ARTICLE HISTORY

Received 19 Mar 2022

Revised 13 Aug 2022

Accepted 25 Aug 2022

Background

Conversation is a complex construct comprising multiple skill components. Factors such as the environment, the background and purpose of the interaction, participant roles and the relationship between participants each have the potential to influence the

CONTACT Susan Howell  s.howell.12@ucl.ac.uk  Department of Language and Cognition, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, Tel: +44 207 679 4001

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

conversation course and degree of participation (Ahlsen and Saldert, 2018). The ability of conversation participants to deploy communication devices appropriate for context and purpose in a group interaction requires an array of cognitive and linguistic capabilities. In addition to speech and language skills, these include the cognitive capabilities to keep track of conversation content and assimilate detail, the social cognitive skills to accurately perceive and respond to non-verbal social cues and sensitivity to social nuance (Sohlberg et al., 2019), the pragmatic understanding to apply those skills in different contexts (Adams, 2005) and the emotional and behavioural control to respond adaptively whilst also adhering to social rules (Dahlberg et al., 2006). These rules will differ according to the group purpose and environment but the overarching aim is for participants to achieve mutual understanding (Wilkinson, 1999). A positive outcome is likely to be characterised by a shared contribution to the group conversation in which everyone is engaged. Communication behaviours with the potential to undermine mutual understanding include under- and over-participation. In the field of acquired brain injury (ABI), examples include conversation excess that displays a lack of acknowledgement of others, conversation insufficiency that may include failure to initiate or respond to an utterance when selected to do so (e.g. via a question), or the under- or over-use of eye gaze in a way that infringes social norms (Hartley and Jensen, 1991; Sim et al, 2013).

Although these factors make conversation challenging to measure, there is consensus on the measurement approach. The INCOG guidelines for cognitive communication disorders following traumatic brain injury recommend measurement of social communication outcomes at the level of participation in real-world settings (Togher, Wiseman-Hakes et al., 2014). Doedens and Meteyard (2020) provide recommendations for the assessment of real-world communication for aphasic adults, derived from a theoretical framework of situated language use. Methods of conversation evaluation need to be valid, reliable and sensitive to the conversation construct in order to profile capability and to capture change. Conversation Analysis, for example, provides a qualitative evaluation of everyday interaction that is sensitive to the natural context, but it is a research method not a clinical measure of change (Best et al., 2016). Quantitative measures such as conversation rating scales, on the other hand, are feasible and accessible for clinicians but sensitivity to the construct of conversation is dependent on psychometric properties and reliable definitions of the behaviours under investigation. A valid and sensitive measure of interactional behaviour needs a broad reach in order to capture the variability in both content and participant contribution that is inherent in everyday conversational interaction. It also needs to be sensitive to change across the communication contexts relevant to different individuals and the encounters that make up their everyday social lives, encompassing both dyadic and group interactions. Measurement of group interaction is a new field of investigation (Howell et al., 2020a).

A full review of the available measures is beyond the scope of this paper and has been comprehensively addressed by previous authors. For a review of available measures for conversation assessment for people with TBI, see Keegan et al. (2022).

Reliability of global observational methods

Given the complexity of conversation evaluation, it is no surprise that measurement reliability (or the degree to which evaluation findings are consistent across independent raters) can be difficult to achieve. The reliability of global observational methods to

quantitatively measure conversation of people with communication difficulties has previously been examined (Eriksson et al., 2014). Conversation rating scales typically require a rater to assign a value along a linear scale (e.g. on a continuum from normal to impaired) to a perceived behaviour (e.g. clarity of expression or cohesiveness). However, these interactional behaviours are complex to define across contexts and human raters bring an inherent bias to their perceptual judgement.

The Measure of Participation in Conversation (MPC) (Kagan et al., 2004), is a global observational measure that has been shown to be reliable in studies of adults with aphasia (Kagan et al., 2001; Kagan et al., 2004; Eriksson et al., 2014). It is made up of two scales that measure interaction and transaction in conversation. Correll, van Steenbrugge and Scholten (2010) examined inter-rater reliability across MPC samples lasting 3 minutes, 5 minutes, 10 minutes and 20 minutes. They found no significant difference in the reliability of independent judges across sample lengths indicating that increasing the length of the sample appears not to increase reliability. The Adapted Measure of Participation in Conversation (Adapted MPC) (Togher et al., 2010) has been shown to be reliable with adults following acquired brain injury. Reliability has been demonstrated across different interaction types and sample lengths of 5 minutes (Rietdijk et al., 2020; Togher et al., 2013) and 10 minutes (Behn et al., 2012). In these studies, the measure was used to evaluate dyadic conversations, and although inter-rater reliability was demonstrated, lengthy training was required to achieve it, e.g. Behn et al. (2012) report a training schedule of more than 40 hours. The Adapted MPC has more recently been tested using 10-minute group samples and a 2.5 hour rater training schedule (Howell et al., 2020b) and reliability was moderate to good.

Methods to improve judgement consistency include standardising the assessment context and task to influence the ease of evaluation, e.g. watching a video and re-telling to a communication partner; but this can be at the expense of the ecological validity of natural conversation (Saldert et al., 2018). Modifying the structure and design of a rating scale to simplify scoring by merging interaction behaviours is another approach (e.g. Saldert et al., 2013) although this risks compromising the validity and sensitivity of the tool (Eriksson et al., 2014). Training for raters can improve consistency, but this can be time intensive (Rietdijk et al., 2020) and findings may not hold up in clinical work with untrained professionals. Measurement reliability also rests on the properties of the tool. Adequate item definition is essential for consistent rating but rater interpretation of concepts and capabilities (e.g. 'competence') may also differ across rehabilitation stages and contexts, resulting in judgement inconsistencies (Horton et al., 2016).

Reliability of behavioural frequency measures

An alternative to rating scales is a more analytical approach to evaluating interaction, whereby the frequency of defined behaviours is tallied. There is a long history of using this measurement approach in discourse analysis. Depending on the target behaviours, the way they are classified, and the interaction context, it may be the case that a specified behaviour does not arise in the conversation or be recorded in a time-limited sample, or it may be functional in some scenarios but not in others (Eriksson et al., 2014). Further, outcomes are dependent on the psychometric properties of the measure. Results need to be consistent irrespective of who is administering the measure. Evaluation of the

psychometric properties of a communication behavioural coding system is therefore essential, of which inter-rater reliability is a cornerstone.

The Interactional Network Tool (INT) (Howell 2018) is a new frequency measure designed to evaluate group interaction behaviour. Its theoretical foundation draws on the proposition that communication competence is influenced by the interactional behaviour of other people and the environment (Bandura, 1971; Lave and Wenger, 1991; Vygotsy, 1978). Previous discourse analyses have used measures of participation and conversation share as indicators of the ability to modify behaviours in response to the communication behaviour of others (Gordon, Tranel and Duff, 2014; Gordon, Rigon and Duff, 2015). The INT coding scheme draws on discourse models that use initiation and response categories (Coulthard, 1984; Eggins and Slade, 1997) and linguistic and non-linguistic behaviours, including eye gaze (Trower et al., 1978) as the basis for analysis. Its methodological approach draws on social network theories and analysis methods to evaluate participant contribution and to visualise the spread of participant contribution across the interaction (Scott, 2017; Wasserman and Faust, 1994).

Behaviours are first coded from a filmed interaction, and entered directly into the INT software as sequences of turns between speakers. The INT software generates a matrix of relational contacts from these spreadsheet data. Figure 1a shows an interaction matrix transformed in Figure 1b into graph form, showing the interactive relationships of the group participants.

The arrows indicate direction. ‘Group’ is also included on the graph as a destination node, to indicate interactions directed to the group as a whole. As the number of verbal and non-verbal initiations and responses increase in frequency, the lines and arrows increase in thickness.

Howell et al. (2020a) tested the feasibility of the measurement approach coding 10-minute samples from peer group conversations between adults following ABI. This study looked at the overall reliability of initiations (total) and responses (total). In the evaluation, inter-rater reliability was excellent for initiation and good for response judgments. However, an evaluation of individual codes showed poor reliability in the category of

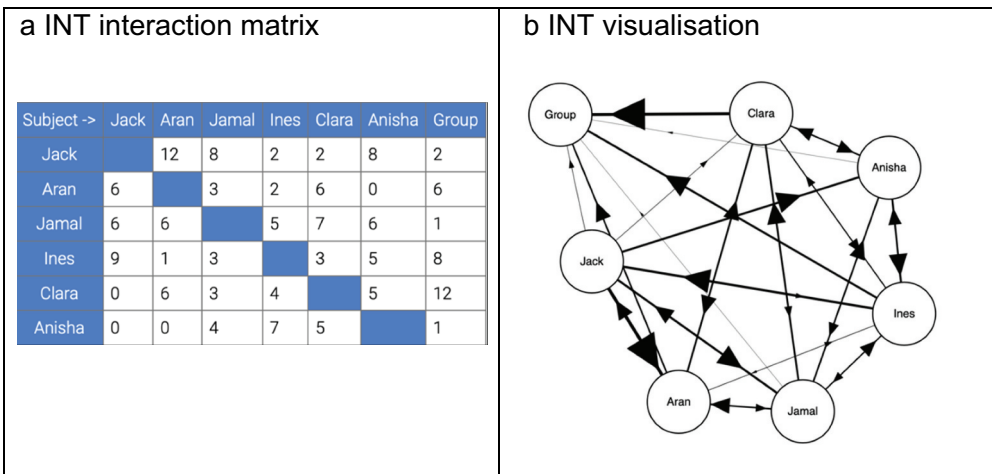


Figure 1. The INT interactional matrix and graph visualisation.

non-linguistic initiations and responses, specifically eye gaze, gesture and facial expression. In addition, 14 INT codes proved to be taxing for raters, which has the potential to undermine measurement accuracy and consistency.

Assessing measurement scale reliability

Intraclass correlation coefficients (ICC) (Shrout and Fleiss, 1979) are the recommended method for assessing measurement scale reliability for continuous variables in the COnsensus-based Standards for the selection of the health Measurement INstruments (COSMIN) risk of bias tool (Mokkink et al., 2020). An ICC value between 0 and 1 indicates the degree of susceptibility to measurement error as a result of variation in human judgement. However, the use of descriptors such as 'satisfactory' or 'moderate' agreement are not used consistently and standards of interpretation vary. According to ICC interpretation guidelines proposed by Cicchetti (1994), values between 0.4 and 0.59 are fair, between 0.6 and 0.74 are considered good, and 0.75 or above is considered excellent. Koo and Li (2016) offer a more stringent guideline whereby values between 0.5 and 0.75 are considered moderate, good reliability falls between 0.75 and 0.9, and above 0.9 is considered excellent. Post (2016) considers a coefficient $<.70$ as indicative of unreliability. Mokkink et al. (2020) recommend reporting 95% confidence intervals in addition to a point estimate of reliability. Further, there are 10 different types of ICC (McGraw and Wong, 1996). Mokkink et al. (2020) recommend reporting details of ICC selection, including the ICC model, type and its definition. These variations potentially result in inconsistent reporting of reliability.

Aims

Sample lengths to achieve a reliable outcome have previously been examined using global observational measures but sample length reliability for coding communication behaviour frequencies awaits investigation. The aims of this investigation are twofold: to examine the reliability of a refined INT coding system, and to evaluate the effects of sample length on reliability.

Method

Participants

Group samples were drawn from a pilot investigation into group interactions for people with ABI. The investigation compared a peer-led group intervention for social communication skills ($n=6$) to a staff-led activity group ($n=6$) (Clinical Trials.gov PRS: NCT02211339). Participants were aged 18 – 70, had sustained a severe ABI and were at least 6 months post injury. See Howell et al. (2020a) for full demographic and profiling variables. A summary of the main participant characteristics is presented in [Table 1](#):

Group activity

Group attendance varied between four and six participants. Conversations in both the intervention and control groups were topic directed (e.g. exercise habits; travel

Table 1. Demographic variables.

Participant	Age (years)	Male/ female	Education (years)	Time post onset (years)	Injury severity/clinical characteristics
Peer-led group					
1	50	M	11	5	ABI: ICH (ruptured AVM). Severe cognitive impairment/behavioural issues
2	57	M	15	1.1	ABI: Hypoxia (multi-organ failure; cardiac arrest). Severe cognitive impairment
3	31	M	11	9.0	TBI: RTA. GCS:6 Severe cognitive impairment
4	45	F	16	6	ABI: Hydracephalic ischemia (intracranial mass lesion). Severe cognitive impairment
5	53	M	12	0.7	ABI: Severe bilateral HSV encephalitis. Severe cognitive impairment
6	39	F	10	24	TBI: RTA. Comatose for several months. Severe cognitive impairment
Mean (SD)	45.8 (9.60)	4/2	12.5 (2.43)	7.63 (8.60)	TBI/ABI: 2/4
Staff-led group					
1	57	M	13	3	TBI: SDH (fall). Severe cognitive impairment
2	33	M	13	2	TBI: SAH; intracerebral haemorrhage (fall). Severe cognitive impairment
3	68	F	16	1	ABI: Hypoxia (cardiac arrest). Severe cognitive impairment
4	49	M	11	1	ABI: Obstructive hydrocephalus. Severe cognitive impairment
5	43	F	15	1	ABI: SAH (Grade 5). Severe cognitive impairment
6	62	M	12	0.5	TBI: SAH and SDH (falls). Severe cognitive impairment
Mean (SD)	52.0 (12.91)	4/2	13.33 (1.86)	1.42 (0.92)	TBI/ABI: 3/3

RTA= road traffic accident; SAH= subarachnoid haemorrhage; SDH= subdural haematoma; ICH= intracranial haemorrhage; HSV = herpes simplex virus; AVM= arteriovenous malformation; GCS= Glasgow Coma Scale.

ambitions). In the intervention group, discussion took place without staff present. In the control group, discussion was facilitated by therapy assistants.

Filming procedure

Group conversations were filmed using a pre-defined protocol in order to capture the interaction from multiple angles. Four tripod-mounted GoPro Hero 3 edition cameras, each attached with a EDITIGE ETM-001 dual microphone, recorded the interaction.

Sampling procedure

Fourteen 20 minute films were randomly selected (30% of total data) and prepared for analysis using Final Cut Pro editing software version 10.5.4 (Apple Inc). Samples were prepared using three-way views of a participant interacting with the rest of the group on one screen. Each clip commenced 10 minutes either side of the half way point in the conversation. Sampling from the same time point is in accordance with accepted protocols to guard against sample selection bias (Correll, van Steenbrugge and Scholten, 2010). Each began with a new sentence and/or new idea and ended at a point of natural pause. Pseudonyms were used to label each clip, and all clips were then copied in random sequence onto an encrypted hard drive.

Rating Procedure

The INT coding system

Refinements to the original coding system reduced the number of codes from 14 to seven. This process was achieved by combining the existing codes into code categories. The new version includes the fourteen original interaction behaviours, combined under six codes with eye gaze as an additional response. See [Table 2](#).

The code definitions are designed to have a broad reach. For example, code 1 (linguistic initiations to one other) records all behaviours initiated linguistically (e.g. both questions and statements) between the participant and the interlocutor(s), and code 4 records a linguistic response to one or more participants or to the group as a whole. Coding procedures require raters to focus on a named participant and to code the initiation and response behaviours for that participant and the interlocutors to whom they are directed or from whom they are received. These are then tallied within the software, and the transformation of these data into graph form provides insights into adaptive behaviour.

The raters

Two raters, both experienced speech and language therapists familiar with the INT coding scheme, independently coded the 14 films. The raters familiarised themselves with the coding system over two 60 minute meetings and used video clips unconnected to the study for coding practice. All discrepancies arising in the practice tasks were discussed and resolved between the raters.

Data analysis procedure

The reliability of each of the seven new codes was calculated using intra-class correlations (ICCs) (Shrout and Fleiss, 1979) across four sample lengths (5, 10, 15, 20 minutes). Inter-rater reliability was measured using a two-way random-effects model in order to generalise findings to coders in clinical practice. The option of ‘single’ (rather than ‘average’)

Table 2. The INT coding system.

(1) Initiation to one other – linguistic: – Verbal – Writing – Signing – AAC	(2) Initiation to group – linguistic: – Verbal – Writing – Signing – AAC	(3) Initiation (non-linguistic): – Point – Reach – Gesture – Draw – Facial expression – Eye gaze (people and objects)
Responses		
(4) Response – linguistic: – Verbal – Writing – Signing – AAC	(5) Response (non-linguistic): – Point – Reach – Head shake/nod – Gesture – Draw – Facial expression	(6) Other voiced response (7) Response eye gaze

measures was chosen to evaluate generalisability to a single coder working in clinical practice. 'Consistency' (rather than 'absolute agreement') was selected because consistency in the number of frequencies tallied across raters is more important than absolute agreement.

Interpretation guidelines from Koo and Li (2016) have been followed in this report. In addition to a point estimate, confidence intervals are also reported.

A one-way repeated measures analysis of variance was used to compare the means of the four sample lengths. All statistical analyses were computed using the IBM SPSS software platform (version 27.0).

Results

Inter-rater reliability

ICC calculations and their 95% confidence intervals across the four sample lengths and the interaction code type are presented in Table 3.

ICC point estimates for the majority of the interaction codes range from moderate to excellent. The confidence intervals are wide, although only one (code 3 non-linguistic initiation) dropped into minus figures in the 5 minute sample length.

We conducted a further analysis of non-linguistic initiations in order to establish the reliability of the individual components that make up this code i.e. point, reach, gesture, draw; facial expression; eye gaze to people; eye gaze to objects. Reliability was calculated from the data in the 20 minute sample in order to maximise the chances of capturing the occurrence of the target behaviours within a longer time limited sample. ICC point estimates are shown in Table 4.

The findings reported in Table 4 indicate that further definition of non-linguistic initiation behaviour for facial expression and eye gaze is required to improve rater accuracy.

Table 3. Intra-class correlations (ICC 3,1) with confidence intervals (CI) 95% across four sample lengths for interaction codes 1-7.

Interaction code	5 minutes		10 minutes		15 minutes		20 minutes	
	Single measures	95% CI	Single measures	95% CI	Single measures	95% CI	Single Measures	95% CI
1. Linguistic initiation to one other	.754	.392 - .914	.686	.265 - .887	.738	.362 - .908	.810	.508 - .935
2. Linguistic initiation to group	.731	.347 - .905	.799	.485 - .931	.788	.461 - .927	.775	.434 - .922
3. Non-linguistic initiation	.458	-.073 - .787	.586	.103 - .845	.632	.175 - .865	.568	.076 - .838
4. Linguistic response	.969	.908 - .990	.958	.876 - .986	.961	.882 - .987	.897	.711 - .966
5. Non-linguistic response	.842	.579 - .946	.878	.664 - .959	.872	.650 - .957	.888	.689 - .963
6. Other voiced response	.864	.630 - .954	.830	.552 - .942	.830	.552 - .942	.893	.700 - .964
7. Response eye gaze	.618	.152 - .859	.689	.271 - .888	.785	.453 - .925	.784	.453 - .925

Table 4. Reliability evaluation of the non-linguistic initiation behaviours combined in code 3.

3. Initiations (non-linguistic)	
Point, reach, gesture, draw	.893
Facial expression	.465
Eye gaze to people	.492
Eye gaze to objects	.276

Sample length evaluation

Sample lengths were evaluated by comparing the means of the four sample lengths.

The one-way repeated measures analysis of variance found no difference between the four sample lengths ($F(3, 18) = 2.06, p = 0.14$).

This finding was further explored using INT data from one randomly selected participant (Erin) to check consistency across the four sample lengths. [Figure 2](#) shows the total number and proportion of initiation and response behaviours across the four sample lengths for Erin and her three co-participants. The network graphs provide a visual representation showing the profile of connections for all interaction types between Erin and co-participants.

The table of interaction frequencies represents the total number of initiation and response counts. These data show that the proportional contribution of Erin and her three co-participants in the interaction was consistent across the sample lengths. The interpretation provided by the network graphs sets the distribution of the interactional behaviours (initiations and responses) in context and shows a similar balance of participation across the sample lengths.

Discussion

In this study we have addressed two research questions. We have evaluated the reliability of a modified coding system for the INT. We have also evaluated the effects of sample length on reliability.

Previous researchers (e.g. Eriksson et al., 2014) have identified a difference between first order conversation measures where the frequency of defined behaviours is tallied, and second order global observational measures where raters are required to assign a value to a scale, based on their perceptual judgement. Both offer a quantitative approach, but global observational measures have previously been shown to be more sensitive to the conversation construct than behavioural frequency measures (Kagan et al., 2004; Off, Rogers and Alarcon, 2006; Correll et al., 2010). The INT differs from frequency measures that tally specific communication behaviours (such as the number of speaking turns). Its coding system is made up of initiations and responses between participant and interlocutors, and code definition is intentionally broad in order to capture participatory behaviours that are sensitive to the behaviour of others in the group. The behavioural frequencies provide a quantitative measure of participation and the software transforms these data into a visual representation. Comparison over time provides insight into adaptive behaviour, such as changes to patterns of dominance or

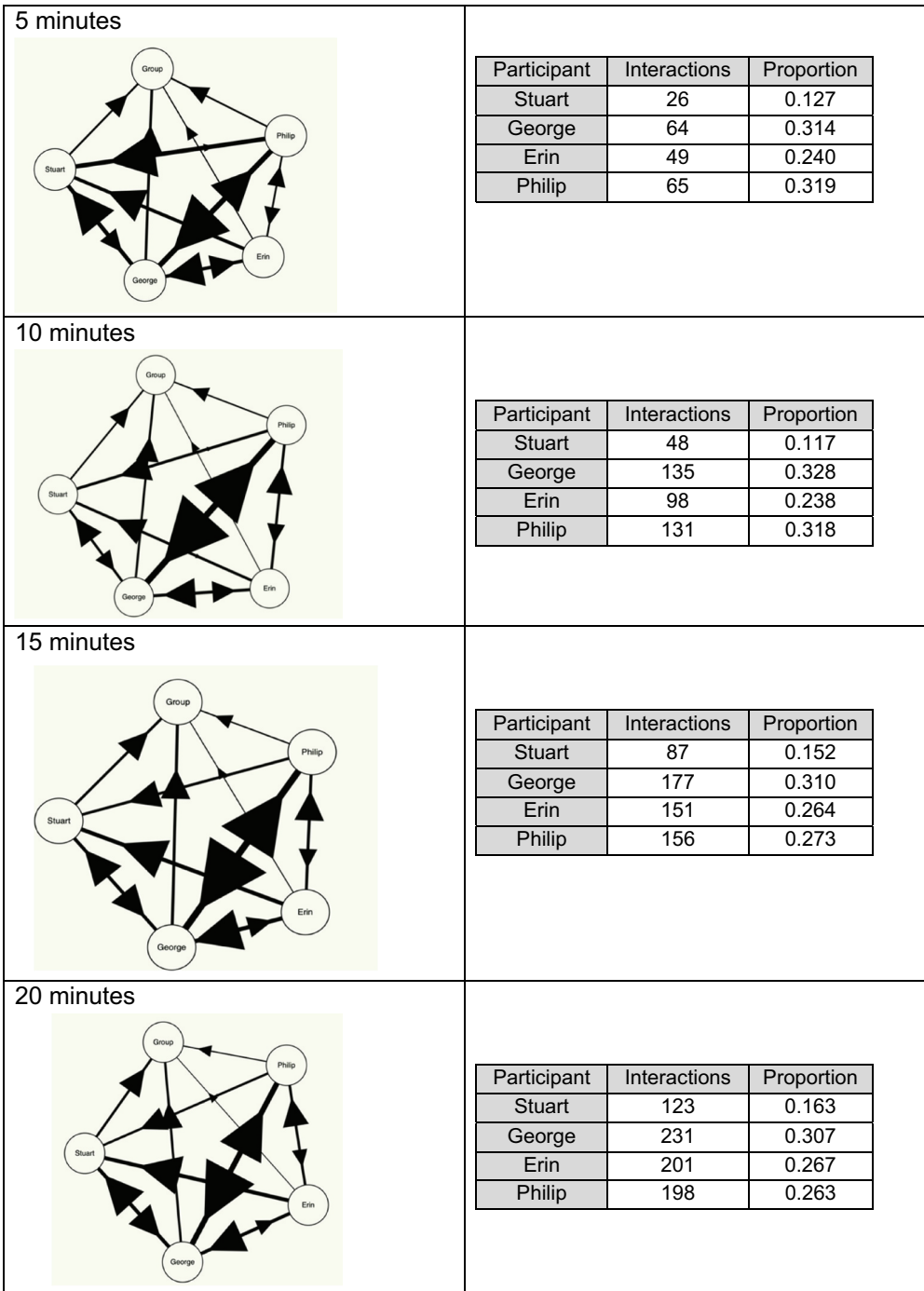


Figure 2. An illustration of INT initiation and response frequencies and proportions across four sample lengths for Erin and her co-participants.

under-participation. Importantly, raters are not required to make subjective judgements on the quality of the interactional behaviours and the INT makes no distinction between functional or dysfunctional content. This lends weight to Eriksson et al. (2014) who argue for a range of assessment methods to evaluate conversational interaction, both quantitative and qualitative, depending on the construct to be measured.

The new coding structure for the INT combined the previous 14 codes into seven code categories. This process drew on established theoretical models of social communication (Trower et al., 1978) and discourse analysis models comprising initiation and response categories to evaluate interaction (Coulthard, 1984; Eggins and Slade, 1997). Despite the attention to category definition, reliability of code 3 (non-linguistic initiations) was poor over all time samples.

Eye gaze initiations, defined as a cue to signal to other people (Argyle and Cook, 1976) were difficult to reliably code. Gaze poses a unique measurement challenge, not least because people will almost always be looking somewhere. Most recent research relies on eye-tracking technologies for directional accuracy (e.g. Auer 2021). Accuracy of interpretation presents a further challenge because gaze is multi-functional, perhaps resulting in a more impressionistic rating. Auer (2021) showed gaze at the end of a speaker turn to be the strategy most frequently employed for next speaker selection. In our investigation with ABI samples, eye gaze initiations often occurred simultaneously with other non-linguistic initiations. The seven item coding system addresses this issue by combining behaviours. This ensures that the initiating behaviour is recorded only once, as facial expression, gesture and an eye gaze cue can happen concurrently. By contrast, eye gaze responses (to a speaker or other participant behaviour) were more reliable. Eye gaze initiations require further definition to ensure that frequency counts tally contributions to participation rather than disengagement (for example looking at an object outside of the conversational context).

Findings from this reliability evaluation of a seven-item coding system for the INT indicate that ratings can be based on 5, 10, 15 or 20 minute samples of everyday conversation. Trained users of the INT in clinical practice can therefore achieve a reliable result using short samples. This has practical benefits for clinical practice as shorter samples mean that data collection and measurement procedures are more time efficient. Longer samples may help less experienced users to achieve the reported levels of reliability.

As other authors have pointed out, short samples may be untypical of the conversation. However, there is no agreed sampling protocol for either dyadic or group conversation. Previous analyses of dyadic conversation have sampled from different points. For example, Best et al. (2016) sampled from 5 minutes into a conversation of approximately 20 minutes in length. Correll et al. (2010) sampled from the beginning, middle and the end to enable judgements on the initiation and conclusion of the conversation. In order to capture a representative snapshot of group conversation, samples in this investigation were taken from the mid-point of conversations lasting up to 1 hour.

ICCs are the recommended method to evaluate outcome measurement reliability. On the one hand values are easy to interpret, falling between 0 (random agreement) and 1 (perfect agreement), but differing interpretation guidelines on the magnitudes of

agreement make comparison across studies difficult. We acknowledge that many studies in the ABI field using Cicchetti's more lenient interpretation guidelines were published before the arrival of more stringent interpretations proposed by Koo and Li (2016) and Post (2016). It is clear that irrespective of a descriptor such as fair or moderate, the degree of measurement error at 0.4 – 0.6, for example, makes reproducibility less convincing. However, with the exception of non-linguistic initiations (code 3), the point estimates for the INT over 20 minutes range from good to excellent reliability using the more stringent interpretation. Modifications to the coding instructions for code 3 indicate the likelihood of stronger correlations between therapists using the measure to evaluate short sample lengths.

Implications for clinical practice, study limitations and future directions

The INT is a new way of measuring group communication behaviour. This paper documents the latest refinements to the INT, which simplify its use in clinical practice, and its reliability has now been demonstrated across short sample lengths. Although the data were drawn from a group intervention for people with ABI, the intention is that the INT will hold relevance for a variety of applications and clinical groups.

The raters in this evaluation were speech and language therapists and experienced users of the INT. Establishing reliability following training for novel raters will be informative. User training for the INT needs further investigation.

As this paper shows, in addition to evaluating the quality of a measurement tool in terms of outcome reproducibility and consistency, measures of reliability are also a means to refine an instrument. In addition to modifications through code definition and evaluation with new clinical populations, future directions also include training as a means to improve rater reliability. Further validation across other clinical populations using the seven item coding system is also now indicated.

Acknowledgements

The authors thank Sprechen – the Software Studio for their support with the INT.

Disclosure statement

The authors report no conflicts of interest.

ORCID

Susan Howell PhD  <http://orcid.org/0000-0002-8329-7529>
Suzanne Beeke PhD  <http://orcid.org/0000-0002-6772-2820>
Emma Louise Sinnott  <http://orcid.org/0000-0002-4673-2448>
Rosemary Varley PhD  <http://orcid.org/0000-0002-1278-0601>
Tim Pring PhD  <http://orcid.org/0000-0002-3671-7471>

References

- Adams, C., 2005, Social communication intervention for school-age children: Rationale and description. *Seminars in Speech and Language*, 26(3), 181–188. <https://doi.org/10.1055/s-2005-917123>
- Ahlsen, A., and Saldert, C. (2018). Activity-based communication analysis – focusing on context in communication partner training. *Aphasiology*, 32:10, 1194–1214, DOI: [10.1080/02687038.2018.1464645](https://doi.org/10.1080/02687038.2018.1464645)
- Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press.
- Auer, P. (2021). Turn-allocation and gaze: A multimodal revision of the “current-speaker-selects-next” rule of the turn-taking system of conversation analysis. *Discourse Studies*, 23(2), pp. 117–140. doi: [10.1177/1461445620966922](https://doi.org/10.1177/1461445620966922).
- Bandura, A. (1971). *Social learning theory*. New York: General Learning Press.
- Behn, N., Togher, L., Power, E. and Heard, R. (2012). Evaluating communication training for paid carers of people with traumatic brain injury. *Brain Injury*, 26(13–14), pp. 1702–1715. doi: [10.3109/02699052.2012.7](https://doi.org/10.3109/02699052.2012.7).
- Best, W., Maxim, J., Heilemann, C., Beckley, F., Johnson, F., Edwards, S.I., Howard, D. and Beeke, S. (2016). Conversation therapy with people with aphasia and conversation partners using video feedback: a group and case series investigation of changes in interaction. *Frontiers in Human Neuroscience*, 7(10), p. 562. doi: [10.3389/fnhum.2016.00562](https://doi.org/10.3389/fnhum.2016.00562).
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), pp. 284–290. doi: [10.1037/1040-3590.6.4.284](https://doi.org/10.1037/1040-3590.6.4.284).
- Correll, A., van Steenbrugge, W. and Scholten, I. (2010). Judging conversation: how much is enough?. *Aphasiology*, 24(5), pp. 612–622. doi: [10.1080/02687030902732752](https://doi.org/10.1080/02687030902732752).
- Coulthard, M. (1984). Conversation analysis and social skills training. In Trower, P. (ed.), *Radical Approaches to Social Skills Training*. The University of Michigan: Croom Helm.
- Dahlberg, C., Hawley, L., Morey, C., Newman, J., Cusick, C.P. and Harrison-Felix, C., 2006, Social communication skills in persons with post-acute traumatic brain injury: three perspectives. *Brain Injury*, 20(4), pp. 425–435. doi: [10.1080/02699050600664574](https://doi.org/10.1080/02699050600664574).
- Doedens, W.J. and Meteyard, L. (2020). Measures of functional, real-world communication for aphasia: a critical review. *Aphasiology*, 34 (4), 492–514. doi:[10.1080/02687038.2019.1702848](https://doi.org/10.1080/02687038.2019.1702848)
- Eggins, S. and Slade, D. (2005). *Analysing Casual Conversation*. London, Equinox
- Eriksson, K., Bergström, S., Carlsson, E., Hartelius, L., Johansson, C., Schwarz, A., & Saldert, C. (2014). Aspects of rating communicative interaction: Effects on reliability and agreement. *Journal of Interactional Research in Communication Disorders*, 5(2), 245–267. <https://doi.org/10.1558/jircd.v5i2.245>
- Gordon, R.G., Rigon, A. and Duff, M.C. (2015). Conversational synchrony in the communicative interactions of individuals with traumatic brain injury. *Brain Injury*, 29 (11), pp. 1300–1308. doi: [10.3109/02699052.2015.1042408](https://doi.org/10.3109/02699052.2015.1042408).
- Gordon, R. G., Tranel, D. and Duff, M. C. (2014). The physiological basis of synchronizing conversational rhythms: the role of the ventromedial prefrontal cortex. *Neuropsychology* 28 (4), pp. 624–630. doi: [10.1037/neu0000073](https://doi.org/10.1037/neu0000073).
- Hartley, L.L. and Jensen, P.J., 1991, Narrative and procedural discourse after closed head injury. *Brain Injury*, 5(3), pp. 267–285. doi: [10.3109/02699059109008097](https://doi.org/10.3109/02699059109008097).
- Horton S., Clark A., Barton G., Lane, K. and Pomeroy, V.M. (2016). Methodological issues in the design and evaluation of supported communication for aphasia training: a cluster- controlled feasibility study. *BMJ Open* 6:e011207. doi:[10.1136/bmjopen-2016-011207](https://doi.org/10.1136/bmjopen-2016-011207)
- Howell, S. (2018). Measuring outcomes from a peer-led social communication skills intervention for adults following acquired brain injury. Thesis (PhD), UCL (University College London). <https://discovery.ucl.ac.uk/id/eprint/10059713>
- Howell, S., Beeke, S., Pring, T. and Varley, R. (2020a). Measuring outcomes of a peer-led social communication skills intervention for adults with acquired brain injury: A pilot investigation. *Neuropsychological Rehabilitation* 31(7): 1069–1090. doi: [10.1080/09602011.176892](https://doi.org/10.1080/09602011.176892)

- Howell, S., Varley, R., Sinnott, E. L., Pring, T., & Beeke, S. (2020b). Measuring group social interactions following acquired brain injury: an inter-rater reliability evaluation. *Aphasiology*, 1–13. <https://doi.org/10.1080/02687038.2020.1836315>
- Kagan, A., Black, S.E., Duchan, J.F., Simmons-Mackie, N. and Square, P. (2001). Training volunteers as conversation partners using ‘Supported Conversation for Adults with Aphasia’ (SCA): A controlled trial. *Journal of Speech, Language and Hearing Research* 44 (3): 624–638. doi: 10.1044/1092-4388(2001/051).
- Kagan, A., Winckel, J., Black, S., Duchan, J. F., Simmons-Mackie, N. and 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. DOI:10.1310/CL3V-A94A-DE5C-CVBE.
- Keegan, L.C., Behn, N., Power, E., Howell, S., and Rietdijk, R. (2022). Assessing Conversation after Traumatic Brain Injury. In C. Coehlo, L.R. Cherney and B.B. Shadden (Eds.), *Discourse Analysis in Adults With and Without Communication Disorders: A Resource for Clinicians and Researchers*. Plural Publishing.
- Koo, T. K. and Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2): 155–163. doi: 10.1016/j.jcm.2016.02.012.
- Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press. doi: 10.2307/2804509.
- McGraw K.O. and Wong SP (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1(1):30–46. <https://doi.org/10.1037/1082-989x.1.1.30>
- Mokkink, L.B., Boers, M., van der Vleuten, C.P.M., Bouter, L.M., Alonso, J., Patrick, D.L., de Vet, H.C.W. and Terwee, C.B. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Medical Research Methodology* 20 (293)
- Off, C., Rogers, M. and Alarcon, N. (2006). Three Methods of Quantifying the Quality of Communication in Aphasia. *Clinical Aphasiology Conference, Ghent, Belgium, 29 May- 2nd June 2006*. <http://aphasiology.pitt.edu/1747/>
- Post, M.W. (2016). What to do with “moderate” reliability and validity coefficients? *Archives of Physical Medicine and Rehabilitation*, 97(7), pp.1051–1052. <http://doi.org/10.1016/j.apmr.2016.04.001>
- Rietdijk, R., Power, E., Brunner, M., & Togher, L. (2020). The reliability of evaluating conversations between people with traumatic brain injury and their communication partners via videoconferencing. *Neuropsychological Rehabilitation*, 30(6) 1074–1091, <https://doi.org/10.1080/09602011.2018.1554533>
- 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. <https://doi.org/10.1080/02687038.2012.710317>
- Saldert, C., Jensen, L.R., Blom Johansson, M and Simmons-Mackie, N. (2018). Complexity in measuring outcomes after communication partner training: alignment between goals of intervention and methods of evaluation. *Aphasiology*, 32:10, 1167–1193. DOI: 10.1080/02687038.2018.1470317
- Scott, J. (2017). *Social Network Analysis*. 4th ed. London: SAGE Publications Ltd.
- Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), pp. 420–428. doi: 10.1037/0033-2909.86.2.420.
- Sim, P., Power, E. and Togher, L. (2013). Describing conversations between individuals with traumatic brain injury (TBI) and communication partners following communication partner training: using exchange structure analysis. *Brain Injury*, 27(6), pp. 717–742. doi: 10.3109/02699052.2013.775485.
- Sohlberg, M.M., MacDonald, S., Byom, L., Iwashita, H., Lemoncello, R., Meulenbroek, P., Ness, B. and O’Neil-Pirozzi, T.M. (2019). Social communication following traumatic brain injury part I: State-of-the-art review of assessment tools, *International Journal of Speech-Language Pathology*, 21:2, 115–127, DOI: 10.1080/17549507.2019.1583280
- Togher, L., McDonald, S., Tate, R., Power, E. and Rietdijk, R. (2013). Training communication partners of people with severe traumatic brain injury improves everyday conversations: a multicenter single blind clinical trial. *Journal of Rehabilitation Medicine*, 45(7), pp. 637–645. doi: 10.2340/16501977-1173

- Togher, L., Power, E., Tate, R., McDonald, S. and 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), pp. 914–927. doi: [10.1080/02687030903422478](https://doi.org/10.1080/02687030903422478).
- Togher, L., Wiseman-Hakes, C., Douglas, J., Stergiou-Kita, M., Ponsford, J., Teasell, R., Bayley, M. and Turkstra, L.S. (2014). INCOG recommendations for management of cognition following traumatic brain injury, part IV: cognitive communication. *Journal of Head Trauma Rehabilitation*, 29(4), pp. 353–368. doi: [10.1097/HTR.0000000000000071](https://doi.org/10.1097/HTR.0000000000000071)
- Trower, P., Bryant, B. and Argyle, M. (1978). *Social Skills and Mental Health*. Hove, UK: Routledge.
- Vygotsky, L.S. (1978). *Mind in Society: the Development of Higher Psychological Processes*. Cambridge, Massachusetts: Harvard University Press.
- Wallace, S. J., Worrall, L.E., Rose, T and Le Dorze, G. (2018). Discourse measurement in aphasia research: have we reached the tipping point? A core outcome set . . . or greater standardisation of discourse measures? *Aphasiology*, 32(4), pp. 479–482. doi: [10.1080/02687038.2017.1398811](https://doi.org/10.1080/02687038.2017.1398811).
- Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Wilkinson, R. 1999, Sequentiality as a problem and resource for intersubjectivity in aphasic conversation: analysis and implications for therapy. *Aphasiology*, 13 (4–5), pp. 327–343. <https://doi.org/10.1080/026870399402127>