Interrater reliability functions in CAT

"CAT provides feedback on interrater reliability for each single code of the observational measure. For example, interrater reliability calculations for a rather detailed coding scheme with twenty codes would result in a total of twenty parameters. A code-specific interrater reliability allows the researcher to inspect which particular codes in the coding scheme are harder/easier for coders to detect. There are multiple reasons for this: Some codes may have unclear definitions, other codes may be harder to distinguish from conceptually similar codes, and some codes may just be harder to observe than others because their require higher levels of observer inference (e.g., "asking a question" might be easier to recognize than "feelings of unease").

The session-based interrater reliability feedback implemented in CAT allows researchers to detect such *trouble-making codes* and gives them a more focused approach to coder training. That is, by inspecting code-specific interrater reliability estimates, the researcher can decide which codes need more attention in subsequent coder discussions, whether codes need to be re-defined, or whether semantically similar codes should be combined altogether. We designed this feature because we believe that interrater reliability calculations should not be considered as an afterthought in observational research but rather at the very beginning. Moreover, session-based interrater reliability feedback allows the researcher to monitor the quality of incoming data on a session-by-session basis. A session constitutes a longer observation period such as a one-hour workplace meeting.

Interrater reliability values reported in CAT are calculated based on intraclass correlations (ICCs; McGraw & Wong, 1996) following recommendations reported in

Hallgren (2012). Users can request estimates as soon as two independent" (Klonek, Meinecke, Hay & Parker, 2020, p. 22-23).

Preliminary comments on computing IRR metrics

Interrater reliability (IRR) is a core criterium when evaluating observational coding schemes (Bakeman & Quera, 2011; Hallgren, 2012; Seelandt, 2019). There are a number of statistical indices that can be used to assess the IRR of a coding scheme (Klonek et al., in press). CAT has some built-in functions to calculate Intraclass correlations (ICC). On a general note, we want to stress that there are many decisions that you will have to make and that will affect how and whether you can/will compute IRR (and ICCs). For example:

- What is your planned number of double-coded sessions (our general rule of thumb is to code at least 5 sessions or at least 20% of the final corpus); for more details see Klonek, Quera, & Kauffeld, 2015, p. 288):

"Sample size for reliability analysis was chosen a priori following two guidelines: Bakeman, Deckner, and Quera (2005) recommended sampling between 15% and 20% of a corpus to check reliability using the kappa coefficient. Furthermore, the minimum sample size for calculation of ICCs lies around five (Yoder & Symons, 2010), while ten or more sessions will result in more robust results (Bakeman & Quera, 2011)"

    - Relatedly, are you using a fully crossed or partly crossed coding design (cf., Hallgren, 2012)?

- What is the length of the double-coded sessions (2 minutes or 4 hour-long sessions?)? – The answer to this question is important to select an appropriate "Grouping Interval" in the ICC calculation in CAT

- What is your number of coders who are coding the same session (i.e., inter-rater reliability) or is there only a single coder who is repeatedly coding the same session (intra-rater reliability)?
- Are you coding single events with nominal categories or using rating scales (e.g., Likert scales)?
- Are you double-coding based on live-observation or media/video-recordings?

This just shows that IRR is a complex issue and any choices with respect to the aforementioned issues will affect how or whether (at all) you can calculate a statistical IRR parameter. In sum, if your goal is to use CAT to calculate ICCs, we recommend to get in contact with one of the CAT developers (Dr. Florian Klonek; florian.klonek@curtin.edu.au) and lay out your approach. This can ensure that you can use CAT for your intended research.

What is required to compute ICC?

As a broader rule of thumb, you need to ensure that you have at least double-coded a single (or ideally multiple) sessions (either by the same or by multiple coders).

Where can I find ICC analysis?

Unfortunately, the ICC function is still a very hidden in the depth of the program (we will change this in the future). Apologies in advance!

To find ICC analyses:

Go to the section "**Measures**".

Click **Option → Feedback →** "**+ create feedback**" (*you need to name feedback file*)→ go on "**data source**" and select your data (e.g., "**filter by recorder**", which will allow you to obtain all sessions that have been double-coded from multiple coders/recorders. Use the logical operator "**OR**" to connect data from multiple coders) → you have to click the red button "**apply filters**" (!); the green button shows you how many events have been selected → in the

section “Feedback content” on the right panel click on the button “+**ADD**” and select “**Data Visualisation**” → in the section “**Visualisation Type**” select “**Data visualization: Code frequency & Interrater reliability feature**” → in the “**Visualization Structure**” select “**X = Inter-rater reliability, Y = Shown Category**”

The next choices depend a bit on your measure and research question:

- If you have multiple classes in your measure, you need to select the correct class in the section “**Data Category/Rating System**”; furthermore, you can select specific codes/categories (which may be useful if you have many codes in your coding system).
- Depending on your data collection approach, select “live coding” or “media file coding”
- We recommend to click “double-code only”. As a result, CAT will select the sessions that have been double-coded by the the second coder.
- In “Select Double-code Source Session”, the sessions that have been coded by coder 1
- In “Select Double-code Session”, the sessions that have been coded by coder 2
- In “Grouping Interval” select the time-interval that is used to calculate summary scores for nominal categories. We developed this feature so that researchers can calculate ICCs based on coding a single session (usually, calculation of ICCs requires multiple sessions). The “Grouping Interval” splits your session into multiple intervals (the length is defined by the user) and then groups the summary counts for a code for each interval.

  The shorter your observation interval for the entire session (e.g., 1 hour vs. 24 hours), the shorter your “grouping interval” needs to be (e.g., 10 min vs. 1 hour). In this example (10 min of 1 hour), CAT treats a 1-hour observation as 10

observations from two observers and compares their alignment. Keep in mind that shorter "grouping interval" also give you less opportunity to sample relevant behavior. That is, how much relevant behavior can actually occur in a 10 minute interval? Shorter intervals will most likely have values of zero for your codes. However, if you use a highly micro-coding system (coding every few seconds), then a time-interval of 10 minutes might be sufficient.

Statistical Calculation of IRR in CAT

The calculation of ICCs (and their variants) assume that ratings from multiple observers for a set of targets (i.e., observation intervals) are composed of a true score component and measurement error component. This can be rewritten from equation 1 in the form

$$X_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$$

where X_{ij} is the frequency of a behavior code for a time interval i provided by observer j, μ is the mean of the true score for this variable X, r_i is the deviation of the true score from the mean, c_j estimates any systematic deviation of observer j, rc_{ij} represents the interaction between observer and deviation for each interval, and e_{ij} is the measurement error. Hence, the session-based ICCs for each code is calculated based on a "time interval" x "observer" matrix. The observation interval length is determined by the researcher.

To illustrate: A researcher wants to know the inter-rater reliability for a code A based on 5-minute resolution for a session with a 50 min. length. The matrix for the ICC calculation is based on ten repeated observations (10 rows (intervals) x 2 observer matrix). If observer 1 has coded 3 instances of behavior A during the first 5 min interval and observer 2 has coded behaviour A 5 times (during the same interval), the observed frequencies for A for the first interval are 3 and 5 for observer 1 and 2, respectively.

CAT provides single-measure ICCs (i.e., ICC(A,1) and ICC(C,1)). The single-measure ICCs are more conservative reliability estimates as they tell the researcher if the coded behaviours from a single observers can be used for further analyses (the average-measure ICC(k) tend to be higher than single-measures ICC(1), Hallgren et al., 2012).

Two types of ICCs are calculated, a consistency based (also called relative) ICC(R) and an absolute ICC(A). Both ICCs are calculated based on a two-way mixed ANOVA model. The relative based ICC(R) provides feedback on whether two (or more) observers' frequency scores are similar in relative rank order. The absolute based ICC(A) is more conservative and indicates whether absolute values are similar between observers (Hallgren, 2012). Researchers can decide which ICC variant is most suitable for their specific research question. If they want to know if the frequencies for each code are similar in absolute value, then absolute ICC(A) should be inspected. If researchers want to know if frequencies for each code are similar in rank order, then consistency should be inspected. To illustrate these differences, consider the following example in which two coders made observations during five consecutive sessions The frequency counts of code A assigned by Coder 1 are generally low (2, 3, 6, 7, 10) whereas Coder 2 assigned code A to a larger extent (8, 9, 12, 13, 16). Because the frequencies are perfectly ordered in rank, the relative ICC(R) in value in this example reaches a maximum (ICC(R) = 1.00). In contrast, the ICC(A) variant pays attention to absolute agreement and would be rather low in this example (ICC(A) = .36). IRR feedback provides a preliminary reliability analysis and we recommend to code multiple sessions (i.e., at least five, ideally ten sessions, see Bakeman & Quera, 2011). (Klonek, Meinecke, Hay, & Parker, 2020, p. 23)

References

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Klonek, F.E., Meinecke, A., Hay, G., & Parker, S. (*in press*). Capturing team dynamics in the wild: The communication analysis tool. *Small Group Research.*

Klonek, F. E., Quera, V., & Kauffeld, S. (2015). Coding interactions in Motivational Interviewing with computer-software: What are the advantages for process researchers?. *Computers in Human Behavior, 44*, 284-292.

Seelandt, J. C. (2018). Quality control: Assessing reliability and validity. In E. Brauner, M. Boos, & M. Kolbe (Eds.), *The Cambridge handbook of group interaction analysis* (pp. 227–244). Cambridge University Press.