**Working title:**

**Reliability and Validity of smartphone-based cognitive-motivational experiments in alcohol use order**

**Authors**: Hilmar Zecha, Maria Waltmannb, c, Ying Leea, d, e , Markus Reichertf, g, h, Ulrich Ebner-Priemerg, h, Michael Rappi, Andreas Heinzj, Ray Doland, e, Michael N. Smolkaa,b,c, Lorenz Deserno a,b,d,e

Maria Garbusow, Michael Marxen, Heike Tost,

**Affiliations:**

a Department of Psychiatry, Technische Universität Dresden, Dresden, Germany

b Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Centre of Mental Health, University of Würzburg, Margarete-Höppel-Platz 1, 97080 Würzburg, Germany

c Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1, 04103 Leipzig, Germany

d Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom

e The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, United Kingdom

f Department of eHealth and Sports Analytics, Faculty of Sport Science, Ruhr-Universität Bochum (RUB), Bochum, Germany

g Mental mHealth Lab, Institute of Sports and Sports Science, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

h Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

i Department for Social and Preventive Medicine, University of Potsdam, Potsdam, Germany

j Department of Psychiatry and Psychotherapy, Charité Campus Mitte, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

**Method**

**General procedures.**

As part of a German research consortium on addiction at three sites (Technical University Dresden, Charité Berlin, and Central Institute of Health Mannheim; Heinz et al., 2020), a smartphone-based longitudinal Eocological Momentary Assessment (EMA) of up to 1 year was performed with a range of subjective reports. Once per month, patients performed four cognitive-motivational tasks on the smartphone taken from the Great Brain Experiment (GBE) app (Brown et al., 2014, see below for details). Before starting this EMA study, patients underwent extensive clinical and neurocognitive assessments, which we refer to as basic assessment (see Heinz et al., 2020). In the current study, we rely on data of cognitive-motivational tasks and subjective reports from the smartphone-based EMA study as well as data from the basic assessment. Patients also participate in multiple subprojects of the consortium (see Heinz et al., 2020), which are not subject to the present study.

**Participants & Instruments.**

The study procedure was approved by the review boards of the local ethics committee at Heidelberg University (2018-621N-MA), Charité – Universitätsmedizin Berlin (EA1/212/18), and Technical University Dresden (EK 459112018). All participants gave written consent before participating in the study. Participants with mild to moderate Alcohol Use Disorder (AUD) were included in the study at all three sites. Mild to moderate AUD was defined as the presence of at least two AUD criteria. Participants were recruited through flyers and advertisements. Telephone screenings were conducted to check inclusion and exclusion criteria. Exclusion criteria were: clinical indication for detoxification treatment, insufficient knowledge of the German language, seeking a therapeutic intervention, MRI contraindications, medical history of DSM-5 bipolar disorder, psychotic disorder, schizophrenia or schizophrenic spectrum disorder, or current use of drugs or medication nor substance dependence thereof other than alcohol, nicotine, or cannabis, as well as medical history of severe head injury, or other severe central nervous system disorders.

If included, participants were invited to an extensive basic assessment (see Heinz et al., 2020) including a retrospective questionnaire about alcohol consumption (Quantity-Frequency Questionnaire). This questionary gives a summary of participants typical alcohol consumption (drinking days, standard drinks per day), which we used in this study. A detailed description of this questionnaire can be found in supplemental material.

During this appointment, which was either conducted inside the laboratory or online via video chat, the app for running the EMA study (Movisens app; movisens GmbH, Germany; Reichert et al., 2021) as well as a customized version of the GBE app for assessment of the four cognitive-motivational tasks (see below) were installed either on participants’ own phone or on a study phone.

As pre-registered (Zech et al., 2021), data from 300 participants was analyzed for the present study. Participant characteristics can be found in Table 1.

|  |  |
| --- | --- |
| Participant characteristics (N = 300) | Mean (SD) [range] or percentage |
| Age (years) | 37.9 (13.0) [17–65] |
| Women | 113 (37.7%) |
| AUD criteria | 3.93 (1.51) [2–9] |
| Drinking days previous 3 months | 49.80 (21.71) [4–90] |
| Standard drinks per drinking day previous 3 months | 6.05 (3.54) [0.15–30.70] |
| Standard drinks per day previous 3 months | 3.13 (2.00) [0.04–12.69] |
| Mean alcohol consumption in first EMA month (g) | 37.06 (25.10) [0–138] |
| Median alcohol consumption in first EMA month (g) | 27.66 (29.41) [0-160] |

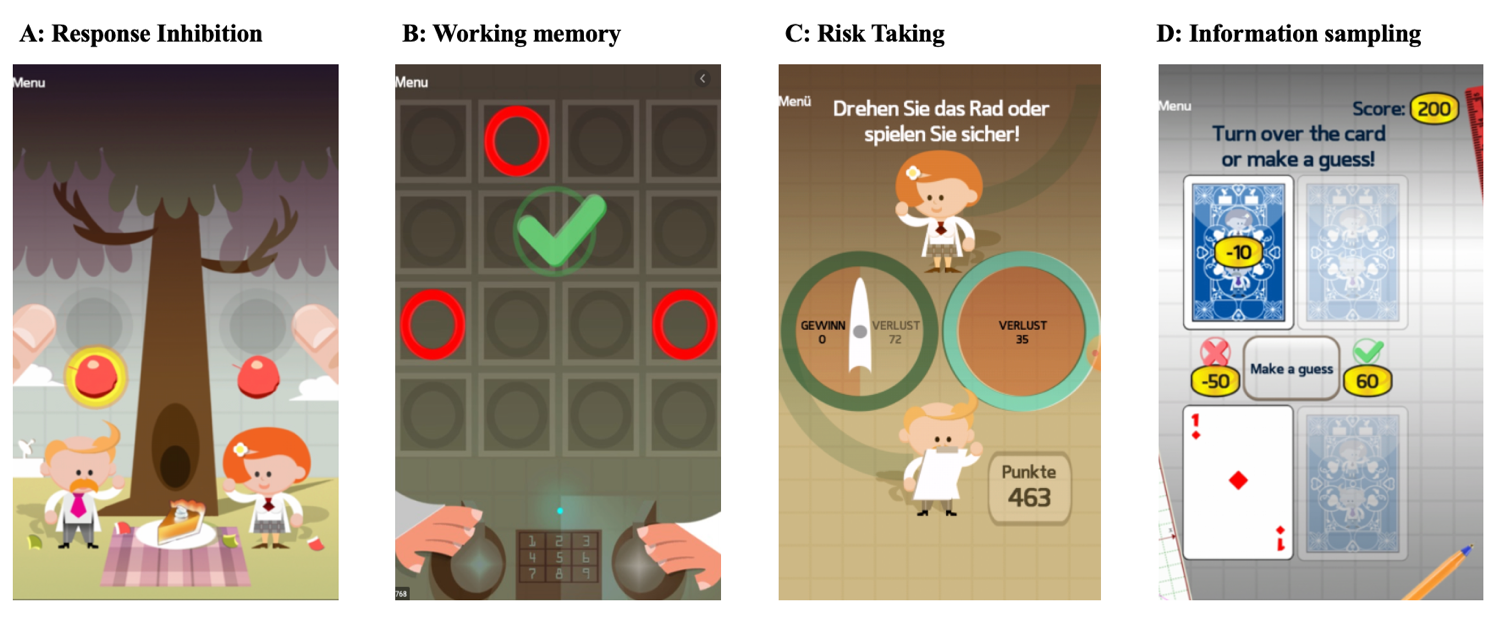
**Table 1. Participant characteristics.** This table shows participant characteristics of the sample.

**EMA Procedures.**

The 1-year EMA study was running entirely through the interface of the Movisense app. The GBE app customized for this study to assess four cognitive-motivational tasks (see subsequent section) could not be used by the participants themselves. Instead, use of the customized GBE app was only initiated through the Movisens app. During each GBE-session, the four tasks were completed in random order. On the first Monday after the basic assessment, the EMA study started. On this day, participants completed two GBE-sessions right after each other. Next, participants indicated during which weekday and at which time of the day they are least likely to consume alcohol. Throughout the following year, participants were notified every four weeks on the indicated weekday and time to complete one GBE-session. Throughout the same period, participants also completed several subjective reports on their smartphone every two days. In the present study, we include reports on alcohol consumption. For that purpose, participants were asked to indicate how many drinks from a list of commonly consumed alcoholic drinks they consumed the last day and the day before the last day. This provided a daily measure of alcohol consumption in g. Wording of the question can be found in supplemental material.

**Smartphone-based cognitive-motivational tasks.**

***Stop Signal Task.*** During the Stop Signal Task (SST; Smittenaar et al., 2015), participants tapped left or right on their smartphone screen to collect fruits falling a tree (Fig. 1A). Each trial began with two fruits hanging at the top of the screen for one to three seconds (randomly selected from a uniform distribution). Next, one of the fruits fell down and passed over one of two circles indicating the time during which participants should collect the fruit through tapping (Go-Trials with a response window spanning from 500 to 800 ms after stimulus onset). In 12 of 32 trials (37.5%), the falling fruit turned brown, indicating that it was rotten and should not be collected (stop trials). At the beginning of each session the delay after which the fruit turned brown (stop signal delay; SSD) was 350 milliseconds (ms). This delay changed according to staircase procedure (Verbruggen et al., 2019): it increased by 50 ms after each successful stop trial (rendering the subsequent stop trial more difficult) and decreased by 50 ms after each unsuccessful stop trial (rendering the subsequent stop trial easier).

**

**Fig. 1.** Illustration of smartphone-based tasks used in this study. The first panel from the left shows the stop signal task, the second shows working memory task, the third shows the risk taking task, the fourth shows the information sampling task.

***Working memory task***. During the working memory task (WMT, McNab et al., 2015) participants were asked to remember the positions of two up to 12 red circles presented on a 4 x 4 grid (see Fig. 1B). The task involved four conditions: In the ‘*long no distractor’* conditioncircles were presented for two seconds (encoding phase), then disappeared for one second (maintenance phase), before participants had to tap on their no-longer visible locations. In the ‘short *no-distractor’* condition, patterns were presented for one instead of two seconds. In the ‘*encoding-distractor’* condition, two yellow distractors were presented together with the red circles during the encoding phase. In the ‘*delayed-distractor’* condition, the same two yellow distractors were presented but during the maintenance phase.

Each condition started with three circles in trial one. If participants failed to respond correctly, two circles were presented in the second trial. If participants failed at this level, the condition was terminated. If a trial was completed correctly, the number of red circles in the corresponding condition increased by one in the next trial. If participants failed in a trial (from level four onwards) the level was repeated for once. If they failed again the condition was terminated. A maximum of eight trials was completed for each condition.

***Risk taking task***. During the risk taking task (RTT; Rutledge et al., 2014), participants repeatedly chose between a certain outcome and a gamble, with 50/50 probabilities of the two outcomes (see Fig. 1C). The task involved three conditions: In the ‘*gain’* conditionparticipants chose between either a certain gain or to gamble for a larger gain against 0 points. In the ‘*loss’* conditionparticipants chose between either a certain loss or to gamble for 0 points against a larger loss. In the ‘*mixed’* condition, participants chose between a certain amount of 0 points or to gamble for a gain against a loss amount. Each condition consisted of ten trials. In each trial, a certain amount was first randomly chosen with replacement from a fixed list of outcomes Gamble amounts were then calculated by multiplying the certain amount with a randomly chosen multiplier from another fixed list (for details, see Rutledge et al., 2014, supplementary materials). The task also involved current mood ratings (“How happy are you at this moment?”; rating line with endpoints “very happy” and “very unhappy”) which were presented after every two to three trials but are not subject to the currently reported reliability analysis.

***Information sampling task.*** During the information sampling task (Hunt et al., 2016) participants were presented with four playing cards in rows of two and had to choose the row with the largest sum of card values (see Fig. 1D). Each of the 21 trials began with all cards face down. Participants could invest points to turn over one card at a time to sample information with increasing costs for each additional card (zero points for the first card, 10 for the first card, 15 for the third, and 20 for the fourth card). Before turning over a card, participants could also chose to guess at no cost which row had the largest value. A choice at this stage would be a gamble (called a guess in the task) at 50/50. Participants won 60 points if this guess was correct and lost 50 points if the guess was incorrect. If turning over one or multiple cards, the costs for information sampling reduced the total win. Card values were sampled randomly with replacement from a discrete uniform distribution with integers ranging from one to 10.

**Data Analysis.**

The main goals of this study was to assess the tasks’ reliability. Where possible—we first assessed the tasks’ split-half reliability—or the consistency with which a task measures its construct within one measurement session. Next, we assessed the tasks’ test–retest reliability—or the consistency with which a task measures its construct between two measurement sessions. While assessing the tasks’ test–retest reliability, we compared two approaches of analyzing task data—the traditional *aggregation* approach, and an alternative *prediction* approach, which has recently been shown to improve reliability in other cognitive tasks (for details, see below; Waltmann et al., 2021). Finally, we also explored the effect of retest-period on test–retest reliability.

**Aggregation vs. prediction approaches**

For each task, we compared two approaches of analyzing task data: The first approach, which we subsequently call *aggregation,* is traditionally used to analyze task data. In this approach, summary scores are first created by aggregating data for each session of each participant. Next, these summary scores are used for inference, for example to calculate test-retest reliabilities. According to Haines et al. (2021) one problem of this approach is that it assumes that scores are estimated without measurement error. This, in turn, leads to ignoring uncertainty during inference, which, for example, attenuates test-retest reliability. A second problem is that this method assumes that person-level parameters are distributed uniformly across an interval that spans beyond a reasonable range of task scores. This is because, knowledge about scores from other participants or scores from other sessions of the same participant is not integrated in estimating individual session scores. Prior research shows that integrating such information into individual score estimation yields more precise scores (Efron & Morris, 1977; Gelman, 2006; Williams et al., 2020; as cited in Haines et al., 2021).

The second analysis approach, which we subsequently call *prediction*, overcomes both problems of the aggregation approach. Instead of first calculating summary scores and using them in a second step for inference, the prediction approach performs inference directly based on trial-level data. This allows it to carry within-session uncertainty into the inference step, thus increasing test-retest reliability. It also allows it to use information from other participants and sessions in each individual session score estimation, which increases precision. In this study, we implemented this approach by translating traditional aggregation-based analyses into analyses using hierarchical mixed models. Hierarchical mixed models allow to analyze data at the trial-level while still accounting for the participant and session structure of the data (for details about these translations for each task, see below).

To verify that this translation did not substantially change task scores, we also included a third analysis approach, in which we used mixed models to aggregate data. Here, first we ran mixed models separately on data from each session in order to predict participant summary scores for each session. Next, similar to the aggregation method, we entered these scores into the second inference stage (i.e., to calculate test-retest reliabilities).

In the following, we will describe for each task, how traditional, aggregation-based scores are traditionally created and how we translated these calculations into trial-level analyses using hierarchical mixed models.

**Task measures.**

***Stop Signal Task.*** The primary outcome of the stop signal task is the stop signal reaction time (SSRT), which captures the latency in inhibiting the response (Verbruggen et al., 2019). As this latency cannot be observed directly, it has to be estimated from the go-reaction times (RTs), stop signal delays (SSDs), and stop trial accuracy. Specifically, SSRTs are estimated by first integrating the RT distribution and finding the point at which the integral equals the probability of responding correctly after a stop signal. Next the mean SSD is subtracted from the result of this calculation to yield the SSRT. This method does not lend itself well to modelling at the trial level. To generate mixed models to estimate SSRTs, we therefore instead used an often implemented simplified method of calculating SSRTs, in which the SSRTs are calculated by subtracting the mean SSD from the mean RT. Predicted session scores from this model correlated highly with scores calculated by the integration method (*r* = .93; see supplementary materials).

***Working memory task.*** The primary outcome of the working memory task is the maximum level, participants reach within each condition. To create equivalent scores in a mixed model, we first imputed missing data from levels that participants did not reach by setting the accuracy for these levels to zero. Next, we predicted trial-level accuracy using binomial mixed models. Predicted session scores from these models correlated highly with traditionally calculated maximum level reached scores (*r*s > .99; see supplementary materials).

***Risk taking task.*** The primary outcome of the risk taking task is the percentage gambles for each condition. This score can easily be calculated with binomial mixed models that predict gambling at a trial level. Predicted session scores from these models correlated highly with traditionally aggregated scores (*r*s > .99; see supplementary materials).

***Information sampling task.*** The primary outcome of the information sampling task is average information oversampling in each session. Oversampling for each trial is calculated as the difference between turned around cards and the optimal number of card turns, derived from a normative dynamic programming model (Hunt et al., 2016). The dynamic programming model calculates optimal information sampling by calculating the expected value of every possible action (seeking more information vs guessing) for each step of the task (see supplementary materials; Hunt et al., 2016). We calculated model-derived scores by predicting oversampling using a mixed model. The predicted scores from this model correlated highly with traditionally aggregated scores (*r* > .99; see supplementary materials).

**Reliability.**

We assessed both split-half and test-retest reliability of each task. Split-half reliability was assessed based on Spearman-Brown-corrected correlations within each session (based on odd-even splits). Note that for the working memory task and for the SSRT, split-half reliabilities could not be computed because these tasks are adaptive. Therefore, splitting the task into two halves is not appropriate (Draheim et al., 2020). Test-retest reliability was calculated based on intra-class correlation coefficients (ICCs) based on data from the first two measurement sessions. To calculate ICCs directly from mixed models, we followed the method recently described by Brown et al. (2020), which calculates reliabilities based on variance components extracted from mixed models. Waltmann et al. (2021) recently showed that this method yields more conservative and more accurate reliabilities than alternative methods (e.g. first predicting sessions scores and calculating reliabilities based on these predictions).

To investigate whether increased retest-periods lead to decreased reliability, we also calculated test-retest reliabilities for longer retest-periods of one to six months.

**Validity.**

To validate the tasks, we correlated all task outcomes with each other and with several indicators of alcohol consumption. These included participants AUD scores, their self-reported retrospective alcohol consumption over the three months before the assessment (number of drinking days, standard drinks per typical drinking day, and standard drinks per day), and their average (mean and median) daily consumption reported during EMA in the first month after the assessment.

|  |  |
| --- | --- |
| Smartphone Task variable | Lab-based Validation tasks |
| Working memory capacity (WMC) | Digit Span Forward and Backward Task  N-Back |
| Stop signal reaction times (SSRTs) | Stop Signal Task  Stroop Task |
| Risk aversion for gains | Value-based decision making battery  Cambridge Gambling task |
| Risk aversion for losses | Value-based decision making battery |
| Loss aversion | Value-based decision making battery |
| Information oversampling relative to optimal | Value-based decision making battery |

**Table. Lab-based validation tasks.**

**Results**

**Split-half reliability**

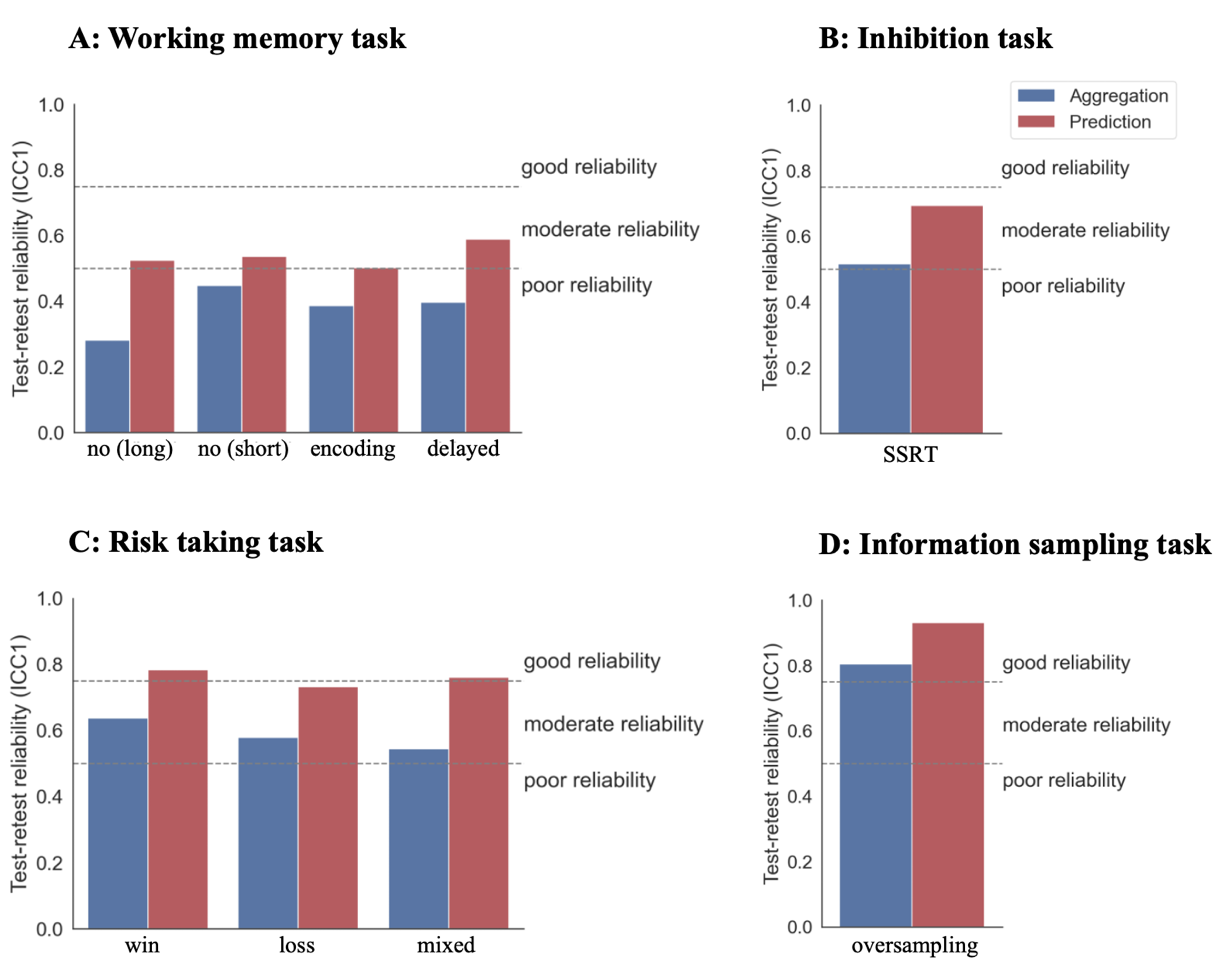
For the risk taking task only split-half reliabilities for the win and loss conditions were adequate (*rsb win session 1* = .84; *r\_sb\_win\_session\_2* = .91; *r\_sb\_loss\_session\_1* = .77; *r\_sb\_loss\_session\_2* = .82; *r\_sb\_mixed\_session\_1* = .67; *r\_sb\_mixed\_session\_2* = .71). For the information sampling task, split-half reliabilities were adequate (*r\_sb\_session\_1*  = .86; *r\_sb\_session\_2*  = .86). Split-half reliabilities for the working memory task and for the SST could not be calculated (see methods section).

**Test-retest reliability**

Test-retest reliability increased for all tasks, when calculating scores based on hierarchical mixed models compared to two-stage summary methods (aggregation). The stop signal task had moderate reliability when scores were calculated based on aggregation (ICC1 = .52; for all ICCs, see Table 2 and Fig. 2), but good reliability when scores were calculated based on mixed models (ICC1 = .69). The working memory task had poor reliability in all conditions when scores were calculated based on aggregation (ICC1s < .46), but reliability increased to moderate levels when scores were calculated based on mixed models (ICC1 ranging from .50 to .59). The risk taking task had moderate reliability in all conditions when scores were calculated based on aggregation (ICC1s ranging from .54 to .64). Reliability increased to moderate to good reliability when scores were calculated based on mixed models (ICC1s ranging from .73 to .76). Finally, the information sampling task sampling task had good reliability when scores were calculated based on aggregation (ICC1 = .80), which further improved when scores were calculated based on mixed models (ICC1 = .93).

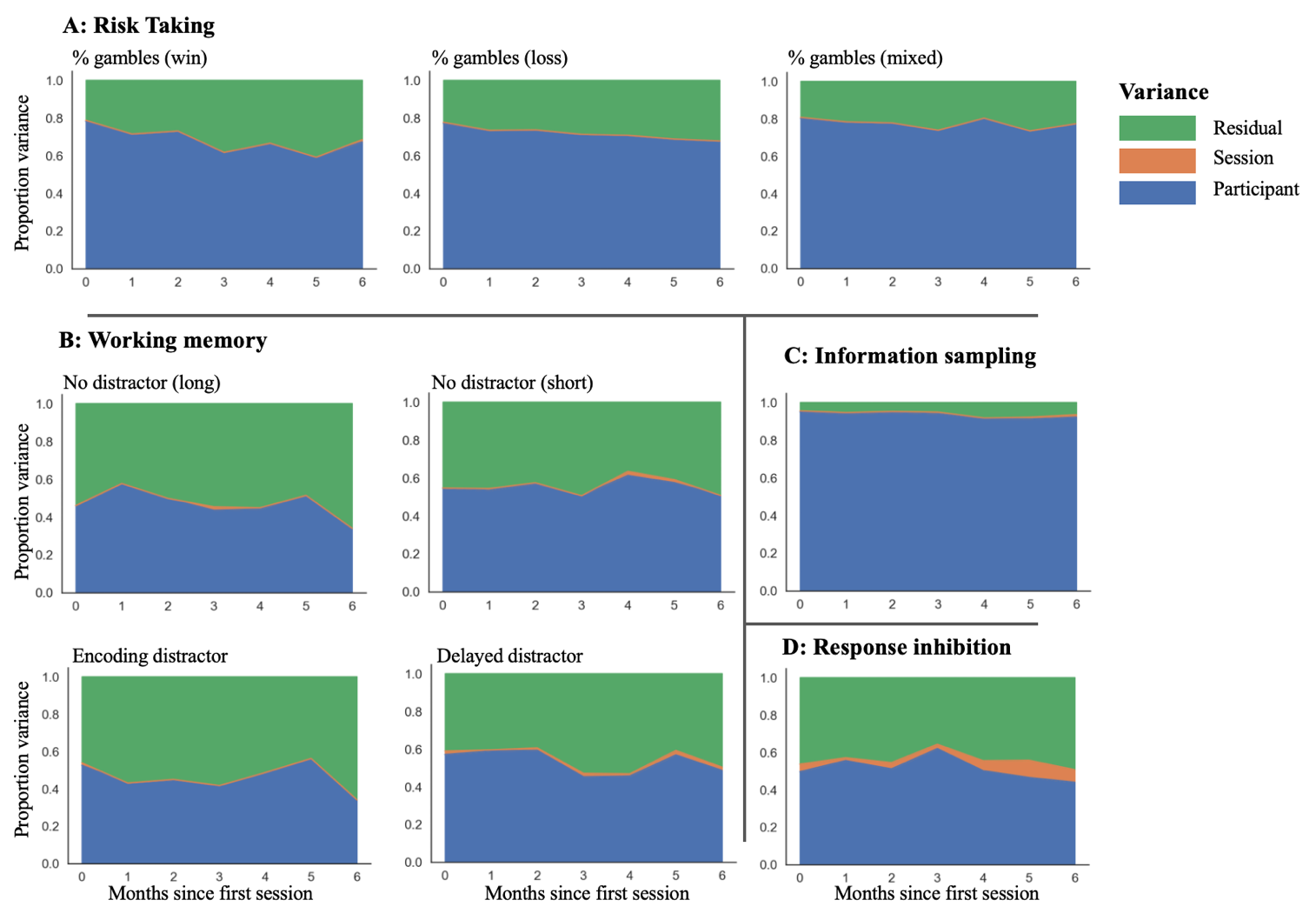
|  |  |  |  |
| --- | --- | --- | --- |
|  | Aggregation | | Prediction |
| Task measure | ICC(1) | ICC(2) | ICC(1) |
| **Response inhibition task** |  |  |  |
| SSRT | .52 | .53 | .69 |
| **Working memory task** |  |  |  |
| No distractor (long) | .28 | .28 | .52 |
| No distractor (short) | .45 | .45 | .54 |
| Encoding distractor | .39 | .39 | .50 |
| Delayed distractor | .40 | .40 | .59 |
| **Risk taking task** |  |  |  |
| Win | .64 | .64 | .78 |
| Loss | .58 | .59 | .73 |
| Mixed | .54 | .55 | .76 |
| **Information sampling task** |  |  |  |
| Oversampling | .80 | .81 | .93 |

**Table 2. Test-retest reliabilities by task measure and analysis approach.** This table shows test-retest reliabilities (ICCs) for the different task measures and analysis approaches. Note that for the prediction method, only ICC1s could be calculated.

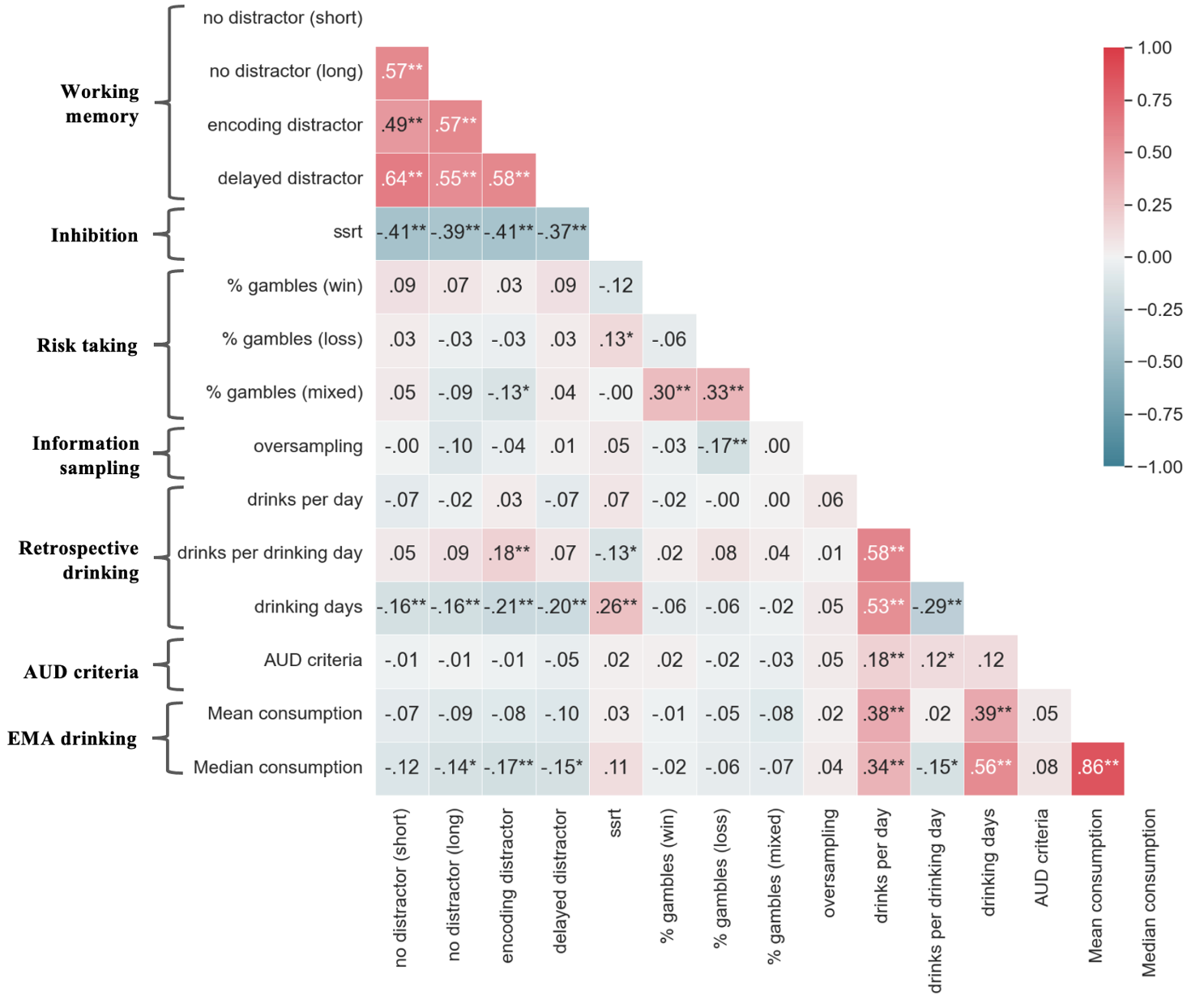
****

**Fig. 2.** Test-retest reliabilities (ICC1s) for the four tasks split by score calculation approach (blue: aggregation; red: prediction).

**Effect of retest period on test-retest reliability**

****

**Fig. 3.** Changes of variance components over time. This figure shows changes of variance components over time for the risk taking task (panel A), the working memory task (panel B), the information sampling task (panel D), and the response inhibition task (panel D). Blue areas reflect between-participant variance, orange areas within-participant, session variance, and green areas residual variance.

**Validity**

**Fig. 4.** Spearman correlations between different task outcomes and measures of alcohol consumption. This figure shows Spearman correlations between task measures and measures of drinking (\* = p < .05; \*\* = p < .01; \*\*\* = p < .001).

**References**

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), Selected Papers of Hirotugu Akaike (pp. 199–213). New York: Springer New York.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D. (2005). Fitting linear mixed models in R. *R News*, *5*, 27–30.

Beck, A., Wüstenberg, T., Genauck, A., Wrase, J., Schlagenhauf, F., Smolka, M. N., ... & Heinz, A. (2012). Effect of brain structure, brain function, and brain connectivity on relapse in alcohol-dependent patients. *Archives of General Psychiatry*, *69*, 842–852.

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan, R. J. (2014). Crowdsourcing for cognitive science–the utility of smartphones. *PloS one*, *9*(7), e100662.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2020). A toolbox approach to improving the measurement of attention control. Journal of *Experimental Psychology: General*. Advance online publication.

Ekhtiari, H., Victor, T. A., & Paulus, M. P. (2017). Aberrant decision-making and drug addiction—how strong is the evidence? *Current Opinion in Behavioral Sciences, 13*, 25–33.

GBD 2016 Alcohol Collaborators (2018). Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *392*(10152), 1015-1035.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91.

Heinz, A., Kiefer, F., Smolka, M. N., Endrass, T., Beste, C., Beck, A., ... & Spanagel, R. (2020). Addiction Research Consortium: Losing and regaining control over drug intake (ReCoDe)—From trajectories to mechanisms and interventions. *Addiction Biology*, *25*, e12866.

Hunt, L. T., Rutledge, R. B., Malalasekera, W. N., Kennerley, S. W., & Dolan, R. J. (2016). Approach-induced biases in human information sampling. *PLoS Biology*, *14*, e2000638.

Kievit, R. A., Brandmaier, A. M., Ziegler, G., Van Harmelen, A. L., de Mooij, S. M., Moutoussis, M., ... & Lindenberger, U. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117.

Konova, A. B., Lopez-Guzman, S., Urmanche, A., Ross, S., Louie, K., Rotrosen, J., & Glimcher, P. W. (2020). Computational markers of risky decision-making for identification of temporal windows of vulnerability to opioid use in a real-world clinical setting. *JAMA Psychiatry, 77*, 368–377.

Koob, G. F. & Volkow, N. D. (2010). Neurocircuity of addiction. *Neuropsychopharmacology, 25,* 217–238.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... & 23andMe Research Team. (2018). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature Genetics*, 1112.

Liu, M., Jiang, Y., Wedow, R. *et al.* (2019) Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*,237–244.

McNab, F., Zeidman, P., Rutledge, R. B., Smittenaar, P., Brown, H. R., Adams, R. A., & Dolan, R. J. (2015). Age-related changes in working memory and the ability to ignore distraction. *Proceedings of the National Academy of Sciences*, *112*, 6515–6518.

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 638–645.

Perrez, M., Schoebi, D., & Wilhelm, P. (2000). How to assess social regulation of stress and emotions in daily family life? A computer‐assisted family self‐monitoring system (FASEM‐C). *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice, 7*, 326–339.

Powell, D. J., McMinn, D., & Allan, J. L. (2017). Does real time variability in inhibitory control drive snacking behavior? An intensive longitudinal study. *Health Psychology*, *36*, 356–365.Reichert, M., Gan, G., Renz, M., Braun, U., Brüßler, S., Timm, I., ... & Meyer-Lindenberg, A. (2021). Ambulatory assessment for precision psychiatry: Foundations, current developments and future avenues. *Experimental Neurology*, *345*, 113807.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., ... & Posthuma, D. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, 912-919.

Smittenaar, P., Rutledge, R. B., Zeidman, P., Adams, R. A., Brown, H., Lewis, G., & Dolan, R. J. (2015). Proactive and reactive response inhibition across the lifespan. *PLoS One*, *10*, e0140383.

Spanagel R., Noori H. R., Heilig M. (2014). Stress and alcohol interactions: Animal studies and clinical significance. *Trends in Neurosciences*,37, 219–227.

Sobell, L. C., Ellingstad, T. P., & Sobell, M. B. (2000). Natural recovery from alcohol and drug problems: Methodological review of the research with suggestions for future directions. *Addiction, 95*, 749–764.

Stavro, K., Pelletier, J., & Potvin, S. (2013). Widespread and sustained cognitive deficits in alcoholism: a meta‐analysis. *Addiction Biology, 18*, 203–213.

Verbruggen, F., Aron, A. R., Band, G. P., Beste, C., Bissett, P. G., Brockett, A. T., ... & Boehler, C. N. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *eLIFE*, *8*, e46323.

World Health Organization (2019). *Global Status Report on Alcohol and Health 2018*. World Health Organization.

**Appendix A – Drinks in drinking questions**

**Drinks listed in retrospective questions**

Biermischgetränk (übliches Glas), Bier (kleines Glas), Bier (übliches Glas), Starkbier (kleines Glas), Starkbier (übliches Glas), Weinmischgetränk (kleines Glas), Weinmischgetränk (übliches Glas), Weinmischgetränk (übliches Flasche), Wein, Sekt/Perlwein (kleines Glas), Wein, Sekt/Perlwein (übliches Glas), Wein, Sekt/Perlwein (übliches Flasche), Aufgespritzter Wein, Sherry (kleines Glas), Aufgespritzter Wein, Sherry (übliches Glas), Aufgespritzter Wein, Sherry (übliches Flasche), Süßer Likör (kleines Glas), Süßer Likör (übliches Glas), Süßer Likör (übliches Flasche), Schnaps (kleines Glas), Schnaps (doppelter Shot/Whisky-Glas), Schnaps (übliches Flasche), Shazam

**Drinks listed in ema questions**

Kleines Bier, Mittleres Bier, Großes Bier, Kleiner Weißwein, Mittlerer Weißwein, Flasche Weißwein, Kleiner Rotwein, Mittlerer Rotwein, Flasche Rotwein, Sekt, Flasche Sekt, Likör-Wein , Kleiner Likör , Großer Likör , Likör süß , Kleine Spirituose , Große Spirituose , Spirituose , Kleine Spirituose