**Measuring self-regulation in the field: Reliability and Validity of smartphone-based cognitive-motivational experiments in alcohol use order**

**Authors**: Hilmar Zech\* a, b, Maria Waltmannb, c, Ying Leea , Markus Reichertd, e, f, Ulrich Ebner-Priemere, f, Michael Rappg, Andreas Heinzh, Ray Dolani, j, Michael N. Smolkaa, Lorenz Desernoa,b,c

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

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

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

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

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

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

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

\*Paste corresponding author name(s) here.

**Email:**  [hilmar.zech@tu-dresden.de](mailto:hilmar.zech@tu-dresden.de) / [deserno\_l@uwk.de](mailto:deserno_l@uwk.de)

**Author Contributions:** Paste the author contributions here.

**Competing Interest Statement:** Disclose any competing interests here.

**Classification:** Paste the major and minor classification here. Dual classifications are permitted, but cannot be within the same major classification.

**Keywords:** Paste the keywords here. There should be at least three and no more than five.

**Abstract**

Experimental measures of self-regulation are thought for long to play an integral role in our scientific understanding of loss of control behaviors, for example in addiction. Yet, experimental task measures cannot be easily deployed where addictive behaviors actually occur—in people’s real lives outside of the laboratory. Moreover, self-regulation tasks have recently come under criticism for their poor test–retest reliability and construct validity. Together, these two disadvantages of existing tasks might explain why ecological validity of task-based self-regulation measures is low. To overcome this problem, we assessed the reliability and construct validity of four smartphone-based tasks designed to measure different aspects of self-regulation by drawing upon measures of cognitive control and decision-making in a large (N=300) sample of psychiatric patients with alcohol use disorder and associated comorbidities. We show that the four tasks have moderate to good reliability, which can be further improved to good to excellent levels if data from separate measurements sessions was modeled jointly. We show that the data is described for by four factors, which cover one cognitive control factor separately from two-three distinct aspects of decision-making. This work indicates that self-regulation can be measured experimentally with sufficient reliability and validity in addiction. This represents a critical milestones towards longitudinal experimental studies in the field in addiction research but also in psychiatry and psychology more generally.

**Introduction**

The ability to control behavior in service of long term-goals (self-regulation) is related to a range of important life outcomes, including wealth, health, and public safety (Moffit et al, 2011). Given its importance for personal and societal outcomes, interest in self-regulation has increased across research domains, including economics, psychology, and cognitive neuroscience. Different research domains have used different methodologies to study the topic, ranging from survey measures, to neurological measures and experimental tasks (Eisenberg et al., 2019).

Experimental tasks play an especially important role in researching self-regulation. Self-regulation is an umbrella term encompassing a range of distinct capacities and tasks excel at isolating distinct processes that cannot be easily tapped by survey measures (Falk and Heckman, 2009). Consequently, tasks have already been successfully used to capture distinct self-regulation capacities, including working memory, inhibition, and adaptive decision making (ref). Moreover, tasks can be modelled using rigorous mathematical models (Haines et al., 2021) and resulting outcomes can be connected to neurological markers. This effort could ultimately lead to a better understanding of self-regulation failures, and lead to targeted treatments of conditions, such as addictions, which are characterized by deficient self-regulation.

Yet, although tasks excel at rigorously measuring distinct, latent capacities underlying self-regulation, research linking task measures to real-life outcomes has so far been largely unsuccessful. For example, in a recent study, Eisenberg et al. (2019) assessed the validity of 22 surveys and 37 task measures of self-regulation and found that while surveys modestly predict real-world outcomes related to self-regulation, tasks show largely no relationship to real-world outcomes. This finding mirrors a more widely held opinion that experimental tasks lack “realism” and generalizability (Falk and Heckman, 2009).

Here, we argue that tasks do not inherently lack real-world relevance, but rather that methodological shortcomings in how tasks are traditionally used can explain their perceived lack of realism. Specifically, it has recently been shown that task measures largely lack reliability (Enkavi et al., 2019; Hedge et al., 2018). Reliability is an important psychometric quality, which refers to the stability with which a task can measure between-participant differences. It is especially important when relating a measurement to other measurements (e.g., self-regulation to addiction)—because (low) reliability limits the possible observed correlation between two outcomes (Spearman, 1904). For example, should working memory and addiction severity *actually* have a high correlation of .8, but we would only measure working memory with a low reliability of .3 (reliability ranges from 0 to 1), the observed correlation between the two measures would decrease to .44. As most traditional measures of self-regulation have been shown to suffer from low reliability, it is possible that the failure to link task measures. to real-life outcomes is not explained by an actual lack of a relationship, but by the task’s low reliability.

An additional disadvantage of experimental tasks as opposed to surveys is that it is difficult to deploy tasks where self-regulation actually matters—outside the laboratory in people’s real lives. Most tasks were designed to run on laboratory computers, requiring specialized software and sometimes specialized hardware, that most people do not have at home. This makes it difficult to link task measures to acute self-regulation failures, which happen rarely (e.g., binge drinking) and usually in specific environments outside the laboratory (e.g., in the bar; Zech et al., 2022, accepted). Moreover, unlike survey measures, task measures cannot measure self-regulation capacities retrospectively which makes it even more difficult to link task measures to real-life events. Finally, running tasks in the laboratory restricts researchers access to large samples and to people for whom self-regulation failures especially matter, for example people suffering from addictions.

Here, we show that all three problems of tasks can be overcome by moving task measures to mobile platforms such as smartphones and modelling task data in a way that makes use of longitudinal data generated by these tasks. We tested the reliability and construct validity of four smartphone-based tasks in a large (*N* = 600) sample of participants suffering from mild to moderate alcohol use disorder. The tasks were designed to capture core components of self-regulation. We show that—in our sample—task reliability was moderate to good and tasks had expected construct validity. We further show that making use of the unique longitudinal data generated by smartphone-based tasks by modelling data from different sessions jointly further improves the tasks reliability to good to excellent levels. Together our results demonstrate that smartphone-based tasks can overcome the weaknesses of existing self-regulation tasks, potentially opening the road to more ecologically valid task-based research and exciting future task-based interventions.

**Results**

***Reliability.***For this manuscript format (Intro – Results etc.), you need to write section that introduce the main characteristics of the design and methods (without any details), really only a few sentence – but this is really necessary / crucial – otherwise no one can read and understand the following….

*Split-half reliability.* For the RTT, split-half reliabilities for the win and loss gambles were adequate (*rsb win session 1* = .84; *rsb\_win session 2* = .91; *rsb loss session1* = .77; *r\_sb\_loss\_session\_2* = .82) but lower for mixed gambles (*rsb\_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 analyzed due reasons inherent to their adaptive task design (see methods section).

*Test-retest reliability.*Test-retest reliability increased for all tasks, when calculating scores based on the prediction approach using hierarchical mixed models compared to the aggregation (two-stage summary) approach (see Table 1 and Figure 1).

The SST had moderate reliability when scores were calculated based on aggregation (ICC1 = .52; for all ICCs, but good reliability when scores were calculated based on mixed models (ICC1 = .69).

The WMT 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 RTT 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 IST 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).

***Construct validity.*** Factor analyses on average scores from both baseline sessions indicated that either a three or four factor solution best described the data.

**Discussion**

Paste your discussion here.

**Materials and Methods**

**General procedure.** This study was part of a larger German research consortium on addiction at three sites (Technical University Dresden, Charité Berlin, and Central Institute of Health Mannheim), in which a smartphone-based longitudinal Ecological Momentary Assessment (EMA) of up to one year was performed with a range of subjective reports. In addition to subjective reports, patients performed four cognitive-motivational tasks on the smartphone once per month. These tasks were taken from the Great Brain Experiment (GBE) app (Brown et al., 2014, see below for details). Before starting the EMA study, patients underwent extensive clinical and neurocognitive assessments (see Heinz et al., 2020). During this assessment 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. On the first Monday following the assessment, participants completed each smartphone task twice. Participants also participated in multiple subprojects of the consortium (see Heinz et al., 2020), which are not subject to the present study.

**Participants.** 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. For study inclusion at all three sites, participants had to fulfill criteria of mild to moderate Alcohol Use Disorder (AUD). According to DSM V, 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 before study inclusion/exclusion. 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. As pre-registered (Zech et al., 2021), data from 300 participants was analyzed for the present study. Participants age ranged from 17 to 65 years (*M* = 37.9, *SD* = 13.0) and 113 participants (37.7%) reported to be female. Participants AUD criteria ranged from 2 to 9 (*M* = 3.93, *SD* = 1.51).

**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.

**Reliability.** The first goal of this study was to assess the smartpone 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 more traditional *aggregation* approach, and an alternative *prediction* approach, which has recently been shown to improve reliability in other cognitive tasks (for details, see below; Brown, 2020; Haines, 2021; 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, can attenuate 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 reliable scores (Efron & Morris, 1977; Gelman, 2006; Williams et al., 2020; as cited in Haines et al., 2021).

The alternative 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 all available trial-level data. This allows it to carry, firstly, within-session uncertainty into the inference step and, secondly, to use information from other participants and sessions in each individual session score estimation. Both of these aspect improved test-retest reliability in previous work. We implemented this approach using hierarchical mixed models specifically designed to model each task’s outcome measure (for details see supplementary materials). Hierarchical mixed models allow to analyze data at the trial-level while still accounting for the participant and session structure of the data. We validated that mixed model based scores did not substantially differ from aggregation-based scores when modelled for each session separately (see supplementary materials).

***Reliability assessment.*** Firstly, 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). Qualitative interpretations of split-half reliabilities are given in line with Nunnally and Bernstein (1994; split-half reliabilities above .8 were labelled as adequate). Secondly, 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. Qualitative interpretations of test-retest reliabilities are given in accordance with Koo and Li (2016): ICCs less than .5 were being interpreted as “poor”, ICCs between .5 and .75 as “moderate”, ICCs between .75 and .9 as “good”, and ICCs above .9 as “excellent”.

**Factor analysis**.

Please continue

**Acknowledgments**

Paste your acknowledgments here.

**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.

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the reliability of computational analyses: Model-based planning and its relationship with compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*, 601–609.

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.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163.

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

Kuitunen‐Paul, S., Rehm, J., Lachenmeier, D. W., Kadrić, F., Kuitunen, P. T., Wittchen, H. U., & Manthey, J. (2017). Assessment of alcoholic standard drinks using the Munich composite international diagnostic interview (M‐CIDI): An evaluation and subsequent revision. *International Journal of Methods in Psychiatric Research*, *26*, e1563.

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.

Nunnally, J., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

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.

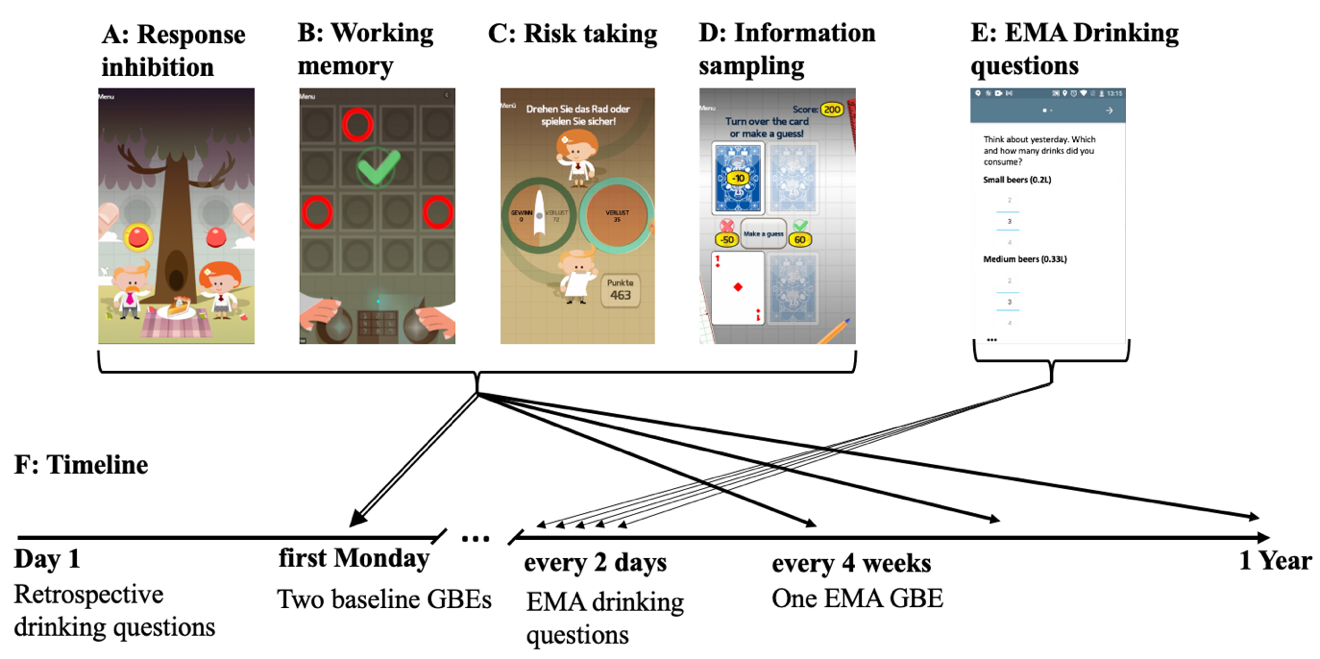
Sobell, L. C., Ellingstad, T. P., & Sobell, M. B. (2000). Natural recovery from alcohol and drug problems: Methodological review of the research with suggestions for future directions. *Addiction, 95*, 749–764.

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

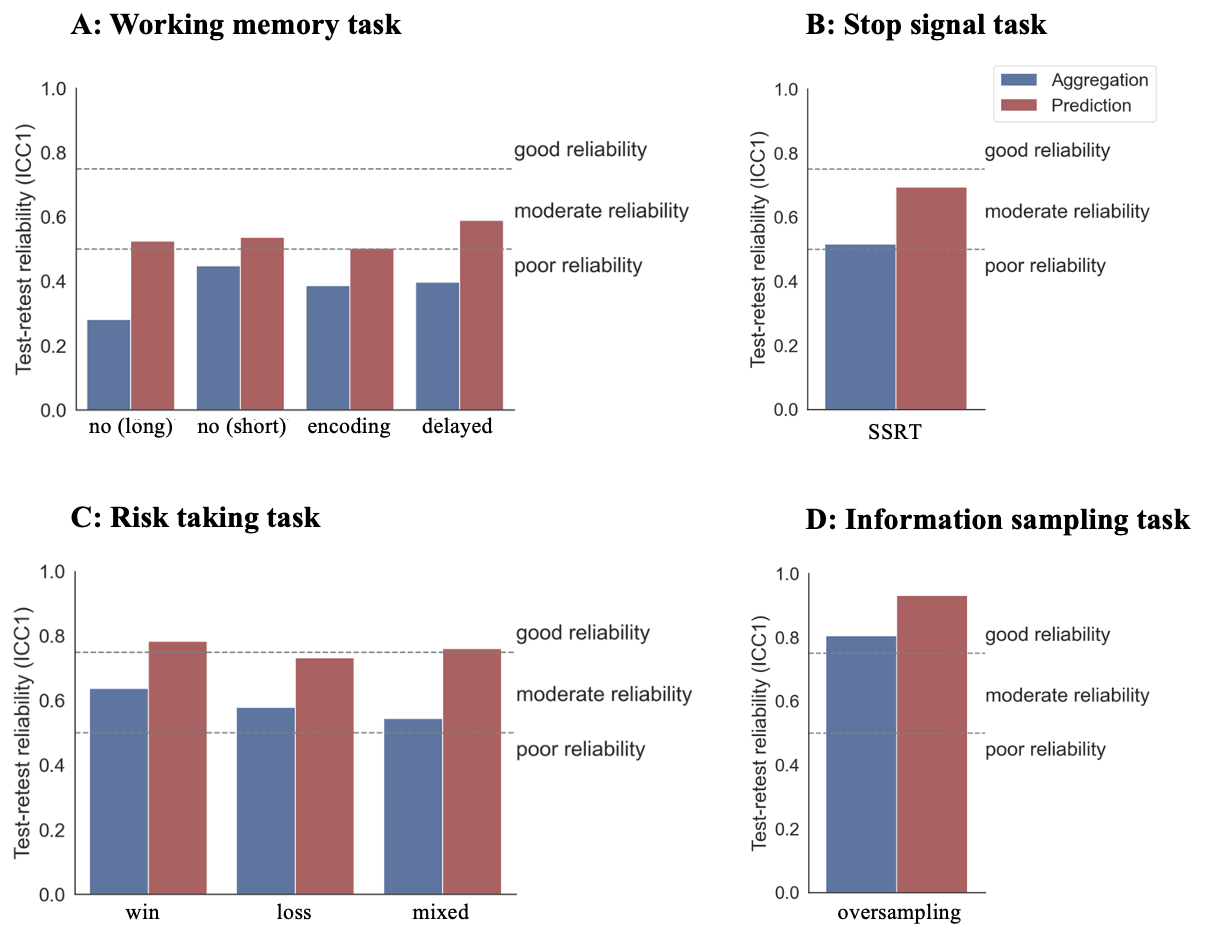
Verbruggen, F., Aron, A. R., Band, G. P., Beste, C., Bissett, P. G., Brockett, A. T., ... & Boehler, C. N. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *eLIFE*, *8*, e46323.

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

**Figures and Tables**

**

**Figure 1.** Illustration of study timeline and smartphone-based tasks from a customized version of the Great Brain Experiment (GBE) app used in this study. All tasks were translated to German. Panel A shows the stop signal task, Panel B the working memory task, Panel C the risk taking task, Panel D shows the information sampling task, Panel E the EMA drinking questions, and Panel F the study timeline

****

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

**Figure 3.** Factor loadings and clusters

**Table 1.** Test-retest reliabilities (ICCs) for the different task measures and analysis approaches. Note that for the prediction method, only ICC1s could be calculated.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |