Internal consistency and test–retest reliability of the P3 event-related potential (ERP) elicited by alcoholic and non-alcoholic beverage pictures

Roberto U. Cofresí1, Thomas M. Piasecki1, Greg Hajcak2, Bruce D. Bartholow1

1Department of Psychological Sciences, University of Missouri, Columbia, Missouri, USA
2Departments of Psychology and Biomedical Sciences, Florida State University, Tallahassee, Florida, USA

Correspondence
Roberto U. Cofresí, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211, USA. Email: cofresir@missouri.edu

Funding information
Funding for the project and manuscript preparation was provided by NIH grant R01 AA025451 (BDB, TMP). At various stages of his involvement, RUC was supported by the local T32 (AA013526), an NIH Diversity Supplement to the parent R01 (AA025451-04S1), and the University of Missouri Department of Psychological Sciences Mission Enhancement Post-Doctoral Fellowship Fund. Funding agencies had no role in the research or manuscript preparation.

Abstract
Addiction researchers are interested in the ability of neural signals, like the P3 component of the ERP, to index individual differences in liability factors like motivational reactivity to alcohol/drug cues. The reliability of these measures directly impacts their ability to index individual differences, yet little attention has been paid to their psychometric properties. The present study fills this gap by examining within-session internal consistency reliability (ICR) and between-session test–retest reliability (TRR) of the P3 amplitude elicited by images of alcoholic beverages (Alcohol Cue P3) and non-alcoholic drinks (NADrink Cue P3) as well as the difference between them, which isolates alcohol cue-specific reactivity in the P3 (ACR-P3). Analyses drew on data from a large sample of alcohol-experienced emerging adults (session 1 N = 211, 55% female, aged 18–20 yr; session 2 N = 98, 66% female, aged 19–21 yr). Evaluated against domain-general thresholds, ICR was excellent ($M \pm SD; r = 0.902 \pm 0.030$) and TRR was fair ($r = 0.706 \pm 0.020$) for Alcohol Cue P3 and NADrink Cue P3, whereas for ACR-P3, ICR and TRR were poor ($r = 0.370 \pm 0.071; r = 0.201 \pm 0.042$). These findings indicate that individual differences in the P3 elicited by cues for ingested liquid rewards are highly reliable and substantially stable over 8–10 months. Individual differences in alcohol cue-specific P3 reactivity were less reliable and less stable.

The conditions under which alcohol/drug cue-specific reactivity in neural signals is adequately reliable and stable remain to be discovered.

Keywords
addiction, alcohol/alcoholism, ERPs, individual differences, P300/LPP

1 INTRODUCTION

There has long been interest among addiction researchers in the ability of event-related potentials (ERPs) to index individual differences in addiction liability factors (Kamarajan & Porjesz, 2015; Kinreich et al., 2021; Rangaswamy & Porjesz, 2014). One of the most common ERP-based measures of addiction risk is P3 amplitude reduction (P3-AR) observed during various cognitive tasks, particularly the “rotated heads” mental rotation oddball task (Begleiter et al., 1984; Iacono et al., 2002). Since its discovery in the context of alcoholism risk (see early review in Porjesz & Begleiter, 1981), a large body of evidence has established that individual differences in the P3-AR reflect a genetically based, heritable endophenotypic vulnerability for externalizing behavior and disorders, including...
excessive substance use (Carlson et al., 2007; Gilmore et al., 2010; Iacono et al., 2002, 2003; Patrick et al., 2006); for meta-analytic reviews, see: Euser et al. (2012), Gao and Raine (2009), Hamidovic and Wang (2019).

Recently, there has been growing interest in the possibility that enhancement of various ERP components elicited by alcohol and drug-related cues could index risk that is more specific to alcohol and drug use. In particular, researchers have focused on neurocognitive processes related to the salience of alcohol and drug-related cues (Littet al., 2012), such as selective attention (e.g., Dickter et al., 2014; Krocce et al., 2018; Petit et al., 2012; Shin et al., 2010) and incentive-motivational value (e.g., Dewese et al., 2018; Dunning et al., 2011; Fleming et al., 2021; Garland et al., 2019; Minnix et al., 2013; Piasecki et al., 2017). Of special interest are the P3 and LPP components, established over half a century of work in experimental psychophysiology as indicators of extrinsic and intrinsic incentive-motivational value attributed to the eliciting stimulus (e.g., Begleiter et al., 1983; Codispoti et al., 2021; Dewese et al., 2016; Franken et al., 2011; Schindler & Straube, 2020; Schupp et al., 2000; for review, see: Hajcak & Foti, 2020).

In particular, enhanced P3/LPP response to alcohol-related relative to non-alcohol cues (alcohol cue reactivity P3/LPP; henceforth: the ACR-P3) has been posited as an indicator of individual differences in the attribution of incentive-motivational value, an aspect of emotional significance, to alcohol-related cues (e.g., Fleming et al., 2021; Herrmann et al., 2001; Krocce et al., 2018). Enhanced ACR-P3 is associated with heavier and more hazardous alcohol use (Herrmann et al., 2001; Krocce et al., 2018; Petit et al., 2013) as well as lower self-reported sensitivity to the acute effects of alcohol (Bartholow et al., 2007, 2010), especially its sedative-like effects (Martins et al., 2019). Lower sensitivity to alcohol itself is associated with heavier and more hazardous alcohol use including use-related negative consequences and alcohol use disorder (AUD) symptoms (Bailey & Bartholow, 2016; Bartholow et al., 2007, 2010; Davis et al., 2021; Fleming & Bartholow, 2014; Fleming et al., 2021; Hone et al., 2017; Martins et al., 2019; Trela et al., 2016), providing converging evidence for the association between enhanced ACR-P3 and AUD risk. Enhanced ACR-P3 also predicts heavier alcohol use prospectively (Bartholow et al., 2007) and differentiates individuals with AUD from those without (Namkoong et al., 2004).

Despite growing interest in the association of the ACR-P3 with heightened risk for alcohol misuse and addiction, its measurement reliability has not been examined. Yet, a measure cannot be valid if it is not reliable (Cronbach & Meehl, 1955; Kline, 1998; Nunnally & Bernstein, 1994). Reliability captures the level of consistency or stability of a measure and is quantified in terms of internal consistency and test–retest reliability (TRR) (Kline, 1998). Internal consistency reliability (ICR) refers to consistency or stability within an assessment (e.g., similarity between scores from different subsets of trials), and is sensitive to random error variance plus error variance unique to different trials (e.g., fatigue effects on later trials in an assessment). TRR refers to consistency or stability between assessment sessions, and is sensitive to random error variance plus unique error variance influencing the different assessments (e.g., factors shared by all trials within an assessment). TRR can depend on person characteristics (e.g., age, sex, gender, education, ability, effort), the amount of time between assessments, and contextual differences between assessments (e.g., affective differences, practice or carryover effects, developmental stage effects).

Lack of attention to these basic psychometric issues is a growing concern for individual differences neuroscience (see Baldwin, 2017; Clayson et al., 2019; Clayson & Miller, 2017; Hajcak et al., 2017; Hajcak et al., 2015; Herting et al., 2018; Infantolino et al., 2018; Patrick et al., 2019; Thigpen et al., 2017). Researchers often assume that if a given measure has shown robust within-person effects across multiple studies, then it must be reliable—and, therefore, can function well as an index of individual differences (see Hajcak et al., 2017; Infantolino et al., 2018). This is a highly problematic assumption insofar as measures can produce robust within-person effects, but fail to reliably differentiate individuals, either because of poor ICR or failure to capture true score variability differences across individuals.

In general, poor reliability impacts not only the ability of neural measures to index individual differences, but also the generalizability and reproducibility of findings using specific neural measures (see Baldwin, 2017). Poor reliability can affect both the magnitude and direction (sign) of the observed association between any two measures (Gelman & Carlin, 2014). Poor reliability also can limit the statistical power to detect between-subject effects (e.g., high vs. low risk group differences, between-subject experimental manipulation effects) (Hajcak et al., 2017; Humphreys, 1993; Kanyongo et al., 2007; Williams et al., 1995).

Consequently, the current study had two goals. Its primary goal was to examine reliability of the P3 response to images of alcoholic beverages (Alcohol Cue P3) and the P3 response to images of non-alcoholic drink cues (NADrink Cue P3) as well as the difference between these

---

1In fact, the maximum possible magnitude of the observed association between any two measures is defined by the square root of the product of their reliabilities (Baugh, 2002; Kline, 1998; Nunnally & Bernstein, 1994).
P3 responses (i.e., the ACR-P3). An important secondary goal was to estimate the minimum number of artifact-free trials required for reliable measurement, and whether more reliable scores are obtained from single electrodes or averaging across the electrode cluster over which the P3 is maximal.

2 | METHOD

2.1 | Participants

Data in this report are taken from two laboratory sessions completed as part of a large, ongoing longitudinal study. Potential participants were recruited from the community (see Supplemental Information for recruitment strategies) to complete an eligibility screening survey via REDCap (Harris et al., 2009). Of the 1220 individuals who had completed the screener, 882 were determined to be eligible and invited to enroll in the study (see Supplemental Information for inclusion-exclusion criteria). Of these individuals, 211 had completed the first laboratory session and 98 had completed the second laboratory session. See Table 1 for participants’ sociodemographic characteristics.

2.2 | Materials

2.2.1 | Picture-viewing task

Participants completed a picture-viewing task similar in structure to tasks in our previous studies (Bartholow et al., 2007, 2010, 2018; Martins et al., 2019). There were 400 picture presentations: 80% consisted of non-beverage neutral pictures (Neutral) and 20% of trials consisted of beverage pictures (10% alcoholic beverage [Alcohol], 10% non-alcoholic drink [NADrink]). Neutral pictures were drawn from the Internal Affective Picture System (IAPS) (Lang et al., 2008) and represented images rated as low in arousal and near the scale midpoint in valence. Alcohol and NADrink pictures were drawn from the “passive” subset (displaying only the bottle and/or empty/full glass on a bland white background) of the Amsterdam Beverage Picture Set (ABPS) (Pronk et al., 2015), and supplemented with pictures of four alcoholic beverages taken by a local professional photographer (based on pretest data indicating favored alcoholic drinks among the population from which the sample was drawn; pictures displayed only the

<table>
<thead>
<tr>
<th>Table 1 Participant characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1 (N = 211)</td>
</tr>
<tr>
<td><strong>M (SD)</strong></td>
</tr>
<tr>
<td><strong>Age, yr</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
</tr>
<tr>
<td><strong>Hispanic</strong></td>
</tr>
<tr>
<td><strong>Race</strong></td>
</tr>
<tr>
<td><strong>American Indian/Alaskan Native</strong></td>
</tr>
<tr>
<td><strong>Native Hawaiian/Pacific Islander</strong></td>
</tr>
<tr>
<td><strong>Asian</strong></td>
</tr>
<tr>
<td><strong>Black</strong></td>
</tr>
<tr>
<td><strong>White</strong></td>
</tr>
<tr>
<td><strong>Multiple selected</strong></td>
</tr>
<tr>
<td><strong>None selected</strong></td>
</tr>
<tr>
<td><strong>Handedness</strong></td>
</tr>
<tr>
<td><strong>Right dominant</strong></td>
</tr>
<tr>
<td><strong>Undergraduate student</strong></td>
</tr>
</tbody>
</table>

Note: Demographic information was collected at screening.

Due to the global COVID19 pandemic, we were unable to conduct laboratory sessions between 03/15/2020 and 08/14/2020. Since 08/14/2020, data collection has been severely limited due to University of Missouri policies meant to mitigate the spread of COVID19. Therefore, and given that power calculations indicated that the sample is large enough for current purposes (i.e., using G∗Power 3.1, we determined that for 80% power to detect |r| ≤ 0.10 using a two-sided t-test against the null hypothesis that |r| = 0 with 5% Type 1 error, we would need N ≥ 779, but we would need only N = 191 to detect |r| = 0.20, N = 120 to detect |r| = 0.25, N = 82 to detect |r| = 0.30, N = 44 to detect |r| = 0.40, and N = 26 to detect |r| = 0.50, which means that the current N for either session 1 or 2 was sufficient to detect medium and large associations, and that the current N for session 1 is sufficient to detect small-to-medium, medium, and large associations), we decided to conduct the current analyses based on the sample as of 03/21/2021.

Note: Due to the global COVID19 pandemic, we were unable to conduct laboratory sessions between 03/15/2020 and 08/14/2020. Since 08/14/2020, data collection has been severely limited due to University of Missouri policies meant to mitigate the spread of COVID19.
Participants were instructed to press one button whenever they saw an alcoholic beverage and a different button whenever they saw a non-alcoholic beverage. Other technical details are presented in Supplemental Information.

2.2.2 EEG acquisition

The electroencephalogram (EEG) was recorded from 32 sintered Ag/AgCl ring electrodes (10–20 system layout) embedded in an elastic fabric cap with adjustable chinstraps (BrainCap; EASYCAP, LLC, Herrshing, Germany). Electrodes were filled with Abralyt HiCl (EASYCAP, LLC) using plastic syringes (and blunt tip needles when hair was thick). Electrodes with impedances ≤10 kΩ were accepted for recording. Impedances were monitored across tasks and adjusted as needed. Data were acquired with a Grael v2 EEG amplifier and Curry 8 EEG acquisition software (both from Compumedics Neuroscan, LLC, Charlotte, NC). The EEG was sampled at 512 Hz and referenced to the right mastoid channel (M2) online; a ground electrode was placed at FPz. The Grael v2 amplifier hardware contains a DC-coupled high-pass filter and applies a 3 dB anti-aliasing low-pass filter online (effective recording bandwidth at 512 Hz sampling rate = 0 to 143 Hz).

2.2.3 EEG preprocessing

After acquisition, each participant’s data underwent a standardized offline pre-processing pipeline implemented in EEGlab (Delorme & Makeig, 2004) and ERPlab (Lopez-Calderon & Luck, 2014). The beginning (including instructions and practice trials), the break between task blocks 1 and 2, and the end of the continuous EEG recording were removed. EEG data were then re-referenced to an average of the two mastoids and resampled at 256 Hz. DC bias was removed. An Infinite Impulse Response (IIR) Butterworth bandpass filter was applied (half-amplitude cutoffs: 0.1–30 Hz; filter order: 2; filter roll-off: 12 dB/oct). Sinusoidal noise (e.g., AC power line fluctuations, fluorescent lighting hum) was attenuated using the CleanLine plug-in for EEGlab (Mullen, 2012). Using session notes and the CleanLine plug-in, “bad” (e.g., excessively noisy) electrodes were identified and removed. Independent components analysis (ICA) was conducted on continuous EEG data from the remaining electrodes. The ADJUST plug-in for EEGlab (Mignon et al., 2011) was used to identify and remove ICs corresponding to blinks and eye movements as well as other artifacts (e.g., EKG). After removal of artifact ICs (Median ± IQR number of artifact ICs removed per participant in session 1 or 2.5 ± 4), previously “bad” electrodes were interpolated using the spherical spline method in EEGlab (Median ± IQR number of electrodes interpolated per participant in session 1 or 2.1 ± 2). Next, EEG data at every electrode were segmented into stimulus-locked epochs (∼100 to 1000 ms). Epochs on which an incorrect response was registered were discarded (M ± SD% of all epochs per participant in session 1 or 2.2.12 ± 3.34). Finally, moving peak-to-peak thresholds (±75 µV, window: 100 ms, step: 50 ms) and point-to-point difference thresholds (±20 µV) were applied to identify artefactual voltage deflections at any electrode for a given epoch. Table 2 provides the number of retained epochs per participant for each picture type by lab session. The processed epoch (trial) × electrode × time (ms) × picture-type data, that is, the single-trial ERPs, for each person were then exported for P3 scoring, visualization, and analysis in R version 3.6.0 using the base library (R Core Team, 2019) and the following packages: erpR (Arcara & Petrova, 2014), ggplot2 (Wickham, 2009), and psych (Revelle, 2018). Table 2 also provides the standardized measurement error (SME) for the P3 scores.

Of the 211 participants who completed the first lab session, ERPs were derived for 210 (1 participant’s continuous EEG data could not be segmented due to equipment malfunction). Of the 98 participants who completed the second lab session, ERPs were derived for 97 (one participant’s continuous EEG data could not be segmented due to equipment malfunction).

2.2.4 P3 scoring

For each picture type, per participant, the time-window mean amplitude of the P3 component was measured from all retained epochs on all available single electrodes as well as on the averaged signal across nine electrodes over the scalp region where the component was maximal. Additionally, following (Luck et al., 2021), the standardized measurement error (SME) for time-window mean amplitude was computed for each picture type per participant. The post-stimulus time-window for P3 amplitude measurement and the scalp region over which the P3

---

3Budweiser can, Coors Light can, Natural Light can, and Jack Daniel’s bottle alongside a filled shot glass.

4Inclusion of categorization error trials had little to no effect on P3 mean amplitudes or their psychometric properties, but that may have been due to the very low rate of errors in beverage categorization in the present dataset. For the same reason, there was no need to exclude participants based on excessive categorization errors. Given that, broadly speaking, erroneous response trials differ from correct response trials in fundamental ways, including neurocognitive determinants, we recommend discarding categorization error trials.
was maximal were determined by visual inspection of the data from both sessions separately, following a collapsed localizer approach (Luck, 2014). Specifically, the waveform for a collapsed grand average ERP representing both types of oddball pictures was inspected to determine the best time-window for P3 mean amplitude measurement. In both sessions, the oddball P3 was maximal over O, PO, and P electrodes and occurred during the 300–700 ms post-stimulus period. Figure 1a shows the scalp topography for the grand average ERPs separated by picture type and Figure 1b,c shows the ERP waveforms. Alcohol cue-specific P3 reactivity (ACR-P3) was isolated using both the within-person difference score (i.e., Alcohol P3 mean amplitude—NADrink mean amplitude) and the residual score approach (i.e., residuals from regressing Alcohol P3 mean amplitude on NADrink P3 mean amplitude).

### 2.2.5 Analytic approach

Following recent work on ERP psychometrics (Brunner et al., 2013; Hämmerer et al., 2013; Ip et al., 2018), we estimated the reliability of person-level mean P3 mean amplitude (at a single electrode [PZ] or the averaged cluster of nine electrodes over which the P3 was maximal [P3, PZ, P4, P7, P8, PO7, PO8, O1, O2]) using Pearson $r$. Specifically, ICR of person-level P3 mean amplitude within each session was estimated as the $r$ for person-level means based on an odd/even split of the artifact-free (i.e., retained) epochs. ICR coefficients were adjusted for task length (Brown, 1910; Spearman, 1910). TRR of P3 mean amplitudes between sessions (8–10 month retest interval) was estimated as the $r$ for person-level means from session 1 and 2. ICR and TRR also were estimated for person-level $T$-tests with two-tailed $p$-values obtained for all $r$s, and Fischer’s $r$-to-$z$ transformation (Fisher, 1921) was used to obtain 95% confidence limits.

To determine the overall level of ICR and TRR, we first computed $r$s using person-level means based on all retained epochs from all participants (see Table 2). To determine the minimum number of retained epochs required for different levels of reliability, we then recomputed $r$s using person-level means based on $n$-many retained epochs and only participants with $n$-many retained epochs, with $n$ determined by the design (e.g., because there were only 40 alcohol beverage picture targets within each session, $n$ could be 1–20 per odd/even split-half for ICR and 1–40 per session for TRR).

To qualify observed levels of measurement reliability, we applied thresholds based on both domain-general guidelines (Nunnally & Bernstein, 1994; Shrout, 1998; Shrout & Fleiss, 1979) and recent work in ERP psychometrics (Brunner et al., 2013; Clayson & Larson, 2013;
Figure 1  Scalp topography and timecourse of the event-related potential (ERP) response to alcoholic and non-alcoholic beverage picture oddball stimuli and neutral picture standard stimuli in each laboratory session. (a) Positivity over occipitoparietal scalp for Alcohol and NADrink picture types visible in the scalp maps at 400 and 600 ms post-stimulus corresponds to the P3 response. On those scalp maps, the unfilled rectangle identifies the cluster of nine electrodes (PZ, P3, P4, P7, P8, PO7, PO8, O1, O2) that captured the maximal P3 response. (b–c) Parietal cluster refers to nine-electrode occipitoparietal cluster (PZ, P3, P4, P7, P8, PO7, PO8, O1, O2). Thin black line at the center of each colorful, thicker line represents the M across persons for the indicated picture type and the thickness of the colorful line represents ±1 SEM. Yellow rectangle drawn in each plot represents the time-window chosen for P3 mean amplitude measurement on all available electrodes and trials. (a–c) Data represent N = 210 for session 1, and N = 97 for session 2.
et al., 2017; Huffmeijer et al., 2014; Ip et al., 2018; Rentzsch et al., 2008). Specifically, we defined “poor” reliability as $r \leq 0.69$, “fair” reliability as $r = 0.70–0.79$, “good” reliability as $r = 0.80–0.89$, and “excellent” reliability $r \geq 0.90$. These thresholds were applied to qualify both ICR and TRR; however, we recognize that lower thresholds may be more suitable for qualifying TRR to the extent that the construct being measured is theorized to be more state- than trait- like (see Chmielewski & Watson, 2009; Watson, 2004). Similarly, different thresholds may be necessary when qualifying the reliability (both forms) of difference and residual scores given known lower reliability relative to constituent scores (Clayson et al., 2021; Meyer et al., 2017; Perkins et al., 2017).

### 2.3 Procedure

Participants were asked to abstain from alcohol use for 24 hr prior to their scheduled laboratory sessions. Upon arrival, participants provided informed consent, and breath alcohol concentration (BrAC) was measured using an Alco-Sensor IV (Intoximeters, St. Louis, MO) to confirm sobriety (i.e., BrAC = 0.000 g%). Two participants had to be rescheduled because they arrived with non-zero BrAC. Participants were then prepared for EEG recording (30–45 min) as described in Light et al. (2010). Participants then completed the picture-categorization task (20–25 min) followed by two other behavioral tasks not reported here. After these tasks, EEG recording electrodes were removed and participants were shown to a restroom where they could wash the recording gel out of their hair. Other procedures taking place during the lab sessions are described in Supplemental Information.

Session 1 and 2 were scheduled to take place 8–10 months apart. No attempts were made to try to match day of the week, time of day, research assistants, or recording suite (one of two identically equipped suites was used) between the two sessions.

### 3 RESULTS

Picture-viewing task behavioral performance descriptive and basic inferential statistics are presented in Supplemental Information alongside overall ICR and TRR for behavioral performance measures. In sum, across sessions, categorization accuracy was relatively similar for Alcohol and NADrink cues, but Alcohol cues were correctly categorized more quickly than NADrink cues. Within-person changes in performance over time did not interact with cue type, $F \leq 1.40$, $p \geq .199$, $\eta^2 \leq 0.006$. Overall ICR and TRR for categorization accuracy were uniformly poor whereas overall ICR and TRR for correct categorization response time mirrored ICR and TRR for P3 mean amplitudes.

P3 mean amplitude descriptive statistics are presented in Table 2, whereas basic inferential statistics are presented in Supplemental Information. There was a significant main effect of cue type on cue-elicited P3 mean amplitudes at each session whether scored from either the single electrode or cluster, $F \geq 226$, $p < .001$, $\eta^2 \geq 0.753$, such that: (i) both Alcohol P3 and NADrink P3 were larger than the Neutral P3 ($p < .001$, $d = 1.315–1.977$); and (ii) the Alcohol P3 was larger than NADrink P3 ($p < .001$, $d = 0.224–0.364$; see also Figure 1b,c). Thus, a large within-person oddball versus standard stimulus difference and a small within-person Alcohol versus NADrink cue difference were observed. Importantly, for neither Parietal Cluster- nor PZ-based P3 scores across the two sessions was there either significant main effect of session, $F \leq 1.70$, $p \geq .193$, $\eta^2 \leq 0.001$, or an interaction of cue type with session, $F \leq 1.10$, $p \geq .334$, $\eta^2 \leq 0.002$, indicating little to no within-person change in P3 scores.

### 3.1 Internal consistency reliability (ICR; within-session)

Overall, there was good ICR for PZ- and Parietal Cluster-based Alcohol and NADrink Cue P3 scores from session 1 (Table 3; see also Figure 2a,b). The ACR-P3 difference and residual scores alike had poor ICR, but ICR was higher for the residual score than the difference score (Table 3; see also Figure 2c,d). In general, ICR tended to be higher for cluster-based than PZ-based scores (Table 3). Alcohol Cue P3 scores from PZ exhibited good ICR with 9–10 trials, and with as few as six trials when
using the cluster (Figure S1a). NADrink Cue P3 scores from PZ exhibited good ICR with six to seven trials, and with as few as six trials when using the cluster (Figure S1b). Accordingly, only 11–13 trials were needed for cluster-based Alcohol and NADrink Cue P3 scores to exhibit excellent ICR, 16–18 were necessary for counterpart PZ-based scores to exhibit excellent ICR (Figure S1a,b). ACR-P3 difference scores and residual scores, whether based on PZ or the Parietal Cluster, continued to exhibit poor ICR until 17–18 trials, at which point ICR became fair (Figure S1c,d). With the exception of the latter, similar results were obtained for session 2 (Figures S2–S3 and Table S4), providing cross-validation for most findings.

FIGURE 2 Relationship between scores drawing on odd/even trials in session 1 for (a) Alcohol Cue P3, (b) NADrink Cue P3, (c) ACR-P3 difference score, and (d) ACR-P3 residual score. (a–d): Parietal cluster refers to nine-electrode occipitoparietal cluster. Regression line and 95% confidence intervals shown. Points are unique participants. Data represent $N = 210$.
TABLE 4  Overall long-term test–retest reliability of oddball P3 measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>N</th>
<th>PZ</th>
<th>Parietal Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol P3</td>
<td>96</td>
<td>0.683 (0.559, 0.777)**</td>
<td>0.719 (0.605, 0.803)**</td>
</tr>
<tr>
<td>NADrink P3</td>
<td>96</td>
<td>0.695 (0.575, 0.876)**</td>
<td>0.726 (0.615, 0.809)**</td>
</tr>
<tr>
<td>ACR-P3 difference score</td>
<td>96</td>
<td>0.221 (0.021, 0.403)†</td>
<td>0.152 (–0.050, 0.342)</td>
</tr>
<tr>
<td>ACR-P3 residual score</td>
<td>96</td>
<td>0.248 (0.050, 0.427)†</td>
<td>0.183 (–0.018, 0.370)</td>
</tr>
</tbody>
</table>

Note: N = 96 out 97 because one of the session 2 completers was the participant whose session 1 electroencephalogram data could not be segmented. Parietal Cluster refers to person-level average across nine-electrode occipitoparietal cluster. Test–retest reliability (TRR) coefficient shown is the Pearson correlation coefficient for sessions (1/2). The 95% confidence interval for each TRR coefficient is shown in parentheses.

*p < .05; **p < .01; ***p < .001.

3.2 | Test–retest reliability (TRR; between-session)

Overall, there was fair TRR for Parietal Cluster-based Alcohol and NADrink Cue P3 scores, and poor-to-fair TRR for corresponding PZ-based P3 scores (Table 4; see also Figure 3a,b). For the ACR-P3 difference and residual scores, TRR was uniformly poor (Table 4; see also Figure 3c,d). Cluster-based Alcohol Cue P3 scores exhibited fair TRR with 18+ trials/session and good TRR with 32+ trials/sessions (Figure S4a). PZ-based Alcohol Cue P3 scores exhibited fair TRR with 17+ trials/session and good TRR with 39+ trials/session (Figure S4a). Cluster-based NADrink Cue P3 scores exhibited fair TRR with 15+ trials/session and good TRR with 30+ trials/sessions (Figure S4a). Cluster-based NADrink Cue P3 scores exhibited fair TRR with 20+ trials/session and good TRR with 34+ trials/session TRR (Figure S4b). In contrast, cluster- and PZ-based ACR-P3 difference and residual scores continued to exhibit poor TRR no matter the number of trials/session (Figure S4c,d).

4 | DISCUSSION

4.1 | Measurement reliability of the Alcohol Cue P3, NADrink Cue P3, and ACR-P3

The current findings concerning the measurement reliability of the Alcohol Cue P3 and NADrink Cue P3 are consistent with previous findings for other P3/LPP responses (e.g., Fabiani et al., 1987; Huffmeijer et al., 2014; Ip et al., 2018; Moran et al., 2013; Rietdijk et al., 2014; Sinha et al., 1992; Thigpen et al., 2017), including the “gold standard” P3-based measure of individual differences in addiction risk/externalizing proneness, the P3-AR (e.g., Carlson & Iacono, 2006; Perkins et al., 2017; Yoon et al., 2015). The Alcohol Cue P3 and NADrink Cue P3 alike exhibited good ICR, in keeping with similarly high ICR in an independent sample with a different stimulus set (Martins et al., 2019). Furthermore, the Alcohol Cue P3 and NADrink Cue P3 alike exhibited fair TRR over 8–10 months. Additionally, there was no evidence for within-person change in the Alcohol Cue P3 scores or NADrink Cue P3 scores across the two sessions.

Our findings concerning the measurement reliability of the ACR-P3 difference and residual scores also are consistent with reports concerning other ERP component difference scores (e.g., Bress et al., 2015; Clayson et al., 2021; Ethridge & Weinberg, 2018; Joyner et al., 2019; Kappenman et al., 2014, 2015; Klawohn et al., 2020; Luking et al., 2017; Olvet & Hajcak, 2009; Perkins et al., 2017; Weinberg & Hajcak, 2011). Higher ICR for ACR-P3 residual relative to difference scores is in keeping with previous psychometric work focused on other ERP components (Bress et al., 2015; Clayson et al., 2021; Ethridge & Weinberg, 2018; Klawohn et al., 2020; Luking et al., 2017; Meyer et al., 2017). Nonetheless, the ICR and TRR of ACR-P3 difference scores and residual scores alike were “poor” (r ≤ 0.69) when evaluated against domain-general thresholds (Nunnally & Bernstein, 1994; Shrout, 1998; Shrout & Fleiss, 1979).

It is important to note that although the absolute amount of reliable variance in difference and residual scores tends be lower than for constituent scores (as a function of the correlation between constituent scores)7 (e.g., Bress et al., 2015; Clayson et al., 2021; Luking et al., 2017; Meyer et al., 2017; Olvet & Hajcak, 2009; Perkins et al., 2017); but see (Moran et al., 2013; Weinberg & Hajcak, 2011), the reliable variance in difference and residual scores may be sufficient for individual differences research—especially if a large proportion of this reliable variance relates to other, relevant individual difference measures (e.g., Bress et al., 2015; Foti et al., 2014; Foti & Hajcak, 2009; Hajcak et al., 2017; Joyner et al., 2019; Klawohn et al., 2020; Liu et al., 2014; Meyer et al., 2017; Moser et al., 2013; Perkins et al., 2017; Yancey

---

7The constituent scores here, Alcohol Cue P3 and NADrink Cue P3, were highly correlated. In session 1, PZ-based scores: r(208) = 0.848, 95% CI (0.805, 0.882), p < .001; Parietal Cluster-based scores: r(208) = 0.884, 95% CI (0.851, 0.910), p < .001. In session 2, PZ-based scores: r(95) = 0.877, 95% CI (0.822, 0.916), p < .001; Parietal Cluster-based scores: r(95) = 0.890, 95% CI (0.840, 0.925), p < .001.
et al., 2016). Thus, difference and residual scores may have modest reliability yet sufficient validity (see Hajcak et al., 2017; Patrick et al., 2019). Ultimately, the nature of variance isolated by these scores may depend on the nature of the constituent scores and may need to be determined empirically.

Discrepant psychometric properties for constituent and difference scores also have been reported for fMRI BOLD responses (Infantolino et al., 2018; Luking et al., 2017), including alcohol cue-specific BOLD reactivity in the mesocorticolumbic system (Bach et al., 2021). Bach and colleagues suggested that one way...
to overcome the limited reliability of difference scores could be to reconceptualize the constituent scores, with the alcohol cue constituent score serving as a measure of individual differences in alcohol cue incentive salience and the non-alcohol cue constituent score serving as a measure of the stability of general cue-induced neural response. One important limitation of this approach is that a large portion of variance in the constituent scores is unrelated to the target construct. An alternative approach may be to re-conceptualize ERP constituent scores or difference scores as “items” rather than stand-alone “tests” of the target construct (see Patrick et al., 2013). Single items contain construct-relevant variance but tend to have inadequate reliability (Borsboom, 2005; Harman, 1967). Aggregating items that each contain construct-related variance may result in a (multi-item) test with greater reliability than any given item, and thereby, greater validity for indexing the target construct. This approach has been applied to other “psychoneurometric” constructs (e.g., Palumbo et al., 2020; Patrick et al., 2013; Venables et al., 2018; Yancey et al., 2016) and shows promise as a way to enhance clinical assessments (Patrick et al., 2019).

Whatever approach is taken to overcome the issue of reliability versus validity in measurement, to understand how the Alcohol Cue P3 or ACR-P3 might change across time it could be helpful to consider its similarities with, and differences from, the NADrink Cue P3. Like the NADrink Cue P3 and P3/LPP responses to cues for other ingested natural (non-drug) rewards, the Alcohol Cue P3 or ACR-P3 (and analogous measures for other drug cues) is theorized to reflect associative learning (e.g., Blechert et al., 2016; Christoffersen et al., 2017; Deweese et al., 2016; Littel & Franken, 2012; Viemose et al., 2013) and to be sensitive to current motivational states. Whereas all three measures might be sensitive to hunger and thirst (e.g., Nijs et al., 2008; Stockburger et al., 2009; Zoon et al., 2018), the Alcohol Cue P3 or ACR-P3 might be uniquely sensitive to alcohol-craving induction (e.g., McDonough & Warren, 2001; Parvaz et al., 2016).

Moreover, unlike the NADrink Cue P3, the Alcohol Cue P3 and ACR-P3 are theorized to reflect the pathophysiology of addiction, especially sensitized salience. The Incentive Sensitization Theory of Addiction (Berridge & Robinson, 2016; Robinson & Berridge, 1993) posits that repeated drug use induces adaptations in the neural circuits that mediate attribution of incentive salience (i.e., motivational significance) to cues, resulting in sensitized drug cue salience. Thus, the stability of the ACR-P3 across long retest intervals could be affected by changes in alcohol involvement, which entail changes in alcohol-related associative learning and, theoretically, changes in the neurocircuitry of incentive salience attribution. Over long retest intervals, real change in ACR-P3 could reflect either a change in a psychological process (e.g., salience attribution) or a change in the neural circuits giving rise to that process (e.g., due to alcohol-induced adaptations). Future substantive research could probe whether changes in alcohol involvement during the interval between lab sessions accounts for between-session change in the Alcohol Cue P3 or ACR-P3. Following Bach et al. (2021), between-session change in the NADrink Cue P3 or similar P3/LPP responses could be used to monitor changes in overall responsivity to appetitive cues. Nonetheless, if there is interest in measuring the Alcohol Cue P3 or ACR-P3 across the lifespan, additional psychometric work will be needed to verify its TRR across longer retest intervals as well as its ICR in different populations, especially clinical ones (e.g., treatment seeking versus non-seeking individuals with AUD).

Another important area for future research is to determine the extent to which Alcohol Cue P3 or ACR-P3 serves as an effective measure of individual differences in alcohol cue incentive salience. The validity of the Alcohol Cue P3 and ACR-P3 has been established primarily by the way of differences between groups varying in alcohol use and related phenotypes (Bartholow et al., 2007, 2010; Herrmann et al., 2001; Namkoong et al., 2004). The Alcohol Cue P3 has been shown to index variation across individuals based on low sensitivity to alcohol (Martins et al., 2019), but not alcohol use per se (Kang et al., 2021); additional studies are necessary, especially for ACR-P3. Furthermore, to our knowledge, only one study has tested the utility of the Alcohol Cue P3 or ACR-P3 as a continuous predictor of future alcohol use behavior (Bartholow et al., 2007). Additional

---

8Enhancement of these ingested natural reward cue-elicited P3 responses is associated with binge/emotional eating (e.g., Versace et al., 2016; Wolz et al., 2017) and obesity (e.g., Nijs et al., 2008, 2010, similar to how enhancement of the ACR-P3 is associated with AUD risk.

9Enhancement of the P3/LPP elicited by other drug-related cue differentiates current users from never-users and former-users (e.g., Dunning et al., 2011; Littel & Franken, 2007; McDonough & Warren, 2001; Minnix et al., 2013; Robinson et al., 2015); for meta-analytic review see: Littel et al. (2012), suggesting that P3 responses to other drug cues index the same addiction liability factors (e.g., incentive salience) as the ACR-P3.

10Behavioral and neurobiological evidence from preclinical and human laboratory studies supports the idea that alcohol-induced neuroadaptations are able to promote progressive sensitization of alcohol cue incentive salience (Cofresí et al., 2019).
studies on the predictive utility of the Alcohol Cue P3 or ACR-P3 also are warranted, particularly those using a prospective, developmental approach to characterize the co-occurrence of alcohol use onset and ACR-P3 variation.

Future prospective studies also could characterize differences between the Alcohol Cue P3 (or ACR-P3) and the P3-AR within an ontogenetic framework (Perkins et al., 2020; Senner et al., 2015). The P3-AR reflects a heritable (Carlson & Iacono, 2006), domain-general cognitive deficit that increases risk for externalizing psychopathology, including substance use, as a premorbid liability (Harper et al., 2021; Joyner et al., 2020; Perlman et al., 2013). In contrast, the Alcohol Cue P3 is theorized to be an alcohol-specific indicator reflecting consequences of alcohol use-related reinforcement learning processes (Bartholow et al., 2007, 2010). Accordingly, whereas P3-AR is observable prior to and predicts the onset of substance involvement (Harper et al., 2021; Iacono et al., 2003; Perlman et al., 2013), in theory the Alcohol Cue P3 should not differentiate individuals’ alcohol use trajectories prior to onset of use. Also, unlike P3-AR, which is largely unaffected by substance use (Joyner et al., 2020; Perlman et al., 2009), increasing alcohol involvement should be expected to exacerbate Alcohol Cue P3/ACR-P3. In contrast to both Alcohol Cue P3 and P3-AR, the NADrink Cue P3 and related P3/LPP responses, which can be used as a constituent component for the ACR-P3 (Martins et al., 2021; Versace et al., 2017), can reflect either a premorbid liability (e.g., reward deficiency syndrome; Blum et al., 1996) or an acquired response representing consequences of pathological reward learning (e.g., from heavy drinking). A longer-term prospective design in which these measures are acquired both before and after alcohol use onset (e.g., in adolescents) could help determine their relative utility for understanding AUD liability and consequences.

4.2 | Factors affecting reliability of the Alcohol Cue P3, NADrink Cue P3, and ACR-P3

4.2.1 | Single electrode versus electrode cluster-based P3 scores

With respect to whether the Alcohol Cue P3, NADrink Cue P3, and ACR-P3 should be measured from a single electrode in the cluster of electrodes over which the P3 is maximal or as an average across that cluster, our findings suggest that it depends. Consistent with previous psychometric studies of stimulus-elicited P3 amplitude modulations (e.g., Fabiani et al., 1987; Ip et al., 2018), we found that the cluster-based Alcohol Cue P3 and NADrink Cue P3 scores exhibited higher ICR and TRR than single electrode-based scores. In contrast, whereas ACR-P3 scores exhibited higher ICR when derived from cluster- than single electrode-based constituent scores, the opposite was true for TRR. Nonetheless, standardized measurement error (Luck et al., 2021) was lower for cluster- than single electrode-based Alcohol Cue P3, NADrink Cue P3, and ACR-P3 scores. Based on our findings we would advise future researchers to use cluster-based scores.

4.2.2 | Number of trials contributing to P3 score

Our findings also provide estimates of the minimum numbers of artifact-free trials needed to adequately measure the Alcohol Cue P3, NADrink Cue P3, and ACR-P3. Using cluster-based scores (given superior psychometric performance), good ICR was obtained with as few as six trials and excellent ICR with as few as 12 trials for the Alcohol Cue P3 and NADrink Cue P3. These minimum trial counts are similar to previous estimates for excellent ICR of the P3/LPP response to affective pictures in general (Moran et al., 2013). TRR for Alcohol Cue P3 and NADrink Cue P3 scores was fair with 15–18 trials/test, and good with 30–32 trials/test. In contrast, the ACR-P3 score required at least 17 trials (of each constituent score) to exhibit fair ICR. TRR for ACR-P3 scores was poor but appeared to increase slowly as the number of trials/test increased, suggesting that fair TRR might be achieved at a number of trials/test >40 (which was the theoretical maximum in our study). Based on these findings, we advise future researchers to use at least six trials to score the Alcohol Cue P3 and NADrink Cue P3 when engaged in exploratory or preliminary research, at least 16 trials per assessment when engaged in confirmatory research, and at least 32
trials per assessment when engaged in clinical research or research with an extensive longitudinal component. Researchers interested in using the ACR-P3 score are advised to use at least 17 trials of each constituent score in exploratory or preliminary research and 40 or more trials of each constituent score per test. It is important to note here that although poor reliability limits statistical power, the number of trials needed for reliable (stable) measurement is not necessarily the number of trials needed for optimal statistical power to detect a between-subject (e.g., risk group) or within-subject effect (e.g., cue category) in task-level (omnibus) analyses (see (Boudewyn et al., 2018).

4.3 Limitations of current study

The current study cannot speak to the psychometric properties of P3 responses to alcohol cues in other tasks, including variants of the picture-viewing paradigm. The current findings also cannot speak to the psychometric properties of P3 responses to other drug cues. It will be important for other addiction scholars to evaluate the psychometric properties of the P3/LPP responses to other drug cues. Additionally, the current study used alcohol and non-alcohol cues that were low in affective arousal (Pronk et al., 2015). Using more affectively arousing or alcohol craving-inducing alcohol cues may be one way to increase the amount of construct-relevant variance in the ACR-P3 difference/residual scores, especially given the sensitivity of the P3/LPP response to the arousal dimension of affect (Hajcak & Foti, 2020).

The sample used for the current study also limits the generalizability of the findings. Samples from other age or sociodemographic populations for whom familiarity with specific alcohol cues might be different, or whose alcohol involvement is more (or less) problematic, could yield different findings. Research has shown that members of different racial and ethnic groups tend to experience different alcohol marketing and advertising (Alaniz, 1998), with specific brands and product types targeting ethnic minority communities (McKee et al., 2011). Our sample was predominantly Non-Hispanic White men and women attending a major public university in the midwestern US, and our stimuli were customized accordingly (based on pretest data). It will be important for future researchers to assess measurement reliability in samples representing other cultural sub-groups in the US (e.g., emerging adults outside of a post-secondary educational setting, Hispanic individuals, Non-Hispanic Black individuals, transgender individuals).

Moreover, as noted previously, the extent to which the current findings would generalize to younger samples or to individuals in whom alcohol involvement or neurodevelopment changes dramatically between assessments remains to be determined. By design, the current sample was relatively homogenous in age at each assessment, and although session 2 took place 8–10 months after session 1, participants were still in the same stage of neural and psychological development (i.e., emerging adulthood). It will be important for future researchers to consider the age or developmental stage of participants, especially in cross-sectional studies.

5 CONCLUSION

The P3 response to visual cues for alcohol and other drugs may be a trait-like neural measure well-suited for individual differences research, but care must be taken to ensure measurement reliability. Adequate measurement will increase statistical power to detect effects of interest as well as the generalizability and reproducibility of scientific discoveries as addiction researchers adopt stimulus-elicited P3 response measurement paradigms to index incentive salience attribution to alcohol and drug-related cues.

ACKNOWLEDGEMENTS

BDB and TMP designed the project and procured its funding. BDB and GH formulated the research questions. BDB, GH, RUC, and TMP wrote the manuscript together. RUC collected, processed, and analyzed the data, and prepared the figures and tables. The authors are grateful to the study coordinators (Haley Benson, Karen Yates, Dr. Sandie Keerstock), numerous undergraduate research assistants, and former graduate research assistants (especially Dr. Jorge Martins) that helped collect and archive data for this project while working in the Social Cognition and Addiction Neuroscience Laboratory (SCANlab).

CONFLICTS OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Roberto U. Cofresi: Formal analysis; Investigation; Visualization; Writing-original draft; Writing-review & editing. Thomas M. Piasecki: Funding acquisition; Resources; Supervision; Writing-review & editing. Greg Hajcak: Conceptualization; Writing-review & editing. Bruce D. Bartholow: Conceptualization; Funding acquisition; Project administration; Resources; Supervision; Writing-review & editing.
ETHICAL APPROVAL
All procedures were approved by the University of Missouri Institutional Review Board.

ORCID
Roberto U. Cofresí https://orcid.org/0000-0003-1131-6142
Thomas M. Piascik https://orcid.org/0000-0002-2793-2049
Greg Hajcak https://orcid.org/0000-0002-5159-7180
Bruce D. Bartholow https://orcid.org/0000-0002-9234-6417

REFERENCES


and Biobehavioral Reviews, 36(8), 1803–1816. https://doi.org/10.1016/j.neubiorev.2012.05.001


**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**TABLE S1** Average behavioral performance scores by session

**TABLE S2** Internal consistency reliability of behavioral performance scores by session

**TABLE S3** Test-retest reliability of behavioral performance scores

**TABLE S4** Overall internal consistency reliability of odd-ball P3 measures in session 2

**FIGURE S1** Internal consistency reliability coefficient in session 1 as a function of number of trials contributing to (a) Alcohol Cue P3 score, (b) NADrink Cue P3 score, (c) ACR-P3 difference score, and (d) ACR-P3 residual score

**FIGURE S2** Relationship between scores drawing on odd/even trials in session 2 for (a) Alcohol Cue P3 score, (b) NADrink Cue P3, (c) ACR-P3 difference score, and (d) ACR-P3 residual score

**FIGURE S3** Internal consistency reliability coefficient in session 2 as a function of number of trials contributing to (a) Alcohol Cue P3 score, (b) NADrink Cue P3 score, (c) ACR-P3 difference score, and (d) ACR-P3 residual score

**FIGURE S4** Test-retest reliability coefficient as a function of number of trials contributing to (a) Alcohol Cue P3 Score, (b) NADrink Cue P3, (c) ACR-P3 difference score, and (d) ACR-P3 residual score

**How to cite this article:** Cofresí, R. U., Piasecki, T. M., Hajcak, G., & Bartholow, B. D. (2021). Internal consistency and test–retest reliability of the P3 event-related potential (ERP) elicited by alcoholic and non-alcoholic beverage pictures. *Psychophysiology, 00*, e13967. https://doi.org/10.1111/psyp.13967