Received: 16 June 2021 | Revised: 18 September 2021 | Accepted: 26 October 2021

DOI: 10.1111/psyp.13967

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

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

**Roberto U. Cofresí**[1] | **Thomas M. Piasecki**[1] | **Greg Hajcak**[2] | **Bruce D. Bartholow**[1]

[1]Department of Psychological Sciences, University of Missouri, Columbia, Missouri, USA

[2]Departments 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 (Littel et al., 2012), such as selective attention (e.g., Dickter et al., 2014; Kroczek et al., 2018; Petit et al., 2012; Shin et al., 2010) and incentive-motivational value (e.g., Deweese 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; Deweese 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; Kroczek et al., 2018). Enhanced ACR-P3 is associated with heavier and more hazardous alcohol use (Herrmann et al., 2001; Kroczek 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 & Patrick, 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).[1] 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

[1]In 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.[2] 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.

**TABLE 1** Participant characteristics

| | Session 1 ($N = 211$) | Session 2 ($N = 98$) |
|---|---|---|
| | *M* (*SD*) | *M* (*SD*) |
| Age, yr | 19.48 ± 0.73 | 20.42 ± 0.85 |
| | *n* (%) | *n* (%) |
| Female | 115 (55) | 65 (66) |
| Ethnicity | | |
| Hispanic | 11 (5) | 5 (5) |
| Race | | |
| American Indian/ Alaskan Native | 1 (<1) | 0 (0) |
| Native Hawaiian/Pacific Islander | 0 (0) | 0 (0) |
| Asian | 6 (3) | 2 (2) |
| Black | 8 (4) | 5 (5) |
| White | 186 (88) | 88 (90) |
| Multiple selected | 10 (5) | 3 (3) |
| None selected | 0 (0) | 0 (0) |
| Handedness | | |
| Right dominant[a] | 184 (87) | 90 (92) |
| Undergraduate student[b] | 203 (96) | 96 (98) |

*Note:* Demographic information was collected at screening.

[a]Right hand dominance was defined as an Edinburgh Handedness Inventory short-form score of 61 or above (Veale, 2014).

[b]Undergraduate student was defined as being enrolled in a 4-year college program (BA/BS-granting institution). Of the 8 participants who were not undergraduate students at screening, 5 were enrolled in a 2-year college program (AA/AS-granting institution), 2 were attending high school or working toward a high school equivalency credential (e.g., GED), and 1 was not enrolled in any form of schooling.

### 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.[3] 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),[4] 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

[2]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 $|r| \leq 0.10$ using a two-sided *t*-test against the null hypothesis that $|r| = 0$ with 5% Type 1 error, we would need $N \geq 779$, 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.

[3]IAPS image codes: 1122, 1350, 1616, 1670, 1675, 1903, 1908, 1935, 1947, 5040, 5120, 5130, 5390, 5395, 5471, 5500, 5510, 5520, 5530, 5531, 5532, 5533, 5534, 5535, 5740, 6150, 7002, 7003, 7004, 7006, 7010, 7011, 7012, 7014, 7016, 7017, 7018, 7019, 7020, 7021, 7025, 7026, 7030, 7032, 7033, 7034, 7036, 7037, 7038, 7039, 7040, 7041, 7043, 7045, 7050, 7052, 7053, 7055, 7056, 7059, 7090, 7140, 7161, 7175, 7180, 7205, 7217, 7224, 7234, 7287, 7290, 7491, 7495, 7705, 7950, 9360, 9469.

[4]ABPS image codes: SDC10695, SDC10709, SDC10716, SDC10917, SDC11010, SDC11069, SDC10744, SDC10804, SDC10808, SDC10815, SDC10821, SDC10825, SDC10836, SDC10858, SDC10946, SDC10967.

beverage on a bland white background as in the ABPS).[5] 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 (Mognon 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 ($M \pm SD$% of all epochs per participant in session 1 or 2:2.12 ± 3.34).[6] 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

[5]Budweiser can, Coors Light can, Natural Light can, and Jack Daniel's bottle alongside a filled shot glass.

[6]Inclusion 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.

**TABLE 2** Average number of retained (artifact-free) epochs (trials), P3 mean amplitude scores, and standardized measurement error by picture type across participants

| | Session 1 ($N$ = 210) | | | | | Session 2 ($N$ = 97) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Picture type | Number of retained epochs | PZ mean amplitude (μV) | PZ SME (μV) | Parietal Cluster mean amplitude (μV) | Parietal Cluster SME (μV) | Number of retained epochs | PZ mean amplitude (μV) | PZ SME (μV) | Parietal Cluster mean amplitude (μV) | Parietal Cluster SME (μV) |
| Alcohol | 34.36 (7.37) | 12.85 (5.50) | 1.81 | 11.68 (4.57) | 1.33 | 33.92 (7.20) | 13.42 (6.05) | 2.00 | 11.73 (4.79) | 1.34 |
| NADrink | 34.01 (7.16) | 10.85 (5.52) | 1.90 | 10.04 (4.40) | 1.35 | 34.04 (7.35) | 11.96 (6.24) | 1.94 | 10.66 (4.68) | 1.31 |
| Neutral | 278.02 (63.57) | 1.68 (3.28) | 0.68 | 4.95 (2.59) | 0.51 | 282.94 (57.58) | 2.19 (3.23) | 0.65 | 4.85 (2.29) | 0.46 |

*Note: M* (*SD*) are shown except for the standardized measurement error (SME) columns. Following Luck et al. (2021), SME was computed as each participant's standard deviation of P3 time-window mean amplitude across single trials divided by the square root of the number of single trials, and aggregated across participants as the root mean square (RMS) of SME. The maximum number of Alcohol, NADrink, and Neutral pictures (trials) per session was 40, 40, and 320, respectively. Parietal Cluster refers to person-level average across nine-electrode occipitoparietal cluster.

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;

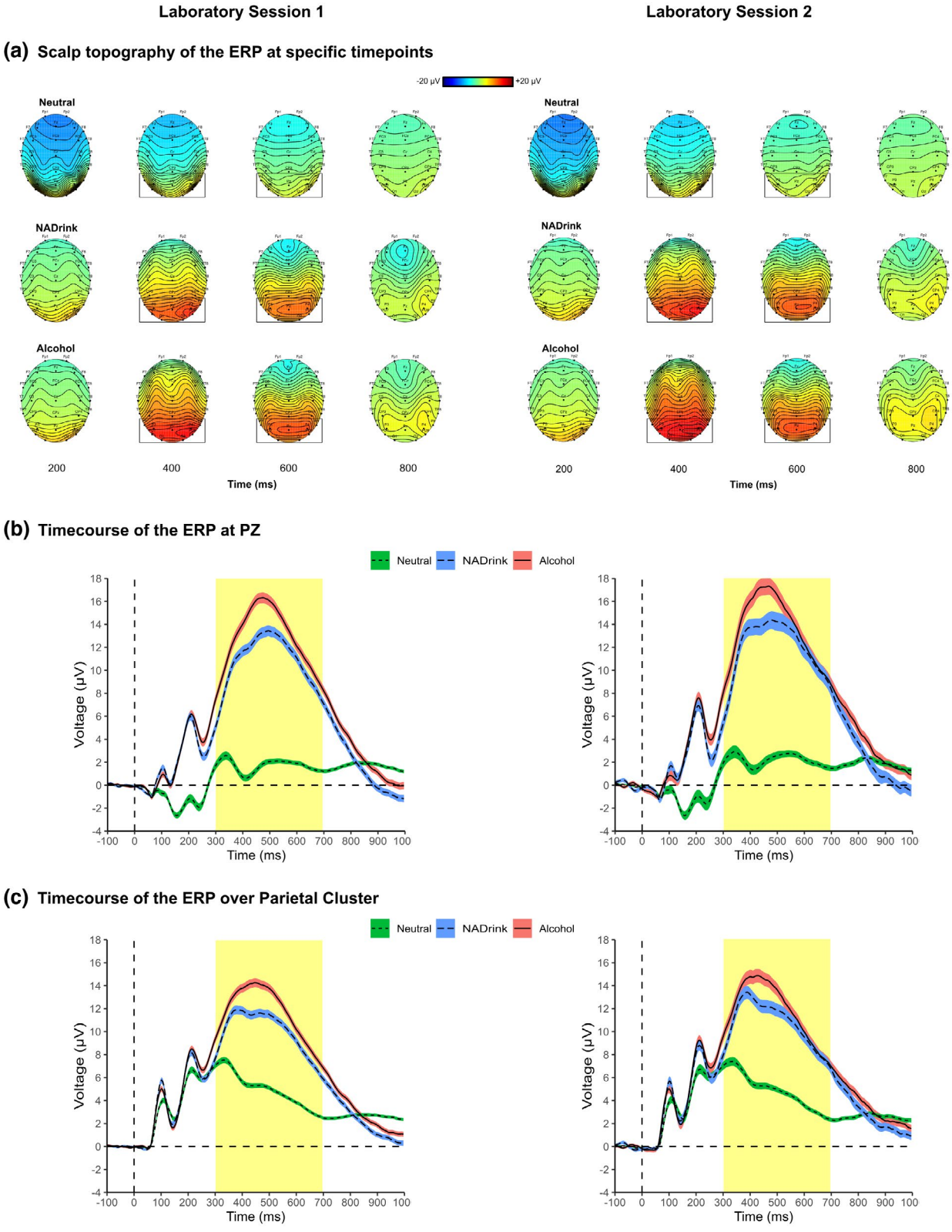


**FIGURE 1** Scalp topography and timecourse of the event-related potential (ERP) response to alcoholic and non-alcoholic beverage picture oddball stimuli and neutral picture standard stimuli in each laboratory session. (a) Positivity over occipitoparietal scalp for Alcohol and NADrink picture types 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

Hajcak 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.

**TABLE 3** Overall internal consistency reliability of oddball P3 measures in session 1

| Measure | *N* | PZ | Parietal Cluster |
|---|---|---|---|
| Alcohol P3 | 210 | 0.861 (0.821, 0.892)*** | 0.902 (0.874, 0.925)*** |
| NADrink P3 | 210 | 0.861 (0.821, 0.892)*** | 0.895 (0.865, 0.919)*** |
| ACR-P3 difference score | 210 | 0.301 (0.173, 0.419)*** | 0.324 (0.198, 0.441)*** |
| ACR-P3 residual score | 210 | 0.411 (0.292, 0.518)*** | 0.413 (0.295, 0.520)*** |

*Note:* Parietal Cluster refers to person-level average across nine-electrode occipitoparietal cluster. ICR coefficient shown is the Pearson correlation coefficient for split-halves (even/odd) adjusted using the Spearman-Brown prophecy formula. The 95% confidence interval for each ICR coefficient is shown in parentheses.

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

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

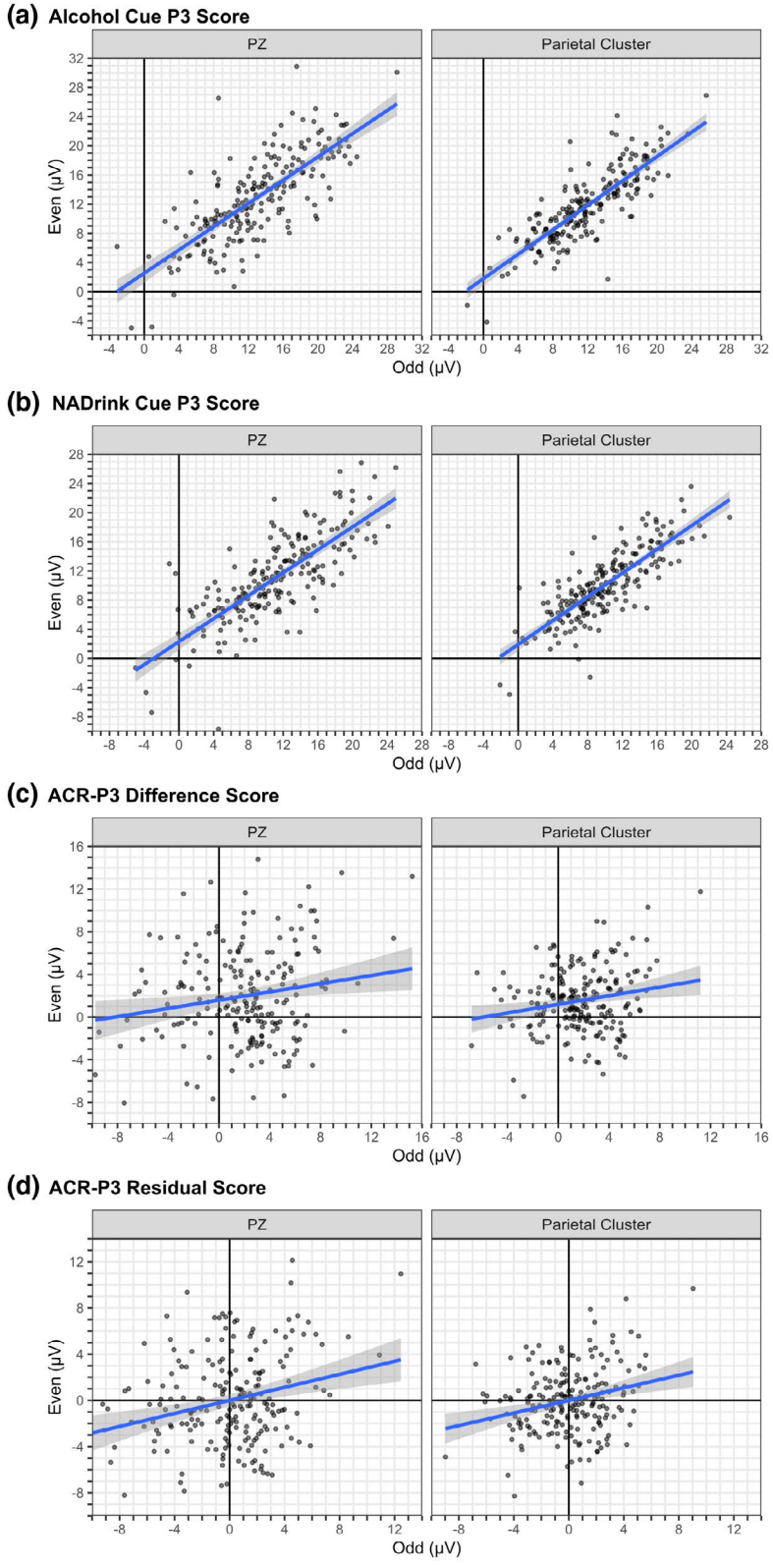


**FIGURE 2** Relationship between scores drawing on odd/even trials in session 1 for (a) Alcohol Cue P3, (b) NADrink Cue P3, (c) ACR-P3 difference score, and (d) ACR-P3 residual score. (a–d): Parietal cluster refers to nine-electrode occipitoparietal cluster. Regression line and 95% confidence intervals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

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.

**TABLE 4** Overall long-term test–retest reliability of oddball P3 measures

| Measure | *N* | PZ | Parietal Cluster |
|---|---|---|---|
| Alcohol P3 | 96 | 0.683 (0.559, 0.777)*** | 0.719 (0.605, 0.803)*** |
| NADrink P3 | 96 | 0.695 (0.575, 0.786)*** | 0.726 (0.615, 0.809)*** |
| ACR-P3 difference score | 96 | 0.221 (0.021, 0.403)* | 0.152 (−0.050, 0.342) |
| ACR-P3 residual score | 96 | 0.248 (0.050, 0.427)* | 0.183 (−0.018, 0.370) |

*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 S4b). PZ-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" ($r \leq 0.69$) 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

[7]The 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), $p < .001$.

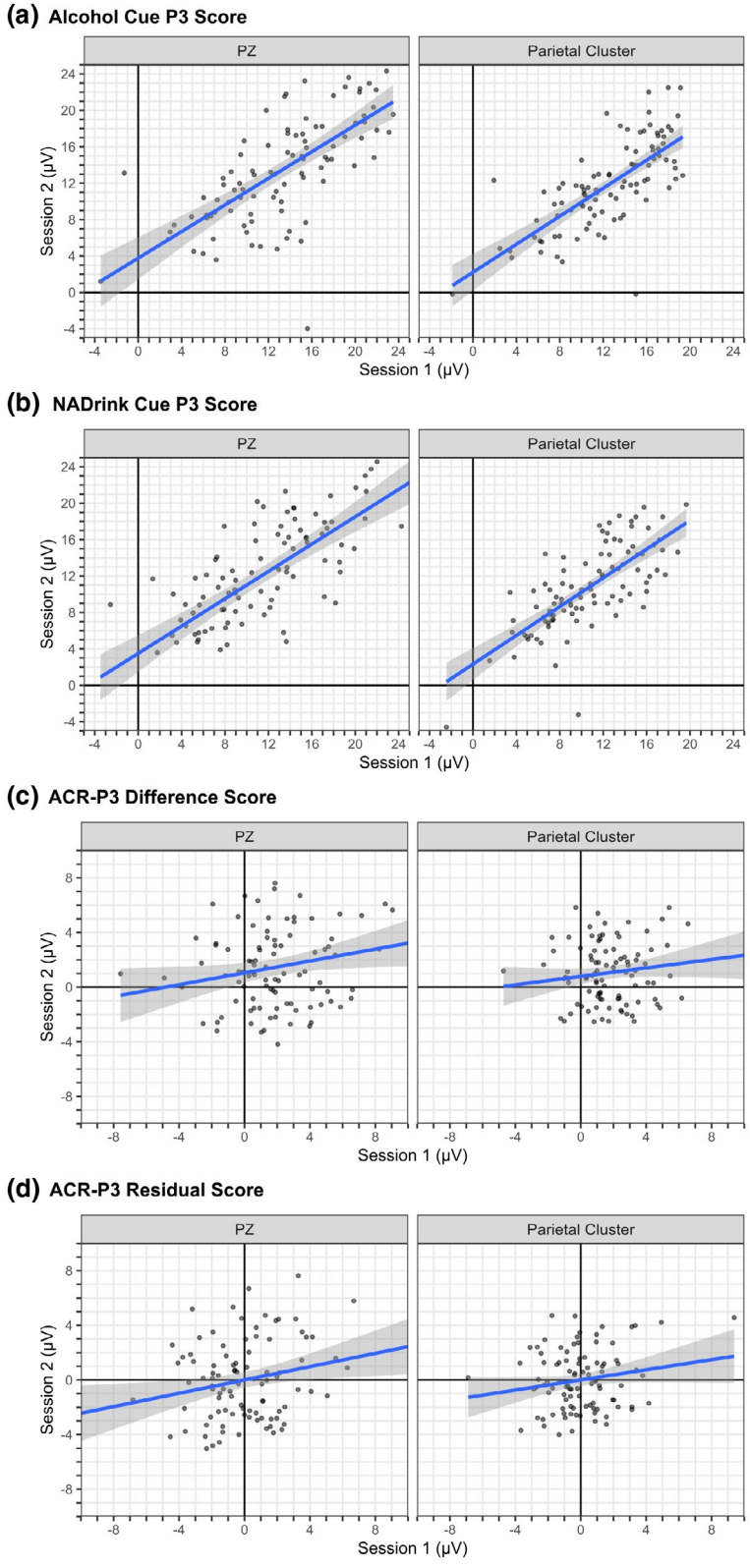


**FIGURE 3** Relationship between session 1 and session 2 for (a) Alcohol Cue P3, (b) NADrink Cue P3, (c) ACR-P3 difference score, and (d) ACR-P3 residual score. (a–d): Parietal cluster refers to nine-electrode occipitoparietal cluster. Regression line and 95% confidence intervals shown. Points are unique participants. Data represent $N = 96$

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 mesocorticolimbic 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,[8] the Alcohol Cue P3 or ACR-P3 (and analogous measures for other drug cues[9]) 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.[10] 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

[8]Enhancement 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.

[9]Enhancement 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.

[10]Behavioral 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).[11] 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.

[11]Since its heritability is unknown, it is unclear whether individuals who have not yet experienced alcohol/drug pharmacodynamics will exhibit the ACR-P3. However, it is possible that non-experiential learning about alcohol and drugs (e.g., alcohol/drug use-outcome expectancies) is sufficient to imbue to alcohol and drug-related cues with some motivational significance prior to direct experiential learning. Thus, a P3 response to alcohol/drug cues can be expected among those who have not yet been exposed to alcohol/drugs. This possibility is reinforced by previous work showing that never-smokers exhibit an LPP response to tobacco smoking-related visual cues (e.g., Minnix et al., 2013; Robinson et al., 2015), although the meaning and significance of alcohol/drug cues among never-users may be driven by the defensive motivational system rather than the appetitive motivational system.

## 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. Cofresí:** 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. Piasecki* 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

Alaniz, M. L. (1998). Alcohol availability and targeted advertising in racial/ethnic minority communities. *Alcohol Research and Health*, *22*(4), 286–289.

Arcara, G., & Petrova, A. (2014). *erpR: ERP analysis, graphics and utility functions*. https://r-forge.r-project.org/projects/erpr/

Bach, P., Reinhard, I., Koopmann, A., Bumb, J. M., Sommer, W. H., Vollstädt-Klein, S., & Kiefer, F. (2021). Test–retest reliability of neural alcohol cue-reactivity: Is there light at the end of the magnetic resonance imaging tube? *Addiction Biology*, January, 1–12. https://doi.org/10.1111/adb.13069

Bailey, K., & Bartholow, B. D. (2016). Alcohol words elicit reactive cognitive control in low-sensitivity drinkers. *Psychophysiology*, *53*(11), 1751–1759. https://doi.org/10.1111/psyp.12741

Baldwin, S. A. (2017). Improving the rigor of psychophysiology research. *International Journal of Psychophysiology*, *111*, 5–16. https://doi.org/10.1016/j.ijpsycho.2016.04.006

Bartholow, B. D., Henry, E. A., & Lust, S. A. (2007). Effects of alcohol sensitivity on P3 event-related potential reactivity to alcohol cues. *Psychology of Addictive Behaviors*, *21*(4), 555–563. https://doi.org/10.1037/0893-164X.21.4.555

Bartholow, B. D., Loersch, C., Ito, T. A., Levsen, M. P., Volpert-Esmond, H. I., Fleming, K. A., Bolls, P., & Carter, B. K. (2018). University-affiliated alcohol marketing enhances the incentive salience of alcohol cues. *Psychological Science*, *29*(1), 83–94. https://doi.org/10.1177/0956797617731367

Bartholow, B. D., Lust, S. A., & Tragesser, S. L. (2010). Specificity of P3 event-related potential reactivity to alcohol cues in individuals low in alcohol sensitivity. *Psychology of Addictive Behaviors*, *24*(2), 220–228. https://doi.org/10.1037/a0017705

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*(2), 254–263. https://doi.org/10.1177/0013164402062002004

Begleiter, H., Porjesz, B., Bihari, B., & Kissin, B. (1984). Event-related brain potentials in boys at risk for alcoholism. *Science*, *225*(4669), 1493–1496. https://www.jstor.org/stable/1693558

Begleiter, H., Porjesz, B., Chou, C. L., & Aunon, J. I. (1983). P3 and stimulus incentive value. *Psychophysiology*, *20*(1), 95–101. https://doi.org/10.1111/j.1469-8986.1983.tb00909.x

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist*, *71*(8), 670–679. https://doi.org/10.1037/amp0000059

Blechert, J., Testa, G., Georgii, C., Klimesch, W., & Wilhelm, F. H. (2016). The Pavlovian craver: Neural and experiential correlates of single trial naturalistic food conditioning in humans. *Physiology and Behavior*, *158*, 18–25. https://doi.org/10.1016/j.physbeh.2016.02.028

Blum, K., Cull, J. G., Braverman, E. R., & Comings, D. E. (1996). Reward deficiency syndrome. *American Scientist*, *84*(2), 132–145.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), 1–16. https://doi.org/10.1111/psyp.13049

Bress, J. N., Meyer, A., & Proudfit, G. H. (2015). The stability of the feedback negativity and its relationship with depression during childhood and adolescence. *Development and Psychopathology*, *27*(4), 1285–1294. https://doi.org/10.1017/S0954579414001400

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Brunner, J. F., Hansen, T. I., Olsen, A., Skandsen, T., Håberg, A., & Kropotov, J. (2013). Long-term test-retest reliability of the P3 NoGo wave and two independent components decomposed from the P3 NoGo wave in a visual Go/NoGo task. *International Journal of Psychophysiology*, *89*(1), 106–114. https://doi.org/10.1016/j.ijpsycho.2013.06.005

Carlson, S. R., & Iacono, W. G. (2006). Heritability of P300 amplitude development from adolescence to adulthood. *Psychophysiology*, *43*(5), 470–480. https://doi.org/10.1111/j.1469-8986.2006.00450.x

Carlson, S. R., McLarnon, M. E., & Iacono, W. G. (2007). P300 amplitude, externalizing psychopathology, and earlier- versus later-onset substance-use disorder. *Journal of Abnormal Psychology*, *116*(3), 565–577. https://doi.org/10.1037/0021-843X.116.3.565

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*(1), 186–202. https://doi.org/10.1037/a0015618

Christoffersen, G. R. J., Laugesen, J. L., Møller, P., Bredie, W. L. P., Schachtman, T. R., Liljendahl, C., & Viemose, I. (2017). Long-term visuo-gustatory appetitive and aversive conditioning potentiate human visual evoked potentials. *Frontiers in Human Neuroscience*, *11*(September), 1–16. https://doi.org/10.3389/fnhum.2017.00467

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2021). Evaluating the internal consistency of subtraction-based and residualized difference scores: Considerations for psychometric reliability analyses of event-related potentials. *Psychophysiology*, *58*(4), 1–14. https://doi.org/10.1111/psyp.13762

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), 1–17. https://doi.org/10.1111/psyp.13437

Clayson, P. E., & Larson, M. J. (2013). Psychometric properties of conflict monitoring and conflict adaptation indices: Response time and conflict N2 event-related potentials. *Psychophysiology*, *50*(12), 1209–1219. https://doi.org/10.1111/psyp.12138

Clayson, P. E., & Miller, G. A. (2017). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*, *111*, 57–67. https://doi.org/10.1016/j.ijpsycho.2016.09.005

Codispoti, M., Micucci, A., & De Cesarei, A. (2021). Time will tell: Object categorization and emotional engagement during processing of degraded natural scenes. *Psychophysiology*, *58*(1), 1–16. https://doi.org/10.1111/psyp.13704

Cofresí, R. U., Bartholow, B. D., & Piasecki, T. M. (2019). Evidence for incentive salience sensitization as a pathway to alcohol use disorder. *Neuroscience and Biobehavioral Reviews*, *107*, 897–926. https://doi.org/10.1016/j.neubiorev.2019.10.009

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Davis, C. N., Piasecki, T. M., Bartholow, B. D., & Slutske, W. S. (2021). Effects of alcohol sensitivity on alcohol-induced blackouts and passing out: An examination of the alcohol sensitivity questionnaire among underage drinkers. *Alcoholism: Clinical & Experimental Research*, March, *45*, 1149–1160. https://doi.org/10.1111/acer.14607

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Deweese, M. M., Codispoti, M., Robinson, J. D., Cinciripini, P. M., & Versace, F. (2018). Cigarette cues capture attention of smokers and never-smokers, but for different reasons. *Drug and Alcohol Dependence*, *185*(February), 50–57. https://doi.org/10.1016/j.drugalcdep.2017.12.010

Deweese, M. M., Robinson, J. D., Cinciripini, P. M., & Versace, F. (2016). Conditioned cortical reactivity to cues predicting cigarette-related or pleasant images. *International Journal of Psychophysiology*, *101*, 59–68. https://doi.org/10.1016/j.ijpsycho.2016.01.007

Dickter, C. L., Forestell, C. A., Hammett, P. J., & Young, C. M. (2014). Relationship between alcohol dependence, escape drinking, and early neural attention to alcohol-related cues. *Psychopharmacology (Berl)*, *231*(9), 2031–2040. https://doi.org/10.1007/s00213-013-3348-6

Dunning, J. P., Parvaz, M. A., Hajcak, G., Maloney, T., Alia-Klein, N., Woicik, P. A., Telang, F., Wang, G.-J., Volkow, N. D., & Goldstein, R. Z. (2011). Motivated attention to cocaine and emotional cues in abstinent and current cocaine users—An ERP study. *European Journal of Neuroscience*, *33*(9), 1716–1723. https://doi.org/10.1111/j.1460-9568.2011.07663.x

Ethridge, P., & Weinberg, A. (2018). Psychometric properties of neural responses to monetary and social rewards across development. *International Journal of Psychophysiology*, *132*(July, 2017), 311–322. https://doi.org/10.1016/j.ijpsycho.2018.01.011

Euser, A. S., Arends, L. R., Evans, B. E., Greaves-Lord, K., Huizink, A. C., & Franken, I. H. A. (2012). The P300 event-related brain potential as a neurobiological endophenotype for substance use disorders: A meta-analytic investigation. *Neuroscience and Biobehavioral Reviews*, *36*(1), 572–603. https://doi.org/10.1016/j.neubiorev.2011.09.002

Fabiani, M., Gratton, G., Karis, D., & Donchin, E. (1987). Definition, identification and reliability of the P300 component of the event related brain potential. *Advances Psychophysiology*, *2*, 1–78.

Fisher, R. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32. https://doi.org/10.1093/biomet/9.1-2.22

Fleming, K. A., & Bartholow, B. D. (2014). Alcohol cues, approach bias, and inhibitory control: Applying a dual process model of addiction to alcohol sensitivity. *Psychology of Addictive Behaviors*, *28*(1), 85–96. https://doi.org/10.1037/a0031565

Fleming, K. A., Cofresí, R. U., & Bartholow, B. D. (2021). Transfer of incentive salience from a first-order alcohol cue to a novel second-order alcohol cue among individuals at risk for alcohol use disorder: Electrophysiological evidence. *Addiction*, *116*(7), 1734–1746. https://doi.org/10.1111/add.15380

Foti, D., Carlson, J. M., Sauder, C. L., & Proudfit, G. H. (2014). Reward dysfunction in major depression: Multimodal neuroimaging evidence for refining the melancholic phenotype. *NeuroImage*, *101*, 50–58. https://doi.org/10.1016/j.neuroimage.2014.06.058

Foti, D., & Hajcak, G. (2009). Depression and reduced sensitivity to non-rewards versus rewards: Evidence from event-related potentials. *Biological Psychology*, *81*(1), 1–8. https://doi.org/10.1016/j.biopsycho.2008.12.004

Franken, I. H. A., van Strien, J. W., Bocanegra, B. R., & Huijding, J. (2011). The p3 event-related potential as an index of motivational relevance: A conditioning experiment. *Journal of Psychophysiology*, *25*(1), 32–39. https://doi.org/10.1027/0269-8803/a000030

Gao, Y., & Raine, A. (2009). P3 event-related potential impairments in antisocial and psychopathic individuals: A meta-analysis. *Biological Psychology*, *82*(3), 199–210. https://doi.org/10.1016/j.biopsycho.2009.06.006

Garland, E. L., Atchley, R. M., Hanley, A. W., Zubieta, J. K., & Froeliger, B. (2019). Mindfulness-oriented recovery enhancement remediates hedonic dysregulation in opioid users: Neural and affective evidence of target engagement. *Science Advances*, *5*(10), 1–13. https://doi.org/10.1126/sciadv.aax1569

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gilmore, C. S., Malone, S. M., Bernat, E. M., & Iacono, W. G. (2010). Relationship between the P3 event-related potential, its associated time-frequency components, and externalizing psychopathology. *Psychophysiology*, *47*(1), 123–132. https://doi.org/10.1111/j.1469-8986.2009.00876.x

Hajcak, G., & Foti, D. (2020). Significance?... Significance! Empirical, methodological, and theoretical connections between the late positive potential and P300 as neural responses to stimulus significance: An integrative review. *Psychophysiology*, *57*(7), 1–15. https://doi.org/10.1111/psyp.13570

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*, *126*(6), 823–834. https://doi.org/10.1037/abn0000274

Hajcak, G., & Patrick, C. J. (2015). Situating psychophysiological science within the Research Domain Criteria (RDoC) framework. *International Journal of Psychophysiology*, *98*(2), 223–226. https://doi.org/10.1016/j.ijpsycho.2015.11.001

Hamidovic, A., & Wang, Y. (2019). The P300 in alcohol use disorder: A meta-analysis and meta-regression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *95*(July), e109716. https://doi.org/10.1016/j.pnpbp.2019.109716

Hämmerer, D., Li, S. C., Völkle, M., Müller, V., & Lindenberger, U. (2013). A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. *Psychophysiology*, *50*(1), 111–123. https://doi.org/10.1111/j.1469-8986.2012.01476.x

Harman, H. H. (1967). *Modern factor analysis*. University of Chicago Press.

Harper, J., Malone, S. M., & Iacono, W. G. (2021). Parietal P3 and midfrontal theta prospectively predict the development of adolescent alcohol use. *Psychological Medicine*, *51*(3), 416–425. https://doi.org/10.1017/S0033291719003258

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Herrmann, M. J., Weijers, H. G., Wiesbeck, G. A., Böning, J., & Fallgatter, A. J. (2001). Alcohol cue-reactivity in heavy and light social drinkers as revealed by event-related potentials. *Alcohol & Alcoholism*, *36*(6), 588–593. https://doi.org/10.1093/alcalc/36.6.588

Herting, M. M., Gautam, P., Chen, Z., Mezher, A., & Vetter, N. C. (2018). Test-retest reliability of longitudinal task-based fMRI: Implications for developmental studies. *Developmental Cognitive Neuroscience*, *33*(June 2017), 17–26. https://doi.org/10.1016/j.dcn.2017.07.001

Hone, L. S. E., Bartholow, B. D., Piasecki, T. M., & Sher, K. J. (2017). Women's alcohol sensitivity predicts alcohol-related regretted sex. *Alcoholism: Clinical and Experimental Research*, *41*(9), 1630–1636. https://doi.org/10.1111/acer.13447

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R. A., & van IJzendoorn, M. H. (2014). Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiology and Behavior*, *130*(2014), 13–22. https://doi.org/10.1016/j.physbeh.2014.03.008

Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement*, *17*(1), 11–14. https://doi.org/10.1177/014662169301700102

Iacono, W. G., Carlson, S. R., Malone, S. M., & McGue, M. (2002). P3 event-related potential amplitude and the risk for disinhibitory disorders in adolescent boys. *Archives of General Psychiatry*, *59*(8), 750–757. https://doi.org/10.1001/archpsyc.59.8.750

Iacono, W. G., Malone, S. M., & McGue, M. (2003). Substance use disorders, externalizing psychopathology, and P300 event-related potential amplitude. *International Journal of Psychophysiology*, *48*(2), 147–178. https://doi.org/10.1016/S0167-8760(03)00052-7

Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage*, *173*(June 2017), 146–152. https://doi.org/10.1016/j.neuroimage.2018.02.024

Ip, C.-T., Ganz, M., Ozenne, B., Sluth, L. B., Gram, M., Viardot, G., l'Hostis, P., Danjou, P., Knudsen, G. M., & Christensen, S. R. (2018). Pre-intervention test-retest reliability of EEG and ERP over four recording intervals. *International Journal of Psychophysiology*, *134*(2018), 30–43. https://doi.org/10.1016/j.ijpsycho.2018.09.007

Joyner, K. J., Bowyer, C. B., Yancey, J. R., Venables, N. C., Foell, J., Worthy, D. A., Hajcak, G., Bartholow, B. D., & Patrick, C. J. (2019). Blunted reward sensitivity and trait disinhibition interact to predict substance use problems. *Clinical Psychological Science*, *7*(5), 1109–1124. https://doi.org/10.1177/2167702619838480

Joyner, K. J., Yancey, J. R., Venables, N. C., Burwell, S. J., Iacono, W. G., & Patrick, C. J. (2020). Using a co-twin control design to evaluate alternative trait measures as indices of liability for substance use disorders. *International Journal of Psychophysiology*, *148*, 75–83. https://doi.org/10.1016/j.ijpsycho.2019.11.012

Kamarajan, C., & Porjesz, B. (2015). Advances in electrophysiological research. *Alcohol Research: Current Reviews*, *37*(1), 53–87.

Kang, D., Fairbairn, C. E., Lee, Z., & Federmeier, K. D. (2021). The effect of acute alcohol intoxication on alcohol cue salience: An event-related brain potential study. *Psychology of Addictive Behaviors*, 1–11. https://doi.org/10.1037/adb0000779

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 81–90. https://doi.org/10.22237/jmasm/1177992480

Kappenman, E. S., Farrens, J. L., Luck, S. J., & Proudfit, G. H. (2014). Behavioral and ERP measures of attentional bias to threat in the dot-probe task: Poor reliability and lack of correlation with anxiety. *Frontiers in Psychology*, *5*(December), 1–9. https://doi.org/10.3389/fpsyg.2014.01368

Kappenman, E. S., MacNamara, A., & Proudfit, G. H. (2015). Electrocortical evidence for rapid allocation of attention to threat in the dot-probe task. *Social Cognitive and Affective Neuroscience*, *10*(4), 577–583. https://doi.org/10.1093/scan/nsu098

Kinreich, S., Meyers, J. L., Maron-Katz, A., Kamarajan, C., Pandey, A. K., Chorlian, D. B., Zhang, J., Pandey, G., Subbie-Saenz de Viteri, S., Pitti, D., Anokhin, A. P., Bauer, L. O., Hesselbrock, V. M., Schuckit, M. A., Edenberg, H. J., & Porjesz, B. (2021). Predicting risk for alcohol use disorder using longitudinal data with multimodal biomarkers and family history: A machine learning study. *Molecular Psychiatry*, *26*(4), 1133–1141. https://doi.org/10.1038/s41380-019-0534-x

Klawohn, J., Meyer, A., Weinberg, A., & Hajcak, G. (2020). Methodological choices in event-related potential (ERP) research and their impact on internal consistency reliability and individual differences: An examination of the error-related negativity (ERN) and anxiety. *Journal of Abnormal Psychology*, *129*(1), 29–37. https://doi.org/10.1037/abn0000458.supp

Kline, P. (1998). *The new psychometrics: Science, psychology and measurement*. Taylor & Frances/Routledge.

Kroczek, A. M., Haeussinger, F. B., Hudak, J., Vanes, L. D., Fallgatter, A. J., & Ehlis, A.-C. (2018). Cue reactivity essentials: Event-related potentials during identification of visual alcoholic stimuli in social drinkers. *Journal of Studies on Alcohol and Drugs*, *79*(1), 137–147. https://doi.org/10.15288/jsad.2018.79.137

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. (Technical Report A-8). University of Florida.

Light, G. A., Williams, L. E., Minow, F., Sprock, J., Rissling, A., Sharp, R., Swerdlow, N. R., & Braff, D. L. (2010). Electroencephalography (EEG) and event-related potentials (ERPs) with human participants. *Current Protocols in Neuroscience*, *52*(Suppl), 1–24. https://doi.org/10.1002/0471142301.ns0625s52

Littel, M., Euser, A. S., Munafò, M. R., & Franken, I. H. A. (2012). Electrophysiological indices of biased cognitive processing of substance-related cues: A meta-analysis. *Neuroscience*

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

Littel, M., & Franken, I. H. A. (2007). The effects of prolonged abstinence on the processing of smoking cues: An ERP study among smokers, ex-smokers and never-smokers. *Journal of Psychopharmacology*, *21*(8), 873–882. https://doi.org/10.1177/0269881107078494

Littel, M., & Franken, I. H. A. (2012). Electrophysiological correlates of associative learning in smokers: A higher-order conditioning experiment. *BMC Neuroscience*, *13*(1), 1–13. https://doi.org/10.1186/1471-2202-13-8

Liu, W., Wang, L., Shang, H., Shen, Y., Li, Z., Cheung, E. F. C., & Chan, R. C. K. (2014). The influence of anhedonia on feedback negativity in major depressive disorder. *Neuropsychologia*, *53*, 213–220. https://doi.org/10.1016/j.neuropsychologia.2013.11.023

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(1 April), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2014). *An introduction to the event-related potential technique*. https://mitpress.mit.edu/books/introduction-event-related-potential-technique

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, *58*(6), 1–15. https://doi.org/10.1111/psyp.13793

Luking, K. R., Nelson, B. D., Infantolino, Z. P., Sauder, C. L., & Hajcak, G. (2017). Internal consistency of functional magnetic resonance imaging and electroencephalography measures of reward in late childhood and early adolescence. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(3), 289–297. https://doi.org/10.1016/j.bpsc.2016.12.004

Martins, J. S., Bartholow, B. D., Lynne Cooper, M., Irvin, K. M., & Piasecki, T. M. (2019). Interactive effects of naturalistic drinking context and alcohol sensitivity on neural alcohol cue-reactivity responses. *Alcoholism: Clinical and Experimental Research*, *43*(8), 1777–1789. https://doi.org/10.1111/acer.14134

Martins, J. S., Joyner, K. J., McCarthy, D. M., Morris, D. H., Patrick, C. J., & Bartholow, B. D. (2021). *Differential brain responses to alcohol-related and natural rewards is associated with alcohol use and problems: Evidence for reward dysregulation*. Submitted.

McDonough, B. E., & Warren, C. A. (2001). Effects of 12-h tobacco deprivation on event-related potentials elicited by visual smoking cues. *Psychopharmacology (Berl)*, *154*(3), 282–291. https://doi.org/10.1007/s002130000647

McKee, P., Jones-Webb, R., Hannan, P., & Pham, L. (2011). Malt liquor marketing in inner cities: The role of neighborhood racial composition. *Journal of Ethnicity in Substance Abuse*, *10*(1), 24–38. https://doi.org/10.1080/15332640.2011.547793

Meyer, A., Lerner, M. D., De Los Reyes, A., Laird, R. D., & Hajcak, G. (2017). Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches. *Psychophysiology*, *54*(1), 114–122. https://doi.org/10.1111/psyp.12664

Minnix, J. A., Versace, F., Robinson, J. D., Lam, C. Y., Engelmann, J. M., Cui, Y., Brown, V. L., & Cinciripini, P. M. (2013). The late positive potential (LPP) in response to varying types of emotional and cigarette stimuli in smokers: A content comparison. *International Journal of Psychophysiology*, *89*(1), 18–25. https://doi.org/10.1016/j.ijpsycho.2013.04.019

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Moran, T. P., Jendrusina, A. A., & Moser, J. S. (2013). The psychometric properties of the late positive potential during emotion processing and regulation. *Brain Research*, *1516*(2013), 66–75. https://doi.org/10.1016/j.brainres.2013.04.018

Moser, J. S., Moran, T. P., Schroder, H. S., Donnellan, M. B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: A meta-analysis and conceptual framework. *Frontiers in Human Neuroscience*, *7*(August), 1–19. https://doi.org/10.3389/fnhum.2013.00466

Mullen, T. (2012). *CleanLine*. NeuroImaging Tools & Resources Collaboratory. https://www.nitrc.org/projects/cleanline/

Namkoong, K., Lee, E., Lee, C. H., Lee, B. O., & An, S. K. (2004). Increased P3 amplitudes induced by alcohol-related pictures in patients with alcohol dependence. *Alcoholism: Clinical and Experimental Research*, *28*(9), 1317–1323. https://doi.org/10.1097/01.ALC.0000139828.78099.69

Nijs, I. M. T., Franken, I. H. A., & Muris, P. (2008). Food cue-elicited brain potentials in obese and healthy-weight individuals. *Eating Behaviors*, *9*(4), 462–470. https://doi.org/10.1016/j.eatbeh.2008.07.009

Nijs, I. M. T., Muris, P., Euser, A. S., & Franken, I. H. A. (2010). Differences in attention to food and food intake between overweight/obese and normal-weight females under conditions of hunger and satiety. *Appetite*, *54*(2), 243–254. https://doi.org/10.1016/j.appet.2009.11.004

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill Inc.

Olvet, D. M., & Hajcak, G. (2009). Reliability of error-related brain activity. *Brain Research*, *1284*, 89–99. https://doi.org/10.1016/j.brainres.2009.05.079

Palumbo, I. M., Perkins, E. R., Yancey, J. R., Brislin, S. J., Patrick, C. J., & Latzman, R. D. (2020). Toward a multimodal measurement model for the neurobehavioral trait of affiliative capacity. *Personality Neuroscience*, *3*, 1–12. https://doi.org/10.1017/pen.2020.9

Parvaz, M. A., Moeller, S. J., & Goldstein, R. Z. (2016). Incubation of cue-induced craving in adults addicted to cocaine measured by electroencephalography. *JAMA Psychiatry*, *73*(11), 1127–1134. https://doi.org/10.1001/jamapsychiatry.2016.2181

Patrick, C. J., Bernat, E. M., Malone, S. M., Iacono, W. G., Krueger, R. F., & McGue, M. (2006). P300 amplitude as an indicator of externalizing in adolescent males. *Psychophysiology*, *43*(1), 84–92. https://doi.org/10.1111/j.1469-8986.2006.00376.x

Patrick, C. J., Iacono, W. G., & Venables, N. C. (2019). Incorporating neurophysiological measures into clinical assessments: Fundamental challenges and a strategy for addressing them. *Psychological Assessment*, *31*(12), 1512–1529. https://doi.org/10.1037/pas0000713

Patrick, C. J., Venables, N. C., Yancey, J. R., Hicks, B. M., Nelson, L. D., & Kramer, M. D. (2013). A construct-network approach to bridging diagnostic and physiological domains: Application to assessment of externalizing psychopathology. *Journal of Abnormal Psychology*, *122*(3), 902–916. https://doi.org/10.1037/a0032807

Perkins, E. R., Joyner, K. J., Patrick, C. J., Bartholow, B. D., Latzman, R. D., DeYoung, C. G., Kotov, R., Reininghaus, U., Cooper, S. E., Afzali, M. H., Docherty, A. R., Dretsch, M. N., Eaton, N. R., Goghari, V. M., Haltigan, J. D., Krueger, R. F., Martin, E. A., Michelini, G., Ruocco, A. C., … Zald, D. H. (2020). Neurobiology and the hierarchical taxonomy of psychopathology: Progress toward ontogenetically informed and clinically useful nosology. *Dialogues in Clinical Neuroscience*, *22*(1), 51–63. https://doi.org/10.31887/DCNS.2020.22.1/eperkins

Perkins, E. R., Yancey, J. R., Drislane, L. E., Venables, N. C., Balsis, S., & Patrick, C. J. (2017). Methodological issues in the use of individual brain measures to index trait liabilities: The example of noise-probe P3. *International Journal of Psychophysiology*, *111*, 145–155. https://doi.org/10.1016/j.ijpsycho.2016.11.012

Perlman, G., Johnson, W., & Iacono, W. G. (2009). The heritability of P300 amplitude in 18-year-olds is robust to adolescent alcohol use. *Psychophysiology*, *46*(5), 962–969. https://doi.org/10.1111/j.1469-8986.2009.00850.x

Perlman, G., Markin, A., & Iacono, W. G. (2013). P300 amplitude reduction is associated with early-onset and late-onset pathological substance use in a prospectively studied cohort of 14-year-old adolescents. *Psychophysiology*, *50*(10), 974–982. https://doi.org/10.1111/psyp.12081

Petit, G., Kornreich, C., Maurage, P., Noël, X., Letesson, C., Verbanck, P., & Campanella, S. (2012). Early attentional modulation by alcohol-related cues in young binge drinkers: An event-related potentials study. *Clinical Neurophysiology*, *123*(5), 925–936. https://doi.org/10.1016/j.clinph.2011.10.042

Petit, G., Kornreich, C., Verbanck, P., & Campanella, S. (2013). Gender differences in reactivity to alcohol cues in binge drinkers: A preliminary assessment of event-related potentials. *Psychiatry Research*, *209*(3), 494–503. https://doi.org/10.1016/j.psychres.2013.04.005

Piasecki, T. M., Fleming, K. A., Trela, C. J., & Bartholow, B. D. (2017). P3 event-related potential reactivity to smoking cues: Relations with craving, tobacco dependence, and alcohol sensitivity in young adult smokers. *Psychology of Addictive Behaviors*, *31*(1), 61–72. https://doi.org/10.1037/adb0000233

Porjesz, B., & Begleiter, H. (1981). Human evoked brain potentials and alcohol. *Alcoholism: Clinical and Experimental Research*, *5*(2), 304–317. https://doi.org/10.1111/j.1530-0277.1981.tb04904.x

Pronk, T., van Deursen, D. S., Beraha, E. M., Larsen, H., & Wiers, R. W. (2015). Validation of the Amsterdam beverage picture set: A controlled picture set for cognitive bias measurement and modification paradigms. *Alcoholism: Clinical and Experimental Research*, *39*(10), 2047–2055. https://doi.org/10.1111/acer.12853

R Core Team. (2019). *R: A language and environment for statistical computing*. https://www.r-project.org/

Rangaswamy, M., & Porjesz, B. (2014). Understanding alcohol use disorders with neuroelectrophysiology. In E. V. Sullivan, & A. Pfefferbaum (Eds.), *Handbook of Clinical Neurology* (vol *125*, pp. 383–414). Elsevier.

Rentzsch, J., Jockers-Scherübl, M. C., Boutros, N. N., & Gallinat, J. (2008). Test-retest reliability of P50, N100 and P200 auditory sensory gating in healthy subjects. *International Journal of Psychophysiology*, *67*(2), 81–90. https://doi.org/10.1016/j.ijpsycho.2007.10.006

Revelle, W. (2018). *psych: Procedures for psychological, psychometric, and personality research*. https://cran.r-project.org/package=psych

Rietdijk, W. J. R., Franken, I. H. A., & Thurik, A. R. (2014). Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe. *PLoS One*, *9*(7), 3–9. https://doi.org/10.1371/journal.pone.0102672

Robinson, J. D., Versace, F., Engelmann, J. M., Cui, Y., Slapin, A., Oum, R., & Cinciripini, P. M. (2015). The motivational salience of cigarette-related stimuli among former, never, and current smokers. *Experimental and Clinical Psychopharmacology*, *23*(1), 37–48. https://doi.org/10.1037/a0038467

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *8*, 247–291.

Schindler, S., & Straube, T. (2020). Selective visual attention to emotional pictures: Interactions of task-relevance and emotion are restricted to the late positive potential. *Psychophysiology*, *57*(9), 1–14. https://doi.org/10.1111/psyp.13585

Schupp, H. T., Cuthbert, B. N., Bradley, M. M., Cacioppo, J. T., Tiffany, I., & Lang, P. J. (2000). Affective picture processing: The late positive potential is modulated by motivational relevance. *Psychophysiology*, *37*(2), 257–261. https://doi.org/10.1017/S0048577200001530

Senner, N. R., Conklin, J. R., & Piersma, T. (2015). An ontogenetic perspective on individual differences. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 1–9. https://doi.org/10.1098/rspb.2015.1050

Shin, E., Hopfinger, J. B., Lust, S. A., Henry, E. A., & Bartholow, B. D. (2010). Electrophysiological evidence of alcohol-related attentional bias in social drinkers low in alcohol sensitivity. *Psychology of Addictive Behaviors*, *24*(3), 508–515. https://doi.org/10.1037/a0019663

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317. https://doi.org/10.1191/096228098672090967

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. http://www.ncbi.nlm.nih.gov/pubmed/18839484. https://doi.org/10.1037/0033-2909.86.2.420

Sinha, R., Bernardy, N. C., & Parsons, O. A. (1992). Long-term test-retest reliability of event-related potentials in normals and alcoholics. *Biological Psychiatry*, *32*(11), 992–1003. https://doi.org/10.1016/0006-3223(92)90060-D

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Stockburger, J., Schmälzle, R., Flaisch, T., Bublatzky, F., & Schupp, H. T. (2009). The impact of hunger on food cue processing: An event-related brain potential study. *NeuroImage*, *47*(4), 1819–1829. https://doi.org/10.1016/j.neuroimage.2009.04.071

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138. https://doi.org/10.1111/psyp.12629

Trela, C. J., Piasecki, T. M., Bartholow, B. D., Heath, A. C., & Sher, K. J. (2016). The natural expression of individual differences in self-reported level of response to alcohol during ecologically assessed drinking episodes. *Psychopharmacology (Berl)*, *233*(11), 2185–2195. https://doi.org/10.1007/s00213-016-4270-5

Veale, J. F. (2014). Edinburgh handedness inventory—Short form: A revised version based on confirmatory factor analysis. *Laterality*, *19*(2), 164–177. https://doi.org/10.1080/1357650X.2013.783045

Venables, N. C., Foell, J., Yancey, J. R., Kane, M. J., Engle, R. W., & Patrick, C. J. (2018). Quantifying inhibitory control as externalizing proneness: A cross-domain model. *Clinical Psychological Science*, *6*(4), 561–580. https://doi.org/10.1177/2167702618757690

Versace, F., Engelmann, J. M., Deweese, M. M., Robinson, J. D., Green, C. E., Lam, C. Y., Minnix, J. A., Karam-Hage, M. A., Wetter, D. W., Schembre, S. M., & Cinciripini, P. M. (2017). Beyond cue reactivity: Non-drug-related motivationally relevant stimuli are necessary to understand reactivity to drug-related cues. *Nicotine and Tobacco Research*, *19*(6), 663–669. https://doi.org/10.1093/ntr/ntx002

Versace, F., Kypriotakis, G., Basen-Engquist, K., & Schembre, S. M. (2016). Heterogeneity in brain reactivity to pleasant and food cues: Evidence of sign-tracking in humans. *Social Cognitive and Affective Neuroscience*, *11*(4), 604–611. https://doi.org/10.1093/scan/nsv143

Viemose, I., Møller, P., Laugesen, J. L., Schachtman, T. R., Manoharan, T., & Christoffersen, G. R. J. (2013). Appetitive long-term taste conditioning enhances human visually evoked EEG responses. *Behavioural Brain Research*, *253*, 1–8. https://doi.org/10.1016/j.bbr.2013.06.033

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319–350. https://doi.org/10.1016/j.jrp.2004.03.001

Weinberg, A., & Hajcak, G. (2011). Longer term test-retest reliability of error-related brain activity. *Psychophysiology*, *48*(10), 1420–1425. https://doi.org/10.1111/j.1469-8986.2011.01206.x

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer.

Williams, R. H., Zimmerman, D. W., & Zumbo, B. D. (1995). Impact of measurement error on statistical power: Review of an old paradox. *The Journal of Experimental Education*, *63*(4), 363–370. https://doi.org/10.1080/00220973.1995.9943470

Wolz, I., Sauvaget, A., Granero, R., Mestre-Bach, G., Baño, M., Martín-Romera, V., De Las, V., Heras, M., Jiménez-Murcia, S., Jansen, A., Roefs, A., & Fernández-Aranda, F. (2017). Subjective craving and event-related brain response to olfactory and visual chocolate cues in binge-eating and healthy individuals. *Scientific Reports*, *7*(February), 1–10. https://doi.org/10.1038/srep41736

Yancey, J. R., Venables, N. C., & Patrick, C. J. (2016). Psychoneurometric operationalization of threat sensitivity: Relations with clinical symptom and physiological response criteria. *Psychophysiology*, *53*(3), 393–405. https://doi.org/10.1111/psyp.12512

Yoon, H. H., Malone, S. M., & Iacono, W. G. (2015). Longitudinal stability and predictive utility of the visual P3 response in adults with externalizing psychopathology. *Psychophysiology*, *52*(12), 1632–1645. https://doi.org/10.1111/psyp.12548

Zoon, H. F. A., Ohla, K., de Graaf, C., & Boesveldt, S. (2018). Modulation of event-related potentials to food cues upon sensory-specific satiety. *Physiology and Behavior*, *196*(December 2017), 126–134. https://doi.org/10.1016/j.physbeh.2018.08.020

## 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 oddball 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. (2022). Internal consistency and test–retest reliability of the P3 event-related potential (ERP) elicited by alcoholic and non-alcoholic beverage pictures. *Psychophysiology*, 59, e13967. https://doi.org/10.1111/psyp.13967