PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2017), 601–607. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12813

Reliability of the electrocortical response to gains and losses in the doors task

AMANDA R. LEVINSON, BRITTANY C. SPEED, ZACHARY P. INFANTOLINO, AND GREG HAJCAK

Department of Psychology, Stony Brook University, Stony Brook, New York, USA

Abstract

The ability to differentiate between rewards and losses is critical for motivated action, and aberrant reward and loss processing has been associated with psychopathology. The reward positivity (RewP) and feedback negativity (FN) are ERPs elicited by monetary gains and losses, respectively, and are promising individual difference measures. However, few studies have reported on the psychometric properties of the RewP and FN—crucial characteristics necessary for valid individual difference measures. The current study examined the internal consistency and 1-week test-retest reliability of the RewP and FN as elicited by the doors task among 59 young adults. The RewP, FN, and their difference score (ΔRewP) all showed significant correlations between Time 1 and Time 2. The RewP and FN also achieved acceptable internal consistency at both time points within 20 trials using both Cronbach's α and a generalizability theory-derived dependability measure. Internal consistency for ΔRewP was notably weaker at both time points, which is expected from two highly intercorrelated constituent scores. In conclusion, the RewP and FN have strong psychometric properties in a healthy adult sample. Future research is needed to assess the psychometric properties of these ERPs in different age cohorts and in clinical populations.

Descriptors: EEG/ERPs, RewP, FN, Doors task

The ability to respond differentially to positive outcomes (i.e., rewards) compared to negative outcomes (i.e., losses) represents a critical mechanism of goal-oriented behavior and learning. Alterations in the ability to detect and process rewards and losses have been associated with multiple forms of psychopathology. For example, increased sensitivity to reward has been associated with substance abuse disorders (Baker, Wood, & Holroyd, 2016; Di Chiara & Bassareo, 2007) and attention deficits (Holroyd, Baker, Kerns, & Müller, 2008), whereas a decreased sensitivity to reward has been associated with depression (Forbes & Dahl, 2005). Recent research has suggested that dysfunctional reward and loss processing may actually be a precursor to the development of psychopathology (Admon et al., 2012; Nelson et al., 2013; Treadway, Buckholtz, & Zald, 2013). Thus, there is increasing effort to better understand individual differences in neural response to reward and loss.

The ERPs following the presentation of rewards and losses are characterized by a relative positivity and negativity, respectively, maximal at frontocentral sites between 250–350 ms after feedback. The relative negativity following losses has been referred to as the *feedback negativity* (FN), whereas the relative positivity following gains has been described as the *reward positivity* (RewP). In light of recent evidence suggesting that variability in the ERP difference between gains and losses may be driven by neural response to rewards, we refer to the gain-loss difference as the Δ*RewP* (Bress & Hajcak, 2013; Carlson, Foti, Mujica-Parodi, Harmon-Jones, & Hajcak, 2011; Foti, Weinberg, Dien, & Hajcak, 2011; Holroyd, Pakzad-Vaezi, & Krigolson, 2008).

The most common laboratory tasks used to elicit neural responses to gains and losses are simple guessing tasks in which outcomes are equally probable. One example is the doors task (Proudfit, 2015), in which participants are presented with an image of two doors and asked to select one; feedback indicating either monetary gain or loss is subsequently delivered with equal probability. The simplicity of the doors task has made it feasible to use in various clinical populations (e.g., Gong et al., 2014; Horan, Foti, Hajcak, Wynn, & Green, 2012) and in children (e.g., Kessel et al., 2016; Kujawa, Proudfit, & Klein, 2014), while still being a potent probe in healthy adults (Weinberg, Riesel, & Proudfit, 2014). Furthermore, this task has repeatedly been used in studies demonstrating associations between ERP measures and individual differences. For instance, studies employing the doors task have found that a blunted ΔRewP is associated with greater symptoms of depression (e.g., Bress & Hajcak, 2013; Bress, Meyer, & Proudfit, 2015; Bress, Smith, Foti, Klein, & Hajcak, 2012; Foti & Hajcak, 2009, 2010; Foti, Kotov, Klein, & Hajcak, 2011). Indeed, one study even found that a blunted ΔRewP in adolescent girls, aged 15–17, predicted the development of first-onset depressive episodes and increases in depressive symptoms at 2-year follow-up (Bress, Foti,

The authors would like to thank Emily Hale-Rude and Elizabeth Parisi for their tireless work in collecting this data. We would also like to thank our research participants for making this research possible by contributing their time and effort.

Address correspondence to: Amanda R. Levinson, Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500, USA. E-mail: amanda.levinson@stonybrook.edu

Kotov, Klein, & Hajcak, 2013; Nelson, Perlman, Klein, Kotov, & Hajcak, 2016). Taken together, these studies indicate that ERP activity recorded during the doors task may be a clinically useful probe for the assessment of individual differences in reward processing reflected in the ERP.

To date, few studies have reported on the psychometric properties of the FN or RewP. One study reported good short-term test-retest reliability that was robust to varying psychological conditions (Segalowitz et al., 2010), and another found good internal consistency within 20 trials for young adults and within 50 trials for older adults (Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011). However, both studies employed more complicated tasks than the doors task to elicit the FN and RewP. Only one study—conducted in girls aged 8–13—has reported on the psychometrics of ERPs from the doors task and found moderate test-retest reliability over 2 years for the RewP and FN (rs = .67 and .64, respectively), but relatively poor test-retest reliability of ΔRewP (r = .18; Bress et al., 2015).

The current study aimed to fill in basic gaps in our understanding of the psychometric properties of these ERP measures. Specifically, we employed a commonly available sample (i.e., college students) and a popular simple task (i.e., the doors guessing paradigm) to assess psychometrics. We expect that, with this straightforward design, we will find that these ERP components reach acceptable reliability levels. The current study examined the internal reliability and shorter-term (i.e., 1 week) test-retest reliability of the FN, RewP, and ΔRewP elicited in the doors task in young adults. Additionally, the psychometric properties of FN and RewP were examined as a function of the number of trials contributing to ERP averages (Marco-Pallares et al., 2011)—these analyses were intended to provide guidance regarding how many trials are necessary for a reliable FN and RewP.

Method

Study Design

In this study, participants completed a computer-based guessing task, with task administrations spaced approximately 1 week apart (mean time between testing = 6.81 ± 1.24 days). EEG was continuously recorded during task administration.

Participants

Sixty-eight Stony Brook University undergraduates were drawn from the psychology department research participation pool, which consisted of 1,528 undergraduates (65.7% female; M_{age} = 20.3; SD_{age} = 2.66; 35.0% Asian, 8.4% African American, 40.0% Caucasian, 16.6% other). Of the 68 participants who enrolled, eight did not return for the second testing session and one was excluded from analysis due to poor EEG data quality. Thus, a total of 59 participants were included in the final analysis. Participants received research participation credit and were compensated for their participation ($25 for completing both visits).

Tasks and Materials

The doors task. In the doors task, the image of two doors are presented on a computer screen at the beginning of each trial; participants are instructed to select one of the doors by clicking either the left or right mouse button. Participants are told that one door will result in winning money and the other will result in losing money. In order to increase the salience of the task, participants are ensured that they will receive their total winnings at the end of the task. The task included three blocks of 20 trials (i.e., 60 trials total). The blocks were separated by participant-timed breaks, during which the instructions "Pause—Click mouse when ready to continue" appeared on the screen until the participant clicked the left or right mouse button. There were an equal number of win and loss trials (30 each) that occurred in a randomized order. The sequence and timing of the stimuli in each trial were as follows: (1) a fixation cross is presented for 500 ms, (2) an image of two doors is presented until the participants make their selection by clicking the left or right mouse button, (3) a fixation cross is presented for 1,500 ms, (4) an upward green arrow or a downward red arrow is presented for 2,000 ms to indicate monetary gain or loss, respectively, (5) a fixation cross is presented for 1,500 ms, (6) the words "Click for next round" appear on the screen until the participant clicks either mouse button. Because monetary losses are experienced as twice as valuable as monetary gains (Tversky & Kahneman, 1992), gains were worth $.50 whereas losses were worth $.25 to equate the subjective value of gains and losses. The total duration of the task ranged between 5 and 7 min depending on how long participants took between trials and blocks of trials. Finally, all participants were paid their actual winnings at the completion of the task (i.e., $7.50), as well as a $10.00 bonus for completing both lab visits.

Psychophysiological Recording and Data Reduction

EEG was continuously recorded from a custom 34-electrode elastic cap configured according to the 10/20 system, using the ActiveTwo BioSemi System (BioSemi, Amsterdam, The Netherlands). Additional data were recorded from the mastoids, and the horizontal and vertical electrooculogram was collected from electrodes placed 1 cm from the outer corners of the eyes and 1 cm above and below the right eye, respectively. Recordings were amplified at the electrode with a gain of one, and the data were digitized at 24-bit resolution with a sampling rate of 1024 Hz using a low-pass fifth-order sinc filter with a half-power cutoff of 204 Hz. EEG was measured online with respect to a common mode sense active electrode, forming a monopolar channel. BrainVision Analyzer (Brain Products, Munich, Germany) was used for offline analysis. Data were rereferenced to the average of the recordings from the left and right mastoid channels, and a band-pass filter with cutoffs of 0.1 Hz and 30 Hz was applied. Ocular artifacts were corrected using the Gratton and colleagues procedure (Gratton, Coles, & Donchin, 1983). Physiological artifacts were corrected using a semiautomated procedure with a maximum allowed voltage step of 50 μV between sampling points, a maximal voltage difference of 300 μV in a given trial, and a maximum allowed voltage of .5 μV within an interval of 100 ms. Remaining artifacts were rejected manually based on visual inspection of the data.

The EEG was segmented into feedback-locked epochs from −200 to 800 ms separately for reward and nonreward feedback. The baseline period (i.e., average activity 200 ms before feedback presentation) was subtracted from all data points. The FN and RewP were quantified as the mean activity from 250 to 350 ms after loss and gain feedback onset, respectively, at FCz. Separate loss and gain averages were created across all trials, as well as split-half averages comparing odd trials and even trials. Moreover, analyses also focused on psychometric properties of the FN and RewP as increasing trials were included in averages. For these analyses, the FN and RewP were scored on each trial. For example,

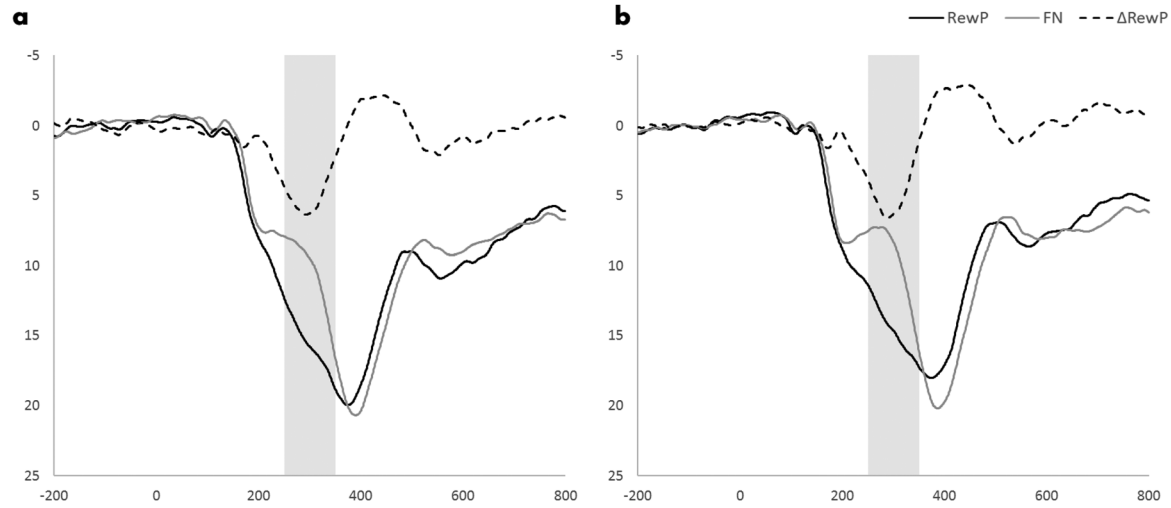

Figure 1. This figure presents grand-averaged waveforms at electrode site FCz at Time 1 (a) and Time 2 (b) of the reward positivity (RewP; solid black lines), feedback negativity (FN; solid gray lines), and their difference wave (ΔRewP; dashed black lines) calculated as RewP minus FN. The vertical axis presents amplitude (in μV) and the horizontal axis presents time (in ms), where 0 ms represents the onset of feedback.

internal reliability of the FN for the 10-trial average was evaluated in terms of Cronbach's alpha considering only the first 10 loss trials with good (i.e., artifact-free) data. Finally, we also report on internal reliability and test-retest reliability of the ΔRewP, defined as the RewP minus the FN.

Data analysis. The majority of our statistical analyses were conducted using SPSS (version 22.0). Using best existing estimates of the reliability of these ERP components, we conducted power analyses and found that our sample of 59 participants provides sufficient power for the following psychometric analyses. Test-retest reliability was assessed using Pearson's *r*, a measure of interindividual stability, and interclass coefficients (ICCs), a measure of score agreement. For ICCs, we used a two-way, mixed-effects model (Model 3 in Shrout & Fleiss, 1979) with the more conservative measure of absolute agreement. We also report 95% confidence intervals (CIs) for ICCs. Additionally, we report generalizability theory (G theory) measures of overall dependability that were computed using the MATLAB (version R2016a) toolbox, ERA Toolbox (Clayson & Miller, 2016a).

Internal consistency of the RewP and the FN was examined using two approaches derived from classical test theory. First, we computed split-half reliability by calculating the correlation between averages based on odd- and even-numbered trials, corrected using the Spearman-Brown prophecy formula (Nunnally, Bernstein, & Berge, 1967). One of the benefits of this approach is that it includes all available data. That is, all available data are included in either the odd or even averages. The drawback of split-half reliability analyses, however, is that it is specific to one possible way of splitting the data (i.e., odd-numbered vs. even-numbered trials). Therefore, we also computed Cronbach's α, which is roughly equivalent to the mean of all possible split-half correlations. The primary drawback of Cronbach's α is that it requires all participants to have the same number of trials. Thus, as trial count increases, some participants are excluded from analyses. All participants had a minimum of 25 good gain and loss trials. Overall, we present mean amplitudes and internal reliabilities (Cronbach's αs) at Time 1 and Time 2, and test-retest reliability of the RewP and FN from Time 1 to Time 2 (i.e., Pearson's *r*; mean time between testing = 6.81 ± 1.24 days), both overall and as a function of increasing trials. The latter analyses provide a sense of how many trials are required to achieve good psychometric values for these ERP components. Recent work suggests that psychometric properties of ERPs may be better understood using G theory estimates of dependability (Clayson & Larson, 2013; Clayson & Miller, 2016b), as opposed to the more traditional classical test theory measures. G theory analyses are able to simultaneously consider multiple sources of error variance, and may be a useful tool in understanding the contribution of each of these possible sources to our data. Thus, for comparison, we included these measures.

The overall internal reliability of the difference score (ΔRewP) was estimated using an adjusted α formula (Furr & Bacharach, 2013). Standard measures of reliability (i.e., split-half or Cronbach's α) have been suggested to be inappropriate for difference scores because the reliability of difference scores is influenced by the reliabilities, variances, and intercorrelations of the two contributing measures (in this case the RewP and the FN). In contrast, the Furr and Bacharach (2013) adjusted α formula[1] accounts for these factors, providing a more accurate assessment of internal reliability.

Results

Cross-Sectional Psychometrics of RewP and FN

ERP waveforms are depicted in Figure 1. At both Time 1 and Time 2, the mean amplitudes of the RewP and FN gradually stabilized over the course of the task, leveling off after 10–20 trials (Figure 2). Classical test theory-derived measures showed that the RewP and FN achieved good to excellent internal consistency at both Time 1 and Time 2 (see Table 1), as assessed using both split-half reliability (*r*s ranged from .71–.89) and Cronbach's α (αs ranged from .81–.88; Gliem & Gliem, 2003). Similarly, generalizability

1. The formula for adjusted alpha proposed by Furr & Bacharach (2013) is

$$R_{\mathrm{d}} = \frac{s_{X_{\mathrm{o}}}^2 R_{XX} + s_{Y_{\mathrm{o}}}^2 R_{YY} - 2r_{X_{\mathrm{o}}Y_{\mathrm{o}}} s_{X_{\mathrm{o}}} s_{Y_{\mathrm{o}}}}{s_{X_{\mathrm{o}}}^2 + s_{Y_{\mathrm{o}}}^2 - 2r_{X_{\mathrm{o}}Y_{\mathrm{o}}} s_{X_{\mathrm{o}}} s_{Y_{\mathrm{o}}}}.$$

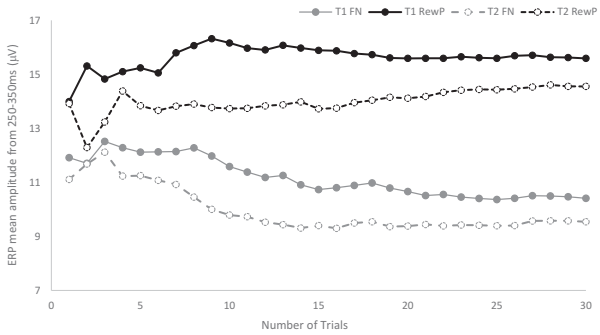

Figure 2. This figure presents the average amplitude (in μV) of the reward positivity (RewP) and the feedback negativity (FN) as a function of the number of trials. Four separate lines are presented to show data for (1) the FN at Time 1 (solid gray line), (2) the RewP at Time 1 (solid black line), (3) the FN at Time 2 (dashed gray line), and (4) the RewP at Time 2 (dashed black line). Time 1 (T1) and Time 2 (T2) are spaced an average of 6.81 ± 1.24 days apart.

theory-derived measures of overall dependability for RewP and FN at each time point (Table 1) ranged from .79 to .88.

At Time 1, an acceptable internal consistency (Cronbach's $\alpha > .7$) was reached and maintained for both the RewP and FN after 10 trials (Figure 3; final values based on 51 participants without missing data). At Time 2, acceptable internal consistency for the RewP (14 trials) and FN (20 trials) required slightly more trials (Figure 3; final values based on the 53 participants without missing data).[2] Similarly, at Time 1, minimum recommended dependability scores of 0.7 or above (Clayson & Miller, 2016b) were reached for FN and RewP after 10 trials, while at Time 2, this threshold was reached for FN after 19 trials and for RewP after 15 trials. ΔRewP showed lower internal consistency than either of the two constituent measures as assessed using adjusted α (Furr & Bacharach, 2013).

Test-Retest Reliability

As shown in Table 2, the FN, RewP, and ΔRewP[3] were all significantly correlated from Time 1 to Time 2. The effect sizes of these correlations were large, moderate, and small, respectively (Cohen, 1992). Similarly, ICCs for the FN, RewP, and ΔRewP were classified as good agreement, moderate agreement, and poor agreement (Portney & Watkins, 2009). In examining test-retest reliabilities of the RewP and FN as a function of increasing trial counts (Figure 4), test-retest reliability of the RewP stabilized at its maximum after approximately 25 trials, while the test-retest reliability of the FN continued to increase over the full 30 trials.

Table 1. *Cross-Sectional Measures of Reliability or Dependability of ERPs to Gains, Losses, and the Difference Score at Time 1 and Time 2*

	Measure	FN	RewP	ΔRewP
Time 1	Split-half	0.89	0.89	–
	Cronbach's α	0.83	0.88	–
	[95% CIs]	[.75, .89]	[0.82, 0.92]	
	Adjusted α	–	–	0.28
	Dependability	0.88	0.87	–
	[95% CIs]	[0.84, 0.92]	[0.83, 0.91]	
Time 2	Split-half	0.71	0.82	–
	Cronbach's α	0.81	0.83	–
	[95% CIs]	[0.73, 0.88]	[0.76, 0.89]	
	Adjusted α	–	–	0.38
	Dependability	0.79	0.83	–
	[95% CIs]	[0.70, 0.86]	[0.76, 0.89]	

Discussion

This study used the doors task to examine the internal consistency and test-retest reliability of the RewP and FN to monetary gains and losses in a sample of young adults. Consistent with previously reported findings (Bress et al., 2015), both the RewP and FN achieved good to excellent internal reliability (all values > .7) within 20 trials of the doors task at both assessments based on split-half reliability, Cronbach's α, and G theory measures of overall dependability (Clayson & Miller, 2016b; Gliem & Gliem, 2003). The overall 1-week test-retest correlations were high for both the RewP and FN and increased as a function of the number of trials, stabilizing after roughly 15 trials. Similarly, test-retest ICCs showed moderate to strong agreement, using the stringent absolute agreement criteria.

For ΔRewP, internal consistency was notably lower than for the FN and RewP at each assessment (values of .284 and .375; Bress et al., 2015; cf. Marco-Pallares et al., 2011). In discussing the adjusted α formula, Furr and Bacharach (2013) highlight that if the constituent measures of a difference score are strongly intercorrelated or if they have unequal variances, the reliability of the difference score will necessarily be adversely affected. Given our particular data set, both of these factors are likely contributing to the poor internal consistency of ΔRewP.[4] Along similar lines, the test-retest reliability of ΔRewP was smaller than the FN and RewP.

A current priority in mental health research is to identify and standardize biomarkers of psychopathology (Insel, 2014). This research may ultimately help with the accurate detection of psychopathology and the planning of more targeted, personalized treatments. As with any individual difference measure, in order to be clinically useful, a biomarker must have good psychometric

2. It is possible that the reliability of these measures may shift over the course of the session for many reasons, including fatigue, changes in attention, or changes in the participant's comfort. In order to assess for these cross-session changes, we calculated Cronbach's αs separately for the first and second half of trials at each session. At Time 1, first-half gains ($\alpha = .80$) and second-half gains ($\alpha = .78$) were very close in reliability values as were first-half losses ($\alpha = .80$) and second-half losses ($\alpha = .74$). At Time 2, the first and second halves of trials differed somewhat more but still fell in a similar range. Specifically, first-half gains ($\alpha = .75$) and second-half gains ($\alpha = .67$) were both near cutoff for acceptability, as were first-half losses ($\alpha = .62$) and second-half losses ($\alpha = .71$).

3. While ERP difference scores are traditionally calculated by subtracting one of the constituent ERPs from the other, recent studies have begun to use residualized scores instead. Residualized scores may provide a more reliable estimate of change than either raw measures (i.e., RewP, FN) or subtraction-based difference scores (i.e., ΔRewP; Cronbach & Furby, 1970; DuBois, 1957; Meyer, Lerner, Reyes, Laird, & Hajcak, in press). Therefore, we also examined the test-retest reliability of the residualized RewP, calculated using linear regression as the residual response to gains adjusting for losses. Similar to our findings using the subtraction-based RewP, this analysis found a small correlation ($r = .22$) between the Time 1 and Time 2 residualized RewP, which trended toward significance ($p < .10$).

4. At both Time 1 and Time 2, gains and losses were highly correlated with one another (rs = .82 and .72, respectively). At Time 1, variances for gain and loss trials were 44.17 and 29.80, respectively, while at Time 2, variances for gain and loss trials were 34.91 and 27.99, respectively.

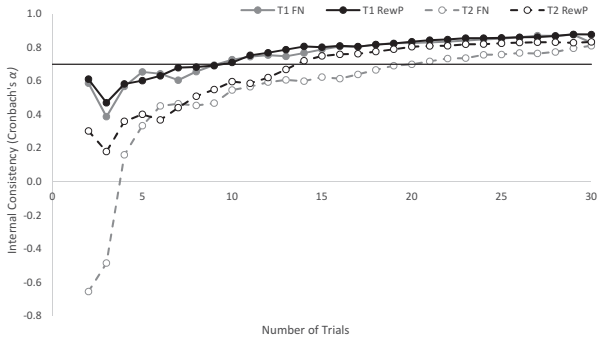

Figure 3. This figure presents the internal consistencies (measured using Cronbach's α) of the reward positivity (RewP) and the feedback negativity (FN) as a function of the number of trials. Four separate lines are presented to show data for (1) the FN at Time 1 (solid gray line), (2) the RewP at Time 1 (solid black line), (3) the FN at Time 2 (dashed gray line), and (4) the RewP at Time 2 (dashed black line). Time 1 (T1) and Time 2 (T2) are spaced an average of 6.81 ± 1.24 days apart. The accepted cutoff in the literature for acceptable reliability (Cronbach's $\alpha \geq 0.7$) is shown as a solid black horizontal line (Gliem & Gliem, 2003).

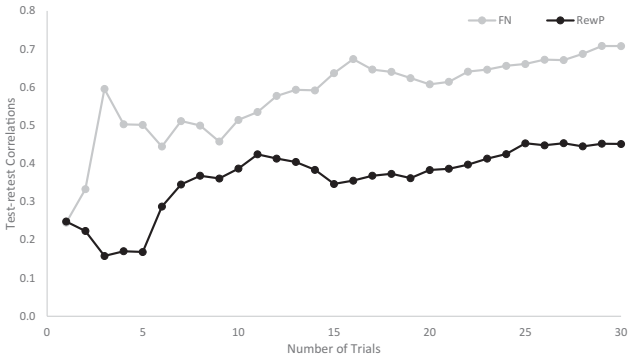

Figure 4. This figure presents the test-retest reliability of the reward positivity (RewP) and the feedback negativity (FN) as a function of the number of trials. Separate lines are shown for the RewP (shown in black) and the FN (shown in gray). Reliability is assessed using Pearson's correlations (*r*) between Time 1 and Time 2. Time 1 and Time 2 are spaced an average of 6.81 ± 1.24 days apart.

properties (Barch & Mathalon, 2011; Cronbach & Meehl, 1955). The reliability of a measure can be thought to index how much true score is contained within that measure—in other words, how much of the variance in the measure is due to the phenomenon being measured, as opposed to error variance. Thus, the internal consistency of a measure also puts a ceiling on how much true score is available to be associated with other variables (e.g., clinical correlates). The RewP, FN, and ΔRewP have all been examined as potential candidate biomarkers of various forms of psychopathology, including depressive symptoms, anxiety symptoms, and behavioral problems (Bress et al., 2012, 2013; Foti & Hajcak, 2009; Kessel et al., 2016; Kessel, Kujawa, Hajcak Proudfit, & Klein, 2015; Thoma, Edel, Suchan, & Bellebaum, 2015; Zhu et al., 2014). The most established and replicated of these associations is between depressive symptoms and the RewP or ΔRewP, but not the FN (Bress et al., 2013; Foti & Hajcak, 2009, 2010; Foti, Kotov et al., 2011). The current study suggests that both FN and RewP have excellent psychometric properties in terms of both cross-sectional and test-retest reliability. As has been found previously in the literature, the difference score (ΔRewP) had lower reliability scores. Thus, difference scores are more constrained in the amount of true score available. A critical issue then is how much reliable variance relates to other clinical variables. In the case of the ΔRewP, although this measure may have less true score variance than its constituent scores, it appears to relate as well or better than the RewP or FN to individual differences in depression. This could be the case if a larger portion of the ΔRewP true score than RewP true score variance related to depression.

In the current study, we also found that increased trials were needed at Time 2 compared to Time 1 in order to reach threshold reliability. This indicates that, in the current dataset, there was an increased error variance (i.e., noise) at Time 2. One possible explanation for this finding is that participants may have been less engaged in the gambling task at second administration. Nevertheless, reliability values at Time 1 and Time 2 are similar enough that this between-session difference may also be spurious. Future research is needed to replicate this difference between first and second task administration.

Internal consistencies and test-retest reliabilities of the RewP and FN in this sample were adequate and similar to previously reported results (Bress et al., 2015; Segalowitz et al., 2010). To date, two papers have previously reported on the internal reliability of ΔRewP, and their results were widely divergent from one another (Bress et al., 2015; Marco-Pallares et al., 2011). In the current study and in line with the previous report that employed the doors task (Bress et al., 2015), ΔRewP had relatively poor internal consistency. In contrast to these findings, Marco-Pallares and colleagues (2011) reported significantly higher internal consistency for ΔRewP, which may be an artifact of the way in which they calculated internal consistency. Specifically, Marco-Pallares et al. calculated consistency (Cronbach's α) by comparing two values: ΔRewP averaged across the first *n* trials (where *n* increased from 1 to 59) and ΔRewP averaged across all 60 trials. However, as the

Table 2. *Means, Standard Deviation, and Test-Retest Reliability for ERP Measures from the Doors Task at Time 1 and Time 2*

	Time 1			Time 2				
		95% CI			95% CI		Time 1 to Time 2 comparison	
	Mean (*SD*)	Lower bound	Upper bound	Mean (*SD*)	Lower bound	Upper bound	Pearson's *r*	ICC [95% CI]
RewP	15.60 (6.78)	13.83	17.36	14.56 (6.04)	12.98	16.13	.45**	.62 [.36, .77]
FN	10.42 (6.87)	8.63	12.21	9.55 (5.18)	8.20	10.90	.71**	.81 [.68, .89]
ΔRewP	5.18 (4.12)	4.11	6.25	5.01 (4.25)	3.90	6.12	.27*	.43 [.04, .66]

Note. Qualitative cutoffs for ICCs are as follows: ICCs > .50: poor agreement, .50 < ICC < .75: moderate agreement, .75 < ICC < .90: good agreement, ICC > .90: excellent agreement (Portney & Watkins, 2009). ICC = intraclass correlation coefficient; CI = confidence interval.
*$p < .05$. **$p < .01$.

subset of trials approaches the full 60 trials, these subaverages become mathematically more similar (i.e., eventually identical) to the overall average. Rather than assessing internal reliability, the measure employed by Marco-Pallares and colleagues (2011) actually provides an estimate of how well subaverages based on fewer trials relate to overall ERP averages.

Taken together, our data indicate that the RewP and FN, as elicited by the doors task, have good psychometric properties in a healthy adult sample within 20 trials. Furthermore, we find that the particular analytical strategy employed did not produce great variation in our data. Specifically, both classical test theory reliability measures and G theory dependability measures produced very similar psychometric profiles of the data. The somewhat weaker psychometric properties of ΔRewP are expected, given that reliabilities are expected to be weaker for difference scores (Furr & Bacharach, 2013; Williams & Zimmerman, 1977). Furthermore, the strong correlations between the two constituent measures (RewP and FN) and the unequal variances of those measures necessarily yield a smaller α for ΔRewP (Furr & Bacharach, 2013). The influence of these factors is apparent in the substantially smaller internal consistency at Time 1, when the RewP and FN were more highly correlated and their variances were more unequal, compared to Time 2. While the findings in the present sample indicate that, in some cases, future research employing the doors task may be justified in reducing participant burden by employing fewer trials, the number of trials required to obtain an acceptably reliable averaged ERP value will vary from sample to sample. Future research might further examine the psychometric properties of these ERPs in clinical samples and in other age groups (children and older adults), as psychometric properties may function differently in these populations. Furthermore, it is recommended that psychometric properties of ERPs (i.e., internal consistency) are reported in each data set and sample, as a standard practice in ERP research, especially for studies examining the relationships between ERPs and other measures of individual differences. Finally, we recommend that future research employing these (or any) ERP components report the number of trials used to compute averaged values, as this study illustrates how the reliability of these measures shifts as a function of trial number.

References

Admon, R., Lubin, G., Rosenblatt, J. D., Stern, O., Kahn, I., Assaf, M., & Hendler, T. (2012). Imbalanced neural responsivity to risk and reward indicates stress vulnerability in humans. *Cerebral Cortex*, *23*, 28–35. doi: 10.1093/cercor/bhr369

Baker, T. E., Wood, J. M., & Holroyd, C. B. (2016). Atypical valuation of monetary and cigarette rewards in substance dependent smokers. *Clinical Neurophysiology*, *127*(2), 1358–1365. doi: 10.1016/j.clinph.2015.11.002

Barch, D. M., & Mathalon, D. H. (2011). Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: Psychometric and quality assurance considerations. *Biological Psychiatry*, *70*(1), 13–18. doi: 10.1016/j.biopsych.2011.01.004

Bress, J. N., Foti, D., Kotov, R., Klein, D. N., & Hajcak, G. (2013). Blunted neural response to rewards prospectively predicts depression in adolescent girls. *Psychophysiology*, *50*(1), 74–81. doi: 10.1111/j.1469-8986.2012.01485.x

Bress, J. N., & Hajcak, G. (2013). Self-report and behavioral measures of reward sensitivity predict the feedback negativity. *Psychophysiology*, *50*(7), 610–616. doi: 10.1111/psyp.12053

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. doi: 10.1017/S0954579414001400

Bress, J. N., Smith, E., Foti, D., Klein, D. N., & Hajcak, G. (2012). Neural response to reward and depressive symptoms in late childhood to early adolescence. *Biological Psychology*, *89*(1), 156–162. doi: 10.1016/j.biopsycho.2011.10.004

Carlson, J. M., Foti, D., Mujica-Parodi, L. R., Harmon-Jones, E., & Hajcak, G. (2011). Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: A combined ERP and fMRI study. *NeuroImage*, *57*(4), 1608–1616. doi: 10.1016/j.neuroimage.2011.05.037

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.

Clayson, P. E., & Miller, G. A. (2016a). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. *International Journal of Psychophysiology*. Advance online publication. doi: 10.1016/j.ijpsycho.2016.10.012

Clayson, P. E., & Miller, G. A. (2016b). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*. Advance online publication. doi: 10.1016/j.ijpsycho.2016.09.005

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi: 10.1037/0033-2909.112.1.155

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68–80. doi: 10.1037/h0029382

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

Di Chiara, G., & Bassareo, V. (2007). Reward system and addiction: What dopamine does and doesn't do. *Current Opinion in Pharmacology*, *7*(1), 69–76. doi: 10.1016/j.coph.2006.11.003

DuBois, P. H. (1957). *Multivariate correlational analysis*. Oxford, England: Harper.

Forbes, E. E., & Dahl, R. E. (2005). Neural systems of positive affect: Relevance to understanding child and adolescent depression? *Development and Psychopathology*, *17*(03), 827–850. doi: 10.1017/S095457940505039X

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

Foti, D., & Hajcak, G. (2010). State sadness reduces neural sensitivity to nonrewards versus rewards. *NeuroReport*, *21*(2), 143–147. doi: 10.1097/WNR.0b013e3283356448

Foti, D., Kotov, R., Klein, D. N., & Hajcak, G. (2011). Abnormal neural sensitivity to monetary gains versus losses among adolescents at risk for depression. *Journal of Abnormal Child Psychology*, *39*(7), 913–924. doi: 10.1007/s10802-011-9503-9

Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principal components analysis and source localization of the feedback negativity. *Human Brain Mapping*, *32*(12), 2207–2216. doi: 10.1002/hbm.21182

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Gliem, R. R., & Gliem, J. A. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales*. Retrieved from https://scholarworks.iupui.edu/bitstream/handle/1805/344/gliem+&+gliem.pdf?sequence=1.

Gong, J., Yuan, J., Wang, S., Shi, L., Cui, X., & Luo, X. (2014). Feedback-related negativity in children with two subtypes of attention deficit hyperactivity disorder. *PLOS ONE*, *9*(6), e99570. doi: 10.1371/journal.pone.0099570

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. doi: 10.1016/0013-4694(83)90135-9

Holroyd, C. B., Baker, T. E., Kerns, K. A., & Müller, U. (2008). Electrophysiological evidence of atypical motivation and reward processing in children with attention-deficit hyperactivity disorder. *Neuropsychologia*, *46*(8), 2234–2242. doi: 10.1016/j.neuropsychologia.2008.02.011

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*(5), 688–697. doi: 10.1111/j.1469-8986.2008.00668.x

Horan, W. P., Foti, D., Hajcak, G., Wynn, J. K., & Green, M. F. (2012). Impaired neural response to internal but not external feedback in schizophrenia. *Psychological Medicine*, *42*(08), 1637–1647. doi: 10.1017/S0033291711002819

Insel, T. R. (2014). The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry. *American Journal of Psychiatry*, *171*, 395–397. doi: 10.1176/appi.ajp.2014.14020138

Kessel, E. M., Dougherty, L. R., Kujawa, A., Hajcak, G., Carlson, G. A., & Klein, D. N. (2016). Longitudinal Associations between preschool disruptive mood dysregulation disorder symptoms and neural reactivity to monetary reward during preadolescence. *Journal of Child and Adolescent Psychopharmacology*, *26*(2), 131–137. doi: 10.1089/cap.2015.0071

Kessel, E. M., Kujawa, A., Hajcak Proudfit, G., & Klein, D. N. (2015). Neural reactivity to monetary rewards and losses differentiates social from generalized anxiety in children. *Journal of Child Psychology and Psychiatry*, *56*(7), 792–800. doi: 10.1111/jcpp.12355

Kujawa, A., Proudfit, G. H., & Klein, D. N. (2014). Neural reactivity to rewards and losses in offspring of mothers and fathers with histories of depressive and anxiety disorders. *Journal of Abnormal Psychology*, *123*(2), 287–297. doi: 10.1037/a0036285

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*(6), 852–860. doi: 10.1111/j.1469-8986.2010.01152.x

Meyer, A., Lerner, M. D., Reyes, A., Laird, R. D., & Hajcak, G. (in press). Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches. *Psychophysiology*.

Nelson, B. D., McGowan, S. K., Sarapas, C., Robison-Andrew, E. J., Altman, S. E., Campbell, M. L., … Shankman, S. A. (2013). Biomarkers of threat and reward sensitivity demonstrate unique associations with risk for psychopathology. *Journal of Abnormal Psychology*, *122*(3), 662–671. doi: 10.1037/a0033982

Nelson, B. D., Perlman, G., Klein, D. N., Kotov, R., & Hajcak, G. (2016). Blunted neural response to rewards as a prospective predictor of the development of depression in adolescent girls. *American Journal of Psychiatry*. Advance online publication. doi: 10.1176/appi.ajp.2016.15121524

Nunnally, J. C., Bernstein, I. H., & Berge, J. M. T. (1967). *Psychometric Theory*. New York, NY: McGraw-Hill.

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459. doi: 10.1111/psyp.12370

Segalowitz, S. J., Santesso, D. L., Murphy, T. I., Homan, D., Chantziantoniou, D. K., & Khan, S. (2010). Retest reliability of medial frontal negativities during performance monitoring. *Psychophysiology*, *47*(2), 260–270. doi: 10.1111/j.1469-8986.2009.00942.x

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi: 10.1037/0033-2909.86.2.420

Thoma, P., Edel, M.-A., Suchan, B., & Bellebaum, C. (2015). Probabilistic reward learning in adults with attention deficit hyperactivity disorder—An electrophysiological study. *Psychiatry Research*, *225*(1), 133–144. doi: 10.1016/j.psychres.2014.11.006

Treadway, M. T., Buckholtz, J. W., & Zald, D. H. (2013). Perceived stress predicts altered reward and loss feedback processing in medial prefrontal cortex. *Frontiers in Human Neuroscience*, *7*(180), 1−10. doi: 10.3389/fnhum.2013.00180

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. doi: 10.1007/BF00122574

Weinberg, A., Riesel, A., & Proudfit, G. H. (2014). Show me the money: The impact of actual rewards and losses on the feedback negativity. *Brain and Cognition*, *87*, 134–139. doi: 10.1016/j.bandc.2014.03.015

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement*, *37*(3), 679–689. doi: 10.1177/001316447703700310

Zhu, C., Yu, F., Ye, R., Chen, X., Dong, Y., Li, D., … Wang, K. (2014). External error monitoring in subclinical obsessive-compulsive subjects: Electrophysiological evidence from a gambling task. *PLOS ONE*, *9*(3), e90874. doi: 10.1371/journal.pone.0090874

(Received April 7, 2016; Accepted November 21, 2016)