Psychological Bulletin
2015, Vol. 141, No. 1, 213–235

0033-2909/15/$12.00 http://dx.doi.org/10.1037/bul0000006

A Neural Reward Prediction Error Revealed by a Meta-Analysis of ERPs Using Great Grand Averages

Thomas D. Sambrook and Jeremy Goslin
University of Plymouth

Economic approaches to decision making assume that people attach values to prospective goods and act to maximize their obtained value. Neuroeconomics strives to observe these values directly in the brain. A widely used valuation term in formal learning and decision-making models is the reward prediction error: the value of an outcome relative to its expected value. An influential theory (Holroyd & Coles, 2002) claims that an electrophysiological component, feedback related negativity (FRN), codes a reward prediction error in the human brain. Such a component should be sensitive to both the prior likelihood of reward and its magnitude on receipt. A number of studies have found the FRN to be insensitive to reward magnitude, thus questioning the Holroyd and Coles account. However, because of marked inconsistencies in how the FRN is measured, a meaningful synthesis of this evidence is highly problematic. We conducted a meta-analysis of the FRN's response to both reward magnitude and likelihood using a novel method in which published effect sizes were disregarded in favor of direct measurement of the published waveforms themselves, with these waveforms then averaged to produce "great grand averages." Under this standardized measure, the meta-analysis revealed strong effects of magnitude and likelihood on the FRN, consistent with it encoding a reward prediction error. In addition, it revealed strong main effects of reward magnitude and likelihood across much of the waveform, indicating sensitivity to unsigned prediction errors or "salience." The great grand average technique is proposed as a general method for meta-analysis of event-related potential (ERP).

Keywords: feedback related negativity (FRN), event-related potential (ERP), reward prediction error (RPE), meta-analysis, great grand average

Supplemental materials: http://dx.doi.org/10.1037/bul0000006.supp

Explaining human behavior under choice requires understanding how humans assign value to goods and actions. This valuation occurs at a nexus of psychological influences running from high level processes such as framing effects and counterfactual comparisons down to basic physiological influences such as satiation. It is likely to be dependent on an individual's knowledge both through conscious extrapolation from experience and simple reinforcement learning.

Early attempts to explain human valuation were aimed at demonstrating that choice was entirely rational, and embodied key axioms of neoclassical economics such as expected utility. This approach employed a black box methodology, observing the "revealed preferences" of outward behavior in favor of the underlying apparatus of valuation, and treating humans only "as if" they computed utilities (Friedman, 1953; Samuelson, 1937). These assumptions have come under attack from the field of behavioral economics, which has succeeded in documenting widespread and consistent deviations from rational choice. A fully psychological approach, behavioral economics has endeavored to open the black box and consider a more varied set of internal representations than the simple axioms of neoclassical economics. This requires extra discriminatory power. However, behavioral economics still largely relies on observing behavior under real or hypothetical choice. There is thus a possibility that the limits of this methodology may ultimately be reached, leaving "too many theories chasing too few data" (Glimcher, Camerer, Fehr, & Poldrack, 2009).

For this reason, the emerging field of neuroeconomics uses the methodologies of neuroscience to test economic theories of human behavior. Because neuroscience can perhaps be characterized as a case of too much data backed up by too little theory, a mutually beneficial relationship might be forged, in which economic theories of human behavior are tested and supported to the degree to which the neural correlates of their terms can be found. This might in turn allow the replacement of the "as if" utilities of neoclassical economics with fully neural descriptions. Paul Glimcher, one of the driving forces behind neuroeconomics has, for example, conjectured that soon enough evidence will have accumulated that we will be able to define subjective value in fully material terms: as action potentials per second, relative to a reference dependent anchoring point given by the baseline firing rate in specific (though as yet unspecified) populations of neurons (Glimcher, 2009).

This article was published Online First December 15, 2014.

Thomas D. Sambrook and Jeremy Goslin, School of Psychology, Cognition Institute, University of Plymouth.

The authors would like to extend their sincere thanks to all researchers who kindly made available their data, thus greatly strengthening the analysis performed in this article.

Correspondence concerning this article should be addressed to Thomas D. Sambrook, School of Psychology, Cognition Institute, University of Plymouth, Drake Circus, Plymouth, Devon, PL4 8AA, U.K. E-mail: tom.sambrook@plymouth.ac.uk

This is a bold stance. To what degree does the current evidence suggest Glimcher's claim, or one like it, might be realized? An undoubted success story in this regard is the literature on single cell activity. This suggests that single cells can indeed code a utility signal which is independent of the stimuli that signal it, and which varies with changes in either of the two determinants of utility: reward magnitude and reward likelihood. Populations of such cells have been shown both for reward prospects (Platt & Glimcher, 1999), and their receipt (Schultz, 2010).

However, such effects need to be demonstrated in larger neural structures if they are to be credible determinants of actual choice behavior, and be accessible by noninvasive techniques suitable for human subjects. Functional MRI (fMRI) has been the dominant methodology here, with over 200 articles on reward valuation published in the last decade, including a number explicitly investigating the terms that underlie behavioral economics' preeminent theory: prospect theory (Tversky & Kahneman, 1992). These studies have shown that activation of certain key areas, particularly the striatum and ventromedial prefrontal cortex, is correlated with the value of anticipated or received rewards. However, individual experiments show wide variations in activated structures, with four recent meta-analyses of the literature (Bartra, McGuire, & Kable, 2013; Diekhof, Kaps, Falkai, & Gruber, 2012; Garrison, Erdeniz, & Done, 2013; Liu, Hairston, Schrier, & Fan, 2011) showing striking disparities in the broad topography of reward processing. fMRI is limited by its poor temporal resolution, particularly with regard to the valuation of outcomes, which, unlike the decisions that precede them, are strongly temporally delimited. For this reason, the event-related potential (ERP) technique, which shows excellent temporal resolution, has a role to play in the investigation of valuation in the human brain.

The purpose of the present article is to assess the evidence that an ERP component known as feedback related negativity performs a neuroeconomic valuation. Although this component has been intensively studied, inconsistencies in its reported behavior have obscured its true nature. It is possible, however, that these inconsistencies actually arise from the diverse ways in which the component is quantified. We develop a novel technique, "great grand averaging," that allows a common quantification of the component to be made, post hoc, to experiments in the existing literature. These are then subjected to meta-analysis.

Feedback-Related Negativity

ERP studies have revealed an electrophysiological component known as feedback-related negativity (FRN) that has been claimed to represent valuation of an outcome. Specifically, it has been claimed by Holroyd and Coles (2002) that this component represents a reward prediction error (RPE), that is, a signed value corresponding to the difference between the amount of reward obtained and the prior expected value of the reward. Expected value refers not to the value of the most likely outcome, but rather to a weighted average of all possible outcomes multiplied by their respective likelihoods, and in this respect is an "average outcome." Positive RPEs are produced by outcomes better than expected value, negative RPEs by those worse.

Much of neuroeconomics (and nearly the entirety of behavioral economics) is concerned with the valuation of prospects *before* their receipt, because this is what is presumed to drive choice. RPEs can be used to investigate this question by holding outcomes constant but varying prospects, with the valuation of a prospect then inferred from the RPE. Furthermore, as RPEs are central to theories of reinforcement learning, they can be used to predict future choice. A positive RPE reinforces the propensity to make the choice that brought it about, a negative RPE promotes the switch to an alternative. The degree of behavioral adjustment should be proportional to the size of the RPE, thus both the RPE's sign and its size are important. Formal models of reinforcement learning (e.g., Sutton & Barto, 1998) use such quantitative RPEs ubiquitously, and have demonstrated power in solving complex problems (producing world class backgammon play, for example), and model learning behavior very effectively.

The FRN is a scalp-recorded electrical potential, strongest at the frontocentral midline, which occurs 200 ms–350 ms after feedback on whether a reward or nonreward is obtained. At minimum, it has been shown to be a very reliable indicator of the valence of an outcome. That is, it can categorically distinguish between positive RPEs and negative RPEs, showing a relatively negative voltage for the latter. However, while this behavior is consistent with an RPE encoding function, Holroyd and Coles' theory requires it to show two further properties beyond this categorical distinction. First, the FRN must be sensitive to how *much* better or worse than expected value an outcome is, that is, the FRN must vary in proportion to the size of the RPE. Moreover, because increases in the size of positive RPEs amount to an improved outcome, but increases in the size of negative RPEs amount to a poorer outcome, if the FRN encodes RPEs on a common scale of reward it should show a Valence $\times$ RPE Size interaction. If it is responsive simply to the main effect of RPE size this suggests an encoding of absolute, or unsigned RPE size, that is, a response to salience. Second, the component should be sensitive to RPE size regardless of how this is determined. It should therefore be modulated by both of the two determinants of RPE size: reward magnitude and reward likelihood.

A large number of studies have tested for these effects, as either a primary or secondary objective. Although their methods vary greatly, broadly, a typical FRN task involves a series of independent trials in which participants are offered a choice of icons to select on a screen, and on each occasion make a selection that they believe will maximize their reward for that trial. After a short delay, feedback is provided on that choice, depicting whether a reward has been obtained or not, or the size of the particular reward. ERPs are time locked to the onset of feedback for each trial and averaged with other trials of that condition for each participant. These individual subject averages are used as data points for statistical tests, and are themselves included in a grand average ERP presented in the published article. Where the FRN's further modulation by RPE size is studied, this variable is most often manipulated as a simple categorical variable of large versus small RPEs. This variable is then crossed with the valence variable. The size of the RPE is varied using either outcome magnitude or outcome likelihood, or occasionally both. Typically, likelihood experiments offer a fixed magnitude reward which is either obtained (positive RPE) or missed (negative RPE) and manipulate RPE size by varying expected value, either by varying the likelihood of reward across blocks, or providing a cue on each trial signaling the likelihood of reward. In contrast, magnitude experiments typically hold expected value constant, often at a value of

zero, give feedback indicating either a gain (positive RPE) or a loss (negative RPE), and vary the magnitude of the outcome.

Note that the "negativity" denoted by the FRN component merely refers to the voltage of the waveform produced by negative RPEs *relative* to that produced by positive RPEs, and should not be taken to imply that negative RPEs have a privileged role in generating this voltage difference. This touches on an important methodological point, that ERP waveforms on their own can be difficult to interpret since their peaks and troughs are the sum of many individual components, some experimental, some incidental. For this reason, some electrophysiological components are described not by measuring deflections on the waveforms of individual conditions, but by those arising on the *difference wave* of two waveforms corresponding to the two levels of an experimental variable. In the case of the FRN, this variable is valence, with the component made apparent in a difference wave created by subtracting the positive RPE waveform from the negative RPE waveform. The simulated data in Figure 1 demonstrates this differencing process, and in doing so also depicts the predictions of Holroyd and Coles' theory.

Figure 1 raises an important issue of nomenclature, as the FRN is typically operationalized as the difference between good (positive RPE) and bad (negative RPE) outcomes, that is, a valence main effect. However, Holroyd and Coles claimed the FRN encoded a quantitative RPE, incorporating RPE size as well as valence, entailing a Valence × RPE Size interaction. To keep the distinction clear, we follow precedent in the present paper by using the term FRN to refer to a component responsive to the main effect of valence. We use the term RPE-FRN to refer to a component responsive to the interaction of valence and RPE size, the hypothesis under test in this meta-analysis. The simplest demonstration in support of Holroyd and Coles would be a single component showing both such effects in the same interval. However, it is also possible that the effects will be asynchronous, suggesting a quantitative RPE encoder of the kind envisaged by Holroyd and Coles accompanied by other components merely coding the sign of an RPE but not its size. In the simulated data of Figure 1, for example, the RPE-FRN in Pane e occupies a briefer interval than the FRNs in Panes c and d. This important distinction between FRN and RPE-FRN notwithstanding, at many points in the forthcoming discussion a point refers equally to both terms. Except in cases where we wish to make a point specific to the Valence × RPE Size interaction we refer simply to "the FRN."

Existing Evidence for Modulation of the FRN by Magnitude and Likelihood

In their original article, Holroyd and Coles (2002) confirmed that the FRN could be modulated by reward likelihood. Although their claim that the FRN constituted an RPE has proven highly influential, at the time of its publication the supporting evidence was limited to this single experiment, with no examination of potential magnitude effects. Now, after more than a decade of research on the FRN, we are in a much better position to assess whether Holroyd and Coles' account is supported by the evidence.

Although it is not an exhaustive review, because it only includes experiments that meet our criteria for the forthcoming meta-analysis, the picture from Appendix 1 would appear to suggest that reward magnitude does not modulate the FRN in the predicted manner. For those studies manipulating magnitude, six experiments showed the expected effect, eight studies reported no effect and three showed the opposite effect (i.e., the FRN was greater for low magnitude outcomes). For those manipulating likelihood, the evidence is stronger, if still not entirely consistent, with 13 studies showing the predicted effect and six reporting no effect.

A similar review by Walsh and Anderson (2012) mirrors this picture. Concerning the likelihood modulator, a simple sign test applied to 25 studies showed a significant effect consistent with an RPE coding. In comparison, magnitude could not be shown to significantly modulate the FRN. The authors argued the absence of magnitude effects could be because the majority of experiments cued participants as to whether an outcome would be high or low magnitude at the beginning of each trial. Thus, magnitude effects in these experiments could have been lost to scaling effects. The two studies in Walsh and Anderson's review that were uncued showed at least partial support for an FRN modulated by magnitude, and on this basis the authors argued support for the Holroyd and Coles theory. However, this very limited sample must be acknowledged to leave any meta-analytical basis for the magnitude modulator unproven.

Problems With FRN Measurement and Implications for Meta-Analysis

The ERP technique's poor spatial resolution means that any individual experiment using ERPs is vulnerable to spurious conclusions arising from overlap of the component under consideration (here, the FRN) with other components occurring in the same temporal interval. The difference wave methodology by which the FRN is best studied will not remove interfering components which also code valence, nor can it remove all the effects of components that are even partially affected by valence. In gathering the data for this meta-analysis, examination of individual studies' waveforms showed a remarkable variability in their character, assumed to arise from differences in task, procedure and stimuli. This suggests that the FRN suffers from serious overlap with unknown components which might, quite incidentally, be responsive to any of the three factors (valence, magnitude, likelihood) under study. Because the sources of the component overlap are unknown in each case, a broad meta-analysis is therefore more robust than any single experiment.

A serious hindrance to meta-analysis, however, is the lack of consistency in how the FRN is quantified. In some articles it is measured by the voltage of a single peak, in others the difference between two peaks, in others the difference of one peak and the average of surrounding peaks, and in others by the mean amplitude in a set interval. Analysis is sometimes conducted on difference waves and sometimes on the original simple waveforms. Perhaps most seriously, the temporal interval in which the FRN is measured varies widely. Appendix 1 shows the quantification of the FRN in each of the studies used in this meta-analysis. It was found that both mean amplitude measures and peak assignment are made in intervals ranging from 50 ms to 150 ms duration, at substantially different latencies, with some studies using intervals that do not even overlap. The 200- to 350-ms interval after feedback, where the bulk of the FRN measures lie, is characterized by a steep, alternating, positive-negative-positive going waveform, and so differences in the interval in which the FRN is measured can have large

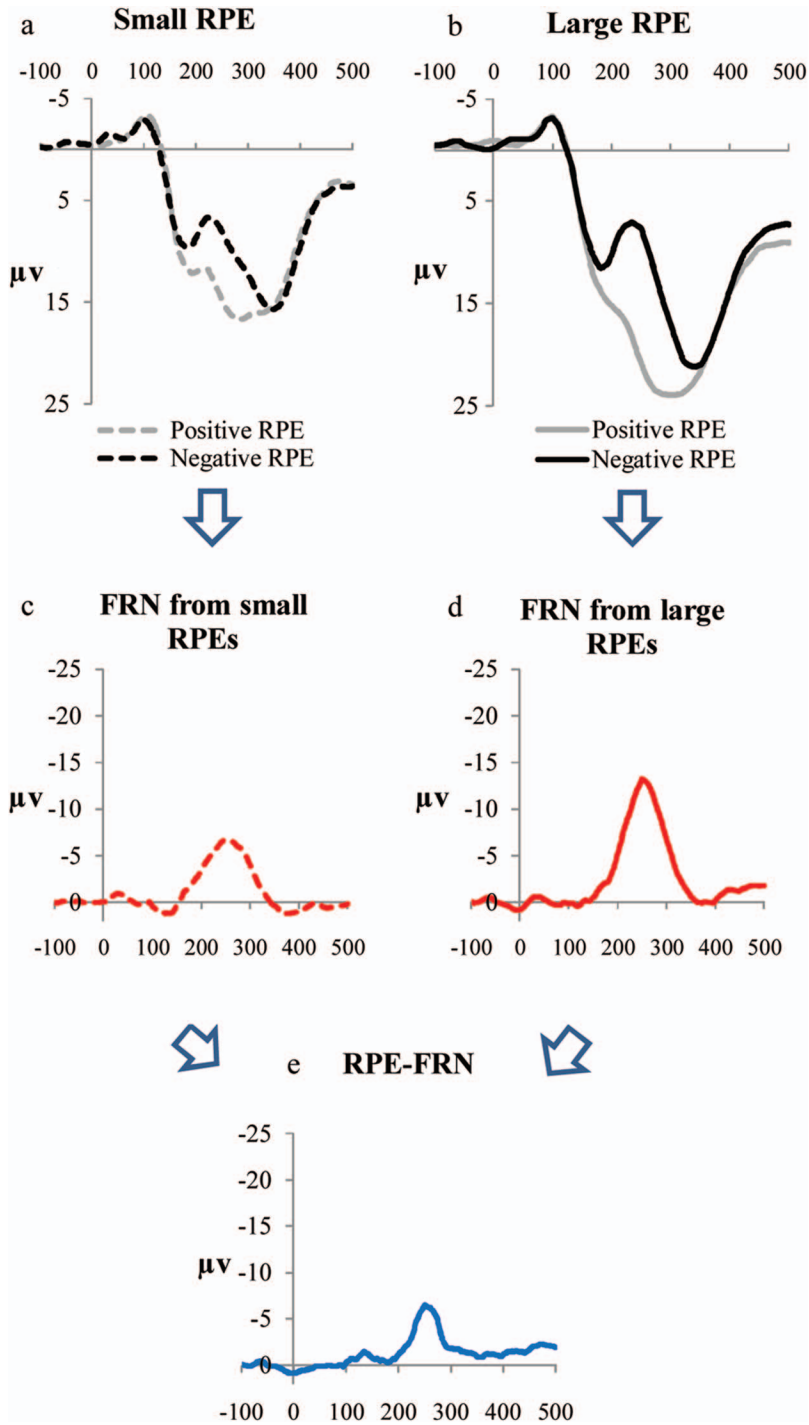

Figure 1. How the FRN is studied. Panes a and b show grand average waveforms for four experimental conditions taken from a 2 × 2 design manipulating valence (positive RPEs vs. negative RPEs) and RPE size (large vs. small). A given experiment would manipulate RPE size using either outcome magnitude or likelihood. The simple waveforms in Panes a and b show complex peaks which are the result of consecutive, overlapping components, many of which are unknown. Difference waves (FRNs), constructed by subtracting the positive RPE outcome waveform from the negative RPE outcome waveform are shown in Panes c and d. These control for components unrelated to valence, allowing the valence effect to be more clearly seen. Comparison of Panes c and d also suggests that the amplitude of the FRN in these data is sensitive to RPE size. This is definitively shown by differencing the difference waves in Pane e. Collectively, the figures represent the prediction of Holroyd and Coles' theory, with Pane a corresponding to low magnitude or likely outcomes and pane b to high magnitude, or unlikely outcomes. See the online article for the color version of this figure.

ramifications, with waveforms from different experiments that look similar on the page producing opposing conclusions once they are quantified and subjected to a statistical test. This may lead to failures of replication. Conversely, it also possible that unexpected effects may go unnoticed as a result of the interval used, leading to successful but unwarranted replication, and an inflating of the apparent robustness of the FRN component and effects associated with it. The consequence for meta-analysis is that the compilation of the *statistical results* of FRN studies would be far more sensitive to idiosyncrasies in how the FRN is quantified than is desirable.

A Novel Method: Great Grand Averages

The problem of miscellaneous measures described above prompted us to use a novel means of meta-analysis. Typically, a meta-analysis uses standardized effect sizes derived from individual research articles as replicates for some statistical test for an effect of interest (most typically whether the effect size is significantly different from zero). A basic assumption of any statistical test, meta-analytic or otherwise, is that all data constitute observations of the same phenomenon. In our case, this phenomenon would consist of neural events comprising activity of the FRN component. As previous discussed, it is not well established whether the neural events measured across the range of previous studies arise exclusively from the FRN. Moreover, the mixture of mean voltage and peak to peak measures, and the wide range of intervals used to quantify the FRN component also raise serious concerns over the equivalence of effect sizes measured across the literature.

Our meta-analysis avoided the miscellaneous measures problem by ignoring the quantification of the FRN the authors of individual articles had used. Instead, we took published grand average waveforms in all experiments that met our criteria, digitized these waveforms to extract their coordinates and averaged these coordinates across experiments to create composite waveforms representing "great grand average" (GGA) waveforms. Although this approach is borne of necessity, there are nevertheless some benefits to bypassing the quantification provided by the original authors and going "upstream" to the published waveforms. One advantage is that a great deal of information is thrown away in the conversion of waveforms to unitary scores of component amplitude, and the GGA technique retains that information up to the point of quantifying the final GGA waveform. This means that effects that are small or lie outside typically analyzed intervals, but which are consistently present, can become noticeable.

A second advantage is the method's potential to reduce the effects of component overlap. In a single ERP experiment, averaging across trials reduces the effect of incidental neural activity that is peculiar to a given trial, thereby accentuating task-related components, which are elicited on every trial. However, this does not help reduce components that happen to be elicited by the task, but are not the subject of the experiment. This causes component overlap, and complicates the measurement of the component under study. Under the GGA technique, some of this component overlap is reduced due to the variations in the tasks used in different experiments. Averaging across experiments reduces the effect of incidental components which are peculiar to a given task, thereby accentuating the component under study, which will be elicited in all tasks. Of course, components other than the one under study which happen to be elicited by the *factors* of an experiment (here valence, magnitude, and likelihood) will not be reduced by the GGA technique.

A third advantage arises from differences in the latency at which the FRN occurs in the pool of experiments used to create the GGA waveforms. This produces "smearing" of the peaks that characterize the feedback-locked ERP, reducing their amplitude, and widening the duration of general positive and negative deflections. For the purposes of the study at hand, we regard such smearing as a methodological strength. This is because it reduces the availability of bespoke intervals in which strong, but likely unreplicable effects can be found. This ensures a fairer test of the RPE account. Thus what is lost in (possibly misleading) peak amplitude is gained in reliability. It must be noted that the advantages of this meta-analysis technique are not specific to the study of the FRN, rather, they are highly generic and could be applied to the meta-analysis of any ERP component.

The disadvantages of the method are that the extraction of the data directly from the waveforms introduces a new source of measurement error, and that disregarding the reported statistics eliminates the only source of information concerning the within-study variance of the studies entering the meta-analysis. These issues are considered empirically later.

Moderators

While the variability of the waveforms produced by FRN experiments has complicated their interpretation and presented methodological challenges, it is possible that some of this variation is systematically related to differences in experimental tasks, and can thus be used to infer properties of the component. We therefore performed the following moderator analyses.

RPE Modulator

Modulator refers to whether outcome magnitude or outcome likelihood was used to manipulate the size of RPEs. While demonstrating that the FRN is a generalized RPE encoder requires that it be responsive to both modulators, and so their effects needed to be established independently, a comparison of those effects is potentially illuminating because evidence that the FRN is a generalized RPE encoder would be bolstered by a relative insensitivity to the source of the RPE size modulation.

Control Over Outcome

The expected value against which an RPE is generated might consist either in the expected value of the preceding stimulus, or in the expected value of the action performed in response to that stimulus (Balleine, Daw, & O'Doherty, 2008). That is to say the RPE might contribute to either Pavlovian or instrumental conditioning. This matter may be addressed by examining the degree to which control over outcome affects FRN amplitude. An RPE used in Pavlovian conditioning, perhaps reflecting a general role in valuation, will occur even when participants passively observe outcomes. In contrast if the FRN is greatest following a meaningful action on the part of the participant this suggests a role in instrumental conditioning.

Magnitude Cueing

Although reward likelihood is fundamentally limited to values between zero and one, there is no such delimitation to reward magnitude. In order to be able to show satisfactory discrimination across a wide range of outcome magnitudes it would appear necessary that the FRN scale its response relative to the range of magnitudes considered available in the immediate context. Such scaling has been shown by Tobler, Fiorillo, and Schultz (2005) in macaque midbrain neurons, which produced equivalent responses to rewards of different magnitudes when the range of magnitude available on that trial was signaled to the subject beforehand. Bunzeck, Dayan, Dolan, and Duzel (2010) found a similar scaling effect in a study of humans using fMRI. Schultz (2009) has suggested that scaling is performed not on the absolute range of outcomes possible in a given context but on their estimated distribution, and that RPEs accordingly represent z scores. Although the question of scaling is of theoretical interest, it is also methodologically important because scaling effects may have been responsible for the absence of FRN sensitivity to magnitude in the literature. This moderator analysis investigated whether effects of magnitude on the FRN were reduced in cued experiments as would be expected if scaling occurred.

Domain

Although the FRN has a well-established sensitivity to valence, i.e. the sign of an RPE, this is formally orthogonal to the domain of the outcome, that is, whether the outcome constitutes an actual monetary loss or gain. For example, losses can still be positive RPEs if they are smaller than expected losses. A number of recent studies have suggested that the FRN is accentuated when measured in the gain rather than loss domain (Kreussel et al., 2012; Kujawa, Smith, Luhmann, & Hajcak, 2013; Mushtaq, Stoet, Bland, & Schaefer, 2013; Sambrook, Roser, & Goslin, 2012; Yu & Zhang, 2014). This suggests the possibility of a neural dissociation of how outcomes are processed in gain and loss domains that is of broad theoretical interest, not least because this reduced sensitivity for outcomes in the loss domain, or "loss indifference," is in direct opposition to the prediction of loss aversion made by prospect theory.

Method

Inclusion and Exclusion Criteria

The independent variables used in this meta-analysis were outcome valence (positive RPE, negative RPE), outcome magnitude (high, low) and outcome likelihood (likely, unlikely), where this final variable refers to prior likelihood of an obtained outcome. For inclusion, an experiment had to contain within it a 2×2 factorial manipulation of valence with respect to either likelihood or magnitude. Where more than two levels of the likelihood or magnitude variable were presented in an article, intermediate ones were ignored in order to maximize contrasts.

The dependent variable differed depending on the particular contrast examined, as detailed in the coding procedures section below. In all cases it was a voltage derived from the differencing of four simple waveforms related to the factorial design described above. Consequently, a key inclusion criterion was that a study must present such a set of simple waveforms. Waveforms had to be plotted for at least 500 ms postfeedback and 100 ms prior, and had to be locked to feedback, not response. Because waveforms were in many cases plotted at only a single electrode, and because variability of the electrode used suggested a broad distribution for the FRN, an experiment was included as long as it presented waveforms at Fz, FCz, Cz or "a frontocentral pool." Variability was minimized by using FCz waveforms where available (even if individual articles reported the FRN to be maximal at a different site), and where they were not, using Fz, Cz, or frontocentral pool in that order of preference.

Studies using populations other than healthy nonolder adults were used only if control data for this population were available and participants had not been selected on the basis of any prescreening (e.g., personality scales). The experiment had to offer monetary rewards conveyed by feedback, although tasks could vary widely, including guessing games, time estimation tasks, and simply passive observation. Experiments could either employ mixed gambles comprised of wins and losses, gain domain gambles where participants either won or failed to win a stake, or loss domain gambles where subjects lost a stake or successfully avoided this. Losses and omitted rewards were classed as negative RPEs, wins and avoided losses were classed as positive RPEs. Where separate waveforms were presented for the portion of an experiment before and after participants learned a rule that allowed them to assess reward likelihood, waveforms for the portion after learning were used, because these could be expected to produce the strongest effect of likelihood on prediction errors. Experiments which manipulated factors other than the three of interest were included, although in some cases waveforms were used at one level of that additional factor, often a control level, if available and appropriate (see Appendix 1).

Experiments were excluded if the factor of magnitude, likelihood, or valence was confounded with another variable. Although experiments manipulating both magnitude and likelihood were acceptable (and in these cases were used twice in the analysis, once for each modulator) they were excluded if these variables confounded each other. This was common in Iowa Gambling Tasks and where participants could genuinely optimize their choice. Experiments where the FRN was a response to observation of another's performance were excluded. Magnitude experiments were considered ineligible if levels of the magnitude variable were blocked, because we expected this would strongly exacerbate scaling effects, with the FRN responding simply to the valence of the outcome at the given level of the stakes in that block (although in fact no otherwise eligible experiments were excluded on this basis). In the case of likelihood experiments, the following criteria were employed. There had to be two levels of the likelihood modulator either side of, and equal distance from 50% probability to avoid confounding likelihood with uncertainty, a property the feedback-locked ERP may be responsive to (Yu, Zhou, & Zhou, 2011). If participants received explicit instruction on probabilities this had to be consistent with real probabilities so that there could be no ambiguity regarding the value of expected value that RPEs were generated with respect to. Experiments were excluded if participants could learn to actively avoid disadvantageous trials (e.g., by knowing which button to press to always ensure $>50\%$ reward probability) because this made unlikely positive RPEs and

likely negative RPEs infrequent, introducing a possible confound, and also leaving participants' motives for their suboptimal choice unclear.

Moderator Analyses

Modulator. Coding of this moderator was straightforward. In those cases (k = 2) where both magnitude and likelihood were manipulated, the two conditions were entered as independent studies for this analysis.

Control over outcome. Operationalizing this moderator was inherently problematic because perception of control is highly variable across people (Langer, 1975). We used three levels of this moderator variable, which we believed would maximize contrasts. Level 1, termed "passive," covered tasks in which participants were given no opportunity to act meaningfully prior to feedback. At the other end of the scale, Level 3, termed "rule implementation" comprised tasks where actions could be performed and where feedback was genuinely (and therefore ultimately visibly) dependent on choice of action. Level 2 was termed "guessing." This level encompassed all tasks in which participants acted but could not actually affect the outcome. This included, for example, cases where participants had to guess the location of a prize, or choose the stakes for a particular trial. It is true that participants might have *believed* that they had a degree of control over the outcome, but information on these beliefs was generally not available. We assumed that where control over an outcome was neither evidently absent (Level 1) nor present (Level 3), participants would experience, on average, some intermediate perception of control.

Domain. A direct comparison of the amplitude of the FRN in loss and gain domains could not be made because only two studies included any pure loss domain trials. However, many studies offered mixed gambles, in which positive RPEs were always gains and negative RPEs always losses, and the loss indifference effect described earlier might be expected to attenuate effects in the loss portion, producing a net reduction of the FRN overall in mixed gambles. Domain was therefore coded with two levels. The first, "gain domain" comprised all cases where the worst possible outcome was no reward. The other level, "mixed domain," comprised cases where monetary losses as well as gains could be incurred.

Magnitude cuing. This analysis applied to magnitude studies only. Cued studies comprised all cases where participants knew the magnitude of the forthcoming feedback but not its valence, uncued studies comprised cases where they knew neither its magnitude nor valence. A single study in which magnitude cuing was manipulated as an independent variable was entered into this analysis as two separate studies.

Search Strategies

Published data. The first author performed the literature search and assessed studies for suitability. A search for English language journal articles and books was performed using the following databases: PsychInfo, PsychBooks, PsychArticles, ERIC, PubMed, and Web of Science. Results were compiled in EndNote. Abstracts, titles, and keywords were searched using the term "feedback negativity" OR "feedback related negativity" OR "feedback error-related negativity" OR "reward positivity" OR "feedback correct related positivity." Duplicates, clearly inappropriate journals and conference abstracts were removed without inspection, as were articles published prior to 1997 (the year of publication of the first FRN article; Miltner, Braun, & Coles, 1997). Two-hundred and 15 papers remained, of which 42 were deemed eligible after checking inclusion and exclusion criteria.

The FRN is sometimes referred to generically as an "error related negativity," even though this term is more commonly used to refer to a waveform locked to subjects' own responses, and indicating internal registration of a known error, rather than a response to external feedback. It is also sometimes referred to generically as a mediofrontal negativity. We conducted a secondary search using the term "error related negativity" OR "mediofrontal negativity" OR "medial frontal negativity." After removing duplicates, duplicates with the earlier search, clearly inappropriate journals and conference proceedings, and articles predating 1997, 1,012 articles remained. The abstracts were scanned for evidence that feedback locked waveforms were studied, producing 125 possible articles, of which four met the criteria for eligibility.

The reference lists of all eligible articles were checked, along with those of two recent reviews of the FRN (San Martin, 2012; Walsh & Anderson, 2012), producing one further eligible article. In total, these search criteria resulted in the inclusion of 47 datasets from published papers in our meta-analysis.

Unpublished data. In an effort to include unpublished data, all first or corresponding authors of the selected articles were contacted with a request for unpublished data. A number of other researchers were also contacted, identified as follows. Articles returned by the searches described above which had been rejected were reexamined, and 154 authors added to a contact list. A search of theses using the ProQuest Dissertations and Theses database and the Ethos database returned 73 hits for the primary search string and 370 for the secondary one. The contents pages of these theses were read online and 17 authors added to the contact list. Abstracts of 56 conference articles, extracted from the searches described earlier, were read, and on this basis eight more authors were added. In the course of contacting authors, a further four suggestions were garnered. One hundred seventy-one of 183 e-mail addresses were successfully obtained by Internet search and these researchers contacted. Responses were obtained from 51 researchers. Four entirely unpublished datasets were retrieved by this process, and one dataset associated with a published article in which the requisite waveforms had not been presented. Three unpublished studies of the authors' own were also added. Therefore, we finally included 55 datasets into our meta-analysis, 47 from published data, eight from unpublished data.

Validation data. As this is the first implementation of the GGA technique, we sought to validate it by comparing its findings with those resulting from conventional meta-analysis based on standardized effect sizes. For a meaningful comparison, it was important that these standardized effect sizes were generated in the same fixed interval of the waveform as that used for the GGA analysis. Effect sizes (or their derivatives) reported in the original articles did not correspond to this, or any fixed interval. It was their variability that prompted development of the GGA technique. To carry out the validation we therefore contacted authors of all the 55 articles used in our GGA analysis with a request for their original data so that we might calculate standardized effect sizes in the designated interval ourselves. This request returned original data for 14 of the 29 magnitude studies and 13 of the 26 likelihood

studies. These studies are hereafter referred to as the validation dataset.

Coding Procedures: Generating Great Grand Averages Waveforms

Digitizing of published waveforms was performed with PlotDigitizer (http://sourceforge.net/projects/plotdigitizer/). Electronic copies of experiments were accessed, and the figures containing the requisite waveforms were enlarged and then opened in the PlotDigitizer software. Digitizing began by using a mouse to calibrate the minimum and maximum values of the *x*- and *y*-axis to the distance they occupied on the screen, thus defining the coordinate space of the area of the figure. The coordinates described by the actual ERPs were extracted by using a mouse to manually lay points along the waveforms at approximately 5 ms intervals. These were then run through a purpose-written program (supplied as a supplementary file) that linearly interpolated coordinates at 1 ms intervals between the existing manually assigned ones. For every waveform undergoing the process, this generated a series of voltage values at discrete 1 ms intervals that made the subsequent process of averaging across studies tractable. The coordinates were immediately replotted to visually check that they corresponded to the original waveform they were taken from to prevent gross errors. All waveforms were digitized twice in this fashion, partly to improve accuracy and partly to allow reliability checks discussed below.

The consequence of this digitizing process was that for each study in the meta-analysis, we were able to recover the data that underlay the four relevant grand average waveforms, plus some measurement error. For the 27 experiments that were also represented in the validation dataset, original data replaced the digitized versions for the bulk of subsequent analysis. In these 27 cases the digitized versions were merely used to assess the degree of digitizing measurement error, as described later.

Coding Procedures: Quantifying the FRN

The grand average waveforms that were recovered by the digitizing process were submitted to the differencing process shown in Figure 1 in order to establish whether an RPE-FRN was present. As noted earlier, the RPE-FRN refers to a component responding to the interaction of RPE size and valence. Such an interaction is present when the difference waves shown in Panes c and d of Figure 1 differ in amplitude. The effect size of the RPE-FRN component was thus the amplitude of the waveform corresponding to the difference of difference waves shown in Pane e of Figure 1, and its significance was based on a comparison of the amplitudes of its constituent difference waves, that is, those corresponding to Panes c and d, with this comparison made across the sample of either likelihood ($k = 26$) or magnitude ($k = 29$) studies.

Difference wave amplitude is typically measured either by using the waveform's peak within a set interval, or its mean amplitude within a usually smaller interval. To provide a robust test of whether an RPE-FRN was present, we used both measures. The interval in which the measures were taken was determined by the average of those intervals used in the original papers. Those studies that used a mean amplitude measure produced an average measurement interval of 228 ms–344 ms, and those using a peak amplitude measure produced an average measurement interval of 128 ms–460 ms.

In addition to exploring the effect of magnitude and likelihood modulators on the FRN, we were interested in the effects of these variables in their own right, that is, their main effects. To study these, rather than differencing the valence variable, it was collapsed out at each level of magnitude and likelihood, allowing the comparison of high and low magnitude waveforms and high and low likelihood waveforms. Thus, in the scheme shown in Figure 1, an average waveform was created in each of Panes a and b and these were then differenced (small RPE–large RPE) to produce an RPE size main effect difference wave.

Statistical Methods

Simple and standardized effect sizes. The differencing process described above was performed on each individual study, generating an effect size for the RPE-FRN, thus allowing a test for the significance of this effect size across the studies that made up the dataset. This process made no use of the standard deviation of the effect size *within a given study*, however, that is, calculated across the subjects of that study, nor could it do so, because the digitizing process only had access to grand average waveforms. As noted, this does not prevent us testing for the significance of the effect across studies, but does prevent the relative weighting of individual studies based upon the variance of their data. This is generally used to down-weight the contribution from studies showing high variability on the basis that their estimate of the effect under question can be assumed to be less reliable. Conventional meta-analysis achieves this weighting up front by using *standardized effect sizes* (often referred to simply as "effect sizes") as the unit of analysis, which down-weight effects when they are underlain by high variability. The standardized effect size metric in which this is most obviously expressed is Cohen's *d*, which is the difference between two scores of interest divided by their pooled standard deviation. Standardized effect sizes can be contrasted with *simple effect sizes* (Baguley, 2009) or "raw mean differences" (Bond, Wiitala, & Richard, 2003) which, as the name suggests, are equivalent to Cohen's *d* without any division by standard deviation. Simple effect sizes are what are produced by the GGA technique and what are used in the GGA meta-analysis presented here.

Both Baguley, 2009 and Bond, Wiitala, and Richard (2003) have argued the virtues of working with simple effect sizes over standardized ones, noting the ease with which they can then be used to practically guide future studies (e.g., in the present case, a simple effect size informs researchers of the size in microvolts that they can expect to be working with) and observing that the standard deviations that are used to calculate Cohen's *d* are themselves subject to the sampling error they purport to correct for. Another reason why standardized effect sizes have become the norm in meta-analyses is that they allow the comparison of scores derived from different scales of measurement, which is not an issue here, where the metric is always voltage. Nevertheless, the use of simple rather than standardized effect sizes is a notable feature of the GGA technique and we later examine its consequences using the validation dataset.

Testing the hypothesis I: *t* tests on GGAs. Our hypothesis was that the FRN would be greater when RPEs were large rather

than small, as described in Panes c and d of Figure 1. Because the criterion for a generalized RPE encoder is that it should be modulated by both reward magnitude and likelihood, these two modulators were tested separately. In each case, a paired samples *t* test was conducted of the amplitude of FRNs constructed from small RPEs versus large RPEs. Four tests were done in total, on peak measures and mean amplitude measures of the magnitude and likelihood modulated FRNs. *T* tests were entirely analogous to those which might be performed on individual FRN experiments but at "one level higher" using grand average data as data points, rather than subject average data.

Because sample size differed over studies, and conventional meta-analysis typically incorporates this information (Field & Gillett, 2010; Hunter & Schmidt, 2004), weighted *t* tests were used. The *t* statistic was calculated with the standard formula for paired samples

$$t = \frac{\overline{X}_D}{\frac{S_D}{\sqrt{N}}}$$

Where $\overline{X}_D$ is the mean difference of the paired samples, and S_D its standard deviation. However, $\overline{X}_D$ was a *weighted* mean difference, calculated from *k* individual study mean differences (x) whose sample size was used as a weight (w), as follows

$$\overline{X}_D = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

The standard deviation of this difference was also weighted, as follows

$$S_D = \frac{\sum_{i=1}^{k} w_i (X_i - \overline{X}_D)}{\sum_{i=1}^{k} w_i}$$

Unless otherwise stated, all statistics performed on GGAs used weighted means and standard deviations.

Sensitivity to publication bias was assessed by inspection of funnel plots followed by trim and fill (Duval & Tweedie, 2000) implemented in R using the metafor package (Viechtbauer, 2010).

Testing the hypothesis II: Data driven cluster randomization of GGAs. The analysis described above provides a fair but straightforward test of the hypothesis because the FRN was quantified in an interval determined a priori by the existing literature. However, it remains possible that this interval is a poor choice, certainly for capturing the RPE-FRN, that is, the response to the interaction of RPE size and valence. We therefore used a second, data driven technique, that examined the full length of waveforms for evidence of an RPE-FRN component. As well as addressing the danger of using the wrong interval, this had the secondary advantage that it could extract the observed interval of the RPE-FRN post hoc. The multiple comparisons resulting from the analysis of the whole waveform were avoided by using the cluster randomization procedure of Maris and Oostenveld (2007). This procedure allows an entire ERP waveform to be analyzed without incurring the excess conservatism of a strict Bonferroni correction for each time point analyzed. It achieves this by recognizing that because voltages are strongly correlated at adjacent time points, the effective number of comparisons being made when an entire waveform is analyzed is much lower than the number of sample points in the waveform. First, *t* tests were performed on the two difference wave amplitudes at each time point, and clusters of time points at which the difference in the difference wave amplitudes was statistically significant ($p < .05$) were marked as being of potential significance. The values of *t* for each time point in these clusters were summed to produce a cluster-level *t* statistic. This was then compared to a probability distribution for such cluster-level t statistics generated by 10,000 runs of a Monte Carlo simulation on null distribution data in the interval occupied by the cluster. This was used to assign a Monte Carlo *p* value to the cluster of significant t values identified at the start of the process.

Heterogeneity of GGAs. Meta-analyses typically report heterogeneity, a measure of the likelihood that the sample effect sizes in the meta-analysis are drawn from more than one population. This is shown by a variance across sample effect sizes which exceeds that expected from the within-study variances. Because within-study variances are unknown under the GGA technique, heterogeneity cannot be measured. It can, however, be implied by demonstration of the significant effect of moderators.

Moderator analysis. This is conventionally performed in conjunction with a standardized effect size based meta-analysis, something we could not do with the GGAs, as we could not compute standardized effect sizes. To test for the effects of moderators, we performed univariate analyses with the moderator as a single categorical independent variable. The dependent variable was the simple effect size of the RPE-FRN. Unweighted effect sizes were used in an ANCOVA analysis with weighting applied using the weighted least squares function. To maximize the power of the moderator analysis, likelihood and magnitude modulated studies were analyzed together. Because validation of the GGA technique (reported later) suggested that mean amplitude measures produced closer estimates to an ideal conventional meta-analysis than peak measures, only mean amplitude measures were used for moderator analysis. Confounding of moderators was checked using contingency coefficients of all possible pairs of the four moderators, and where significant χ^2 values where found, entering the confounding moderators as covariates.

Meta-analysis of validation data. Conventional meta-analysis was performed using standardized effect sizes of the RPE-FRN generated from original data obtained from authors. Differencing of waveforms and calculation of *t* values was performed in the same way as was done for GGAs, with *t* values then converted to Cohen's *d*. A calculation from *t* values was used rather than direct calculation using the mean difference divided by its standard deviation because of problems arising from the standard deviation term of paired samples designs. As Dunlap, Cortina, Vaslow, and Burke (1996) have observed, paired samples designs increase power by reducing the standard deviation term. This makes it easier to detect an effect (e.g., *t* is increased). However, the paired samples design does not change the effect's size, which is what *d* purports to represent. Using the paired samples standard deviation in calculating *d* therefore conflates effect size with effect significance and inflates the estimate of *d*. Because the degree of

this inflated estimate is proportional to the additional power the paired design provides, and this in turn is proportional to the extent to which the paired scores move together, *d* can be corrected by using the correlation coefficient of the two conditions underlying the *t* test. Dunlap et al.'s formula for this unbiased calculation is shown below and was used for calculation of *d*. Note that the *r* term should not be confused with an effect size metric.

$$d = t \cdot \sqrt{\frac{2(1-r)}{n}}$$

Meta-analysis was conducted using the method of Hunter and Schmidt (2004), with studies weighted by their sample size rather than inverse variance, because this allowed the closest comparison with the GGA technique. A random effects model was used, due to concerns over the generalizability of fixed effects models (Field & Gillett, 2010). The meta-analysis produced an estimated effect size, confidence intervals for this estimate, and, most importantly for our validation purposes, a significance test that could be compared with that produced by the GGA technique. Heterogeneity was measured using the Q statistic. Analyses were implemented in the macros provided by Field and Gillett (2010) apart from trim and fill which was implemented in R using the metafor package.

Meta-analysis of published data. Although the GGA technique is premised on the unsuitability of published FRN effect sizes for meta-analysis, we ran a further meta-analysis using published effects for illustrative purposes. The effect size measure used was once again Cohen's *d*. Effect sizes were frequently not reported in the published articles, and where they were it was typically in the form of partial eta squared. Values of *d* were therefore calculated directly from reported test statistics using conventional approximations. Where the reported statistic was *t*, the Dunlap formula above was used with *r* estimated at 0.5: The average correlation found in our validation dataset was in fact 0.49. Where the statistic given was an *F* value, that is, rather than a difference of difference waves, the RPE-FRN effect size was expressed as a Valence x RPE Size interaction, Rosenthal's (1991) conversion was used:

$$d = 2\left(\sqrt{\frac{F}{df_d}}\right)$$

In cases where effects were reported as "nonsignificant" or an inequality based on a canonical value such as $F < 1$ was given, *d* was set to zero. If a noncanonical value of a statistic or *p* value was given (e.g., $p < .06$) this was taken as the actual value. Meta-analysis was then performed as described for the validation data.

Results

Modulation of the FRN by Magnitude and Likelihood

Figure 2 shows simple great grand average waveforms for magnitude and likelihood designs. The underlying data for the digitized grand average waveforms are provided as supplementary information, as are the derived difference waves that follow. These can be interpreted and replotted using the accompanying documentation. Figure 3 depicts the central test of the hypothesis. It can be seen from Figure 3a that the FRN for high magnitude outcomes

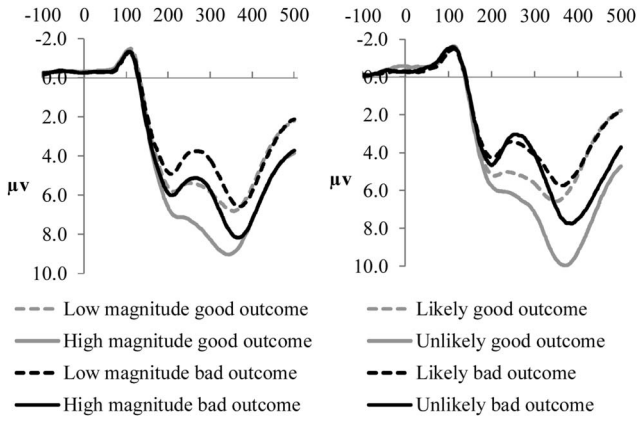

Figure 2. Simple waveforms for (a) magnitude experiments, and (b) likelihood experiments. Only 100 ms of baseline is shown, explaining how the baseline has become negative overall.

is of greater amplitude than the FRN for low magnitude outcomes, suggesting that the FRN is sensitive to outcome magnitude in the manner predicted. This sensitivity is plotted as an RPE-FRN, that is, the difference of the high magnitude difference wave and the low magnitude difference wave. A paired samples *t* test on mean FRN amplitudes in the interval 228 ms–334 ms revealed a significant difference ($M_{low} = -1.52$ μv, $M_{high} = -2.20$ μv, RPE-FRN simple effect size = −.68 μv, $t(28) = -4.41$, $p < .001$). A *t* test on peak FRN amplitudes in the interval 129 ms–447 ms also showed a significant difference ($M_{low} = -2.30$ μv, $M_{high} = -3.11$ μv, RPE-FRN simple effect size = −0.81 μv, $t(28) = -3.11$).

Similar comparisons for the likelihood modulator can be seen in Figure 3b, where it can be seen that, as predicted, the FRN for unlikely outcomes is of greater amplitude than the FRN for likely outcomes, again generating an RPE-FRN. The effect was significant under a mean amplitude measure in the interval 228 ms–334 ms ($M_{likely} = -1.56$ μv, $M_{unlikely} = -3.10$ μv, RPE-FRN simple effect size = −1.54 μv, $t(25) = -5.44$ μv, $p < .001$) and a peak measure in the interval 129 ms–447 ms ($M_{likely} = -2.84$ μv, $M_{unlikely} = -4.65$ μv, RPE-FRN simple effect size = 1.84 μv, $t(25) = -5.62$ μv, $p < .001$). The RPE-FRN simple effect sizes for both modulators under the mean amplitude measure are shown as a forest plot in Figure 4. As a further check, the *t* tests described above were conducted on unweighted scores to ensure that the effects were not unduly affected by a few studies with large sample sizes. All effects remained strongly significant.

The hypothesis was thus supported using a quantification of the FRN based on a priori intervals derived from the literature. The Maris and Oostenveld procedure was then used to more accurately determine the latency of the RPE-FRN specifically. For the magnitude modulator, a single significant cluster of RPE-FRN activity was found (Monte Carlo $p = .0001$), running from 240 ms–341 ms, with the effect greatest at 298 ms (−.91 μv). For the likelihood modulator a single cluster of RPE-FRN activity was found (Monte Carlo $p < .0001$), running from 209 ms to the edge of the measurement interval at 500 ms. The effect was equally great at 274 ms and 352 ms (−1.80 μv) but much more significantly so at the earlier peak: $t(25) = -6.46$.

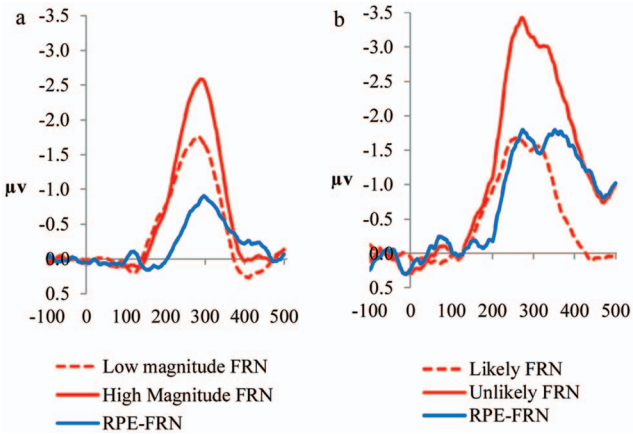

Figure 3. Modulation of the FRN by (a) magnitude, and (b) likelihood. Difference waves (FRNs) are created from negative RPE minus positive RPE waveforms. The RPE-FRN simple effect size is the difference of the two difference waves. See the online article for the color version of this figure.

Publication Bias

Publication bias was assessed by inspection of the funnel plots shown in Figure 5. Because these suggested a small degree of asymmetry, albeit largely among studies with large rather than small sample sizes, we applied a trim and fill procedure. This was implemented by entering the simple effect sizes derived from GGA analyses into a conventional meta-analysis, rebalancing potential asymmetry in the funnel plots by adding additional imputed studies, and then recalculating effect sizes. In fact, this procedure resulted in no additional studies being imputed, leaving effect sizes unchanged, and demonstrating absence of publication bias.

Moderator Analyses of the FRN

χ^2 tests revealed strong associations ($p < .001$) between three of the four moderators: modulator, domain, and control over outcome. To test the effect of each moderator individually, while controlling for the effects of the others, analysis of covariance was used with the confounding moderators entered as covariates. Once again, a mean amplitude measure in the interval 228 ms–334 ms was used.

Modulator. No significant effect of modulator on RPE-FRN simple effect size was found (magnitude: −.72 μv, $k = 27$; likelihood: −1.60 μv, $k = 24$; F(1, 47) = 2.92, $p = .09$). The apparent strong effect of modulator shown by a comparison of the subplots in Figure 3, and the means above, was due to the mediating effect of control over outcome (see below). Because Figure 3 also suggested the possibility that the RPE-FRN of likelihood experiments occupied a longer interval than that for magnitude experiments, this was investigated using a mean amplitude measure in the interval 335 ms–500 ms. The effect of modulator on RPE-FRN in this interval proved to be narrowly nonsignificant (magnitude: −.23 μv, $k = 27$; likelihood: −1.38 μv, $k = 24$; $F(1, 47) = 3.80$, $p = .057$).

Control over outcome. A significant effect of control over outcome was found, with RPE-FRN amplitude increasing as control grew (passive: −.07 μv, $k = 5$; guessing: −.88 μv, $k = 34$; rule implementation: −2.47 μv, $k = 12$; $F(2, 46) = 9.71$, $p < .001$). Post hoc comparisons revealed all pairwise comparisons to be significant ($p < .05$). A significant effect was also found in the later interval of 335 ms–500 ms (passive: −.41 μv, $k = 5$; guessing: −.25 μv, $k = 34$; rule implementation: −2.56 μv, $k = 12$; F(2, 46) = 7.40, $p = .002$). Post hoc comparisons in this interval revealed that rule implementation produced a significantly stronger RPE-FRN than passive or guess designs ($p < .05$), but these two levels did not significantly differ. Waveforms of the RPE-FRN for the three levels (with modulator collapsed out) are shown in Figure 6.

Domain. No effect of domain on the RPE–FRN was found (gain: −1.37 μv, $k = 25$; mixed: −.82 μv, $k = 26$; $F(1, 51) < .01$).

Magnitude cuing. No effect of magnitude cuing on the RPE-FRN was found (cued: −.70 μv, $k = 20$; uncued: −.59 μv, $k = 8$; $F(1, 26) < .1$).

Validation of the GGA Technique

Where an electrophysiological component is quantified in diverse ways in a literature, we have argued that the GGA technique is superior to conventional meta-analysis because it allows quantification to be made in a standardized interval. Nevertheless, the GGA technique suffers two potential drawbacks relative to conventional meta-analysis. The first is that the process of recovering original data from published figures introduces measurement error. The second is that the GGA technique has no access to information on within-study variability and treats each study as equivalent in this regard. In comparison, conventional meta-analysis uses standardized effect sizes which incorporate a measure of this variability, serving to down-weight effects found in studies with high variability. The output of the GGA analyses was therefore compared with the output from analyses performed on the original data obtained directly from authors, allowing us to assess the impact of these potential drawbacks.

Digitizing error. Digitizing error could be easily measured by comparing the digitized data with the original data in the 27 studies of the validation dataset. The difference between the two data sources could either be as a result of the process used to digitize the figures, or discrepancies between the original data and the figures used in publication. To quantify the digitizing error, a second coder (naive to the hypothesis under test) repeated the digitizing process of the original coder for the whole of the validation dataset. This allowed us to assess the degree of error within a single coder (intracoder error), between the two coders (intercoder error), and between the main coder and the original data (coder-original error). Average errors for the RPE-FRN in the critical interval 228 ms–334 ms were as follows. The main coder showed an intracoder error of −.011 μv ($SD = .099$), and the secondary coder −.004 μv ($SD = .028$). Comparison of the two coders' average scores revealed an intercoder error of −.005 μv ($SD = .075$). Comparison of the main coder with original data revealed a coder-original error of .096 μv ($SD = .327$). Intra- and intercoder error was very low suggesting that an accurate digitizing of a published figure is unproblematic. Error rates between the main coder and the original data were higher than between the two coders, implying

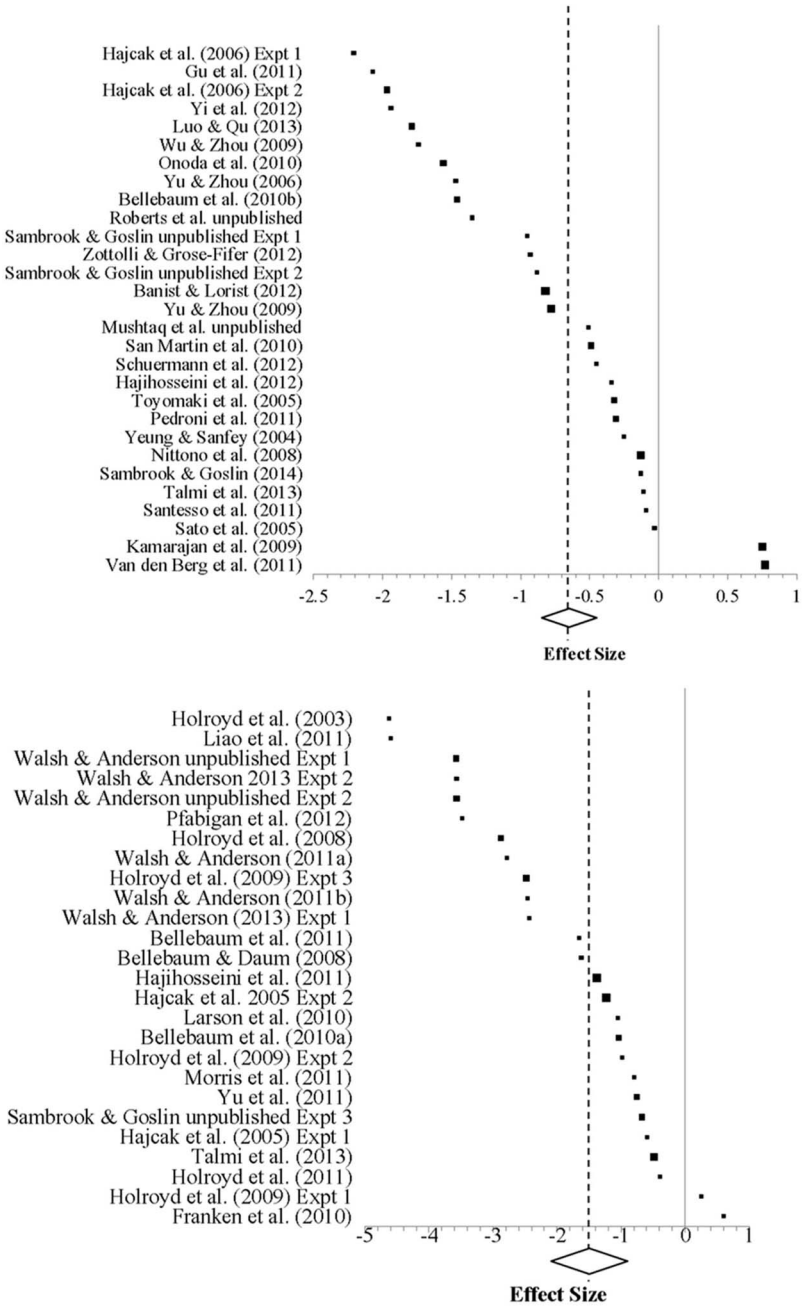

Figure 4. Forest plot showing RPE-FRN simple effect size in (a) magnitude and, (b) likelihood designs. Simple effect size was measured by mean amplitude in the interval 228 ms–334 ms. The size of squares indicates the sample size, which also constituted the weighting in the GGA meta-analysis. The diamond shows average weighted simple effect size and 95% confidence intervals. No confidence intervals could be computed for individual studies because of the GGA technique used.

that the main source of discrepancy lies with the preparation of graphs for publication. Examination on a case by case basis revealed that this was isolated to a few studies that would appear to have used a low-pass filter on the figure, but not the data, or a degree of erroneous vertical or horizontal translation of the waveform of one of the experiment's conditions. However, the amount of coder-original error is nevertheless very modest compared with the average simple effect sizes found in the GGA meta-analysis. Furthermore, it should be stressed that these digitizing errors did not affect the statistical testing applied to GGAs earlier, because original data was used in their stead. They merely give an estimate of the extent of the error in the remaining 27 studies for which no original data was available, and for the use of the technique generally.

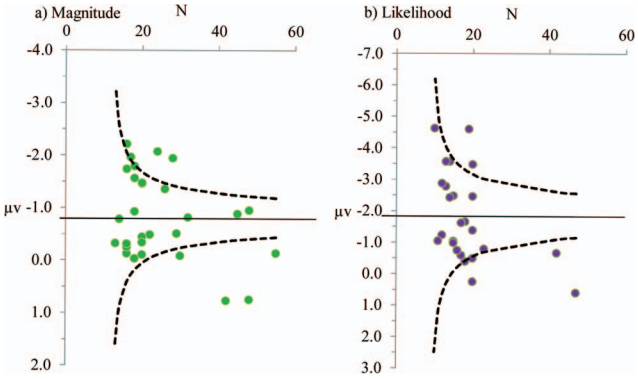

Figure 5. Funnel plots for the unweighted simple effect size of the RPE-FRN under (a) magnitude, and (b) likelihood. Dotted lines represent three standard deviations. See the online article for the color version of this figure.

Meta-analysis of original data. To assess the overall performance of the GGA technique, an "ideal" conventional meta-analysis was conducted using the same interval as used for the GGA meta-analysis, but with standardized effect sizes calculated from the original data in the validation dataset. The results of this meta-analysis were then compared with a GGA meta-analysis run on the same subsample of 27 studies. Results of both meta-analyses are given in Table 1. Quite aside from its role in validating the GGA technique it can be seen that the conventional

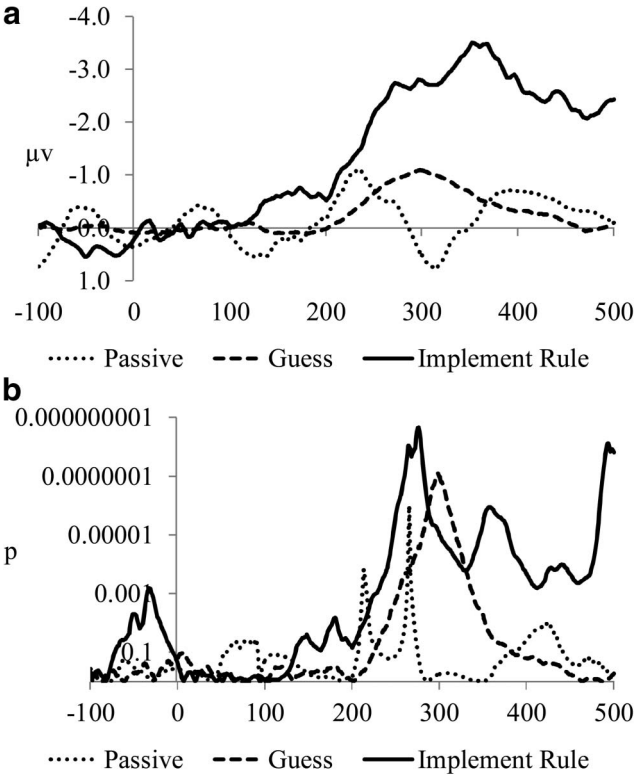

Figure 6. RPE-FRN at different levels of the "control over outcome" moderator: (a) simple effect size, and (b) significance of simple effect size.

Table 1
Results of a Conventional Versus GGA Meta-Analysis

		Conventional meta-analysis on validation set original data				GGA meta-analysis on validation set	
Modulator	Measure	Average standardized effect size	Significance	Heterogeneity	Effect after trim and fill	Average simple effect size (μv)	Significance
Magnitude	Mean	−.49 [−.67, −.31]	$z = 5.36$, $p < .001$	$\chi^2(13) = 11.75$, $p > .05$	−.49	−.79 [−1.06, −.42]	$t = 5.27$, $p < .001$
	Peak	−.51 [−.68, −.34]	$z = 6.01$, $p < .001$	$\chi^2(13) = 12.43$, $p > .05$	−.51	−.89 [−1.34, −.43]	$t = 4.12$, $p < .001$
Likelihood	Mean	−.71 [−.92, −.50]	$z = 6.55$, $p < .001$	$\chi^2(12) = 12.00$, $p > .05$	−.71	−2.17 [−2.75. −1.59]	$t = 6.19$, $p < .001$
	Peak	−.87 [−1.09, −.64]	$z = 7.43$, $p < .001$	$\chi^2(12) = 12.90$, $p > .05$	−.87	−2.31 [−3.09, −1.52]	$t = 4.97$, $p < .001$

meta-analysis also strongly supports this study's hypotheses, showing a significant RPE-FRN effect size under both the magnitude and likelihood modulators. With regard to validating the GGA technique here and more generally, it can be seen that the two meta-analytic methods give very close results in regard to significance testing of the mean amplitude measure. The z statistic from conventional meta-analysis and the t statistic from the GGA technique are very similar under both magnitude (5.36 vs. 5.27) and likelihood (6.55 vs. 6.19) modulators. For the peak measure, the ideal conventional meta-analysis reveals the GGA technique to have been conservative. This is to be expected, as the GGA technique measures the peak amplitude of grand averages rather than participant averages, and thus is subject to greater temporal smearing due to individual differences in latency across participants. Note that while Table 1 reports both average standardized effect size under ideal conventional meta-analysis and average simple effect size under the GGA technique, these should not be directly compared as they are denominated in different units. For GGAs they are measured in microvolts, for the conventional meta-analysis, in standard deviations of microvolts. Effect sizes for individual studies see Table S1 in the online supplemental materials. Note also that the validation dataset can be considered representative insofar as there was no significant difference in the RPE-FRN simple effect size of studies in or out of the validation dataset, $t(53) = 1.54$, $p = .13$.

Meta-analysis of published effect sizes. We also performed a conventional meta-analysis of published effects. As previously stated, we believe this is an unsound meta-analysis because it draws on effect sizes measured in different intervals and from quite different quantifications of the FRN (e.g., mean amplitude, peak of difference wave, peak to peak of simple waves). Nevertheless it is interesting for comparative purposes and furthermore permits a quantifying of the simpler "face value" of accumulated reporting findings regarding likelihood and magnitude modulators. The meta-analysis was performed on a reduced dataset because a number of articles did not report statistics for the RPE-FRN effect (see Appendix 1) The average standardized effect size for the magnitude modulator ($k = 15$) was nonsignificantly different from zero ($d = -.26$ [−.80, .29], z = .914, $p = .361$). The average standardized effect size for the likelihood modulator ($k = 18$) was however significant ($d = -.95$ [−1.34, −.56], z = 4.82, $p < .001$). Standardized effect sizes for individual studies see Table S2 in the online supplemental materials.

Main Effects of Magnitude and Likelihood

Although the principal objective of the study was to test for the existence of an RPE-FRN by examining the FRN's sensitivity to modulation by magnitude and likelihood, a consideration of these modulators' main effects is also valuable in interpreting the postfeedback waveform that FRN studies are likely to generate. Component overlap is an ever-present concern in ERP experiments and we felt it was very possible that an RPE-FRN would be superimposed on other components responding to magnitude, likelihood, or indeed valence, alone. Figure 7 represents all main effects in the form of difference waves. The RPE-FRN, calculated from magnitude and likelihood studies combined, is added for the purposes of comparison. Significance of main effects was determined using the Maris and Oostenveld technique. This revealed a magnitude main effect (Monte Carlo $p < .0001$), such that low magnitude outcomes were associated with a relative negativity in an interval running from 124 ms to the measurement boundary of 500 ms, with the effect greatest at 322 ms (−2.10 μv). Also revealed was a significant main effect of likelihood (Monte Carlo $p < .0001$), such that high likelihood outcomes were associated with a relative negativity in an interval running from 299 ms to the measurement boundary of 500 ms (Monte Carlo $p < .0001$), with the effect strongest at 426 ms (−3.51 μv). Finally there was a main effect of valence (Monte Carlo $p < .0001$), that is, an FRN, in the interval 150 ms–401 ms, with the effect greatest at 276 ms (−2.27 μv).

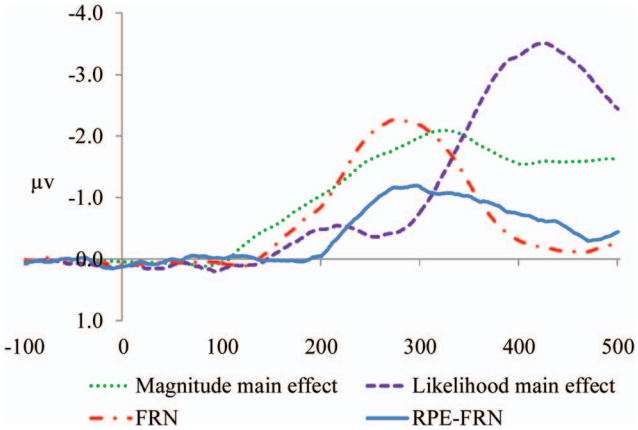

Figure 7. Main effects of magnitude, likelihood and valence (RPE-FRN shown for comparison). See the online article for the color version of this figure.

Discussion

The RPE-FRN and Main Effects of Valence, Magnitude, and Likelihood

Holroyd and Coles (2002) proposed that the FRN encoded an RPE. The results are consistent with this claim. FRNs created from large RPEs were of greater amplitude than those from small RPEs, both when RPE size was modulated by magnitude, and by likelihood. The demonstration that the FRN is responsive to variations in magnitude is important because it is a key requirement of a general RPE encoder, and evidence in previous experiments has largely been against this. The present meta-analysis shows that once quantification of the FRN is standardized, a clear magnitude effect on the FRN can be seen.

A number of recent articles have reported evidence consistent with the FRN constituting an unsigned prediction error or "salience" encoding (Hauser et al., 2014; Oliveira, McDonald, & Goodman, 2007; Talmi, Atkinson, & El-Deredy, 2013; Talmi, Fuentemilla, Litvak, Duzel, & Dolan, 2012). Such a component should show a strong main effect of RPE size (i.e., of likelihood and magnitude) but no main effect of valence, and no interaction of RPE size and valence (i.e., no RPE-FRN), because unsigned prediction errors should be insensitive to valence. The present study refutes this claim. Nevertheless, salience is clearly coded in the postfeedback waveforms, as shown by the strong main effects of likelihood and magnitude in Figure 7, with these main effects

approximately twice the size of the RPE-FRN to which each modulator contributes. The later time course of these effects suggests that they may well be P3 effects.

Regardless of their source, the fact that multiple components contribute to activity at frontocentral electrodes touches on an important conceptual point. This meta-analysis shows that frontocentral activity in the interval in which the FRN is typically measured is responsive to the main effects of magnitude, likelihood, and valence, and also to the interaction of valence with both magnitude and likelihood. It appears that multiple components operate in this interval. However, the debate following the publication of Holroyd and Coles' theory has crystallized around the idea of a single component in this interval, whose character will ultimately be resolved through careful experimentation. In practice, we suspect that the character of the component described by a given experiment as the "FRN" will depend strongly on the interval in which this component is measured. For example, in Talmi, Fuentemilla, Litvak, Duzel, and Dolan (2012) and Hauser et al.'s (2014) articles, evidence was presented favoring a salience account, however measurement was made at the latency of maximal FRN amplitude, that is the maximal main effect of valence. Although this was a pragmatic choice and based on precedent, this latency was nevertheless not necessarily one best suited to demonstrating an RPE-FRN if there was one to be found, since that is shown by a Valence × RPE Size interaction, not a valence main effect. In practice, this resulted in these papers measuring effects at ~220 ms. However, this was prior to the period where the RPE-FRN was observed in this meta-analysis but where salience effects were marked in magnitude and close to significance for likelihood. In the FRN debate generally, we suspect that the sensitivity of the FRN to the key factors that are used to infer its function has depended on the latency of its measurement to a degree which has not been fully appreciated.

It is possible that in the future, separation of these components may be assisted by improved knowledge about their scalp distributions. Because of the limited and variable electrode arrays available, the present meta-analysis cannot offer guidance here. Furthermore, the FRN itself is partly defined by being maximal over frontocentral sites. Given that is a now well-established definition, it is likely any published example of the FRN would also have to demonstrate a frontocentral maximum, in order to be accepted as such. Therefore, any meta-analysis of the FRN would be very likely to reflect this established scalp distribution. In contrast, it seems likely that the later strong likelihood effect, and possibly magnitude effect shown in Figure 7 are P3 effects and would be maximal at more parietal locations.

The RPE-FRN was stronger when participants were engaged in a task over which they had reason to believe they enjoyed some control. In the strongest case of control, where participants implemented a known rule, the RPE-FRN also lasted much longer, as can be seen in Figure 6. These results suggest the possibility that the RPE-FRN might be selectively recruited by the apparatus of instrumental conditioning, rather than acting as a general purpose representation of value. Some caution must be exercised in regard to this finding, first, because subjective involvement was probably lower with reduced control (Yeung, Holroyd, & Cohen, 2005) and, second, because 10 of the 12 studies used for the "rule implementation" level of this moderator came from experiments conducted by just two authors.

The RPE-FRN in magnitude studies was unaffected by whether the magnitude of the forthcoming outcome was cued in advance. As such, it appears that the RPE-FRN does not scale RPEs to the range of outcomes on a given trial. We do not believe this should be regarded as evidence that RPEs are genuinely coded on an absolute scale however, because this would be functionally extremely limited and is biologically implausible. Scaling, or "adaptation," is a ubiquitous feature of sensory processes, allowing, for example, the eye to discriminate luminance over nine orders of magnitude despite only three orders of contrast being available at a given moment. We would expect such a solution to be used for evaluating RPEs, which likewise have a very broad range As such, this moderator analysis suggests that outcomes are not scaled to the range of magnitudes available on a *trial*, but the wider context of the *experiment*. Nevertheless, this is an interesting result, because it suggests that the expected value term against which RPEs are calculated may not simply be inherited from the midbrain dopamine system, or at least those midbrain dopaminergic neurons that have shown strong scaling effects (Tobler et al., 2005), and is thus relevant to the ongoing question of the afferents of the FRN.

Applications of the Present Findings

The FRN is a robustly elicited component, easy to study in human participants, and appears to encode an RPE. It may thus contribute to the daunting task of uncovering the network of neural events that give rise to subjective valuation by humans. Holroyd and Coles' theory of the FRN was focused on its role in reinforcement learning, rather than its role as a general index of subjective value. However, the relationship between reinforcement learning and valuation is close. The information concerning action-reward contingencies that is held in a reinforcement learning system presumably strongly informs the valuation of the actions available to people in a given situation. Thus, if it can be measured (e.g., by the FRN) it is has predictive power for human choice of the kind that neuroeconomics strives to attain.

The nature of the reinforcement learning system underlying the FRN is therefore pertinent. Reinforcement learning falls into two broad classes, model-free and model-based. Model-free reinforcement learning assigns values to actions based on the net reward they can expect to incur, without consideration of the actual outcomes that are produced. The values are updated in light of RPEs, but are termed "habit values" because they encode only the historical value of an action. Such learning is computationally efficient and information poor because the structure of rewards and the probabilities that follow an action is cached into a single value. Model-based reinforcement learning uses a model of the environment which represents actions, rewards, and intermediate states, and calculates values of actions by a tree search of this model. Although more computationally expensive, this can be more quickly updated. A recent review of model-free and model-based reinforcement learning is provided by Walsh and Anderson (2014).

The relevance of this distinction to human choice is that model-based reinforcement learning is likely to be continuous with general cognition (Chater, 2009). Thus the degree to which choice on any one occasion is influenced by wider knowledge, by deliberative reasoning, or by verbal instruction will depend on the degree to which a model-free habitual system or a model-based belief system is dominant at that time. If the FRN can be established as

belonging to one system or the other, it can be used as a much more direct means to investigate the relative contributions of habitual and belief based valuation to behavior, and assist in accounting for variations in both inter- and intraindividual choice that elude the revealed preferences method.

Although the present demonstration that the FRN encodes an RPE places the debate on a much firmer footing, there has nevertheless been limited work on this important question. Hajcak, Moser, Holroyd, and Simons (2007) and Moser and Simons (2009) both showed a relationship of FRN amplitude to RPEs generated against subjective predictions but not to reinforcement history, implying the component might arise from model-based reinforcement learning, whereas Ichikawa, Siegle, Dombrovski, and Ohira (2010) found comparable contributions of subjective prediction and reinforcement history to FRN amplitude. However, Walsh and Anderson (2011b) found persuasive evidence against model-based reinforcement learning. They compared the FRN in cases where participants received verbal instruction on choice-outcome contingencies to cases where they did not. In the instruction condition, participants used this instruction, as shown by their behavior, thus adopting the given "model." However, when unexpected outcomes, that is model-based RPEs, occurred, the FRN was initially insensitive to these. Its sensitivity developed only at the rate shown in the no-instruction condition suggesting it was dependent on a model-free history of reinforcement. A number of other authors have been able to show that FRN amplitude corresponds to the size of RPEs derived from a model-free reinforcement learning algorithm (Chase, Swainson, Durham, Benham, & Cools, 2011; Cohen & Ranganath, 2007; Philiastides, Biele, Vavatzanidis, Kazzer, & Heekeren, 2010). Other evidence for a model-free basis for the FRN comes from the demonstration that dopamine, the neurotransmitter implicated in generating the FRN, promotes model-free rather than model-based reinforcement learning (Wunderlich, Smittenaar, & Dolan, 2012). On current balance the evidence favors the FRN's role in model-free reinforcement learning.

Insofar as model-free reinforcement learning is computationally cheap, it might be expected to occur by default, and indeed, to continue to compute valuations and associated RPEs even when a superior model-based reinforcement learning system was guiding behavior. Bayer and Glimcher (2005), for example, showed that midbrain dopaminergic neurons, which are believed to underlie the FRN, showed firing patterns consistent with a model-free RPE and continued to behave in this fashion even when their effect on behavior was weak. In the case of the FRN itself, the component has in some cases been shown to predict choice in a way that is consistent with reinforcement learning (Cohen & Ranganath, 2007; Van der Helden, Boksem, & Blom, 2010; Yasuda, Sato, Miyawaki, Kumano, & Kuboki, 2004), but in other cases it has not (Mars, De Bruijn, Hulstijn, Miltner, & Coles, 2004; Mas-Herrero & Marco-Pallarés, 2014; San Martín, Appelbaum, Pearson, Huettel, & Woldorff, 2013; Yeung & Sanfey, 2004). In particular, Chase, Swainson, Durham, Benham, and Cools (2011) showed that in a reversal learning task, the nature of which would be expected to engage model-based reinforcement learning, an FRN was observed that was well described by model-free reinforcement learning but which nevertheless did not predict behavior, suggesting it was overridden by a model-based system. Findings such as these suggest that the FRN might be used to predict behavior in situations promoting relatively automatic, fast judgments, what has been described by dual process theories as System 1 (Kahneman, 2003). Such valuation has been underrepresented by the traditional methods of behavioral economics, which rely on stated (rather than observed) preferences in one-shot (rather than repeated) choices, which place prominence on deliberative processing. However, perhaps the most serious challenge that the studies cited above pose for behavioral and neoclassical economics lies in the possibility that rather than value being constructed from multiple terms, as is suggested for example by prospect theory, quite separate independent valuations might be constructed which have differential access to behavior depending on circumstances.

Even while the precise nature of the valuation associated with the FRN remains unresolved, it may nevertheless serve as a biomarker for subjective value. It has been proposed in this regard for a range of psychopathologies such as hypomania and depression (Bress, Smith, Foti, Klein, & Hajcak, 2012; Mason, O'Sullivan, Bentall, & El-Deredy, 2012) 2012) and pathological gambling (Hewig et al., 2010). Furthermore, a number of recent studies have shown that variation in dopaminergic genes affects the component (e.g., Foti & Hajcak, 2012; Marco-Pallarés et al., 2009) raising the possibility that it might be used to investigate the proximate basis of genetic effects on behavior. With the advent of mobile electroencephalography (EEG) setups that can be ready to use within minutes, the FRN may also provide a useful general measure of the subjective value of an outcome even in studies in which the brain is not the principal focus, much as other psychophysiological techniques such as skin conductance and pupillometry are used more broadly. As a dependent variable of subjective value it has a number of advantages over self-report. Asking subjects to report on their valuations brings in extra processes which generally undermines the ecological validity of the study of "online" evaluation. Reported valuations may be subject to demand characteristics because participants are likely to be aware of at least some norms in economic preference, such as avoiding obvious inconsistencies and intransitivities. Self-report may also be affected by what reference point the stated valuation is taken with respect to, which depends in turn on the framing of the question used to prompt self-report.

Alternative Accounts of the FRN

A number of tasks elicit a frontocentral negativity, or N2, at the latency of the FRN (see Folstein & Van Petten, 2008, for a review), and as such, alternative accounts of the FRN exist. One of these is that it is merely an oddball, detecting the unexpectedness of events. This is rather close to the claim that it simply codes salience which has been disconfirmed in this meta-analysis. Attempts to experimentally dissociate the FRN and N2 oddball have met with some success (Holroyd, Pakzad-Vaezi, & Krigolson, 2008; Warren & Holroyd, 2012).

The N2 is believed to indicate activity in the anterior cingulate cortex (Nieuwenhuis, Yeung, Van Den Wildenberg, & Ridderinkhof, 2003; Yeung, Botvinick, & Cohen, 2004). Botvinick, Braver, Barch, Carter, and Cohen (2001) have claimed that the ACC is responsible for cognitive control, becoming active when response conflict occurs, and Brown and Braver (2005) have made the related claim that the anterior cingulate cortex detects the likelihood of errors. Indeed, circumstances in which cognitive control and error likelihood are high do increase N2 amplitude, for example on no-go trials in a go/no-go

task (Folstein & Van Petten, 2008). The theories can account for the FRN's response to reward if nonreward is regarded as an error, thus signaling the need for increased cognitive control. Furthermore, a different component, the error related negativity shares a common scalp distribution with the FRN, and is strongly implicated in these functions, inasmuch as it indicates internal registration of an error. In fact, Holroyd and Coles' theory also specifies a functional relationship between these two components, arguing that they both reflect RPEs arising from a sudden revision of reward expectation, either by external feedback in the case of the FRN or internal monitoring in the case of the error related negativity.

A further alternative account of the FRN is that it is an affective rather than economic response to outcomes (Gehring & Willoughby, 2002; Luu, Tucker, Derryberry, Reed, & Poulsen, 2003).This is rather difficult to disentangle from the RPE account because of the affective nature of reward. However, those studies that have compared affective ratings of outcomes with the FRN amplitudes associated with them have tended to find a poor relationship (Li, Han, Lei, Holroyd, & Li, 2011; Sambrook et al., 2012; Yang, Gu, Tang, & Luo, 2013).

Implications for the Measurement of the RPE-FRN and FRN

We have distinguished between a response simply to valence, the FRN, well established in the literature, and a neural response to the valence and size of an RPE, the RPE-FRN, for which we have presented evidence here. The distinction is important for the testing of Holroyd and Coles' theory. However, it is not widely made in the literature and the comments below apply equally to both FRN and RPE-FRN.

The present meta-analysis revealed a wide variation in methods used to quantify the FRN, and we have noted the role this may play both in failures of replication and inflation of false positives. We have also noted the variability of the waveforms themselves. These two aspects are linked insofar as inconsistencies in FRN quantification possibly reflect the genuine attempt to best tailor analysis to a component of seemingly inconsistent character on an experiment-by-experiment basis. However, if, as we have argued, variability in the waveforms largely reflects the vagaries of component overlap rather than real variability in the FRN, then this latitude in quantification is harmful. For example, in the present meta-analysis, P2 and N2 peaks varied so much in their latency across experiments that while we initially intended to apply the GGA technique to a peak to peak measure, implemented in standardized intervals, we were unable to do so. This illustrates the point that while peaks might provide compelling landmarks by which to detect the FRN in any *individual* study, the lack of consistency *across* studies suggests the benefits of locking FRN quantification to simple waveform peaks may be illusory. The loose relationship between single waveform peaks and the underlying components has been cogently described by Luck (2005).

As such, measures based on difference waves are to be preferred. For the specific case of the RPE-FRN, a measurement interval of 270 ms–300 ms is suggested by the present study since this captures the strongest effects of both magnitude and likelihood and is thus the best estimate of the RPE-FRN's latency. However, the RPE-FRN in individual experiments may be subject to genuine latency differences and so, based on the course of the effect under both modulators, the interval 240 ms–340 ms may be more appropriate. It should be noted that studies which more effectively decompose waveforms into constituent components, for example using principal components analysis, may reveal a rather different latency for the underlying RPE encoder, or encoders. Indeed Figure 7 suggests that such decomposition may well be necessary to fully isolate the individual components.

Evaluation of the GGA Technique

The GGA technique was developed because the great variety in how the FRN was quantified rendered conventional meta-analysis highly problematic. It is worthwhile assessing how this technique performed, partly in judging the present findings, but also for its future use in ERP meta-analysis. First, our concerns regarding the conventional meta-analysis of the FRN using effect sizes derived from diverse quantifications proved justified. When such a meta-analysis was performed it failed to find a significant effect of magnitude on the FRN, despite this effect being strongly present in an ideal conventional meta-analysis on original data. In contrast, the GGA technique was in close agreement. The conclusion we draw from the superior performance of the GGAs is that it is more important to employ an appropriate and consistent quantification of a component than to have access to the measures of within-study variance that use of published statistics provides. Of course the ideal meta-analysis achieved both of these objectives. However, the GGA technique only requires access to published data. This has a great number of advantages. Most importantly, it avoids the large reduction in sample size that reliance on original authors inevitably entails. It substantially reduces the effort required to acquire data and convert it to a common format, and makes no demands at all on the original authors. It removes the uncertainty surrounding the number of studies that the meta-analysis will contain, allowing the viability of the exercise to be assessed in advance. It avoids the danger of bias arising from authors selectively complying with the request for original data depending on what they perceive the meta-analyzer's hypothesis to be. Finally, the technique can be used to guide the development of future work. If an effect, component, or other subset of the ERP in a published study was not selected for analysis within that study, there will not be any effect sizes on which to base a traditional meta-analysis. The GGA technique allows for post hoc exploration of published ERPs, allowing the researcher to approximate the effect sizes of previously disregarded data to guide the design of new empirical study, theory, or analysis technique. It is for this reason that we have made available the grand averages used in this meta-analysis as supplementary files.

The GGA technique has some disadvantages. Simple, rather than standardized effect sizes were used, meaning that the GGA meta-analysis could not down-weight studies with large variance, thus introducing some noise into hypothesis testing. The extent of this can be simply estimated from the validation dataset by calculating the correlation coefficient of the simple effect sizes of the GGA technique and the standardized effect sizes of the ideal conventional meta-analysis: the lower the correlation, the greater the noise introduced by failure to use standardized effect size. The value was $r = .8$, suggesting a moderate degree of noise introduced. This is, however, an overestimate of the problem insofar as standardized effect sizes themselves are not perfect because the standard deviations they are built from are themselves subject to sampling error. The remaining source of noise in the GGA technique consists in deviations between the digitized waveforms used for the meta-analysis and the original data;

however, comparisons with the validation dataset set suggest this is relatively small.

It should be noted that differences between experiments regarding the reference electrode, filters, and baseline do not impact on the GGA technique because all contrasts and simple effect sizes are generated *within-experiment*, and so these extraneous factors can never become confounds for the simple generic reason that they are held constant at the point of generating simple effect sizes. It is true that FRN amplitude itself may be affected by these parameters, and that a poorly chosen reference electrode, for example, might reduce FRN amplitudes overall, concomitantly reduce effect sizes, and assist in rendering a meta-analysis nonsignificant. However, in this regard GGAs do not differ from conventional meta-analysis. We simply note that because reference electrode is held constant within each study it does not confound the simple effect size generated for that individual study, and because this meta-analysis is simply a collation of such simple effect sizes it likewise cannot be confounded by reference electrode.

We propose the GGA technique as a general method for meta-analysis of ERP components, not just the FRN. While inconsistency of measurement has been shown to be a particular problem for the FRN, this is also likely to be true to some degree of other components. Furthermore, even when conventional meta-analysis is applied, we still propose that this be performed in concert with a GGA analysis, to check there is no gross difference in results. As an accompanying method it also has the advantage that it allows the plotting of a waveform to accompany the reported effects. Individual ERP experiments ubiquitously plot an entire waveform despite their reported effects occurring in a small portion of the waveform because it provides a "sanity check" that the ERP shows a representative character and that the interval chosen for analysis is reasonable. A GGA waveform serves the same function in the case of a meta-analysis.

Conclusion

Neuroeconomics attempts to explain valuation by the brain. The present study addressed this question at the relatively large scale of EEG. It found that an easily elicited electrophysiological component, the FRN, behaved in manner consistent with it representing valuation of an outcome. Because of the temporal precision of EEG and the inherent benefits of convergent evidence from different methodologies, it is to be hoped that further study of the FRN will assist in uncovering the full picture of how the brain represents subjective value.

References

Asterisks denote studies in the GGA meta-analysis, double asterisks denote studies in the validation dataset.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100,* 603–617. http://dx.doi.org/10.1348/000712608X377117

Balleine, B. W., Daw, N. D., & O'Doherty, J. P. (2008). Multiple forms of value learning and the function of dopamine. *Neuroeconomics: Decision Making and the Brain, 2008,* 367–385.

**Banis, S., & Lorist, M. M. (2012). Acute noise stress impairs feedback processing. *Biological Psychology, 91,* 163–171. http://dx.doi.org/10.1016/j.biopsycho.2012.06.009

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage, 76,* 412–427. http://dx.doi.org/10.1016/j.neuroimage.2013.02.063

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47,* 129–141. http://dx.doi.org/10.1016/j.neuron.2005.05.020

**Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience, 27,* 1823–1835. http://dx.doi.org/10.1111/j.1460-9568.2008.06138.x

**Bellebaum, C., Kobza, S., Thiele, S., & Daum, I. (2010a). It was not MY fault: Event-related brain potentials in active and observational learning from feedback. *Cerebral Cortex, 20,* 2874–2883. http://dx.doi.org/10.1093/cercor/bhq038

**Bellebaum, C., Kobza, S., Thiele, S., & Daum, I. (2011). Processing of expected and unexpected monetary performance outcomes in healthy older subjects. *Behavioral Neuroscience, 125,* 241–251. http://dx.doi.org/10.1037/a0022536

**Bellebaum, C., Polezzi, D., & Daum, I. (2010b). It is less than you expected: The feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia, 48,* 3343–3350. http://dx.doi.org/10.1016/j.neuropsychologia.2010.07.023

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406–418. http://dx.doi.org/10.1037/1082-989X.8.4.406

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652. http://dx.doi.org/10.1037/0033-295X.108.3.624

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,* 156–162. http://dx.doi.org/10.1016/j.biopsycho.2011.10.004

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307,* 1118–1121. http://dx.doi.org/10.1126/science.1105783

Bunzeck, N., Dayan, P., Dolan, R. J., & Duzel, E. (2010). A common mechanism for adaptive scaling of reward and novelty. *Human Brain Mapping, 31,* 1380–1394. http://dx.doi.org/10.1002/hbm.20939

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience, 23,* 936–946. http://dx.doi.org/10.1162/jocn.2010.21456

Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. *Cognition, 113,* 350–364. http://dx.doi.org/10.1016/j.cognition.2008.06.014

Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *The Journal of Neuroscience, 27,* 371–378. http://dx.doi.org/10.1523/JNEUROSCI.4421-06.2007

Diekhof, E. K., Kaps, L., Falkai, P., & Gruber, O. (2012). The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude - an activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and outcome processing. *Neuropsychologia, 50,* 1252–1266. http://dx.doi.org/10.1016/j.neuropsychologia.2012.02.007

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177. http://dx.doi.org/10.1037/1082-989X.1.2.170

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63,* 665–694. http://dx.doi.org/10.1348/000711010X502733

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology, 45,* 152–170.

Foti, D., & Hajcak, G. (2012). Genetic variation in dopamine moderates neural response during reward anticipation and delivery: Evidence from event-related potentials. *Psychophysiology, 49,* 617–626. http://dx.doi.org/10.1111/j.1469-8986.2011.01343.x

*Franken, I. H., Van den Berg, I., & Van Strien, J. W. (2010). Individual differences in alcohol drinking frequency are associated with electrophysiological responses to unexpected nonrewards. *Alcoholism: Clinical and Experimental Research, 34,* 702–707. http://dx.doi.org/10.1111/j.1530-0277.2009.01139.x

Friedman, M. (1953). The methodology of positive economics. *Essays in Positive Economics,* 3–16.

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 37,* 1297–1310. http://dx.doi.org/10.1016/j.neubiorev.2013.03.023

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295,* 2279–2282. http://dx.doi.org/10.1126/science.1066893

Glimcher, P. W. (2009). *Choice: Towards a standard back-pocket model*. San Diego, CA: Elsevier Academic Press.

Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A. (2009). *Introduction: A brief history of neuroeconomics*. San Diego, CA: Elsevier Academic Press.

**Gu, R., Lei, Z., Broster, L., Wu, T., Jiang, Y., & Luo, Y. J. (2011). Beyond valence and magnitude: A flexible evaluative coding system in the brain. *Neuropsychologia, 49,* 3891–3897. http://dx.doi.org/10.1016/j.neuropsychologia.2011.10.006

*Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology, 42,* 161–170. http://dx.doi.org/10.1111/j.1469-8986.2005.00278.x

*Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology, 71,* 148–154. http://dx.doi.org/10.1016/j.biopsycho.2005.04.001

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology, 44,* 905–912. http://dx.doi.org/10.1111/j.1469-8986.2007.00567.x

**HajiHosseini, A., Rodríguez-Fornells, A., & Marco-Pallarés, J. (2012). The role of beta-gamma oscillations in unexpected rewards processing. *NeuroImage, 60,* 1678–1685. http://dx.doi.org/10.1016/j.neuroimage.2012.01.125

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage, 84,* 159–168. http://dx.doi.org/10.1016/j.neuroimage.2013.08.028

Hewig, J., Kretschmer, N., Trippe, R. H., Hecht, H., Coles, M. G., Holroyd, C. B., & Miltner, W. H. (2010). Hypersensitivity to reward in problem gamblers. *Biological Psychiatry, 67,* 781–783. http://dx.doi.org/10.1016/j.biopsych.2009.11.009

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709. http://dx.doi.org/10.1037/0033-295X.109.4.679

*Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective & Behavioral Neuroscience, 9,* 59–70. http://dx.doi.org/10.3758/CABN.9.1.59

*Holroyd, C. B., Krigolson, O. E., & Lee, S. (2011). Reward positivity elicited by predictive cues. *NeuroReport, 22,* 249–252. http://dx.doi.org/10.1097/WNR.0b013e328345441d

*Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport, 14,* 2481–2484. http://dx.doi.org/10.1097/00001756-200312190-00037

*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,* 688–697. http://dx.doi.org/10.1111/j.1469-8986.2008.00668.x

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Atlanta, GA: Sage.

Ichikawa, N., Siegle, G. J., Dombrovski, A., & Ohira, H. (2010). Subjective and model-estimated reward prediction: Association with the feedback-related negativity (FRN) and reward prediction error in a reinforcement learning task. *International Journal of Psychophysiology, 78,* 273–283. http://dx.doi.org/10.1016/j.ijpsycho.2010.09.001

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review, 93,* 1449–1475. http://dx.doi.org/10.1257/000282803322655392

*Kamarajan, C., Porjesz, B., Rangaswamy, M., Tang, Y., Chorlian, D. B., Padmanabhapillai, A., . . . Begleiter, H. (2009). Brain signatures of monetary loss and gain: Outcome-related potentials in a single outcome gambling task. *Behavioural Brain Research, 197,* 62–76. http://dx.doi.org/10.1016/j.bbr.2008.08.011

Kreussel, L., Hewig, J., Kretschmer, N., Hecht, H., Coles, M. G. H., & Miltner, W. H. R. (2012). The influence of the magnitude, probability, and valence of potential wins and losses on the amplitude of the feedback negativity. *Psychophysiology, 49,* 207–219. http://dx.doi.org/10.1111/j.1469-8986.2011.01291.x

Kujawa, A., Smith, E., Luhmann, C., & Hajcak, G. (2013). The feedback negativity reflects favorable compared to nonfavorable outcomes based on global, not local, alternatives. *Psychophysiology, 50,* 134–138. http://dx.doi.org/10.1111/psyp.12002

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328. http://dx.doi.org/10.1037/0022-3514.32.2.311

*Larson, M. J., Kelly, K. G., Stigge-Kaufman, D. A., Schmalfuss, I. M., & Perlstein, W. M. (2007). Reward context sensitivity impairment following severe TBI: An event-related potential investigation. *Journal of the International Neuropsychological Society, 13,* 615–625. http://dx.doi.org/10.1017/S1355617707070762

Li, P., Han, C., Lei, Y., Holroyd, C. B., & Li, H. (2011). Responsibility modulates neural mechanisms of outcome processing: An ERP study. *Psychophysiology, 48,* 1129–1133. http://dx.doi.org/10.1111/j.1469-8986.2011.01182.x

*Liao, Y., Gramann, K., Feng, W., Deák, G. O., & Li, H. (2011). This ought to be good: Brain activity accompanying positive and negative expectations and outcomes. *Psychophysiology, 48,* 1412–1419. http://dx.doi.org/10.1111/j.1469-8986.2011.01205.x

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 35,* 1219–1236. http://dx.doi.org/10.1016/j.neubiorev.2010.12.012

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

**Luo, Q., & Qu, C. (2013). Comparison enhances size sensitivity: Neural correlates of outcome magnitude processing. *PLoS ONE, 8,* e71186. http://dx.doi.org/10.1371/journal.pone.0071186

Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of

action regulation. *Psychological Science, 14,* 47–53. http://dx.doi.org/10.1111/1467-9280.01417

Marco-Pallarés, J., Cucurell, D., Cunillera, T., Krämer, U. M., Càmara, E., Nager, W., . . . Rodriguez-Fornells, A. (2009). Genetic variability in the dopamine system (dopamine receptor D4, catechol-*O*-methyltransferase) modulates neurophysiological responses to gains and losses. *Biological Psychiatry, 66,* 154–161. http://dx.doi.org/10.1016/j.biopsych.2009.01.006

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190. http://dx.doi.org/10.1016/j.jneumeth.2007.03.024

Mars, R. B., De Bruijn, E. R., Hulstijn, W., Miltner, W. H., & Coles, M. G. (2004). What if I told you: "You were wrong"? Brain potentials and behavioral adjustments elicited by feedback in a time-estimation task. In M. Ullsperger & M. Falkenstein (Eds.), *Errors, conflicts, and the brain. Current opinions on performance monitoring* (pp. 129–134). Leipzig, Germany: MPI of Cognitive Neuroscience.

Mas-Herrero, E., & Marco-Pallarés, J. (2014). Frontal theta oscillatory activity is a common mechanism for the computation of unexpected outcomes and learning rate. *Journal of Cognitive Neuroscience, 26,* 447–458. http://dx.doi.org/10.1162/jocn_a_00516

Mason, L., O'Sullivan, N., Bentall, R. P., & El-Deredy, W. (2012). Better than I thought: Positive evaluation bias in hypomania. *PLoS ONE, 7,* e47754. http://dx.doi.org/10.1371/journal.pone.0047754

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience, 9,* 788–798. http://dx.doi.org/10.1162/jocn.1997.9.6.788

Morris, S. E., Holroyd, C. B., Mann-Wrobel, M. C., & Gold, J. M. (2011). Dissociation of response and feedback negativity in schizophrenia: Electrophysiological and computational evidence for a deficit in the representation of value. *Frontiers in Human Neuroscience, 5,* 123. http://dx.doi.org/10.3389/fnhum.2011.00123

Moser, J. S., & Simons, R. F. (2009). The neural consequences of flip-flopping: The feedback-related negativity and salience of reward prediction. *Psychophysiology, 46,* 313–320. http://dx.doi.org/10.1111/j.1469-8986.2008.00760.x

Mushtaq, F., Stoet, G., Bland, A. R., & Schaefer, A. (2013). Relative changes from prior reward contingencies can constrain brain correlates of outcome monitoring. *PLoS ONE, 8,* e66350. http://dx.doi.org/10.1371/journal.pone.0066350

**Mushtaq, F. M. W. R., Mon-Williams, M., & Schaefer, A. (2014). Unpublished data, University of Leeds and Monash University.

Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective & Behavioral Neuroscience, 3,* 17–26. http://dx.doi.org/10.3758/CABN.3.1.17

*Nittono, H., Otsuka, Y., & Ullsperger, P. (2008). Asymmetrical effects of frequent gains and frequent losses in a gambling task. *NeuroReport, 19,* 1345–1349. http://dx.doi.org/10.1097/WNR.0b013e32830c21a8

Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience, 19,* 1994–2004. http://dx.doi.org/10.1162/jocn.2007.19.12.1994

**Onoda, K., Abe, S., & Yamaguchi, S. (2010). Feedback-related negativity is correlated with unplanned impulsivity. *NeuroReport, 21,* 736–739.

**Pedroni, A., Langer, N., Koenig, T., Allemand, M., & Jäncke, L. (2011). Electroencephalographic topography measures of experienced utility. *The Journal of Neuroscience, 31,* 10474–10480. http://dx.doi.org/10.1523/JNEUROSCI.5488-10.2011

**Pfabigan, D. M., Alexopoulos, J., Bauer, H., & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology, 48,* 656–664. http://dx.doi.org/10.1111/j.1469-8986.2010.01136.x

Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzer, P., & Heekeren, H. R. (2010). Temporal dynamics of prediction error processing during reward-based decision making. *NeuroImage, 53,* 221–232. http://dx.doi.org/10.1016/j.neuroimage.2010.05.052

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400,* 233–238. http://dx.doi.org/10.1038/22268

**Roberts, K., Infantolino, Z., Stanley, E. M., & Simons, R. F. (2013). Unpublished data, University of Delaware.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412984997

**Sambrook, T. D., & Goslin, J. (2014). Mediofrontal event-related potentials in response to positive, negative and unsigned prediction errors. *Neuropsychologia, 61,* 1–10. http://dx.doi.org/10.1016/j.neuropsychologia.2014.06.004

**Sambrook, T. D., & Goslin, J. (2013). Unpublished data Experiment 1, University of Plymouth.

**Sambrook, T. D., & Goslin, J. (2014). Unpublished data Experiment 2, University of Plymouth.

**Sambrook, T. D., & Goslin, J. (2013). Unpublished data Experiment 3, University of Plymouth.

Sambrook, T. D., Roser, M., & Goslin, J. (2012). Prospect theory does not describe the feedback-related negativity value function. *Psychophysiology, 49,* 1533–1544. http://dx.doi.org/10.1111/j.1469-8986.2012.01482.x

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies, 4,* 155–161. http://dx.doi.org/10.2307/2967612

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience, 6,* 304. http://dx.doi.org/10.3389/fnhum.2012.00304

San Martín, R., Appelbaum, L. G., Pearson, J. M., Huettel, S. A., & Woldorff, M. G. (2013). Rapid brain responses independently predict gain maximization and loss minimization during economic decision making. *The Journal of Neuroscience, 33,* 7011–7019. http://dx.doi.org/10.1523/JNEUROSCI.4242-12.2013

*San Martín, R., Manes, F., Hurtado, E., Isla, P., & Ibañez, A. (2010). Size and probability of rewards modulate the feedback error-related negativity associated with wins but not losses in a monetarily rewarded gambling task. *NeuroImage, 51,* 1194–1204. http://dx.doi.org/10.1016/j.neuroimage.2010.03.031

**Santesso, D. L., Dzyundzyak, A., & Segalowitz, S. J. (2011). Age, sex and individual differences in punishment sensitivity: Factors influencing the feedback-related negativity. *Psychophysiology, 48,* 1481–1489. http://dx.doi.org/10.1111/j.1469-8986.2011.01229.x

*Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., & Kuboki, T. (2005). Effects of value and reward magnitude on feedback negativity and P300. *NeuroReport, 16,* 407–411. http://dx.doi.org/10.1097/00001756-200503150-00020

*Schuermann, B., Endrass, T., & Kathmann, N. (2012). Neural correlates of feedback processing in decision-making under risk. *Frontiers in Human Neuroscience, 6,* 204. http://dx.doi.org/10.3389/fnhum.2012.00204

Schultz, W. (2009). Midbrain dopamine neurons: A retina of the reward system. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 321–328). London, UK: Academic Press.

Schultz, W. (2010). Dopamine signals for reward value and risk: Basic and recent data. *Behavioral and Brain Functions, 6,* 24. http://dx.doi.org/10.1186/1744-9081-6-24

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge, MA: MIT Press.

**Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors. *The Journal of Neuroscience, 33,* 8264–8269. http://dx.doi.org/10.1523/JNEUROSCI.5695-12.2013

Talmi, D., Fuentemilla, L., Litvak, V., Duzel, E., & Dolan, R. J. (2012). An MEG signature corresponding to an axiomatic model of reward prediction error. *NeuroImage, 59,* 635–645. http://dx.doi.org/10.1016/j.neuroimage.2011.06.051

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307,* 1642–1645. http://dx.doi.org/10.1126/science.1105370

*Toyomaki, A., & Murohashi, H. (2005). Discrepancy between feedback negativity and subjective evaluation in gambling. *NeuroReport, 16,* 1865–1868. http://dx.doi.org/10.1097/01.wnr.0000185962.96217.36

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

*Van den Berg, I., Franken, I. H. A., & Muris, P. (2011). Individual differences in sensitivity to reward association with electrophysiological responses to monetary gains and losses. *Journal of Psychophysiology, 25,* 81–86. http://dx.doi.org/10.1027/0269-8803/a000032

van der Helden, J., Boksem, M. A. S., & Blom, J. H. G. (2010). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex, 20,* 1596–1603. http://dx.doi.org/10.1093/cercor/bhp224

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48.

**Walsh, M. M., & Anderson, J. R. (2011a). Learning from delayed feedback: Neural responses in temporal credit assignment. *Cognitive, Affective & Behavioral Neuroscience, 11,* 131–143. http://dx.doi.org/10.3758/s13415-011-0027-0

**Walsh, M. M., & Anderson, J. R. (2011b). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 19048–19053. http://dx.doi.org/10.1073/pnas.1117189108

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews, 36,* 1870–1884. http://dx.doi.org/10.1016/j.neubiorev.2012.05.008

**Walsh, M. M., & Anderson, J. R. (2013). Electrophysiological responses to feedback during the application of abstract rules. *Journal of Cognitive Neuroscience, 25,* 1986–2002. http://dx.doi.org/10.1162/jocn_a_00454

Walsh, M. M., & Anderson, J. R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. *Psychological Bulletin, 140,* 466–486. http://dx.doi.org/10.1037/a0033455

**Walsh, M. M., & Anderson, J. (2011). Unpublished data Experiment 1, Carnegie Mellon University.

**Walsh, M. M., & Anderson, J. (2011). Unpublished data Experiment 2, Carnegie Mellon University.

Warren, C. M., & Holroyd, C. B. (2012). The impact of deliberative strategy dissociates ERP components related to conflict processing vs. reinforcement learning. *Frontiers in Neuroscience, 6,* 43. http://dx.doi.org/10.3389/fnins.2012.00043

**Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research, 1286,* 114–122. http://dx.doi.org/10.1016/j.brainres.2009.06.032

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75,* 418–424. http://dx.doi.org/10.1016/j.neuron.2012.03.042

Yang, Q., Gu, R., Tang, P., & Luo, Y. J. (2013). How does cognitive reappraisal affect the response to gains and losses? *Psychophysiology, 50,* 1094–1103. http://dx.doi.org/10.1111/psyp.12091

Yasuda, A., Sato, A., Miyawaki, K., Kumano, H., & Kuboki, T. (2004). Error-related negativity reflects detection of negative reward prediction error. *NeuroReport, 15,* 2561–2565. http://dx.doi.org/10.1097/00001756-200411150-00027

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111,* 931–959. http://dx.doi.org/10.1037/0033-295X.111.4.931

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex, 15,* 535–544.

*Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of Neuroscience, 24,* 6258–6264. http://dx.doi.org/10.1523/JNEUROSCI.4537-03.2004

*Yi, F., Chen, H., Wang, X., Shi, H., Yi, J., Zhu, X., & Yao, S. (2012). Amplitude and latency of feedback-related negativity: Aging and sex differences. *NeuroReport, 23,* 963–969. http://dx.doi.org/10.1097/WNR.0b013e328359d1c4

Yu, R., & Zhang, P. (2014). Neural evidence for description dependent reward processing in the framing effect. *Frontiers in Neuroscience, 8,* 56. http://dx.doi.org/10.3389/fnins.2014.00056

*Yu, R., Zhou, W., & Zhou, X. (2011). Rapid processing of both reward probability and reward uncertainty in the human anterior cingulate cortex. *PLoS ONE, 6,* e29633. http://dx.doi.org/10.1371/journal.pone.0029633

*Yu, R., & Zhou, X. (2006). Brain responses to outcomes of one's own and other's performance in a gambling task. *NeuroReport, 17,* 1747–1751. http://dx.doi.org/10.1097/01.wnr.0000239960.98813.50

*Yu, R., & Zhou, X. (2009). To bet or not to bet? The error negativity or error-related negativity associated with risk-taking choices. *Journal of Cognitive Neuroscience, 21,* 684–696. http://dx.doi.org/10.1162/jocn.2009.21034

*Zottoli, T. M., & Grose-Fifer, J. (2012). The feedback-related negativity (FRN) in adolescents. *Psychophysiology, 49,* 413–420. http://dx.doi.org/10.1111/j.1469-8986.2011.01312.x

(*Appendix follows*)

Appendix

Experiments Used in the Meta-Analysis, Including Whether an Effect Consistent With an RPE-FRN Was Found, FRN Quantification, the Waveform Used if Additional Ones Were Present in the Listed Figure (WAV), the Domain of the Gamble (DOM), and Whether Magnitude of an Outcome Was Cued (CUE)

Experiment	*N*	RPE-FRN[a]	Site	FRN[b]	WAV	FIG	DOM	CUE[c]
				Likelihood modulator				
Bellebaum & Daum (2008)	17	Yes	Pool	Mean amp 220–280	Postlearning	4	Gain	
Bellebaum et al. (2010a)	15	No	Pool	Peak to peak: P2 (150–N2) to N2 (200–340)	Postlearning	4b	Gain	
Bellebaum et al. (2011)	18	Yes	Cz	Peak of diff wave 100–300	Younger participants	3a	Gain	
Franken et al. (2010)	47	—	Fz	Mean amp 200–300		2	Gain	
Hajcak et al. (2005) Experiment 1	17	No	Fz	Peak of diff wave 200–500		1	Gain	
Hajcak et al. (2005) Experiment 2	12	No	Fz	Peak of diff wave 200–500		3	Gain	
Hajihosseini et al. (2012)	20		Fz	Mean amp 100 either side of N2 peak		2a	Mixed	
Holroyd et al. (2003)	10	Yes	FCz	Peak to peak: P2 (160–240) to N2 (P2–325)		2c	Gain	
Holroyd et al. (2008)	12	—	FCz	Peak to peak: P2 (160–240) to N2 (P2–325)	Time-estimation	1b	Gain	
Holroyd et al. (2009) Experiment 1	20	Yes	FCz	Peak of diff wave 0–600		2	Gain	
Holroyd et al. (2009) Experiment 2	15	Yes	FCz	Peak of diff wave 0–600		2	Gain	
Holroyd et al. (2009) Experiment 3	15	Yes	FCz	Peak of diff wave 0–600		2	Gain	
Holroyd et al. (2011)	18	Yes	FCz	Peak of diff wave 200–300	Outcome locked	1a	Gain	
Larson et al. (2007)	11	—	FCz	Peak to peak: P2 (125–325) to N2 (P2–325)	Control	2	Gain	
Liao et al. (2011)	19	Yes	Fz	Peak of diff wave 150–500	Outcome locked	4	Gain	
Morris et al. (2011)	23	No	Cz	Peak of diff wave ~180–300	Passive gambling task	1	Gain	
Pfabigan et al. (2012)	20	No	FCz	Peak to peak: P2 (preceding positive peak) to N2 (200–350)	Second half	1	Mixed	
Sambrook & Goslin (unpublished) Experiment 3	42		FCz	Unpublished			Gain/Loss	
Talmi et al. (2013)	20	No ($p < .06$)	Pool	Amplitudes at sample points 205–250	Reward condition	4[d]	Gain	
Walsh & Anderson (2011a)	13	Yes	FCz	Mean amp of diff wave 200–300		4	Gain	
Walsh & Anderson (2011b)	20	Yes	FCz	Mean amp of diff wave 200–350	No instruction condition	3	Gain	
Walsh & Anderson (2013) Experiment 1	14	Yes	FCz	Mean amp of diff wave 240–400	Standard and novel	6	Gain	
Walsh & Anderson (2013) Experiment 2	14	Yes	FCz	Mean amp of diff wave 240–400	Standard and novel	6	Gain	
Walsh & Anderson, (unpublished) Experiment 1	13		FCz	Unpublished			Gain	
Walsh & Anderson (unpublished) Experiment 2	13		FCz	Unpublished			Gain	
Yu et al. (2011)	16	—	Fz	Mean amp 275–325 and peak to peak (details not given)	Outcome locked 25%/75%	2b,c	Mixed	
				Magnitude modulator				
Banis & Lorist (2012)	32	Wrong way	FCz	Mean amp 230–300/Mean amp 230–300 relative to average of mean amps of P2 (180–225) and P3 (320–390)/Peak to peak P2 (150–230) to N2 (P2–330)	Average of noise	2	Mixed	N
Bellebaum et al. (2010b)	20	Yes	Fz	Peak to peak: P2 (preceding positivity from 150) to N2 (200–350)	Blocks 3–6, 5c vs. 50c	3b	Gain	Y

(*Appendix continues*)

Appendix (*continued*)

Experiment	*N*	RPE-FRN[a]	Site	FRN[b]	WAV	FIG	DOM	CUE[c]
Gu et al. (2011)	24	Yes	Fz	Peak to average peak: P2 (preceding positive peak) to N2 (200–400) to P3 (succeeding positive peak)	Outcome valence subsequently	4b	Mixed	Y
Hajcak et al. (2006) Experiment 1	16	No	Fz	Peak to peak: P2 (150–350) to N2 (P2–350)		1	Mixed	N
Hajcak et al. (2006) Experiment 2	17	No	Fz	Peak to peak: P2 (150–350) to (P2–350)		3	Mixed	N
Hajihosseini et al. (2012)	20	No	Fz	Mean amp 100 either side of N2 peak		2c	Mixed	Y
Kamarajan et al. (2009)	48	Wrong way	FCz	N2 peak 200–275	Sexes averaged	4	Mixed	Y
Luo & Qu (2013)	18	Yes	FCz	Mean amp 200–250	Win/loss at ¥1 vs. ¥40	3a,b	Mixed	Y
Mushtaq et al. (unpublished)	29		FCz	Unpublished			Mixed	Y
Nittono et al. (2008)	16	—	Fz	Peak to peak: P2 (preceding positive peak) to N2 (150–300)	Even cond., −10/−1/+1/+10	1a	Mixed	N
Onoda et al. (2010)	17	—	FCz	Peak of diff wave 250–400		1	Mixed	Y
Pedroni et al. (2011)	16	—	Cz	TANCOVA over entire waveform		From author	Gain	Y
Roberts et al. (unpublished)	26		Fz	Unpublished			Mixed	N
Sambrook & Goslin (2014)	55	Yes	Pool	Correlation of voltage and utility over entire waveform			Gain/Loss	N
Sambrook & Goslin (unpublished) Experiment 1	48		FCz	Unpublished			Gain	Both
Sambrook & Goslin (unpublished) Experiment 2	45		FCz	Unpublished			Gain/Loss	N
San Martin et al. (2010)	22	—	FCz	Mean amp 240–310		4	Mixed	Y
Santesso et al. (2011)	30	Wrong Way	FCz	N2 Peak (200–400)	Adult participants	2	Mixed	Y
Sato et al. (2005)	18	No	Fz	Peak to peak: P2 (150–220) to N2 (P2–325)		1	Mixed	Y
Schuermann et al. (2012)	20	Yes	FCz	Peak to peak: P2 (100–300) to N2 (200–400)		2	Mixed	Y
Talmi et al. (2013)	20	—	Pool	Amps at sample points 205–250	Reward	4[d]	Gain	Y
Toyomaki & Murohashi (2005)	13	—	Fz	Peak to peak: P2 (unspecified) to N2 (unspecified)	−500/−10/+10/+500	2	Mixed	Y
Van den Berg et al. (2011)	42	—	Fz	Peak to peak: P2 (150–350) to N2 (following negative peak)		2	Mixed	N
Wu & Zhou (2009)	16	No	FCz	Mean amp 250–350	Expected magnitudes	1	Mixed	Y
Yeung & Sanfey (2004)	16	—	FCz	Peak to average Peak: P2 (preceding positive peak) to N2 (200–400) to P3 (succeeding positive peak)		2	Mixed	Y
Yi et al. (2012)	28	No	Fz	Peak of N2 (200–400)		4b	Mixed	Y
Yu & Zhou (2006)	20	No	Fz	Mean amp 25 before and after peak of diff wave	Execution	1	Mixed	Y
Yu & Zhou (2009)	14	No	Fz	Mean amplitude 200–300	"To bet" trials	3	Mixed	Y
Zottoli & Grose-Fifer (2012)	18	Yes	FCz	Peak to peak: P2 (150–300) to N2 (200–425)	Adult participants	2a	Mixed	Y

[a] A dash indicates that the RPE-FRN was not reported. [b] Values in parentheses indicate the interval in which peak assignment was made in milliseconds. [c] In cued studies participants know the magnitude of the forthcoming feedback but not its valence, in uncued studies participants knew neither its magnitude nor its valence. [d] Eight waveforms corresponding to the Valence × Magnitude × Likelihood design were given; these were all digitized and the unwanted factor collapsed out by averaging pairs of waveforms.

Received January 17, 2014
Revision received October 5, 2014
Accepted October 7, 2014 ■