NeuroImage 173 (2018) 146–152

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons

Zachary P. Infantolino [a,*], Katherine R. Luking [a], Colin L. Sauder [b], John J. Curtin [c], Greg Hajcak [d]

[a] *Department of Psychology, Stony Brook University, United States*
[b] *Department of Psychiatry, University of Texas Health Science Center at San Antonio, United States*
[c] *Department of Psychology, University of Wisconsin-Madison, United States*
[d] *Department of Psychology and Biomedical Sciences, Florida State University, United States*

ARTICLE INFO

Keywords:
fMRI
Internal reliability
Adolescence
Faces
Amygdala

ABSTRACT

Advances in cognitive and affective neuroscience come largely from within-subjects comparisons, in which the functional significance of neural activity is determined by contrasting two or more experimental conditions. Clinical and social neuroscience studies have attempted to leverage between-subject variability in such condition differences to better understand psychopathology and other individual differences. Shifting from within-to between-subjects comparisons requires that measures have adequate internal consistency to function as individual difference variables. This is particularly relevant for difference scores—which have lower reliability. The field has assumed reasonable internal consistency of neural measures based on consistent findings across studies (i.e., if a within-subject difference in neural activity is robust, then it must be reliable). Using one of the most common fMRI paradigms in the clinical neuroscience literature (i.e., a face- and shape-matching task), in a large sample of adolescents ($N = 139$) we replicate a robust finding: amygdala activation is greater for faces than shapes. Moreover, we demonstrate that the internal consistency of the amygdala in face and shape blocks was excellent (*Spearman-Brown corrected reliability [SB]* $> .94$). However, the internal consistency of the activation *difference* between faces and shapes was nearly zero ($SB = -.06$). This reflected the fact that the amygdala response to faces and shapes was highly correlated ($r = .97$) across individuals. Increased neural activation to faces versus shapes could not possibly function as an individual difference measure in these data—illustrating how neural activation can be robust within subjects, but unreliable as an individual difference measure. Strong and reproducible condition differences in neural activity are not necessarily well-suited for individual differences research—and neuroimaging studies should always report the internal consistency of, and correlations between, activations used in individual differences research.

Introduction

The incredible progress in human affective and cognitive neuroscience has come largely from within-subjects contrasts. In a typical study, neural activation is compared between two conditions, and the resulting neural activity is interpreted in terms of psychological functions that differ between conditions. These within-subjects comparisons have had a major impact on the localization of function. For instance, researchers may parametrically manipulate working memory load to isolate neural activity implicated as memory demands range from low to high (e.g., Rypma et al., 2002). As an additional example germane to the current study, researchers have consistently found that the amygdala is more activated by stimuli that induce fear and other emotions, relative to affectively benign stimuli—amygdala activation seems to be a hallmark of manipulations that increase fear (LeDoux, 2003; Sergerie et al., 2008).

Clinical neuroscientists have attempted to leverage results from these within-subjects studies to better understand psychopathology—an effort to elucidate the neuroscience of individual differences. It certainly stands to reason that if amygdala activation is a hallmark of within-subject manipulations that increases fear, then between-subject variability in amygdala activation may relate to individual differences in the experience of fear and anxiety. Clinical neuroscience studies have, for instance, used an individual's amygdala activation as an individual difference variable—correlating it with other individual difference measures, such as trait anxiety (Stein et al., 2007).

In this way, clinical neuroscience studies shift seamlessly from within-

[*] Corresponding author. Department of Psychology, Stony Brook University, 100 Nicolls Road, Stony Brook, NY 11794, United States.
E-mail address: zinfantolino@gmail.com (Z.P. Infantolino).

https://doi.org/10.1016/j.neuroimage.2018.02.024
Received 7 June 2017; Received in revised form 25 January 2018; Accepted 13 February 2018
Available online 16 February 2018

to between-subjects comparisons: from interpreting increased amygdala activation in one condition, to interpreting *an individual's* amygdala activation. Unfortunately, there are key psychometric issues surrounding individual difference measures that neuroscientific studies of individual differences have, by and large, failed to consider. An individual's score on any measure comprises some combination of true score and error. Thus, an individual's amygdala activation is based on both reliable variance common across trials or blocks and noise. The amount of reliable variance in a measure can be estimated by assessing internal consistency, using split-half reliability, and reflects the degree to which neural measures from half of the task relate to neural measures from the other half of the task.

Clinical neuroscience studies have generally assumed reliability. A task that has demonstrated robust within-subjects differences (i.e., contrast differences or effects) is chosen to study individual differences, with the assumption that the task reflects *reliable* neural activation *because* the activation is consistently demonstrated across multiple studies. However, these are two distinct properties: *robust* neural activation is reflected in means and standard deviations that differ across within-subject conditions whereas *reliable* neural activation implies *score consistency across individuals.* Importantly, according to classical test theory, internal consistency places an upper limit on how well a measure can correlate with other individual difference variables such that the maximum correlation between two measures is defined as the square root of the product of the reliability of the measures. We recently demonstrated how internal consistency of a neural measure limits between-subjects effect sizes (Hajcak et al., 2017). Thus, failure to ensure that measures are reliable can lead to low relationships between those measures, which discourages future studies examining the relationship between the constructs those measures are purported to assess, simply because an unreliable measure was selected.

The issue of internal consistency is especially relevant in the case of difference scores, when within-subjects contrasts are being used as individual difference measures. The variance of a difference score reflects the sum of both scores unique variance and their error. In other words, any true score variance that is shared between conditions is actually eliminated through subtraction, while condition-related error summates. Thus, difference scores tend to have lower internal consistency than constituent condition scores because, in practice, the constituent scores are often correlated, resulting in considerable shared true score variance (Chiou and Spreng, 1996; Willett, 1988; Zumbo, 1999). When the shared true score variance is removed through subtraction, the difference score contains a lower proportion of true score variance to error than the constituent scores and thus has a lower internal consistency. Issues associated with utilizing difference scores as individual difference measures have been discussed previously (e.g., Cronbach and Furby, 1970; Lord, 1956) and are a concern across fields and measurement methods (e.g., Hedge et al., 2017; Meyer et al., 2017; Ross et al., 2015). However, until recently, these concerns have rarely been discussed in social, cognitive, and clinical neuroscience (Luking et al., 2017; Meyer et al., 2017).

The present study sought to examine the internal consistency of amygdala activation elicited using a modified version of an emotional face-matching task that has been used extensively to study individual differences (Hariri et al., 2002). This task produces robust amygdala activation when participants match emotional faces compared to shapes. Moreover, the increased amygdala activation to faces compared to shapes (i.e., difference score) has been examined in relation to individual differences in psychopathology (Kleinhans et al., 2010; Rasetti et al., 2009), genetics (Bertolino et al., 2005; Hariri et al., 2005), Parkinson's disorder (Tessitore et al., 2002), and other self-report individual difference measures (Drabant et al., 2009). The main effect of the task (faces matching vs. shape matching) was examined to confirm that the task elicited a pattern of results in the amygdala that is similar to what has been reported in the literature. We then examined the internal consistency of amygdala activation to faces, shapes, as well as the difference score.

Material and method

Participants

Participants in the present study were females aged 8–14 years who participated in a larger longitudinal study investigating relationships between neural reward response, emotion processing, pubertal development, and emerging risk for psychopathology. The sample was recruited from the Long Island, NY community using online classified advertisements, community postings, local referral sources, and a commercial mailing list targeting homes with an 8-14 year-old girl. Inclusion criteria were English fluency, ability to read and comprehend questionnaires, absence of an intellectual disability, and a biological parent consenting to participate in the study.

Of the 317 adolescent girls that participated in the parent study, 145 completed the emotional face matching task in the MRI scanner, and 139 provided sufficient quality data (exclusions for excessive motion $n = 5$, scores at least three standard deviations from the mean $n = 1$). Girls were aged 8–14 years ($M = 12.65$; $SD = 1.74$) and were 85.6% Caucasian, 5.0% African American, 2.9% Hispanic, 6.5% identified as “Other” or did not answer. The research protocol was approved by the Stony Brook University Institution Review Board.

Faces-matching task

Participants completed an emotional face-matching task adapted from Hariri et al. (2002) that used 16 male and 16 female neutral, fearful, sad, and happy faces selected from the NIMH Child Emotional Faces Picture Set (Egger et al., 2011). Selected facial stimuli had a direct gaze and were from subjects aged 10–16 years ($M = 13.42$). Shape matching was used as a control condition. The task was divided into two runs for data acquisition.

During each trial, a target face or shape was presented at the top of the screen and two faces or shapes were presented at the bottom of the screen. Participants were instructed to select the face or shape at the bottom of the screen that matched the target facial expression or shape at the top of the screen. Participants selected the left or right face or shape using their index or middle finger. Faces and shapes remained onscreen for 5 s, after which a new trial immediately began. Each facial expression was presented in two 20-s blocks for each run, with each block consisting of four trials. Thus, each facial expression was presented a total of 16 times across both runs. Blocks alternated between face- and shape-matching conditions and were counterbalanced. Throughout emotional face-matching blocks, the nonmatching facial expression was always neutral. During neutral face-matching blocks, the nonmatching facial expression was either fearful or happy with equal probability. Once the first run of the task was completed, data collection stopped and the experimenter checked in on the participant. When the participant was ready to continue, the second run of the task and data collection began.

fMRI data acquisition and analysis

MR data were acquired with a 12-channel head coil using a whole-body 3 T S Tim-Trio scanner (Siemens AG, Erlangen, Germany). Gradient fieldmaps were collected to correct for geometric distortions in the functional data caused by magnetic field inhomogeneity (Jezzard and Balaban, 1995). Three hundred and twenty-four T2*-weighted whole-brain volumes were acquired using an echo-planar imaging sequence ($TR = 2000$ ms, $TE = 23$ ms, flip angle $= 83$, slice thickness $= 3.5$ mm, and 0 mm gap).

Data analysis was performed using Statistical Parametric Mapping (SPM8; Wellcome Department of Cognitive Neurology, Institute of Neurology, London, United Kingdom). Standard preprocessing procedures with default parameters, including image realignment corrections for head movements, slice timing corrections for acquisition order, normalization to the Cincinnati Children's Hospital Medical Center pediatric template (Wilke et al., 2002), and spatial smoothing with a Gaussian full-width-at-half-maximum 8 mm filter. ArtRepair version 4 (Mazaika et al., 2009) was used to identify and repair (interpolate) volumes where volume-to-volume motion exceeded 2 mm (1 voxel). Participants were excluded from analyses if greater than 20% of volumes were interpolated (n = 5).

Fixed-effects general linear models (GLMs) were created for each participant. The task was designed so that each emotional expression was presented in one block of trials using male faces and one block of trials using female faces during each run. If blocks were selected for split-half reliability analyses based on block order (i.e., odd/even), which is common, internal consistency would have been confounded with gender because the task was designed such that for a given emotional expression, the same gender appeared first in both runs. To avoid this, split-half reliability analyses relied on splitting the data into A and B blocks that randomized gender across A and B blocks. The gender for A blocks in the first run was randomly selected for each emotional expression for each participant; the other gender was then selected for A blocks in the second run. For example, if female was selected as the gender in A blocks for happy faces in the first run, then male happy faces would be included in A blocks in the second run; in this example, B blocks would then include male happy faces from the first run and female happy faces from the second run. This ensured that A and B blocks each contained both male and female stimuli. Two regressors were created for facial expressions for A and B blocks. Consistent with the random selection of emotional expression blocks to A and B regressors, shape blocks were randomly assigned to A and B regressors for each participant.

Contrasts were created to examine the main effect of faces vs. shapes for A and B blocks, the main effect of faces compared to implicit baseline for A and B blocks, shapes compared to implicit baseline for A and B blocks, the main effect of faces compared to implicit baseline for the entire task, and shapes compared to implicit baseline for the entire task. The latter two contrasts, which collapsed across A and B blocks, were created to examine the relationship between the response to faces compared to implicit baseline and shapes compared to implicit baseline. Because preliminary results suggested that the reliability of amygdala activation to faces was similar across different emotional expressions, the present study combined all emotional expressions. A mask was created for right and left amygdala based on the Harvard-Oxford probabilistic subcortical structural atlas. All voxels within the mask reached significance for the faces vs. shapes contrast after correcting for multiple comparisons across the entire brain using false-discovery rate correction (Genovese et al., 2002). The mean activation from all voxels within the right and left amygdala masks were extracted for each contrast using the MarsBar toolbox (Brett et al., 2002). Values were imported into IBM SPSS Statistics, version 22.0 (IBM, Armonk, N.Y.) for split-half analyses. The present study focuses on the right amygdala, although results were qualitatively identical for the left.

To calculate the split-half reliability of each contrast, the extracted activations were compared on A and B blocks using Pearson correlation coefficients. Given that splitting the data into A and B blocks artificially reduces the number of trials by half, which reduces the internal consistency of a measure, Pearson correlation coefficients were adjusted using the Spearman-Brown prediction formula ($SB = 2r/(1 + r)$) to predict the split-half reliability of each contrast if all trials had been included. The response to faces and shapes, collapsed across A and B blocks, were compared to one another using a Pearson correlation coefficient.

To further visualize the unique and overlapping variance attributable to faces and shapes, the observed score variance of each measure was divided into multiple components. According to classical test theory, the observed score variance is comprised of the true score variance and error variance.

$$s^2_{Faces} = s^2_{T_{Faces}} + s^2_{E_{Faces}}$$

$$s^2_{Shapes} = s^2_{T_{Shapes}} + s^2_{E_{Shapes}}$$

Where s^2_{Faces} and s^2_{Shapes} are the total observed score variances for the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline, respectively; $s^2_{T_{Faces}}$ and $s^2_{T_{Shapes}}$ are the true score variances for the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline, respectively; and $s^2_{E_{Faces}}$ and $s^2_{E_{Shapes}}$ are the error variances for the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline, respectively. The true score variance can be further decomposed into the unique variance associated with each measure and the variance shared between the two measures.

$$s^2_{T_{Faces}} = s^2_{U_{Faces}} + s_{Faces,Shapes}$$

$$s^2_{T_{Shapes}} = s^2_{U_{Shapes}} + s_{Faces,Shapes}$$

Where $s^2_{U_{Faces}}$ and $s^2_{U_{Shapes}}$ are the unique variances for the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline, respectively, and $s_{Faces,Shapes}$ is the shared variance between the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline. The variances in the equations above can be calculated using the following formulas (more details can be found in the Supplemental Materials).

$$s^2_{E_{Faces}} = s^2_{Faces} * (1 - r_{Faces})$$

$$s^2_{E_{Shapes}} = s^2_{Shapes} * (1 - r_{Shapes})$$

$$s_{Faces,Shapes} = s_{Faces} * s_{Shapes} * r_{Faces,Shapes}$$

$$s^2_{U_{Faces}} = \left(s^2_{Faces} * r_{Faces}\right) - s_{Faces,Shapes}$$

$$s^2_{U_{Shapes}} = \left(s^2_{Shapes} * r_{Shapes}\right) - s_{Faces,Shapes}$$

Where r_{Faces} and r_{Shapes} are the split-half reliabilities of the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline, respectively; s_{Faces} and s_{Shapes} are the standard deviations of the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline, respectively, and $r_{Faces,Shapes}$ is the correlation between the response to faces compared to implicit baseline and the response to shapes compared to implicit baseline.

According to classical test theory, the total observed variance of a difference score is defined as the following:

$$s^2_{Faces-Shapes} = s^2_{Faces} + s^2_{Shapes} - 2*s_{Faces,Shapes}$$

Substituting in the three total observed variance components, unique, shared, and noise, for each measure, the formula for the total observed variance of the faces compared to shapes simplifies to the following:

$$s^2_{Faces-Shapes} = s^2_{U_{Faces}} + s^2_{E_{Faces}} + s_{Faces,Shapes} + s^2_{U_{Faces}} + s^2_{E_{Shapes}} + s_{Faces,Shapes} - 2*s_{Faces,Shapes}$$

$$s^2_{Faces-Shapes} = s^2_{U_{Faces}} + s^2_{E_{Faces}} + s^2_{U_{Faces}} + s^2_{E_{Shapes}}$$

The total variance of the response to faces compared to implicit baseline and shapes compared to implicit baseline were compared to the total variance of the response to faces vs. shapes using a Morgan-Pitman test of equal variance in dependent samples (Morgan, 1939; Pitman, 1939).

Results

fMRI activation

Both A and B blocks in the right amygdala exhibited a robust difference in the response to faces vs. shapes, $ts > 12.75$, $ps < .001$, and this difference was similar between A and B blocks, $t\,(137) = -.28$ (Table 1; Fig. 1). Thus, the task performed similarly to previously published papers insofar as matching faces elicited greater amygdala activation than matching shapes. Additionally, this effect was consistent across A and B blocks.

Both A and B blocks demonstrated a robust response to faces relative to the implicit baseline, $ts > 3.74$, $ps < .001$, whereas the response to shapes did not differ from the implicit baseline, $ts < 1.11$ (Table 1; Fig. 1). Similar to the faces vs. shapes contrast, the response to faces relative to the implicit baseline and the response to shapes relative to implicit baseline did not differ between A and B blocks, absolute $ts < .97$. Thus, the difference between the response to faces and shapes appears to be driven by an increase in activation to faces, and this response is consistent across A and B blocks.

Table 1
One-sample and Paired-sample T-tests for Contrasts of Interest.

One-sample T-tests		
Contrast	*t*	*p*
Faces vs. Shapes A	12.76	<.001
Faces vs. Shapes B	16.53	<.001
Faces A	3.75	<.001
Faces B	4.21	<.001
Shapes A	−1.10	.28
Shapes B	-.95	.35
Paired-sample T-test		
Contrast		
Faces vs. Shapes: A vs. B	-.28	.78
Faces: A vs. B	-.96	.34
Shapes: A vs. B	-.40	.69

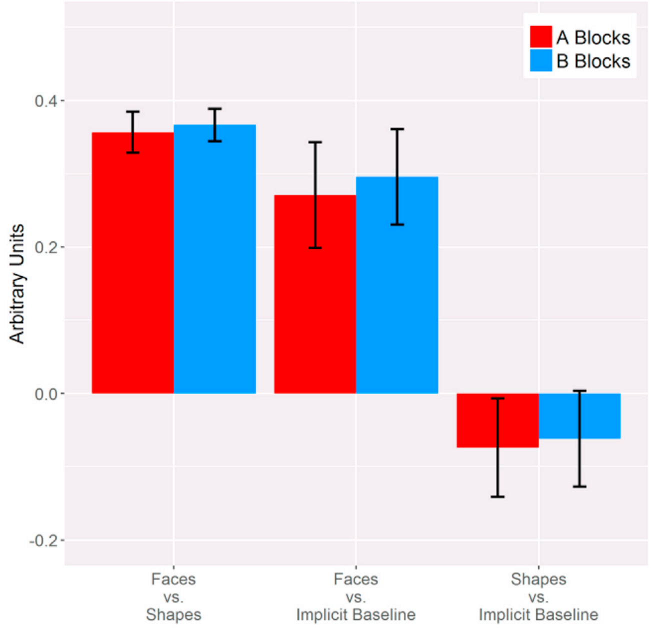

Fig. 1. Bar graphs depicting the response to each contrast for A and B blocks. Error bars represent ± standard error of the mean.

Table 2
Spearman-brown predicted internal consistency coefficients.

Contrast	*SB*
Faces vs. Shapes	-.06
Faces	.97
Shapes	.95

Split-half reliability

The split-half reliability of the faces vs. shapes contrast was poor, $SB = -.06$, whereas the split-half reliability of both the response to faces and shapes relative to baseline were high, $SBs > .94$ (Table 2; Fig. 2). Thus, each individual main effect appears to have excellent internal consistency, whereas the faces vs. shapes contrast has poor internal consistency. This was due to a strong, positive correlation between the response to faces relative to implicit baseline and the response to shapes relative to implicit baseline across all blocks, $r = .97$, $p < .001$ (Fig. 3).

Variance distribution

A list of variance components are visualized in Fig. 4. The majority of reliable variance in the response to faces relative to the implicit baseline and the response to shapes relative to implicit baseline were shared with one another (over 90% for each). This reliable shared variance is removed when amygdala activity to shapes is subtracted from amygdala activity to faces; thus, the total amount of variance in the response to faces vs. shapes is smaller than the variance in the response to faces relative to implicit baseline and the response to shapes relative to implicit baseline ($t(137)s > 97.45$, $ps < .001$). This difference in variance is reflected in the size of the error bars in Fig. 1. Noise variance from the response to faces relative to implicit baseline summates with the noise variance from the response to shapes relative to implicit baseline and accounts for nearly 50% of the variance in the faces vs. shapes difference score. Indeed, the only other contribution to the faces vs. shapes contrast was the small amount of unique variance from the faces vs implicit baseline contrast.

Discussion

In the present study, the amygdala showed greater activation in emotional face-matching blocks than in shape-matching blocks, thus replicating within-subject effects that have been reported in numerous studies (Drabant et al., 2009; Hariri et al., 2002; Kleinhans et al., 2010; Marusak et al., 2013). Moreover, the amygdala response to both face- and shape-matching, relative to implicit baseline, was characterized by extremely high internal consistency. That is, the amygdala response derived from A blocks was highly correlated with the amygdala response to B blocks, across individuals—and this was true for both faces and shapes relative to baseline. However, the internal consistency for the response to faces vs. shapes was nearly zero—the *difference score* derived from A blocks was unrelated to the difference score derived from B blocks. The low internal consistency resulted from the fact that the amygdala response to faces was highly correlated with the amygdala response to shapes—indeed, this correlation approached the level of the internal consistencies. That is, the correlation between the amygdala response to faces and shapes was approximately the same as the correlation between amygdala response to faces on A and B blocks. As evidenced by the division of variance, the vast majority of reliable variance

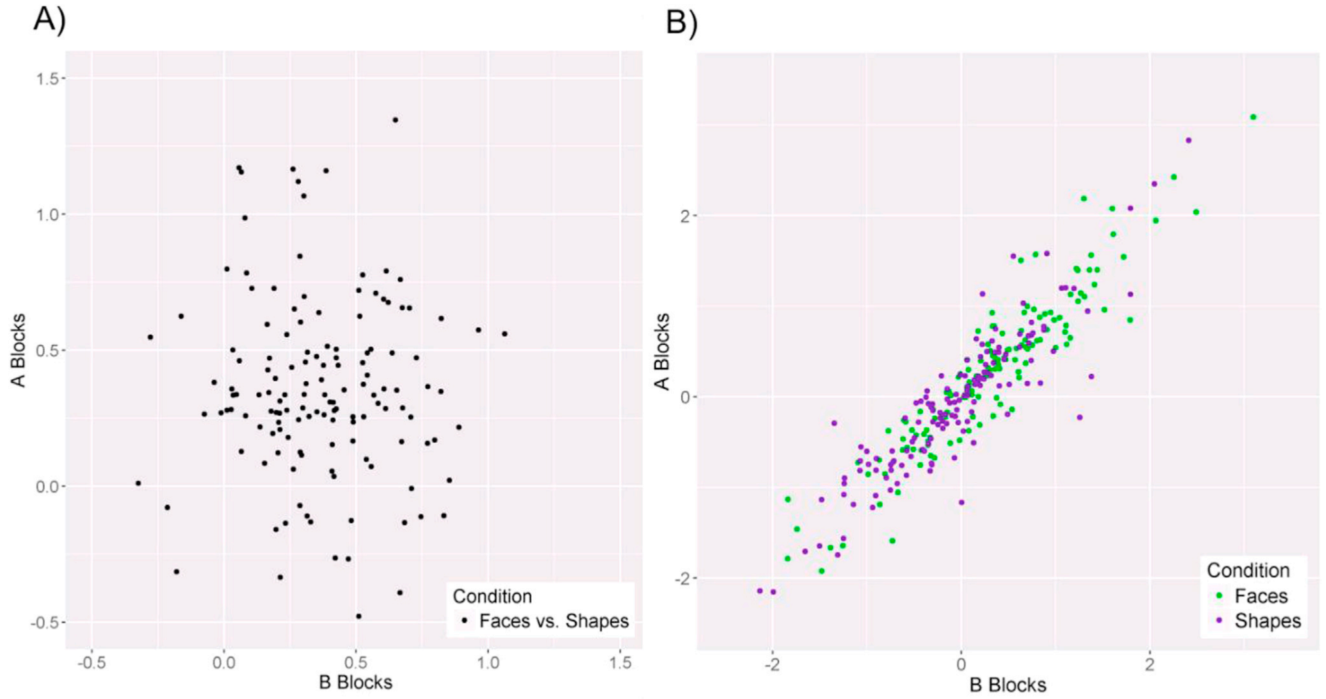

Fig. 2. a) Scatter plot depicting the relationship between A and B blocks for the faces vs. shapes contrast. b) Scatter plot depicting the relationship between the A and B blocks for the faces vs. implicit baseline and shapes vs. implicit baseline contrasts.

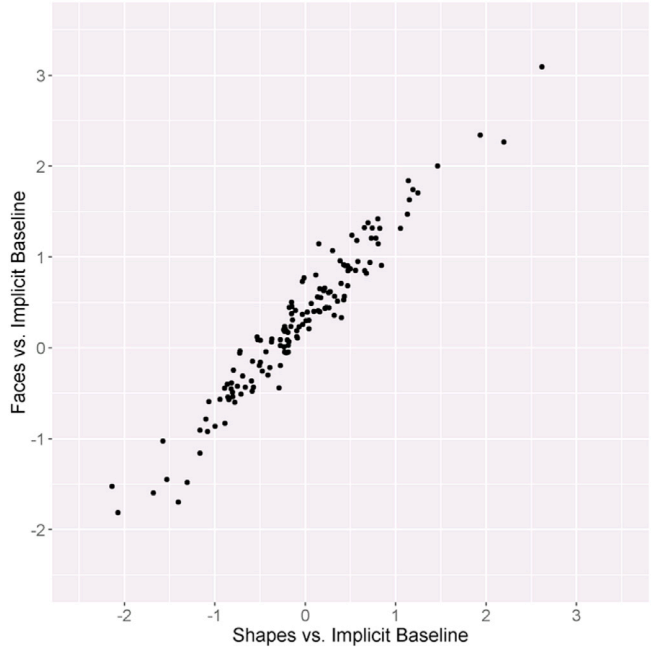

Fig. 3. Scatter plot depicting the relationship between the faces vs. implicit baseline and shapes vs. implicit baseline contrasts across all trials.

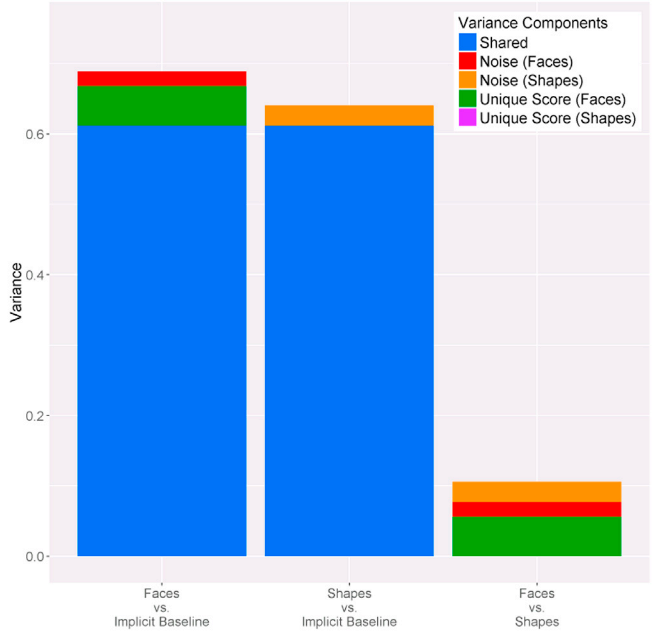

Fig. 4. Stacked bar graphs depicting the variance distribution for each contrast.

was shared between the face and shape conditions, which was removed in the difference score.

Insofar as internal consistency places a limit on validity, these data suggest that the difference score measure (i.e., amygdala response to faces vs shapes) could not possibly function as a valid individual difference measure. We would emphasize that the current results may or may not generalize to other studies that have leveraged this task and used the faces vs shapes contrast as an individual difference measure. Indeed, it would be informative to examine psychometric properties of the amygdala in existing data sets that have found relationships between the faces vs. shapes contrast and other individual difference variables. One possibility is that the response to faces and shapes would be less correlated in older participants, or participants with psychopathology—-which, assuming the internal consistencies of the constituent measures remain constant, would increase the reliable variance in the difference score for these populations.

In a sample largely overlapping with the present study, striatal activation to feedback indicating monetary gain and loss in a guessing task were much more modestly correlated (.40 and .55 for left striatum even and odd trials, respectively, and .36 and .49 for right striatum even and odd trials, respectively)—and the internal consistency for the gain minus loss difference score was quite a bit better (.35 and .33 for left and right ventral striatum, respectively) (Luking et al., 2017). Thus, it is possible to have difference scores with more true score variance—provided the constituent scores are themselves internally consistent and only modestly correlated.

For difference scores that have low internal consistency, one alternative is to use just one of the constituent measures as an individual difference measure, assuming it has adequate reliability. In the current study, either the faces compared to implicit baseline or the shapes compared to implicit baseline contrast could function well as an individual measure. Indeed, the current study suggests that these measures may be nearly redundant (i.e., over 90% of the observed variance in each measure was shared). In other words, the current study suggests that even though the amygdala is more activated by matching faces than shapes, *individual differences* in amygdala activity could be indexed well by having subjects only match shapes. Of course, this calls into question the very construct typically being measured using amygdala activation difference scores (e.g., individual differences in emotional reactivity).

The current study demonstrates the importance of distinguishing between a *robust* within-subject difference (i.e., increased amygdala activation to emotional face-matching versus shape-matching), and a *reliable* individual difference measure. As we have argued elsewhere, an important step for improving the reproducibility and rigor of clinical and social neuroscience is to more thoroughly examine the psychometric properties of neural measures (Hajcak et al., 2017; Patrick and Hajcak, 2016). Furthermore, given changes in the direction of clinical neuroscience research in the past several years, namely the National Institute of Mental Health's Research Domain Criteria (RDoC) initiative, examining the psychometric properties of neural measures is increasingly important. The RDoC initiative seeks to better integrate information across multiple units of analysis (e.g., neural circuits, physiology, behavior, and self-reports) and domains (e.g., positive valence systems, cognitive systems, arousal and regulatory systems). The cells that populate units of analyses across domains are intended to be measures of individual differences. However, nearly all of the psychophysiological and neural measures that are highlighted were originally studied as within-subjects variables. Thus, this new framework for psychopathology research provides an excellent opportunity to emphasize the importance of psychometrically sound clinical neuroscience research.

Moreover, these issues apply more broadly to any neuroscience research that attempts to employ a robust within-subject *difference* as an individual difference variable (i.e., any time a difference score is correlated with another between-subject variable). We suggest that all studies should examine and report on the internal consistency of neural activations, as well as the correlation between activations, in individual differences research. Indeed, knowing the internal consistency of different psychophysiological measures can help guide task design. For example, Luking et al. (2017) found that internal consistency was just as high in the first half of a monetary guessing task as in the whole task, suggesting that the task could be shortened in order to reduce costs without reductions in internal consistency.

The present study benefits from a large sample size, and an analytic approach (i.e., split-half reliability) that although uncommon in fMRI research (c.f. Luking et al., 2017), is easily calculated using existing data sets and task designs. However, there are several limitations. First, by taking steps to avoid confounding internal consistency and effects of gender (ensuring that A and B blocks did not contain only one gender), internal consistency was confounded with the interaction between face gender and run number. Given the high internal consistency for both face and shape contrasts, this does not appear to have been a problem—though future studies might better optimize task design to assess internal consistency. Second, although this task was modeled after the task developed by Hariri et al. (2002), there are a few key differences: adolescent faces were used in the task, additional facial expressions were included, and data were collected across two runs. Though these differences did not appear to negatively impact the reliability of the amygdala response to faces and shapes in isolation, the low internal consistency of the faces vs. shapes contrast in the present study may not generalize to other versions of this task. Finally, the present study examined internal consistency using a sample of adolescent girls aged 8 to 14 and the results may not generalize to males, or females of different ages.

Overall, the present study highlights potential pitfalls of assuming reliability from consistent findings across studies: despite robust within-subjects effects on the amygdala for faces vs shape matching, this difference score itself was unreliable. This was true despite the fact that amygdala response to both faces and shapes was characterized by high internal consistency. The poor internal consistency of the difference score was attributable to the fact that both scores were highly correlated with one another. The authors suggest that future research include internal consistency, and correlations between neural measures used in difference scores, to aid in task design and selection, and in the interpretation of findings related to individual differences.

Acknowledgements and disclosures

This work was supported by the National Institute of Mental Health grant R01 MH097767. We thank the adolescents and parents who contributed their time to this project. We also thank the numerous contributions of the research assistants and staff to this project. All authors report no conflicts of interest.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.02.024.

References

Bertolino, A., Arciero, G., Rubino, V., Latorre, V., De Candia, M., Mazzola, V., Nardini, M., 2005. Variation of human amygdala response during threatening stimuli as a function of 5′ HTTLPR genotype and personality style. Biol. Psychiatr. 57 (12), 1517–1525.

Brett, M., Anton, J.-L., Valabregue, R., Poline, J.-B., 2002, June. Region of interest analysis using an SPM toolbox. In: Poster session presented at the International Conference on Functional Mapping of the Human Brain, Sendai, Japan.

Chiou, J.S., Spreng, R.A., 1996. The reliability of difference scores: a re-examination. J. Consumer Satisfaction, Dissatisfaction Complain. Behav. 9, 158–167.

Cronbach, L.J., Furby, L., 1970. How we should measure "change": or should we? Psychol. Bull. 74 (1), 68–80.

Drabant, E.M., McRae, K., Manuck, S.B., Hariri, A.R., Gross, J.J., 2009. Individual differences in typical reappraisal use predict amygdala and prefrontal responses. Biol. Psychiatr. 65 (5), 367–373.

Egger, H.L., Pine, D.S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K.E., Angold, A., 2011. The NIMH child emotional case picture set (NIMH-CHEFS): a new set of children's facial emotion stimuli. Int. J. Meth. Psychiatr. Res. 20 (3), 145–156.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15 (4), 870–878.

Hajcak, G., Meyer, A., Kotov, R., 2017. Psychometrics and the neuroscience of individual differences: internal consistency limits between-subjects effects. J. Abnorm. Psychol. 126 (6), 823–834.

Hariri, A.R., Drabant, E.M., Munoz, K.E., Kolachana, B.S., Mattay, V.S., Egan, M.F., Weinberger, D.R., 2005. A susceptibility gene for affective disorders and the response of the human amygdala. Arch. Gen. Psychiatr. 62 (2), 146–152.

Hariri, A.R., Tessitore, A., Mattay, V.S., Fera, F., Weinberger, D.R., 2002. The amygdala response to emotional stimuli: a comparison of faces and scenes. Neuroimage 17, 317–323.

Hedge, C., Powell, G., Sumner, P., 2017. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Meth. 1–21. https://doi.org/10.3758/s13428-017-0935-1.

Jezzard, P., Balaban, R.S., 1995. Correction for geometric distortion in echo planar images from B0 field variations. Magn. Reson. Med. 34 (1), 65–73.

Kleinhans, N.M., Richards, T., Weaver, K., Johnson, L.C., Greenson, J., Dawson, G., Aylward, E., 2010. Association between amygdala response to emotional faces and social anxiety in autism spectrum disorders. Neuropsychologia 48 (12), 3665–3670.

LeDoux, J., 2003. The emotional brain, fear, and the amygdala. Cell. Mol. Neurobiol. 23 (4), 727–738.

Lord, F.M., 1956. The measurement of growth. Educ. Psychol. Meas. 16, 421–437.

Luking, K.R., Nelson, B.D., Infantolino, Z.P., Sauder, C.L., Hajcak, G., 2017. Internal consistency of fMRI and EEG measures of reward in late childhood and early adolescence. Biol. Psychiatr.: Cogn. Neurosci. Neuroimag. 2 (3), 289–297.

Marusak, H.A., Carré, J.M., Thomason, M.E., 2013. The stimuli drive the response: an fMRI study of youth processing adult or child emotional face stimuli. Neuroimage 83, 679–689.

Mazaika, P., Hoeft, F., Glover, G.H., Reiss, A.L., 2009. Methods and software for fMRI analysis for clinical subjects (poster presented at). Hum. Brain Mapp. 1–1.

Meyer, A., Lerner, M.D., De Los Reyes, A., Laird, R.D., Hajcak, G., 2017. Considering ERP difference scores as individual difference measures: issues with subtraction and alternative approaches. Psychophysiology 54, 114–122.

Morgan, W.A., 1939. A test for the significance of the difference between the two variances in a Sample from a normal bivariate population. Biometrika 31, 13–19.

Patrick, C.J., Hajcak, G., 2016. RDoC: translating promise into progress. Psychophysiology 53, 415–424.

Pitman, E.J.G., 1939. A note on normal correlation. Biometrika 31, 9–12.

Rasetti, R., Mattay, V.S., Wiedholz, L.M., Kolachana, B.S., Hariri, A.R., Callicott, J.H., Weinberger, D.R., 2009. Evidence that altered amygdala activity in schizophrenia is related to clinical state and not genetic risk. Am. J. Psychiatr. 166 (2), 216–225.

Ross, D.A., Richler, J.J., Gauthier, I., 2015. Reliability of composite task measurements of holistic face processing. Behav. Res. Meth. 47 (3), 736–743.

Rypma, B., Berger, J.S., D'esposito, M., 2002. The influence of working-memory demand and subject performance on prefrontal cortical activity. J. Cognit. Neurosci. 14 (5), 721–731.

Sergerie, K., Chochol, C., Armony, J.L., 2008. The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies. Neurosci. Biobehav. Rev. 32, 811–830.

Stein, M.B., Simmons, A.N., Feinstein, J.S., Paulus, M.P., 2007. Increased amygdala and Insula activation during emotion processing in anxiety-prone subjects. Am. J. Psychiatr. 164 (2), 318–327.

Tessitore, A., Hariri, A.R., Fera, F., Smith, W.G., Chase, T.N., Hyde, T.M., Mattay, V.S., 2002. Dopamine modulates the response of the human amygdala: a study in Parkinson's disease. J. Neurosci. 22 (20), 9099–9103.

Wilke, M., Schmithorst, V.J., Holland, S.K., 2002. Assessment of spatial normalization of whole-brain magnetic resonance images in children. Hum. Brain Mapp. 17 (1), 48–60.

Willett, J.B., 1988. Questions and answers in the measurement of change. Rev. Res. Educ. 15, 345–422.

Zumbo, B.D., 1999. The simple difference score as an inherently poor measure of change: Some reality, much mythology. In: Thompson, B. (Ed.), Advances in Social Science Methodoloy. JAI Press, Greenwich, pp. 269–304.