Test-retest reliability of functional MRI food receipt, anticipated receipt, and picture tasks

Sonja Yokum,[1] *Cara Bohon,*[2] *Elliot Berkman,*[3] *and Eric Stice*[2]

[1]Oregon Research Institute, Eugene, OR, USA; [2]Department of Psychiatry, Stanford University, Stanford, CA, USA; and [3]Department of Psychology, Center for Translational Neuroscience, University of Oregon, Eugene, OR, USA

ABSTRACT

Background: Functional MRI (fMRI) tasks are increasingly being used to advance knowledge of the etiology and maintenance of obesity and eating disorders. Thus, understanding the test-retest reliability of BOLD signal contrasts from these tasks is important.

Objectives: To evaluate test-retest reliability of responses in reward-related brain regions to food receipt paradigms (palatable tastes, anticipated palatable tastes), food picture paradigms (high-calorie food pictures), a monetary reward paradigm (winning money and anticipating winning money), and a thin female model picture paradigm (thin female model pictures).

Method: We conducted secondary univariate contrast-based analyses in data drawn from 4 repeated-measures fMRI studies. Participants (Study 1: $N = 60$, mean [M] age $= 15.2 \pm 1.1$ y; Study 2: $N = 109$, M age $= 15.1 \pm 0.9$ y; Study 3: $N = 39$, M age $= 21.2 \pm 3.7$ y; Study 4: $N = 62$, M age $= 29.7 \pm 6.2$ y) completed the same tasks over 3-wk to 3-y test-retest intervals. Studies 3 and 4 included participants with eating disorders and obesity, respectively

Results: Test-retest reliability of the food receipt and food picture paradigms was poor, with average ICC values ranging from 0.07 to 0.20. The monetary reward paradigm and the thin female model picture paradigm also showed poor test-retest reliability: average ICC values 0.21 and 0.12, respectively. Although several regions demonstrated moderate to good test-retest reliability, these results did not replicate across studies using similar paradigms. In Studies 3 and 4, but not Study 1, test-retest reliability in visual processing regions was moderate to good when contrasting single conditions with a low-level baseline.

Conclusions: Results underscore the importance of examining the temporal reliability of fMRI tasks and call for the development and use of well-validated standardized fMRI tasks in eating- and obesity-related studies that can provide more reliable measures of neural activation. The trials were registered at clinicaltrials.gov as NCT02084836, NCT01949636, NCT03261050, and NCT03375853. *Am J Clin Nutr* 2021;0:1–16.

Keywords: test-retest reliability, reward, food picture, food taste, monetary reward, model, repeated-measures fMRI

Introduction

There has been an explosion of studies searching for neural vulnerability factors that predict overeating and future weight gain, serve as biomarkers for treatment response, and capture neural plasticity resulting from overeating that may maintain obesity (1–3). These studies have advanced etiologic theory and identified intervention targets for obesity and eating disorder prevention programs and treatments, but concerns have been raised about the reproducibility, or reliability of findings from functional MRI (fMRI) studies (4–6).

Prior reviews have found moderate average test-retest reliability for a range of nonfood fMRI paradigms. Bennett and Miller (7) found a mean (M) intraclass correlation coefficient (ICC) = 0.50 (range ICC: 0.17–0.75; range interval: 1–413 d) and Elliott et al., (8) an M ICC = 0.40 (range ICC: –0.04–0.92; range interval: 1–1008 d), though several fMRI tasks analyzed by Elliott et al. (8) showed poor test-retest reliability (M ICC = 0.23). Few studies have evaluated the reliability of fMRI tasks used in eating and obesity-related studies. We used whole-brain analyses to test whether neural responsivity to tastes

Support for this work was provided by NIH grants DK080760, DK092468, DK112762, and MH111782.

Supplemental Figures 1–4, Supplementary Material, and Supplementary Table 1 are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/ajcn/.

Address correspondence to SY (e-mail: sonjas@ori.org).

Abbreviations used: ACC, anterior cingulate cortex; DARTEL, Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra; EPI, echo planar imaging; fMRI, functional MRI; FOV, field of view; ICC, intraclass correlation coefficient; M, mean; MNI, Montreal Neurological Institute; MOG, middle occipital gyrus; MP-RAGE, magnetization-prepared rapid gradient-echo; NAcc, nucleus accumbens; OFC, orbitofrontal cortex; ROI, regions of interest; ROI, region of interest; SPL, superior parietal lobe; SPSS, Statistical Package for the Social Sciences; TE, echo time; TR, repetition time; VAS, visual analog scale.

Received October 13, 2020. Accepted for publication March 5, 2021.

First published online 0, 2021; doi: https://doi.org/10.1093/ajcn/nqab096.

of milkshake predicted future weight gain and then evaluated whether the effects replicated in 10 randomly selected bootstrap samples from the larger dataset (6); results from the whole-brain analyses only replicated in 5 or fewer of the bootstrap samples. In a second study, we attempted to replicate the finding that elevated striatal and lateral orbitofrontal cortex (OFC) response to high-calorie food pictures predicts elevated future weight gain (9); elevated striatal response only predicted weight gain in 2 of the 6 examined datasets. Another team (10) examined the test-retest reliability of fasted-state reward-associated brain responses to a food picture fMRI paradigm in adults with overweight and obesity over 18 d; results showed poor test-retest reliability; *M* ICC = 0.17.

As the associations between fMRI activation and weight gain show limited reproducibility in our studies, it seemed critical to investigate the test-retest reliability of the activation produced by the fMRI tasks. In the present study, we examined test-retest reliability of responses in reward-related brain regions to food receipt paradigms, food picture paradigms, a monetary reward paradigm, and a thin female model picture paradigm (i.e. model picture paradigm). A limitation of studies that examined test-retest reliability of fMRI paradigms is that most used small samples and examined only short test-retest intervals (7, 8). We conducted secondary analyses in data drawn from 4 separate repeated-measures studies with sample sizes ranging from $N = 39$ to 109 in which participants completed the same tasks over 3-wk to 3-y test-retest intervals. Prior studies have examined test-retest reliability of non-fMRI tasks (11) and fMRI tasks (4) over multiyear intervals in children and adolescents. However, no study has examined long-term test-retest reliability of fMRI tasks used in eating and weight-related studies. We therefore examined test-retest reliability of our fMRI tasks over 3-y intervals in 2 studies with adolescents.

Methods

Study 1

Participants.

In total, 162 healthy weight adolescents were recruited from Eugene, Oregon, via advertisements for a 3-y prospective study on neural vulnerability factors that predict weight gain (NCT02084836; see **Supplementary Figure 1** for participant flow chart). The primary aim of Study 1 was to examine neural risk factors that predict future weight gain (12). Exclusion criteria were fMRI contraindicators (e.g. dental braces), a BMI (BMI = kg/m^2) <18 or >25, binge eating or compensatory behavior in the past 3 mo, current use of psychoactive medications or drugs more than once weekly, or Axis I psychiatric disorder in the past year. Participants in this study and Study 2 provided written assent and legal guardians provided written informed consent and participants in Studies 3 and 4 provided written informed consent according to the Declaration of Helsinki. The Oregon Research Institute Institutional Review Board approved this study and the other studies mentioned below. All 162 participants completed baseline scans. We invited participants to repeat the fMRI tasks a second time at 2- or 3-y follow-up when they showed either a ≥3% increase in body fat, a ≥3% decrease in body fat, or less than a 2% change in body fat over follow-up. We selected these definitions of weight gain, weight loss, and weight stability after the first year of the study because we estimated that they would allow us to invite enough participants back to complete a second scan to create large enough cell sizes for testing whether participants who gained weight showed differential changes in the BOLD response to these paradigms than participants who showed weight stability (13). In total, 65 adolescents fitted one of these criteria and were scanned a second time at either 2- or 3-y follow-up. However, data from 5 participants were excluded from analyses due to excessive movement ($n = 1$), incomplete scans ($n = 2$), and acquisition errors ($n = 2$), resulting in $n = 60$ (34 females; *M* age = 15.2 ± 1.1; *M* BMI = 21.1 ± 1.9; self-reported ethnic and racial backgrounds: 11.7% Hispanic American, 3.8% American Indian/Alaska Native, 1.9% Asian American, 75.1% white, 7.5% mixed racial heritage). Of the $n = 60$ who completed a second scan, 39 participants completed the second scan at the 2-y follow-up and 21 at the 3-y follow-up. For the current analyses, we combined the scan data from both follow-ups and assessed test-retest reliability between the baseline scan and the follow-up scan (either 2- or 3-y follow-up).

fMRI paradigms.

Participants completed a food receipt paradigm and a monetary reward paradigm (**Table 1**; [14]). All pictures in this study and in Studies 2–4 were modified for consistent size, resolution, and luminance. The food receipt paradigm assesses the BOLD response to receipt and anticipated receipt of a chocolate milkshake and a tasteless solution. Stimuli were 2 pictures (glasses of milkshake and water, 50 repeats of each) that cued the delivery of either 0.5 cc milkshake or tasteless solution (30 repeats of each). On 40% of the trials, the taste was not delivered following the cue. As no differences were observed between paired and unpaired milkshake and tasteless solution cues, paired and unpaired cues were combined for analyses to increase sensitivity (12). Pictures were presented for 2 s, followed by a jittered blank screen (1–7 s). Taste delivery (30 repeats each) occurred 10 s after cue onset and lasted 5 s, followed by a swallow cue (2 s). The trial ended with a second 1–7 s jitter. Stimuli were presented in 5 runs (9.5 min/run). The order of the 5 runs was randomized over participants.

The monetary reward paradigm assesses the BOLD response to receipt and anticipated receipt of monetary reward. Stimuli consisted of 3 coins (heads or tails). During the task, a coin on the left side of the screen would blink heads (H) or tails (T) 2–4 times for 300 ms per blink before it "landed" on either H or T. After 2 s, a second coin in the middle of the screen blinked 4–6 times before landing on H or T. After 3 s, a third coin blinked 8–10 times on the right side of the screen before landing on H or T. After the presentation of the 3 coins, a message (2–3 s) appeared saying; "You win $3" or "You don't win," followed by a fixation cross (2 s). Participants won $3 each time 3 H or 3 T were displayed. We did not include a motor response because we did not want to confound motor responsivity with activation of reward circuitry to receiving and anticipating receiving monetary reward. In total, there were 16 repeats of a winning coin display (3 H or 3 T in a row), 28 repeats of a potential winning coin display (2 H or 2 T in a row), and 44 reward neutral coin displays (the time the first coin stopped blinking [1 H or 1 T]). Stimuli were presented

TABLE 1 Overview of samples, fMRI tasks, and contrasts[1]

Study	Interval (wk)	Sample	fMRI task	Contrasts
Study 1	$M = 126.8 \pm 25.53$	60 healthy weight adolescents (male and female; mean [*M*] age = 15.2 ± 1.1)	Food receipt; 5 runs	Milkshake glass >water glass; milkshake receipt >tasteless solution receipt
			Monetary reward; 2 runs	Potential winning coin display (2 H or 2 T in a row) >reward neutral coin display (the time the first coin stopped blinking [1 H or 1 T]); winning coin display (third coin stopped blinking and matched previous 2) >reward neutral coin display
Study 2	$M = 153.62 \pm 5.48$	109 healthy weight adolescents (male and female; *M* age = 15.1 ± 0.9)	Food picture; 1 run	Appetizing food pictures >unappetizing food pictures; appetizing food pictures >glasses of water
			Food receipt; 2 runs	High-fat/high-sugar milkshake receipt >tasteless solution receipt; high-fat/low-sugar milkshake receipt >tasteless solution receipt; low-fat/high-sugar milkshake receipt >tasteless solution receipt; low-fat/low-sugar milkshake receipt >tasteless solution receipt
Study 3	$M = 8.74 \pm 1.84$	39 females with eating disorders (*M* age = 21.2 ± 3.7)	Food picture; 1 run	High-calorie binge foods >low-calorie healthy foods
			Model picture; 1 run	Thin models >average-weight models
Study 4	$M = 3.03 \pm 0.72$	62 overweight and obese adults (male and female; *M* age = 29.7 ± 6.2)	Food picture; 1 run	Appetizing high-calorie food >appetizing low-calorie food; appetizing high-calorie food >water glasses

[1]fMRI, functional MRI; H, heads; M, Mean; T, tails.

in 2 runs (8.5 min/run). The order of the 2 runs was randomized over participants. Each participant was presented with the same order of the 2 tasks on both scan days.

Study 2

Participants.

In total, 135 healthy-weight male and female adolescents aged between 14 and 17 y were recruited from Portland, Oregon, via advertisements for a 3-y prospective study investigating neural vulnerability factors that predict weight gain and neural adaptations associated with weight gain (NCT01949636; see **Supplementary Figure 2** for participant flow chart). The primary aim of Study 2 was to examine neural vulnerability factors that predicted future weight gain and neural plasticity of reward and attention circuitry that occurs in response to overeating that leads to weight gain (3, 6). Adolescents were excluded if they reported fMRI contraindicators, BMI <18 or >25, binge eating or compensatory behavior in the past 3 mo, at least weekly use of psychotropic medications or illicit drugs, or an Axis I psychiatric disorder in the past year. Participants were scanned annually over a 3-y period. In total, 109 participants (61 females, *M* age = 15.1 ± 0.9; *M* BMI = 21.2 ± 2.2; self-reported ethnic and racial backgrounds: 6.4% Hispanic American, 2.8% American Indian/Alaska Native, 6.6% Asian American, 8.5% black, 0.9% Native Hawaiian or Other Pacific Islander, 74.7% white) successfully completed all 4 scans (i.e., baseline, and 1-, 2-, and 3-y follow-ups). For the test-retest reliability analyses, we examined the average test-retest reliability across the 4 time points.

fMRI paradigms.

Participants completed a food receipt paradigm and a food picture paradigm (Table 1; [6]). Participants were presented with the same order of the 2 tasks on the scan days. The food receipt paradigm assesses neural response to the receipt of 4 chocolate

milkshakes varying in sugar and fat content (a high-fat/high-sugar milkshake, a high-fat/low-sugar milkshake, a low-fat/high-sugar milkshake, and a low-fat/low-sugar milkshake) and to receipt of a tasteless solution. Participants were informed that they would receive 4 different milkshakes but not about the fat and sugar content of the milkshakes. During the task, pictures of glasses of milkshake or water signaled the delivery of the tastes. All milkshakes were preceded by the same image of a milkshake. Pictures were presented for 1 s, followed by a fixation cross during which the tastes (0.7 cc) were delivered over 5 s and a swallow cue (3 s). The delivery of the tastes occurred in 6 variable-length blocks (1 block presented 4, 5, or 7 events) over 2 runs (32 events of each taste across the 2 runs; 13.5 min/run). Only 1 type of milkshake was delivered per block. After a milkshake block was completed, subjects received a rinse of the tasteless solution followed by a swallow cue (0.5 s) and a jitter (9–11 s). The tasteless solution block followed the same pattern but without a rinse. The order of presentation of blocks was randomized across participants.

The food picture paradigm assesses the BOLD response to pictures of food and glasses of water. Prior to the scan, participants rated how appetizing they found various foods shown in 129 pictures using a visual analog scale (VAS) with the line on the VAS converted to numbers based on length (VAS range: "least appetizing" = −395 to "most appetizing" = 395). Food rated between –395 and –250 was defined as "unappetizing food." Food rated between 250 and 395 was defined as "appetizing food." During the paradigm, each participant was exposed to pictures of food they rated as the most appetizing, pictures of food they rated as least appetizing, and pictures of glasses of water (32 events of each). Participants were asked to imagine tasting and eating the pictured food and to imagine drinking the water. Pictures were presented for 5 s, followed by a jittered fixation cross (2–4 s). Stimuli were presented in 1 16-min run. The order of presentation of pictures was randomized over participants.

Study 3

Participants.

In total, 138 women with Diagnostic and Statisticial Manual of Mental Disorders 5th Edition (DSM-5) eating disorders were recruited from universities and surrounding communities in Oregon and Texas via advertisements inviting women with body image and eating concerns to participate in an eating disorder treatment trial (NCT03261050; see **Supplementary Figure 3** for participant flow chart). The primary aim of Study 3 was to test whether the eating disorder treatment trial reduced activation in regions implicated in reward valuation in response to thin models and high-calorie binge foods relative to a waitlist condition (15). Those who reported a BMI <75% ideal body weight, current acute suicidal ideation, a comorbid psychiatric disorder that would disrupt the group treatment (e.g. bipolar disorder), or serious medical problems (e.g. diabetes) were excluded. Women who reported contraindicators of MRI, recreational drug use, or medication use were excluded from the fMRI scans.

Participants were randomly assigned to the treatment or the waitlist control condition using a random number table. In the 8-session group treatment (referred to as *Body Project Treatment*), participants completed verbal, written, and behavioral activities in which they discuss negative effects of pursuing the thin beauty ideal and engaging in disordered eating behaviors. Participants were scanned prior to treatment and immediately after treatment and a parallel period of time for waitlist controls (an average of 9 wk) to test whether the treatment trial reduces valuation of the thin beauty ideal and high-calorie binge foods, the intervention targets, relative to waitlist controls. For the current study, we investigated 9-wk test-retest reliability of brain activation in waitlist controls. In total, 55 participants were randomized to the waitlist control condition. Of waitlist controls, 43 participants completed the baseline scan, and 39 completed postscans (*M* age = 21.2 ± 3.7; *M* BMI = 25.3 ± 7.2; self-reported ethnic and racial backgrounds: 20.5% Hispanic American; 16.1% Asian American; 6.5% black; 56.9% white).

fMRI paradigms.

Participants completed a food picture paradigm and a model picture paradigm (Table 1; [15]). The food picture paradigm assessed the BOLD response to pictures of high-calorie binge foods (e.g. chocolate cake) and low-calorie foods (fruits and vegetables). Participants were asked to think about how much they wanted to eat the pictured food (20 events of each). The model image paradigm assessed the BOLD response to pictures of thin models and average-weight models. Participants were asked to think about the attractiveness of each model (20 events of each). A panel of research assistants rated food and model pictures for each category. If a picture was rated incorrectly as a binge or healthy food/thin or average-weight model, the picture was not used. The final pictures were chosen based on consensus. For each thin model picture chosen, an average-weight model was chosen that matched on aspects outside of size (e.g. conventional attractiveness, clothing type, race). In both paradigms, pictures were presented for 5 s in a randomized order. A 4–8 s jittered fixation cross occurred between each image. During each paradigm, stimuli were presented in 1 7.5-min scanning run. The order of paradigms and order of presentation of the pictures were randomized over participants. To avoid order habituation, participants were presented with the opposite order of the 2 tasks on their second scan day.

Study 4

Participants.

In total, 179 overweight/obese adults were recruited in Eugene, Oregon, via advertisements for a study evaluating the efficacy of a 4-weekly session multifaceted food response and attention training intervention on weight and body fat (NCT03375853; see **Supplementary Figure 4** for participant flow chart). The primary aim of Study 4 was to examine whether the food response and attention training reduces body fat. Exclusion criteria were a BMI >40, any current use of psychotropic medications or illicit drugs, current Axis I psychiatric disorder, serious medical problems (e.g. diabetes), bariatric surgery, special dietary requirements (e.g. vegan), and participation in a paid weight loss program. Participants who reported contraindicators of MRI were excluded from the fMRI aspect of the study. Participants were randomly assigned to a food response and attention training weight loss intervention or an active control condition

(a parallel generic response and attention training comparison condition involving nonfood pictures) using a random number table. Participants were scanned prior to their first training session and immediately after their last training session (3 wk later) to test whether the food response and attention training weight loss intervention would produce greater reductions in neural responses to food images compared with the control condition. For the current study, we investigated the 3-wk test-retest reliability of brain activation in participants who completed the control condition ($N = 75$). Of the 70 control participants who completed the baseline scan, $n = 62$ completed the postscan (49 females; M age $= 29.7 \pm 6.2$; M BMI $= 31.6 \pm 4.3$; self-reported ethnic and racial backgrounds: 16.1% Hispanic American, 1.9% American Indian/Alaska Native, 5.6% Asian American, 3.7% black, 59.8% white, 13.0% mixed racial heritage).

fMRI paradigm.

Participants completed a food picture paradigm that assessed the BOLD response to pictures of high-calorie foods, low-calorie foods, and glasses of water. Prior to the scan, participants rated the palatability of 100 high-calorie and 100 low-calorie food pictures (16). During the paradigm, participants were exposed to pictures of their highest rated high-calorie foods, pictures of their highest rated low-calorie foods, and pictures of glasses of water (20 events each from the same category; appetizing high-calorie foods, appetizing low-calorie foods, glasses of water). Participants were asked to think about how much they wanted to eat the pictured food. Stimuli were presented in 1 8-min scanning run. Pictures were presented for 5 s, followed by a 4–8 s jittered crosshair. The order of presentation of the pictures was randomized over participants.

fMRI data acquisition.

Study 1 baseline MRI data were acquired on a Siemens Allegra 3 Tesla (3T) scanner. Follow-up scans were primarily performed on the Allegra, but 4 participants were scanned with a Siemens Skyra 3T scanner. Study 2 MRI data were acquired on a Siemens Tim Trio 3T MRI scanner. Studies 3 and 4 MRI data were acquired on a Siemens Skyra 3T MRI scanner. In Studies 1 and 2, functional scans used a T2* weighted gradient single-shot echo planar imaging (EPI) sequence (echo time [TE] = 30 ms, repetition time [TR] = 2000 ms, flip angle = 80°) with an in-plane resolution of 3.0 × 3.0 mm^2 (64 × 64 matrix; field of view [FOV] = 192). To cover the whole brain, 32 4-mm slices (interleaved acquisition, no skip) were acquired along the anterior commissure-posterior commissure line (AC-PC) transverse, oblique plane as determined by the midsagittal section. Structural scans were collected using an inversion recovery T1 weighted sequence (magnetization-prepared rapid gradient-echo [MP-RAGE]) in the same orientation as the functional sequences to provide detailed anatomic images aligned to the functional scans. High-resolution structural MRI sequences (FOV = 256, thickness = 1.0 mm) were acquired. In Studies 3 and 4, functional scans used a T2* weighted EPI plus sequence (72 slices, TE = 25 ms, TR = 2000 ms, flip angle = 90°, matrix size = 100 × 100, voxel size = 2 mm^3, axial slices = 72, FOV = 200, multiband acceleration factor = 3). Structural scans were collected using a high-resolution anatomical T1-weighted MP-RAGE scan (TE = 3.43 ms, TR = 2500 ms, 256 × 256 matrix, voxel size = 1 mm^3, sagittal slices = 176, FOV = 256).

Statistical analysis

fMRI preprocessing.

Neuroimaging data were skullstripped using the Brain Extraction Tool in Functional MRI of the Brain Analysis Group Software library (FSL) and then preprocessed and analyzed using Statistical Parametric Mapping version 12 (SPM12) (Wellcome Department of Cognitive Neurology; http://www.fil.ion. ucl.ac.uk/spm) in Matlab (Mathworks, Inc.). Anatomical images were segmented and normalized to Montreal Neurological Institute (MNI) space with the use of the Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL) toolbox, coregistered to the mean functional image, and segmented into 6 tissue types using unified segmentation approach (17). Functional data were corrected for slice-time acquisition differences (Studies 1–2 only), adjusted for variation in magnetic field distortion using field maps, realigned to the mean functional from that run, coregistered with the anatomical, normalized to MNI space using the DARTEL template and deformation fields output, and smoothed to 6 mm Gaussian full-width-at-half-maximum (FWHM). Best practice has shifted with respect to slice-timing correction over the years these studies were conducted. Previous studies have shown that slice-timing correction in event-related designs results in little to no benefit with a small TR (≤2 s) (18). As a result, functional data from the ongoing Studies 3 and 4 were not corrected for slice-time acquisition. Functional data were assessed to detect spikes in global mean response and motion outliers using the Artifact Detection Toolbox (ART; Gabrieli Laboratory). Head motion >3 mm or degrees in any direction was our a priori exclusion criteria. Motion parameters <3 mm were included as regressors in the design matrix at individual fixed effect analysis. Image volumes where the z-normalized global brain activation showed >1.5 mm of composite (linear plus rotational) movement were flagged as outliers and deweighted during individual-level model estimation. There were no significant differences in scans deweighted between the 4 datasets (Ps >0.48).

fMRI analysis.

At the subject level, the BOLD signal was modeled in a fixed effects analysis with separate regressors modeling each condition of interest for each task. T-maps were constructed for comparisons of activation within participants (see Table 1 for contrasts). For all data, a high-pass filter of 128 s was applied to eliminate low-frequency noise and slow drifts in the signal. First-order autoregressive error was used to correct for serial autocorrelations. To identify brain activation at group level, we calculated separately for each paradigm and each scan 1-sample t-tests, using the contrast images obtained in the single subject analysis as input data. Self-reported hunger prior to the scan, which was measured in all studies, was included as a covariate of no interest in the food-specific fMRI task analyses (see **Supplementary Material** and **Table 2** for more details on hunger ratings). We performed a sensitivity analysis with zBMI as an

TABLE 2 Means and SDs for scan times and hunger ratings for Study 1 ($n = 60$), Study 2 ($n = 109$), Study 3 ($n = 39$), and Study 4 ($n = 62$)[1]

	Study 1[2] Mean ± SD	Study 2[3] Mean ± SD	Study 3[4] Mean ± SD	Study 4[4] Mean ± SD
Time of day scan 1	15:16 ± 2:31	13:05 ± 1:59	12:33 ± 3:15	12:15 ± 2:45
Time of day scan 2	14:27 ± 1:46	13:22 ± 1:55	12:58 ± 2:34	11:49 ± 2:49
Time of day scan 3		12:37 ± 1:57		
Time of day scan 4		13:32 ± 2:27		
Hunger scan 1	7.62 ± 4.36	11.20 + 3.93	8.15 ± 4.62	10.75 ± 4.16
Hunger scan 2	10.35 ± 3.58	11.28 ± 4.17	8.63 ± 4.28	12.02 ± 3.85
Hunger scan 3		11.50 ± 4.05		
Hunger scan 4		12.02 ± 3.62		

[1]Scan time is reported in military time. Participants rated their hunger level on 20-cm crossmodal visual analog scales (VASs). VAS ratings were anchored by 0 (not at all), 10 (neutral), and 20 (never been more hungry).

[2]Paired t-test analyses showed that in Study 1 there was a significant difference in scan time between the baseline scan and follow-up scan ($t(59) = 2.21$, $P = 0.03$) and in hunger ratings between both scans ($t(59) = 3.79$, $P < 0.001$).

[3]In Study 2, there was a significant difference in scan time between the second scan and third scan ($t(108) = 2.82$, $P = 0.006$) and between the third scan and fourth scan ($t(108) = 2.91$, $P = 0.004$), but no significant differences in hunger ratings between scans.

[4]In Study 3 and Study 4, there were no significant differences in scan times and hunger ratings between the 2 scans (Ps > 0.07).

additional covariate in the models in Studies 1 and 2: the test-retest reliability results were unchanged when including zBMI as a covariate. Although we attempted to conduct baseline and follow-up scans at the same time of day, because of scheduling limitations this was not always possible (Table 2).

Regions of interest approach. ICCs, which are considered the preferred method of assessing test-retest reliability of fMRI paradigms (7, 8), can be calculated both on a voxel-by-voxel basis ("Voxel-ICCs") and on a region of interest (ROI) basis. We calculated the ICC on an ROI basis, which is less noisy than voxel-ICCs because regions of interest (ROIs) are averages of individual voxels. We used anatomically defined ROIs using masks from the Wake Forest University PickAtlas toolbox (19). As ROIs, we chose left and right anterior cingulate cortex (ACC), amygdala, caudate, insula, medial orbitofrontal cortex (medial OFC), nucleus accumbens (NAcc), putamen, and thalamus, as these regions have been activated by food picture paradigms (20–22), food receipt paradigms (23–27), and monetary reward paradigms (27–31). We also calculated reliability statistics for 1 functionally defined reference region per contrast (i.e. the peak cluster). For this latter approach, we conducted whole brain analyses with cluster extent thresholds at $P < 0.001$ with clusters being statistically significant corrected for multiple comparisons across the whole brain (determined with the SPM cluster size threshold tool: https://github.com/CyclotronResearchCentre/SPM_ClusterSizeThreshold). We selected regions that were most strongly activated by each of the contrasts (e.g. milkshake cue >tasteless solution cue) at baseline.

Reliability analysis.

Subject-level parameter estimates were extracted for each contrast, ROI, and scanning session and analyzed with the Statistical Package for the Social Sciences 24 (SPSS 24, SPSS Inc.). The data within the ROIs, including those from the reference regions, were extracted using the MARSeille Boîte À Région d'Intérêt (MarsBar) tool (32). We assessed reliability using ICC (1, 3), a 2-way mixed effects ICC, defined by (33) with k referring to the number of repeated scans (i.e. 2 scans in Studies 1, 3, and 4; 4 scans in Study 2). This form of ICC estimates the correlation between the BOLD fMRI signal intensities between scans and has been used previously (5, 10). The effect of scans is assumed to be fixed whereas the effects of subjects are assumed to be random (5). Due to the possibility that participants might habituate to the stimuli over time, we employed a "consistency" measure of ICC, rather than testing the absolute agreement between scan days. That is, a high ICC implies that subjects with a greater BOLD response during the baseline scan (relative to other subjects) also show greater activation during the follow-up scan. We followed the conventional classification of ICCs to quantify the degree of reliability: ICC <0.4 = poor reliability; 0.4–0.75 = moderate to good reliability; >0.75 = excellent reliability (5, 34). A negative ICC is interpreted as a reliability of zero (35). We report P values for all reliable activations and 95% CIs for all ICCs.

Results

Differences in scan times and hunger ratings

In Study 1, there was a significant difference in scan time between the baseline scan (M = 15:16 ± 2:31 h) and follow-up scan (M = 14:27 ± 1:46; $t(59) = 2.21$, $P = 0.03$). There was also a significant difference in hunger prior to the baseline scan ($M = 7.62 \pm 4.36$) versus follow-up scan ($M = 10.35 \pm 3.58$; $t(59) = 3.79$, $P < 0.001$) (Table 2). In Study 2, there was a significant difference in scan time between the second scan (M = 13:22 ± 1:55) and third scan (M = 12:37 ± 1:57; $t(108) = 2.82$, $P = 0.006$) and between the third scan and fourth scan (M = 13:32 ± 2:37; $t(108) = 2.91$, $P = 0.004$), but no significant differences in hunger ratings (Table 2). In Study 3 and Study 4, there were no significant differences in scan times and hunger ratings between the 2 scan days (Table 2; Ps >0.07).

Main activation patterns Study 1

The food receipt paradigm elicited the strongest activation in the left lingual gyrus (MNI coordinates: –24, –84, –12, Z = 5.79, $k = 192$, $P < 0.001$) in response to the contrast

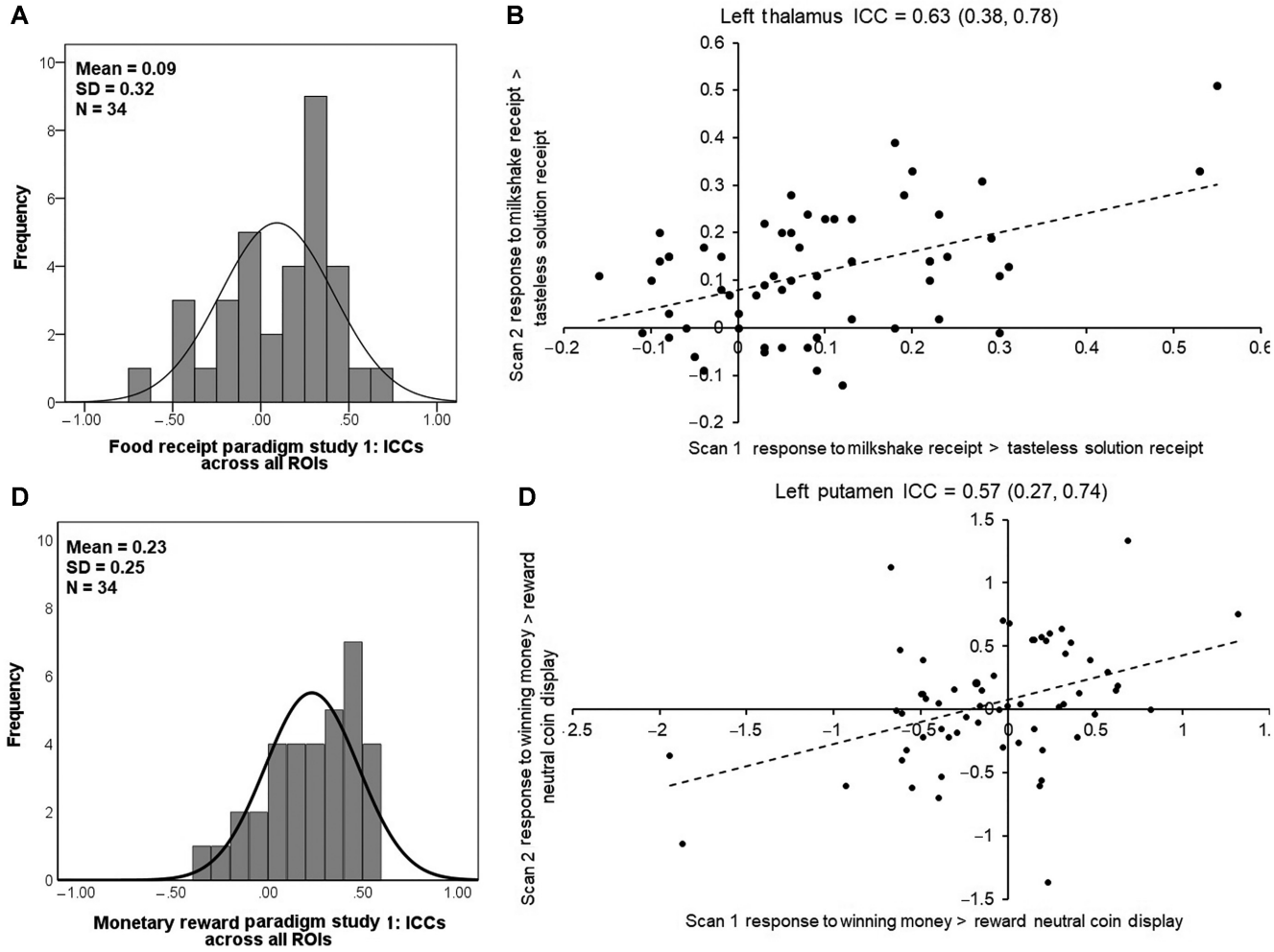

FIGURE 1 (A) Intraclass correlation coefficient (ICC) frequency distributions across all regions (16 anatomical regions and 1 functional reference region per contrast) for the food receipt paradigm in Study 1 ($n = 60$; number [N] of ICC coefficients = 34). (B) Distribution of parameter estimates in Study 1 ($n = 60$) for the left thalamus in response to the contrast milkshake receipt >tasteless solution receipt (food receipt paradigm). When excluding 2 outliers in the left thalamus response (i.e. parameter estimates exceeding 3 SDs from the mean parameter estimate), the ICC was lower but still in the moderate range: ICC = 0.43, $P = 0.02$, 95% CI = 0.04, 0.66. (C) ICC frequency distributions across all regions (16 anatomical regions and 1 functional reference region per contrast) for the monetary reward paradigm in Study 1 (N ICC coefficients = 34). (D) Distribution of parameter estimates in Study 1 ($n = 60$) for the left putamen in response to the contrast winning money >reward neutral coin display (monetary reward paradigm). When excluding 2 outliers in left putamen response, the test-retest reliability remained in the moderate range: ICC = 0.46, $P = 0.01$, 95% CI = 0.08, 0.68. ROI, region of interest.

milkshake glass >water glass and the left postcentral gyrus (MNI coordinates: –51, –12, 39, Z >8, $k = 4394$, $P < 0.001$) in response to the contrast milkshake receipt >tasteless solution receipt. The monetary reward paradigm elicited activation in the left cuneus (MNI coordinates: –3, –72, 3, Z = 6.54, $k = 2127$, $P < 0.001$) in response to the contrast potential win >reward neutral coin display and left cuneus (MNI coordinates: –6, –72, 3, Z = 7.74, $k = 2749$, $P < 0.001$) in response to the contrast winning money >reward neutral coin display.

Main activation patterns Study 2

The food receipt paradigm elicited the strongest activation in the right middle temporal gyrus (MTG MNI coordinates: 51, –24, –12, Z = 4.16, $k = 57$, $P < 0.001$) in response to the contrast high-fat/high-sugar milkshake receipt >tasteless solution receipt, the right precuneus (MNI coordinates: 15, –63, 39, Z = 4.21, $k = 22$, $P < 0.001$) in response to the contrast high-fat/low-sugar milkshake receipt >tasteless solution receipt, the right insula (MNI coordinates: 36, –6, 0, Z = 6.09, $k = 868$, $P < 0.001$) in response to the contrast low-fat/high-sugar milkshake receipt >tasteless solution receipt, and the right precuneus (MNI coordinates: 15, –66, 39, Z = 5.52, $k = 44$, $P < 0.001$) in response to the contrast low-fat/low-sugar milkshake receipt >tasteless solution receipt.

The food picture paradigm elicited activation in the left superior parietal lobe (SPL MNI coordinates: –9, –69, 54, Z = 5.89, $k = 1078$, $P < 0.001$) in response to the contrast appetizing foods >unappetizing foods and the left lingual gyrus (MNI coordinates: –3, –87, –9, Z > 8, $k = 6991$, $P < 0.001$) in response to the contrast appetizing foods >glasses of water.

Main activation patterns Study 3

The food picture paradigm elicited activation in the left cuneus (MNI coordinates: –6, –100, 14, Z = 6.44, $k = 1854$, $P < 0.001$)

TABLE 3 Intrasubject reliability statistics for the food receipt paradigm and monetary reward paradigm in Study 1 ($n = 60$)[1]

	Food receipt paradigm		Monetary reward paradigm	
Contrasts	Milkshake glass >water glass ICC (95% CI)	Milkshake receipt >tasteless solution receipt ICC (95% CI)	Potential win >reward neutral coin ICC (95% CI)	Winning > reward neutral coin ICC (95% CI)
Reference regions (i.e. peak clusters)	L Lingual gyrus –0.01 (–0.70, 0.39)	L Postcentral gyrus –0.39 (–1.33, 0.17)	L Cuneus 0.41 (0.01, 0.65)*	L Cuneus 0.57 (0.28, 0.74)*
ROIs:				
L ACC	–0.06 (–0.77, 0.37)	–0.46 (–1.44, 0.13)	–0.11 (–0.86, 0.34)	–0.12 (–0.88, 0.34)
R ACC	0.01 (–0.66, 0.41)	–0.71 (–1.86, –0.02)	–0.30 (–1.19, 0.23)	–0.38 (–1.33, 0.18)
L Amygdala	0.16 (–0.41, 0.50)	0.38 (–0.03, 0.63)	0.25 (–0.26, 0.56)	0.40 (–0.01, 0.64)*
R Amygdala	–0.06 (–0.78, 0.37)	0.36 (–0.07, 0.62)	0.19 (–0.36, 0.52)	0.36 (–0.08, 0.62)
L Caudate	0.30 (–0.17, 0.58)	0.26 (–0.25, 0.56)	0.28 (–0.21, 0.57)	0.35 (–0.10, 0.61)
R Caudate	0.34 (–0.11, 0.61)	0.22 (–0.31, 0.54)	0.14 (–0.44, 0.49)	0.50 (0.15, 0.70)*
L Insula	0.14 (–0.44, 0.49)	0.42 (0.02, 0.65)*	0.19 (–0.36, 0.20)	0.01 (–0.66, 0.41)
R Insula	–0.30(–1.18, 0.22)	0.38 (–0.04, 0.63)	–0.07 (–0.80, 0.37)	0.02 (–0.65, 0.42)
L mOFC	0.26 (–0.24, 0.56)	0.27 (–0.23, 0.56)	–0.01 (–0.70, 0.40)	0.29 (–0.19, 0.58)
R mOFC	–0.16 (–0.93, 0.31)	0.31 (–0.16, 0.59)	0.32 (–0.14, 0.60)	0.38 (–0.04, 0.63)
L NAcc	–0.17 (–0.96, 0.31)	0.35 (–0.10, 0.61)	0.16 (–0.41, 0.50)	0.44 (0.06, 0.67)*
R NAcc	0.16 (–0.42, 0.50)	–0.06 (–0.77, 0.37)	0.06 (–0.59, 0.44)	0.48 (0.13, 0.70)*
L Putamen	0.07 (–0.57, 0.44)	0.56 (0.26, 0.74)*	0.53 (0.20, 0.72)*	0.57 (0.27, 0.74)*
R Putamen	–0.06 (–0.77, 0.37)	0.26 (–0.23, 0.56)	0.35 (–0.10, 0.61)	0.46 (0.09, 0.68)*
L Thalamus	–0.50 (–1.50, 0.11)	0.63 (0.38, 0.78)*	0.27 (–0.23, 0.56)	0.46 (0.09, 0.68)*
R Thalamus	–0.24 (–1.08, 0.26)	0.48 (0.13, 0.69)*	0.03 (–0.63, 0.43)	0.41 (0.01, 0.65)*
Average ICC per contrast across all regions	–0.01 (–0.69, 0.40)	0.19 (–0.35, –0.52)	0.16 (–0.42, 0.48)	0.31 (–0.19, 0.59)

[1]The intraclass correlation coefficient (ICC) across the 2 time points is reported for each task and for each anatomically defined region of interest (ROI). A negative ICC is interpreted as indicating a reliability of zero (35). *ICC = 0.40–0.75. ACC, anterior cingulate cortex; L, left hemisphere; mOFC, medial orbitofrontal cortex; MTG, middle temporal gyrus; NAcc, nucleus accumbens; R, right hemisphere; ROI, region of interest.

in response to the contrast high-calorie binge foods >low-calorie foods. The model picture paradigm elicited activation in the left inferior occipital gyrus (MNI coordinates: –42, –70, –4, Z = 3.79, k = 13, P <0.001) in response to the contrast thin models >average-weight models.

Main activation patterns Study 4

The food picture paradigm elicited activation in the right middle occipital gyrus (MNI coordinates 30, –82, 8, Z = 4.41, k = 194, P <0.001) in response to the contrast appetizing high-calorie foods >appetizing low-calorie foods and in the left lingual gyrus (MNI coordinates –12, –94, –10, Z >8, k = 5617, P <0.001) in response to the contrast appetizing high-calorie foods >glasses of water.

Test-retest reliability in Study 1

The average test-retest reliability of the food receipt paradigm across all regions was poor: M ICC = 0.09 (**Figure 1**A). The reference regions (i.e. peak clusters) did not exceed an ICC of 0.4 (**Table 3**). Within the ROI, 4 out of the 32 instances showed moderate to good test-retest reliability. All 4 instances occurred in response to the contrast milkshake receipt >tasteless solution receipt (Table 3): left insula (ICC = 0.42, P = 0.02, 95% CI: 0.02, 0.65), left putamen (ICC = 0.56, P = 0.001, 95% CI: 0.26,0.74), left thalamus (ICC = 0.63, P <0.001, 95% CI: 0.38, 0.78; Figure 1B) and right thalamus (ICC = 0.48, P = 0.006, 95% CI: 0.13, 0.69). When excluding 2 outliers in the left thalamus response (i.e. parameter estimates exceeding 3 SDs from the mean parameter estimate), the ICC remained in the moderate to good range: ICC = 0.43, P = 0.02, 95% CI: 0.04, 0.66. Previous studies have found robust neural activation in the amygdala, caudate, putamen, and thalamus in response to food receipt (23, 24, 26). In the current study, the average ICC across these regions was 0.39 (95% CI: –0.12, 0.64).

The average test-retest reliability of the monetary reward paradigm across all regions was likewise poor: M ICC = 0.23 (Figure 1C). However, certain brain regions demonstrated moderate to good reliability. The reference regions exceeded an ICC of 0.4: ICC left cuneus in response to the contrast potential win >reward neutral coin display = 0.41, P = 0.02, 95% CI: 0.01, 0.65; ICC left cuneus in response to the contrast winning money >reward neutral coin = 0.57, P <0.001, 95% CI: 0.28, 0.74 (Table 3). Within the ROIs, 9 out of the 32 instances showed moderate test-retest reliability (Table 3). In response to the contrast potential win >reward neutral coin display, the ICC of the left putamen exceeded 0.4: ICC = 0.53, P = 0.003, 95% CI: 0.20, 0.72). In response to the contrast winning money >reward neutral coin display, the ICCs of the following regions >0.4: left amygdala (ICC = 0.40, P = 0.027, 95% CI: –0.01, 0.64), right caudate (ICC = 0.50, P = 0.005, 95% CI: 0.15,0.70), left NAcc (ICC = 0.44, P = 0.014, 95% CI: 0.06, 0.67), right NAcc (ICC = 0.48, P = 0.007, 95%

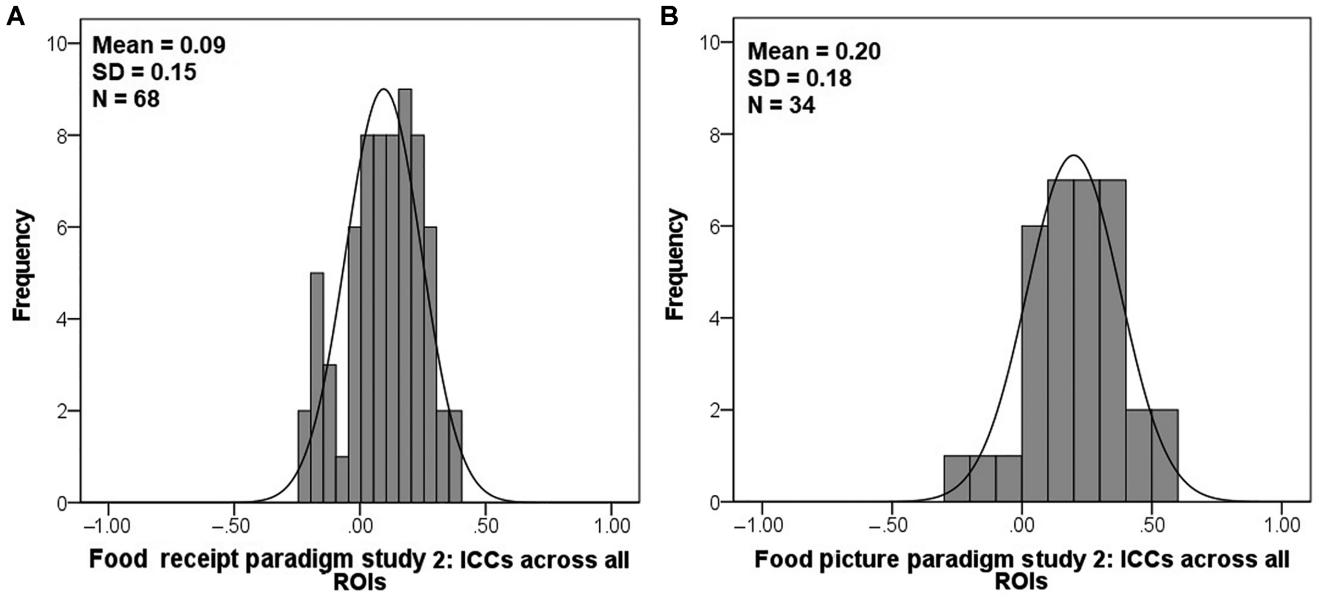

FIGURE 2 Intraclass correlation coefficient (ICC) frequency distributions across all regions (16 anatomical regions and 1 functional reference region per contrast for (A) the food receipt paradigm (N ICC coefficients = 68) and (B) the food picture paradigm (N ICC coefficients = 34) in Study 2 (n = 109). ROI, region of interest.

CI: 0.13, 0.70), left putamen (ICC = 0.57, P <0.001, 95% CI: 0.27, 0.74; Figure 1D), right putamen (ICC = 0.46, P = 0.01, 95% CI: 0.09, 0.68), left thalamus (ICC = 0.46, P = 0.027, 95% CI: 0.09, 0.68), and right thalamus (ICC = 0.41, P = 0.023, 95% CI: 0.01, 0.65). When excluding 2 outliers in left putamen response, the ICC remained >0.4: ICC = 0.46, P = 0.01, 95% CI: 0.08, 0.68. Monetary reward paradigms have shown good to excellent test-retest reliability in the striatum (NAcc M ICC = 0.94: [5]; caudate and putamen M ICC = 0.74 [36]). In the current study, test-retest reliability in the striatum was moderate in response to the contrast winning money >reward neutral coin display (M ICC = 0.47) but poor in response to the contrast potential win >reward neutral coin display (M ICC = 0.25).

Controlling for scanner type (Skyra 3T MRI scanner n = 4) did not change the results. Further, when excluding the MRI data of the 4 subjects that were acquired on a Siemens Skyra 3T MRI scanner, the average test-retest reliability across all regions remained low for both paradigms: food receipt paradigm M ICC = 0.11 and monetary reward paradigm M ICC = 0.21.

Test-retest reliability in Study 2

The average test-retest reliability of the food receipt paradigm and food picture paradigm across the 4 time points was poor: M ICC = 0.09 and M ICC = 0.20 (**Figure 2**A,B), respectively. In the food receipt paradigm, the ICCs of the reference regions and ROIs did not exceed 0.4 (**Table 4**). The average ICCs for the thalamus, striatum, and amygdala (23, 24, 26) were all poor: high-fat/high-sugar milkshake receipt >tasteless solution receipt M ICC = 0.19, high-fat/low-sugar milkshake receipt >tasteless solution receipt M ICC = 0.17, low-fat/high-sugar milkshake receipt >tasteless solution receipt M ICC = 0.12, low-fat/low-sugar milkshake receipt >tasteless solution receipt M ICC = 0.03. In the food picture paradigm, the ICCs of both reference regions and 2 of the ROIs were >0.4. In response to the contrast appetizing food pictures >unappetizing food pictures, the reference region showed moderate to good test-retest reliability (left SPL ICC = 0.52, P <0.001, 95% CI: 0.35, 0.65). The left lingual gyrus (ICC = 0.56, P <0.001, 95% CI: 0.41, 0.69), right ACC (ICC = 0.41, P <0.001, 95% CI: 0.19, 0.57), and left amygdala (ICC = 0.43, P <0.001, 95% CI: 0.23, 0.59) showed moderate to good test-retest reliability in response to the contrast appetizing food pictures >glasses of water (**Table 5**). A previous study (10) found moderate to good test-retest reliability in the OFC (M ICC = 0.44) but poor test-retest reliability in the insula (M ICC = 0.32), amygdala (M ICC = 0.27), caudate (M ICC = 0.08), and putamen (M ICC = –0.15) in response to food pictures. In the current study, these regions showed poor average test-retest reliability (Table 5): appetizing food >unappetizing food M ICC = 0.15, appetizing food >glasses of water M ICC = 0.27.

We also calculated the average test-retest reliabilities across 2 time points (baseline versus 1-y follow-up) and across 3 time points (baseline, 1-y follow-up, and 2-y follow-up) (**Figure 3**A, B). Similar to the results above, the average test-retest reliability across 2 time points (food receipt paradigm: M ICC = 0.04, 95% CI = –0.40, 0.33; food picture paradigm: M ICC = 0.16, 95% CI = –0.21, 0.42) and 3 time points (food receipt paradigm: M ICC = 0.07, 95% CI = –0.29, 0.34; food picture paradigm: M ICC = 0.22, 95% CI = –0.07, 0.45) was poor.

Test-retest reliability Study 3

The average test-retest reliability of the food picture paradigm and model picture paradigm across all regions was M ICC = 0.12 for each of the paradigms (**Figure 4**A,B). The ICCs of the reference regions and ROIs in the food picture paradigm did not exceed 0.4 (**Table 6**). The average test-retest reliability in the OFC, insula, amygdala, caudate, and putamen (10) was

TABLE 4 Intrasubject reliability statistics for the food receipt paradigm in Study 2 ($n = 109$)[1]

Contrasts	HFHS milkshake receipt >tasteless solution receipt	HFLS milkshake receipt >tasteless solution receipt	LFHS milkshake receipt >tasteless solution receipt	LFLS milkshake receipt >tasteless solution receipt
	ICC (95% CI)	ICC (95% CI)	ICC (95% CI)	ICC (95% CI)
Reference regions (i.e. peak clusters)	R MTG 0.01 (–0.35, 0.30)	R Precuneus –0.18 (–0.61, 0.15)	R Insula 0.00 (–0.36, 0.29)	R Precuneus –0.17 (–0.59, 0.17)
ROIs:				
L ACC	0.21 (–0.08, 0.43)	0.37 (0.13, 0.55)	0.10 (–0.22, 0.36)	–0.09 (–0.49, 0.22)
R ACC	0.28 (0.01, 0.48)	0.31 (0.06, 0.51)	0.15 (–0.16, 0.39)	–0.20 (–0.64, 0.14)
L Amygdala	0.01 (–0.35, 0.29)	0.22 (–0.07, 0.44)	0.13 (–0.18, 0.38)	0.13 (–0.19, 0.38)
R Amygdala	–0.04 (–0.41, 0.26)	0.10 (–0.23, 0.36)	–0.18 (–0.61, 0.16)	–0.19 (–0.62, 0.15)
L Caudate	0.23 (–0.06, 0.45)	0.27 (0.00, 0.48)	0.34 (0.09, 0.53)	–0.15 (–0.57, 0.18)
R Caudate	0.20 (–0.09, 0.43)	0.29 (0.04, 0.50)	0.26 (–0.01, 0.47)	–0.04 (–0.42, 0.26)
L Insula	0.12 (–0.20, 0.37)	–0.13 (–0.54, 0.19)	–0.11 (–0.51, 0.21)	0.02 (–0.34, 0.30)
R Insula	0.17 (–0.13, 0.41)	0.13 (–0.19, 0.38)	0.00 (–0.36, 0.29)	–0.22 (–0.67, 0.13)
L mOFC	0.07 (–0.27, 0.33)	0.17 (–0.14, 0.40)	0.10 (–0.23, 0.36)	0.07 (–0.26, 0.34)
R mOFC	0.15 (–0.16, 0.39)	0.20 (–0.10, 0.43)	0.01 (–0.35, 0.30)	–0.04 (–0.42, 0.26)
L NAcc	0.23 (–0.05, 0.45)	0.18 (–0.12, 0.42)	0.06 (–0.29, 0.33)	0.02 (–0.34, 0.30)
R NAcc	0.36 (0.13, 0.54)	0.27 (–0.07, 0.44)	0.25 (–0.03, 0.46)	0.17 (–0.13, 0.41)
L Putamen	0.18 (–0.11, 0.42)	–0.04 (–0.42, 0.26)	0.15 (–0.16, 0.39)	0.07 (–0.27, 0.33)
R Putamen	0.23 (–0.05, 0.45)	0.17 (–0.14, 0.41)	0.05 (–0.29, 0.32)	0.03 (–0.32, 0.31)
L Thalamus	0.26 (–0.01, 0.47)	0.09 (–0.25, 0.35)	0.01 (–0.35, 0.29)	0.07 (–0.28, 0.33)
R Thalamus	0.24 (–0.04, 0.46)	0.16 (–0.15, 0.40)	0.12 (–0.20, 0.37)	0.23 (–0.06, 0.45)
Average ICC per contrast across all regions	0.17 (–0.13, 0.36)	0.14 (–0.16, 0.39)	0.08 (–0.24, 0.35)	–0.01 (–0.39, 0.27)

[1]The intraclass correlation coefficient (ICC) across the 4 time points is reported for each task and for each anatomically defined region of interest (ROI). A negative ICC is interpreted as indicating a reliability of zero (35). ACC, anterior cingulate cortex; HFHS, high-fat/high-sugar; HFLS, high-fat/low-sugar; L, left hemisphere; LFHS, high-fat/low-sugar; LFLS, low-fat/low-sugar; MTG, middle temporal gyrus; mOFC, medial orbitofrontal cortex; NAcc, nucleus accumbens; R, right hemisphere.

M ICC = 0.11. In the model picture paradigm, the ICC of the reference regions did not exceed 0.4 (Table 6). Within the ROIs for the model picture paradigm, 2 out of the 32 instances showed a reliability of >0.4 (Table 6): right medial OFC (ICC = 0.52, $P = 0.014$, 95% CI: 0.08, 0.75) and left putamen (ICC = 0.47, $P = 0.029$, 95% CI: 0.02, 0.72). However, when excluding 2 influential outliers, the ICCs in the right medial OFC (ICC = 0.35, $P = 0.10$, 95% CI: –0.25, 0.66) and left putamen (ICC = 0.28, $P = 0.16$, 95% CI: –0.38, 0.63) fell in the poor range.

Test-retest reliability in ROIs in Study 4

The food picture paradigm showed poor test-retest reliability: *M* ICC = 0.07 (**Figure 5**A). Although the reference regions showed moderate to good test retest reliability (ICC right middle occipital gyrus [MOG] = 0.70, $P < 0.001$, 95% CI: 0.50, 0.82; Figure 5B; ICC left lingual gyrus = 0.56, $P < 0.001$, 95% CI: 0.27, 0.74), the ICCs of the ROIs did not exceed 0.4 (**Table 7**). The average test-retest reliability in the OFC, insula, amygdala, caudate, and putamen (10) showed poor average test-retest reliability: contrast high-calorie >low-calorie foods *M* ICC = 0.03 and contrast high-calorie >glasses of water *M* ICC = 0.12.

Posthoc analyses

For the abovementioned test-retest reliability analyses, we included 2 contrasted conditions (e.g. high-calorie foods >low-calorie foods). It is possible that both conditions change over time due to variation in, for example, psychological and motivational factors, resulting in poor test-retest reliability. Therefore, we conducted exploratory analyses to examine test-retest reliability for contrasts comparing a single event of interest to a low-level baseline condition (e.g. high-calorie food >fixation cross). As our paradigms were designed to compare complex events of interest (e.g. pictures of high-calorie foods) with closely matched control conditions (pictures of low-calorie foods), information on onset and duration of fixation cross was not available for some of the tasks (e.g. food receipt paradigms). We conducted the exploratory analyses for the following tasks: monetary reward paradigm (Study 1; fixation cross $n = 44$), food picture paradigm in Study 3 (fixation cross $n = 40$), the model picture paradigm (Study 3; fixation cross $n = 40$), and the food picture paradigm in Study 4 (fixation cross $n = 60$). We calculated test-retest reliability for the fusiform gyrus and MOG, as these regions are implicated in visual processing (22).

In Study 3, we found moderate to good test-retest reliability for the contrasts high-calorie binge foods >fixation cross (*M* ICC = 0.73), low-calorie foods >fixation cross (*M* ICC = 0.73), thin models >fixation cross (*M* ICC = 0.84), and average-weight models >fixation cross (*M* ICC = 0.77) (see **Supplementary Table 1**). In contrast, the ICCs for the 2 contrasted conditions in these regions were low: high-calorie binge foods >low-calorie foods *M* ICC = 0.01 and thin models >average-weight models *M* ICC = 0.20 (Supplementary Table 1). In Study 4, we found moderate to good test-retest reliability for the contrasts high-calorie foods >fixation cross (*M* ICC = 0.75) and low-calorie

TABLE 5 Intrasubject reliability statistics for the food picture paradigm in Study 2 ($n = 109$)[1]

Contrasts	Appetizing food >unappetizing food ICC (95% CI)	Appetizing food >glasses of water ICC (95% CI)
Reference regions (i.e. peak clusters)	L SPL 0.52 (0.35, 0.65)*	L Lingual gyrus 0.56 (0.41, 0.69)*
ROIs:		
L ACC	0.01 (–0.34, 0.29)	0.35 (0.12, 0.54)
R ACC	0.21 (–0.07, 0.44)	0.41 (0.19, 0.57)*
L Amygdala	0.21 (–0.07, 0.44)	0.43 (0.23, 0.59)*
R Amygdala	0.19 (–0.11, 0.42)	0.32 (0.08, 0.52)
L Caudate	–0.00 (–0.36, 0.28)	0.26 (–0.01, 0.47)
R Caudate	0.08 (–0.25, 0.34)	0.17 (–0.12, 0.41)
L Insula	0.16 (–0.14, 0.40)	0.25 (–0.02, 0.46)
R Insula	0.23 (–0.04, 0.45)	0.19 (–0.11, 0.42)
L mOFC	0.12 (–0.20, 0.37)	0.22 (–0.07, 0.44)
R mOFC	0.35 (0.12, 0.53)	0.17 (–0.13, 0.40)
L NAcc	–0.13 (–0.54, 0.19)	0.35 (0.11, 0.53)
R NAcc	–0.29 (–0.75, 0.07)	0.34 (0.10, 0.53)
L Putamen	0.07 (–0.27, 0.33)	0.35 (0.11, 0.53)
R Putamen	0.07 (–0.26, 0.34)	0.32 (0.08, 0.51)
L Thalamus	0.10 (–0.23, 0.35)	0.22 (–0.06, 0.44)
R Thalamus	–0.05 (–0.43, 0.24)	0.01 (–0.35, 0.29)
Average ICC per contrast across all regions	0.11 (–0.21, 0.36)	0.29 (0.03, 0.49)

[1]The intraclass correlation coefficient (ICC) across the 4 time points is reported for each task and for each anatomically defined region of interest (ROI). A negative ICC is interpreted as indicating a reliability of zero (35). *ICC = 0.40–0.750. ACC, anterior cingulate cortex; L, left hemisphere; mOFC, medial orbitofrontal cortex; NAcc, nucleus accumbens; R, right hemisphere; SPL, superior parietal lobe.

foods >fixation cross (*M* ICC = 0.71) (Supplementary Table 1). Similarly, the test-retest reliability for 2 contrasted conditions in these regions was moderate to good: high-calorie foods >low-calorie foods *M* ICC = 0.69 and high-calorie foods >glasses of water *M* ICC = 0.72 (Supplementary Table 1). In Study 1, the ICCs for the contrasts potential winning >fixation cross (*M* ICC = –0.09) and winning >fixation cross (*M* ICC = –0.11) showed low test-retest reliability in these 2 regions (see Supplementary Table 1). The test-retest reliability for the 2 contrasted conditions in these regions was also poor: potential win >reward neutral coin display *M* ICC = 0.15 and winning >reward neutral coin display *M* ICC = 0.38.

Discussion

The results suggest that reliability of the examined fMRI tasks was poor on average, converging with findings from past studies (7, 8, 10), with the extent of reliability varying between regions and contrasts. In Study 1, we found moderate test-retest reliability in the insula, putamen, and thalamus in response to milkshake receipt. We also found moderate test-retest reliability in the putamen in response to monetary reward anticipation and in the caudate, NAcc, and putamen in response to monetary reward. In Study 2, we found moderate test-retest reliability in the strongest activated reference regions (SPL, lingual gyrus), ACC, and amygdala in response to appetizing food images. In Study 4, we found moderate to good test-retest reliability in the strongest activated reference regions (i.e. visual cortex) in response to high-calorie foods. However, most of the peak test-retest reliability estimates did not replicate across the studies that used similar paradigms. For instance, although several moderate to good ICC values emerged for the ROIs for the food picture paradigm used in Study 2, none of these effects replicated in the food picture paradigms in Study 3 and Study 4. Similarly, although moderate to good ICC values emerged for 4 ROIs for the food receipt paradigm in Study 1, none of these effects replicated in the food receipt paradigm in Study 2. Differences in the paradigms across the studies may have contributed to this variability. However, our results are still problematic for the field even if these differences do account for the variability in reliability. Results suggest that the low test-retest reliability of these fMRI tasks may have contributed to the low reproducibility of activation patterns in these tasks and the lack of associations between fMRI data and individual difference measures (e.g. weight gain) reported previously (6, 9).

Exploratory analyses showed that the food picture fMRI tasks in Studies 3 and 4, but not the monetary reward task in Study 1, exhibited moderate to good test-retest reliability in visual processing regions (i.e. MOG, fusiform gyrus) when comparing brain activity during viewing of individual events to fixation cross. In Study 4, we also found moderate to good test-retest reliability in these regions in response to the high-calorie food >low-calorie food and high-calorie food >glasses of water contrasts. Past studies have found higher reliability in visual processing regions compared with regions involved in more complex cognitive processes (37). Further, it has been posited that contrasts against a low-level baseline show better reliability than

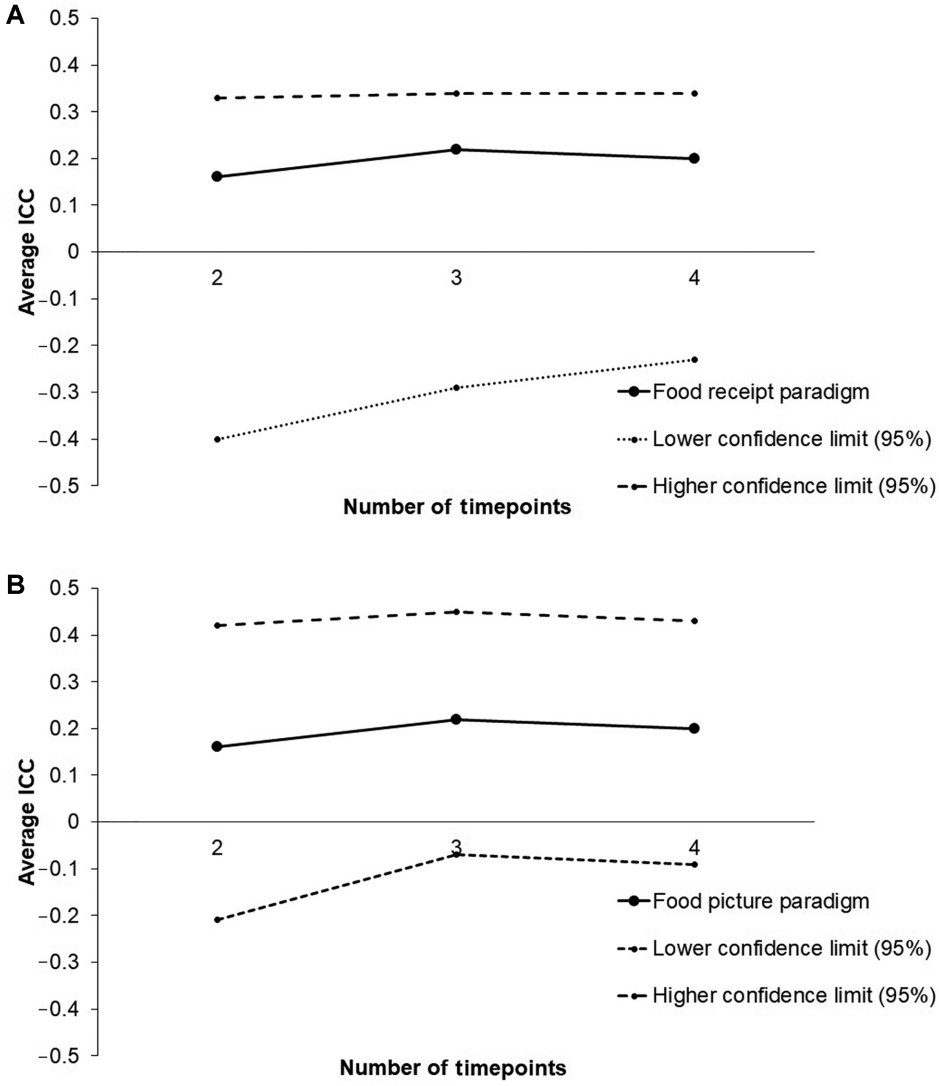

FIGURE 3 Average intraclass correlation coefficient (ICC) across all regions for the food receipt paradigm and food picture paradigm in Study 2 ($n = 109$) across 2 (baseline and 1-y follow-up), 3 (baseline, 1-y follow-up, and 2-y follow-up), and 4 timepoints (baseline, 1-y follow-up, 2-y follow-up, and 3-y follow-up).

contrasts among closely matched conditions because the event of interest and low-level baseline share fewer neurocognitive processes, retaining more variance in the BOLD comparison (5). We see evidence for this in our data. In Study 4, MOG differences in the high-calorie food >fixation cross contrast (baseline and posttest Z-values >8) and the high-calorie food >glasses of water contrast (baseline and posttest Z-values >8) were indeed stronger than those to the high-calorie food >low-calorie food contrast (baseline and posttest Z-values <6.44), which had the lowest reliability among the 3 comparisons. Results suggest that these fMRI paradigms are not inherently unreliable, but that temporal stability of contrasts among the closely matched conditions might be more affected by noise or occasion-specific factors than the contrasts including a low-level baseline.

The poor test-retest reliability observed herein, in conjunction with the poor test-retest reliability for other fMRI paradigms (10), suggests it is important to determine how to increase the test-retest reliability of fMRI paradigms used in eating- and obesity-related studies (38). Design optimization (e.g. optimizing the order and timing of the stimuli) has shown to maximize detection power and estimation efficiency within subjects (39), which may improve test-retest reliability. One study (40) found that block designs and shorter test-retest intervals have higher reliability than event-related designs and longer test-retest intervals. Yet, a recent meta-analysis of the test-retest reliability of fMRI paradigms did not find that reliability correlated with task design, task type, task length, test-retest interval, or population studied (8). Moreover, multiband (the use of multiband excitation pulses to excite and collect multiple slices simultaneously), which increases temporal resolution (41), did not appear to result in higher test-retest reliability in Studies 3 and 4. Friedman and Glover (42) found that increasing the number of functional runs improved the interscanner reliability of their paradigms

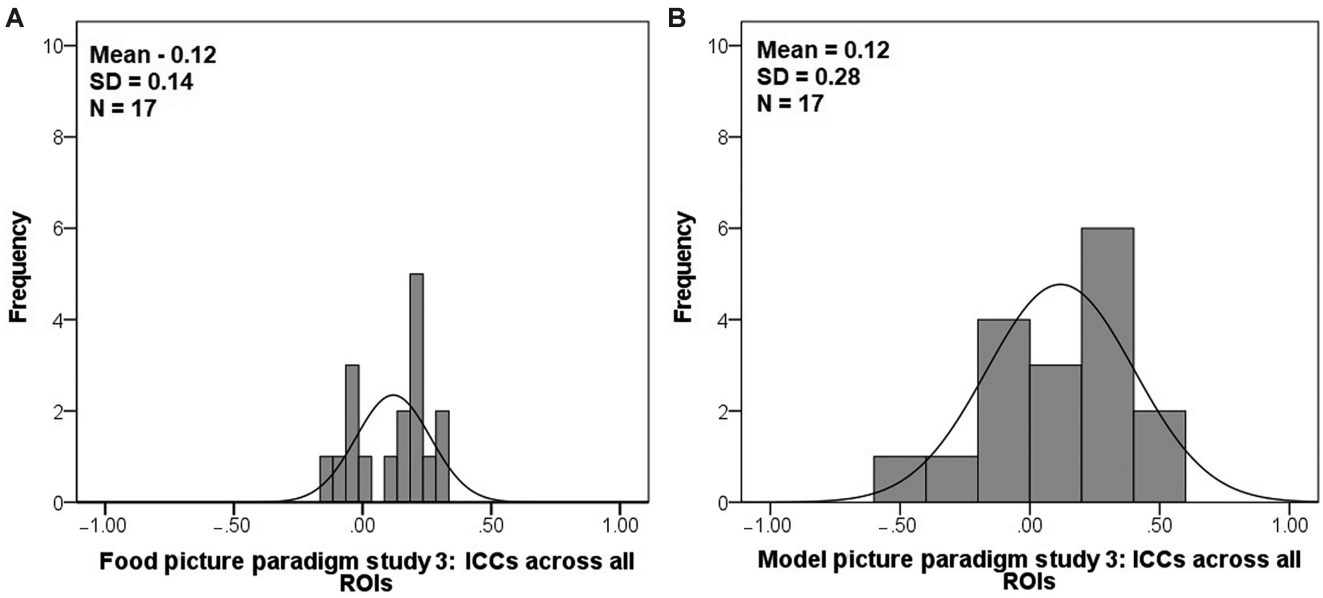

FIGURE 4 Intraclass correlation coefficient (ICC) frequency distributions across all regions 16 anatomical and 1 functional reference region per contrast) for (A) the food picture paradigm (N ICC coefficients = 17) and (B) the model picture paradigm (N ICC coefficients = 17) in Study 3 ($n = 39$). ROI, region of interest.

from an ICC of 0.26 to 0.58 over a 1-d interval, suggesting that this may be a promising way of improving the test-retest reliability of fMRI findings. However, fatigue may increase over time, reducing reliability in the later runs. Research should systematically examine the optimal number of runs to increase reliability of fMRI tasks. Alternative approaches to data analysis might also increase reliability. Similar to past studies (8, 10), we used univariate contrast-based analyses to

TABLE 6 Intrasubject reliability statistics for the food picture paradigm and model picture paradigm in Study 3 ($n = 39$)[1]

Contrasts	Food picture paradigm High-calorie binge foods >low-calorie foods *ICC (95% CI)*	Model picture paradigm Thin models >average-weight models *ICC (95% CI)*
Reference regions (i.e. peak clusters)	L Cuneus 0.24 (–0.45, 0.60)	L Inferior occipital gyrus 0.32 (–0.30, 0.64)
ROIs:		
L ACC	0.23 (–0.47, 0.60)	0.33 (–0.27, 0.65)
R ACC	–0.04 (–0.99, 0.45)	0.17 (–0.59, 0.56)
L Amygdala	–0.14 (–1.18, 0.40)	0.28 (–0.38, 0.62)
R Amygdala	–0.03 (–0.97, 0.46)	–0.06 (–1.02, 0.44)
L Caudate	0.10 (–0.71, 0.53)	0.07 (–0.78, 0.51)
R Caudate	–0.01 (–0.93, 0.47)	–0.28 (–1.45, 0.33)
L Insula	0.19 (–0.55, 0.57)	0.21 (–0.51, 0.59)
R Insula	–0.02 (–0.95, 0.47)	0.38 (–0.17, 0.68)
L mOFC	0.32 (–0.30, 0.64)	0.29 (–0.36, 0.63)
R mOFC	0.33 (–0.28, 0.65)	0.52 (0.08, 0.75)º
L NAcc	–0.08 (–1.05, 0.44)	–0.04 (–0.98, 0.46)
R NAcc	0.21 (–0.50, 0.59)	–0.11 (–1.11, 0.42)
L Putamen	0.19 (–0.54, 0.58)	0.47 (–0.02, 0.72)º
R Putamen	0.15 (–0.62, 0.56)	0.14 (–0.65, 0.55)
L Thalamus	0.21 (–0.50, 0.59)	–0.56 (–1.98, 0.18)
R Thalamus	0.18 (–0.57, 0.57)	–0.12 (–1.14, 0.41)
Average ICC per contrast across all regions	0.12 (–0.68, 0.54)	0.12 (–0.68, 0.54)

[1]The intraclass correlation coefficient (ICC) across the 2 time points is reported for each task and for each anatomically defined region of interest (ROI). Between-day reliability is reported averaging across runs. A negative ICC is interpreted as indicating a reliability of zero (35). ACC, anterior cingulate cortex; L, left hemisphere ; mOFC, medial orbitofrontal cortex; NAcc, nucleus accumbens; R, right hemisphere; º, when excluding 2 influential outliers (parameter estimates exceeding 3 SDs from the mean parameter estimate), the ICC was <0.4.

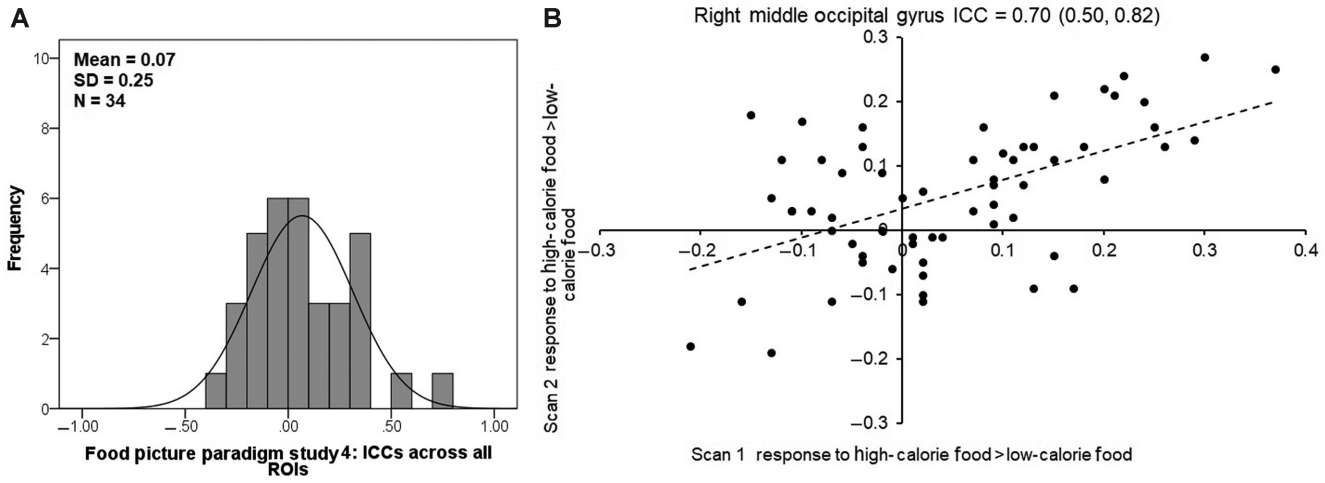

FIGURE 5 Intraclass correlation coefficient (ICC) frequency distributions across all regions 16 anatomical and 1 functional reference region per contrast) for (A) the food picture paradigm in Study 4 ($n = 62$; N ICC coefficients = 34). (B) Distribution of parameter estimates in Study 4 ($n = 62$) for the right middle occipital gyrus in response to the contrast appetizing high-calorie food >appetizing low-calorie food (food picture paradigm). ROI, region of interest.

determine reliability of the fMRI tasks. However, prior studies (8, 43, 44) have shown that other analytic approaches, such as multivoxel pattern analyses, task-based connectivity, or network-specific parameters, increase test-retest reliability. Future research should examine effects of these analytic approaches on reliability of fMRI tasks used in eating and obesity-related studies.

Limitations and future directions

The current research has strengths, including 4 relatively large datasets varying in sample demographics and time intervals and the use of similar paradigms across samples. However, there are also important limitations. First, we conducted analyses in data drawn from fMRI studies that were not designed a priori

TABLE 7 Intrasubject reliability statistics for the food picture paradigm in Study 4 ($n = 62$)[1]

Contrasts	high-calorie foods >low-calorie foods	high-calorie foods >glasses of water
	ICC (95% CI)	*ICC (95% CI)*
Reference regions (i.e. peak clusters)	R Middle occipital gyrus 0.70 (0.50, 0.82)*	L Lingual gyrus 0.56 (0.27, 0.74)*
ROIs:		
L ACC	–0.03 (–0.71, 0.38)	–0.05 (–0.74, 0.37)
R ACC	0.07 (–0.55, 0.44)	–0.08 (–0.81, 0.35)
L Amygdala	0.00 (–0.65, 0.40)	–0.16 (–0.93, 0.30)
R Amygdala	0.08 (–0.53, 0.44)	0.04 (–0.60, 0.42)
L Caudate	–0.24 (–1.06, 0.25)	–0.06 (–0.76, 0.36)
R Caudate	0.03 (–0.62, 0.41)	–0.29 (–1.14, 0.22)
L Insula	–0.33 (–1.21, 0.20)	–0.15 (–0.91, 0.31)
R Insula	0.04 (–0.59, 0.42)	–0.15 (–0.91, 0.31)
L mOFC	0.16 (–0.39, 0.50)	0.29 (–0.19, 0.57)
R mOFC	0.34 (–0.10, 0.60)	0.15 (–0.41, 0.49)
L NAcc	0.29 (–0.18, 0.57)	0.34 (–0.10, 0.60)
R NAcc	0.37 (–0.05, 0.62)	–0.17 (–0.95, 0.29)
L Putamen	0.36 (–0.06, 0.61)	0.30 (–0.16, 0.58)
R Putamen	–0.14 (–0.88, 0.32)	0.15 (–0.41, 0.49)
L Thalamus	0.26 (–0.22, 0.56)	–0.07 (–0.78, 0.36)
R Thalamus	–0.23 (–1.05, 0.26)	–0.07 (–0.78, 0.35)
Average ICC per contrast across all regions	0.10 (–0.49, 0.46)	0.03 (–0.60, 0.42)

[1]The intraclass correlation coefficient (ICC) across the 2 time points is reported for each task and for each anatomically defined region of interest (ROI). Between-day reliability is reported averaging across runs. A negative ICC is interpreted as indicating a reliability of zero (35). *ICC = 0.40–0.75. ACC, anterior cingulate cortex; L, left hemisphere; mOFC, medial orbitofrontal cortex; NAcc, nucleus accumbens, R, right hemisphere.

to investigate test-retest reliability of the fMRI tasks. Additional research is needed to systematically investigate test-retest reliability of the examined fMRI tasks. Second, the test-retest reliability of our fMRI tasks may have been confounded by variation in scan time, scan order, and sample characteristics, such as variation in hunger, motion, cardiac and respiratory signals, endocrine factors, and female menstrual cycle. However, hunger was included as a covariate in the food-related fMRI task analyses, which partially addresses this limitation. Further, there were no significant differences in hunger and in scan time between the baseline and follow-up scans in Studies 3 and 4, suggesting that these variables did not confound those results. Third, Studies 3 and 4 recruited participants for an eating disorder treatment trial and an obesity treatment trial, respectively. Despite the fact that analyses were conducted with participants randomly assigned to the control conditions, it is possible that participants' relation with food and thin-ideal changed from pre to posttest due to regression to the mean because all participants had an eating disorder or were obese at baseline, respectively. Fourth, the test-retest reliability of the fMRI paradigms in the 2 longitudinal adolescent studies (Studies 1 and 2) may have been confounded by age-related development (4) over the 3-y interval, attenuating test-retest reliability. Similar to past developmental fMRI studies (4), we found that reliability of the examined fMRI paradigms varied among regions and contrasts within the tasks. No study has examined test-retest reliability of food receipt and food picture tasks in adolescents over short time intervals. Therefore, it is unknown if the reliability of these tasks would be higher over shorter time intervals. Future research should examine test-retest reliability of fMRI tasks used in eating and obesity-related studies over short and long time intervals in order to advance knowledge regarding how length of time between scans affects reliability of BOLD data. Future research should also compare reliability of the examined fMRI tasks with reliability of behavioral non-fMRI tasks, such as the Food Choice Task (45) which has shown moderate to good test-retest reliability over short time intervals (M ICC 3 d = 0.75; M ICC 1 mo = 0.64). More broadly, behavioral measures such as delay discounting, go/no-go tasks, and stop-signal tasks, have also shown good short- and long-term reliability (46–48).

Conclusions

Overall, the current results suggest that the examined fMRI tasks show poor reliability on average and that the extent of reliability varied strongly between regions and contrasts within the tasks. These findings highlight the importance of repeated-measures fMRI studies, particularly those examining longitudinal effects, to report test-retest reliability of their fMRI paradigms. Further, our findings call for the development and use of well-validated standardized fMRI tasks in eating and obesity-related studies that can provide reliable measures of neural activation.

We would like to thank Paul Novak for his help quality checking the fMRI data. The authors' contributions were as follows—ES and SY: designed and conducted the research; SY: performed statistical analyses; SY, CB, EB, and ES: wrote the manuscript; SY: had primary responsibility for final content; and all authors: read and approved the final manuscript. The authors report no conflicts of interest.

Data Availability

Data described in the manuscript, code book, and analytic code will be made available upon request pending application and approval.

References

1. Stice E, Yokum S. Neural vulnerability factors that increase risk for future weight gain. Psychol Bull 2016;142(5):447–71.
2. Val-Laillet D, Aarts E, Weber B, Ferrari M, Quaresima V, Stoeckel LE, Alonso-Alonso M, Audette M, Malbert CH, Stice E. Neuroimaging and neuromodulation approaches to study eating behavior and prevent and treat eating disorders and obesity. NeuroImage: Clin 2015;8:1–31.
3. Yokum S, Stice E. Weight gain is associated with changes in neural response to palatable food tastes varying in sugar and fat and palatable food images: a repeated-measures fMRI study. Am J Clin Nutr 2019;110(6):1275–86.
4. Herting MM, Gautam P, Chen Z, Mezher A, Vetter NC. Test-retest reliability of longitudinal task-based fMRI: implications for developmental studies. Dev Cogn Neurosci 2018;33:17–26.
5. Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Gerdes AB, Sauer C, Tost H, Esslinger C, et al. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. Neuroimage 2012;60(3):1746–58.
6. Stice E, Yokum S. Relation of neural response to palatable food tastes and images to future weight gain: using bootstrap sampling to examine replicability of neuroimaging findings. Neuroimage 2018;183:522–31.
7. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 2010;1191:1, 133–55.
8. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, Sison ML, Moffitt TE, Caspi A, Hariri AR. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol Sci 2020;31(7):792–806.
9. Yokum S, Gearhardt AN, Stice E. In search of the most reproducible neural vulnerability factors that predict future weight gain: analyses of data from six prospective studies. Soc Cogn Affect Neurosci 2021;nsab013.
10. Drew Sayer R, Tamer GG Jr, Chen N, Tregellas JR, Cornier MA, Kareken DA, Talavage TM, McCrory MA, Campbell WW. Reproducibility assessment of brain responses to visual food stimuli in adults with overweight and obesity. Obesity 2016;24(10):2057–63.
11. Watkins MW, Smith LG. Long-term stability of the Wechsler Intelligence Scale for Children – fourth edition. Psychol Assess 2013;25(2):477–83.
12. Stice E, Burger KS, Yokum S. Reward region responsivity predicts future weight gain and moderating effects of the TaqIA allele. J Neurosci 2015;35(28):10316–24.
13. Stice E, Yokum S. Gain in body fat is associated with increased striatal response to palatable food cues, whereas body fat stability is associated with decreased striatal response. J Neurosci 2016;36(26):6949–56.
14. Stice E, Yokum S, Burger KS, Epstein LH, Small DM. Youth at risk for obesity show greater activation of striatal and somatosensory regions to food. J Neurosci 2011;31(12):4360–6.
15. Stice E, Yokum S, Rohde P, Shaw H, Gau JM, Johnson S, Johns A. Randomized trial of a dissonance-based transdiagnostic group treatment for eating disorders: an evaluation of target engagement. J Consult Clin Psychol 2019;87(9):772–86.
16. Blechert J, Meule A, Busch NA, Ohla K. Food-pics: an image database for experimental research on eating and appetite. Front Psychol 2014;5:617.
17. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26(3):839–51.
18. Parker D, Liu X, Razlighi QR. Optimal slice timing correction and its interaction with fMRI parameters and artifacts. Med Image Anal 2017;35:434–45.
19. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 2003;19(3):1233–9.
20. Neseliler S, Tannenbaum B, Zacchia M, Larcher K, Coulter K, Lamarche M, Marliss EB, Pruessner J, Dagher A. Academic stress and

personality interact to increase the neural response to high-calorie food cues. Appetite 2017;116:306–14.
21. Stoeckel LE, Weller RE, Cook EW 3rd, Twieg DB, Knowlton RC, Cox JE. Widespread reward-system activation in obese women in response to pictures of high-calorie foods. Neuroimage 2008;41(2):636–47.
22. van Meer F, van der Laan LN, Adan RA, Viergever MA, Smeets PA. What you see is what you eat: an ALE meta-analysis of the neural correlates of food viewing in children and adolescents. Neuroimage 2015;104:35–43.
23. Babbs RK, Sun X, Felsted J, Chouinard-Decorte F, Veldhuizen MG, Small DM. Decreased caudate response to milkshake is associated with higher body mass index and greater impulsivity. Physiol Behav 2013;121:103–11.
24. Felsted JA, Ren X, Chouinard-Decorte F, Small DM. Genetically determined differences in brain response to a primary food reward. J Neurosci 2010;30(7):2428–32.
25. Nolan-Poupart S, Veldhuizen MG, Geha P, Small DM. Midbrain response to milkshake correlates with ad libitum milkshake intake in the absence of hunger. Appetite 2013;60(1):168–74.
26. Rudenga KJ, Sinha R, Small DM. Acute stress potentiates brain response to milkshake as a function of body weight and chronic stress. Int J Obes 2013;37(2):309–16.
27. Stice E, Yokum S, Burger K, Epstein L, Smolen A. Multilocus genetic composite reflecting dopamine signaling capacity predicts reward circuitry responsivity. J Neurosci 2012;32(29):10093–100.
28. Elliott R, Newman JL, Longe OA, Deakin JF. Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. J Neurosci 2003;23(1):303–7.
29. Knutson B, Fong GW, Adams CM, Varner JL, Hommer D. Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport 2001;12(17):3683–7.
30. Rademacher L, Krach S, Kohls G, Irmak A, Grunder G, Spreckelmeyer KN. Dissociation of neural networks for anticipation and consumption of monetary and social rewards. Neuroimage 2010;49(4):3276–85.
31. Small DM, Gitelman D, Simmons K, Bloise SM, Parrish T, Mesulam MM. Monetary incentives enhance processing in brain regions mediating top-down control of attention. Cereb Cortex 2005;15(12):1855–65.
32. Brett M, Anton JL, Valabreque R, Poline JB. Region of interest analysis using the MarsBar toolbox [abstract]. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Neuroimage 2002;16(2):S497.
33. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.
34. Fleiss JL. Reliability of Measurement. The Design and Analysis of Clinical Experiments. New York: John Wiley & Sons; 1986. p. 1–32.
35. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19(1):3–11.
36. Keren H, Chen G, Benson B, Ernst M, Leibenluft E, Fox NA, Fox NA, Pine DS, Stringaris A. Is the encoding of Reward Prediction Error reliable during development? Neuroimage 2018;178:266–76.
37. McDermott TJ, Kirlic N, Akeman E, Touthang J, Cosgrove KT, DeVille DC, Clausen AN, White EJ, Kuplicki R, Aupperle RL. Visual cortical regions show sufficient test-retest reliability while salience regions are unreliable during emotional face processing. Neuroimage 2020;220:117077.
38. Smeets PAM, Dagher A, Hare TA, Kullmann S, van der Laan LN, Poldrack RA, Preissl H, Small D, Stice E, Veldhuizen MG. Good practice in food-related neuroimaging. Am J Clin Nutr 2019;109(3):491–503.
39. Durnez J, Blair R, Poldrack R. Neurodesign: optimal experimental designs for task fMRI. BioRxiv 2017(119594):1–21.
40. Bennett CM, Miller MB. fMRI reliability: influences of task and experimental design. Cogn Affect Behav Neurosci 2013;13(4): 690–702.
41. Demetriou L, Kowalczyk OS, Tyson G, Bello T, Newbould RD, Wall MB. A comprehensive evaluation of increasing temporal resolution with multiband-accelerated protocols and effects on statistical outcome measures in fMRI. Neuroimage 2018;176:404–16.
42. Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, Gollub RL, Lauriello J, Lim KO, Cannon T, et al. Test-retest and between-site reliability in a multicenter fMRI study. Hum Brain Mapp 2008;29(8):958–72.
43. Haller SP, Kircanski K, Stoddard J, White LK, Chen G, Sharif-Askary B, Zhang S, Towbin KE, Pine DS, Leibenluft E, et al. Reliability of neural activation and connectivity during implicit face emotion processing in youth. Dev Cogn Neurosci 2018;31:67–73.
44. Noble S, Scheinost D, Constable RT. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. Neuroimage 2019;203:116157.
45. Foerde K, Gianini L, Wang Y, Wu P, Shohamy D, Walsh BT, Steinglass JE. Assessment of test-retest reliability of a food choice task among healthy individuals. Appetite 2018;123:352–6.
46. Anokhin AP, Golosheykin S, Mulligan RC. Long-term test-retest reliability of delayed reward discounting in adolescents. Behav Processes 2015;111:55–9.
47. Hedge C, Powell G, Sumner P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Beha Res Meth 2018;50(3):1166–86.
48. Martinez-Loredo V, Fernandez-Hermida JR, Carballo JL, Fernandez-Artamendi S. Long-term reliability and stability of behavioral measures among adolescents: the Delay Discounting and Stroop tasks. J Adolesc 2017;58:33–9.