Received: 5 March 2024 | Accepted: 18 July 2024

DOI: 10.1002/ejsp.3100

EASP WILEY

RESEARCH ARTICLE

Introducing the brief reverse correlation: An improved tool to assess visual representations

Mathias Schmitz | **Marine Rougier** | **Vincent Yzerbyt**

Institut de recherche en sciences psychologiques, Faculté de psychologie et des sciences de l'éducation, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Correspondence
Mathias Schmitz, Place du Cardinal Mercier 10, B-1348, Louvain-la-Neuve, Belgium.
Email: mathias.schmitz@pucp.pe

Funding information
Fonds de la Recherche Scientifique (FNRS); Mathias Schmitz, Grant/Award Number: 1.A393.17; Marine Rougier, Grant/Award Number: 1.B347.19

Abstract

The reverse correlation is an innovative method to capture visual representations (i.e., classification images, CIs) of social targets. However, this method necessitates many trials to compute high-quality CIs, which poses important practical and economic challenges. We introduce a new version of the reverse correlation method, namely, the Brief-RC. By increasing the number of stimuli (i.e., noisy faces) presented at each trial, the Brief-RC improves the quality of individual (and average) CIs and lowers the overall task length. In two experiments, assessments by external judges confirmed that the new method delivers equally good (Experiment 1) or higher-quality (Experiment 2) outcomes than the traditional method for the same number of trials, time length and number of stimuli. The informational values of CIs were also compared using a more objective metric (infoVal). Because the Brief-RC facilitates the production of higher-quality individual CIs, social psychology researchers may more easily address a series of relevant research questions.

KEYWORDS
Brief-RC, classification images, infoVal, rcirc package, reverse correlation, visual representations

1 | INTRODUCTION

The reverse correlation technique is a method that provides visual proxies of mental representations (Mangini & Biederman, 2004). Specifically, it allows capturing visual representations at a group-level (i.e., from a sample of participants in a condition) as well as at an individual level (i.e., from a single participant). On the one hand, visual representations produced at a group level are of much better quality than those at an individual level (because they rest on a larger number of answers), but they lead to inflated Type I errors (Cone et al., 2020). On the other hand, individual level representations allow for more fine-grained analyses (e.g., correlation with individual level variables), but they require a very large number of trials to achieve good quality outcomes, which entails other issues (i.e., economically costly, time demanding, decreased participants' motivation to complete the task in a conscientious manner; Brinkman et al., 2019b; Todorov et al., 2011). As some researchers have noted, the challenge is now to improve the method to 'generate higher quality outcomes or reduce the number of trials' (Todorov et al., 2011, p. 787). With these concerns in mind, we introduce the Brief Reverse Correlation (Brief-RC). This new method aims to address the issues just mentioned by increasing the number of stimuli at each trial, thereby reducing the overall number of trials and task length while improving the outcome quality.

1.1 | The reverse correlation paradigm

The RC is a method rooted in signal-detection theory that aims to identify the information that underlies perception (Ahumada & Lovell, 1971; Ahumada et al., 1975). Essentially, the method estimates the diagnostic information (i.e., the signal) that drives perception in random variations of the stimulus (Jack & Schyns, 2017). In other words, this technique tries to capture people's expectations about a given target (i.e., the signal, e.g., a happy face in visual perception) by presenting them with noisy stimulus (e.g., a face with random noise added)

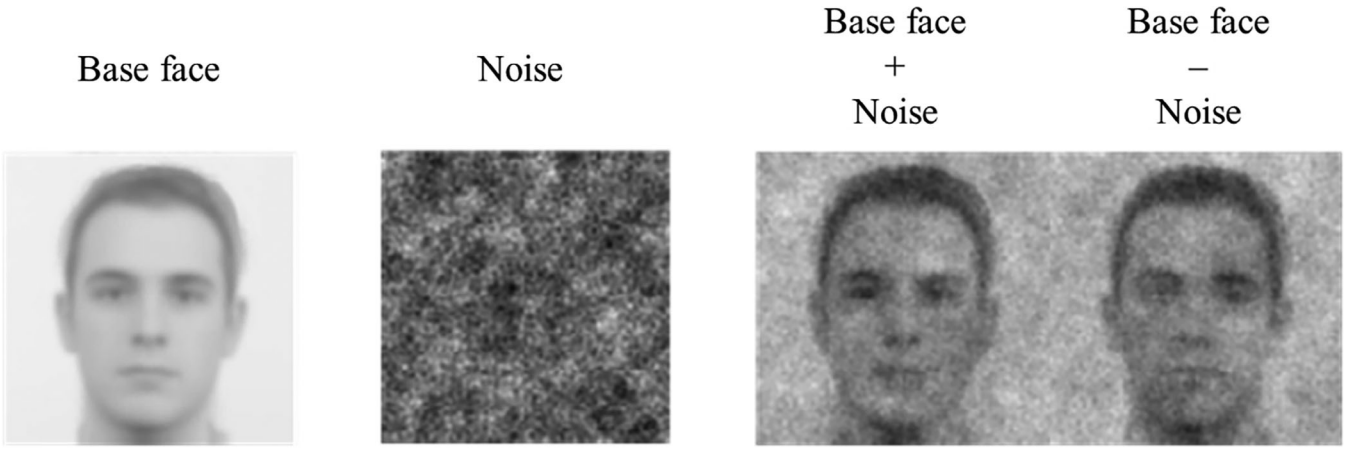

FIGURE 1 Base face (from Experiment 1), example of a random noise pattern, and example of a pair of stimuli (i.e., noisy faces) produced by adding (left) or subtracting (right) the noise pattern from the base face.

and retaining those that, by pure coincidence, happen to match their expectations. For instance, a random noise would likely be selected as a happy face if it slightly distorts the mouth so that it appears smiling (e.g., Kontsevich & Tyler, 2004).

Over the years, the procedure has gained popularity in social psychology and has proven to be particularly useful in identifying the diagnostic features that drive social perception (see Brinkman et al., 2017; Jack & Schyns, 2017; Todorov et al., 2011, 2013). For instance, it contributed to uncover not only facial diagnostic components of social categories such as race, profession and occupation, age and religion but also personality traits and emotions (e.g., Albohn & Adams, 2020; Brooks et al., 2018; Brown-Iannuzzi et al., 2018; Degner et al., 2019; Dotsch et al., 2008; Lloyd et al., 2020; Oliveira et al., 2019; Schmitz et al., 2024). Furthermore, the RC method provides insights into how these visual representations can be distorted by prior knowledge, attitudes, beliefs or behaviours (e.g., Brown-Iannuzzi et al., 2016; Dotsch et al., 2013; Hinzman & Maddox, 2017; Rougier et al., 2021; Young et al., 2013).

Although several implementations of this paradigm exist in the domain of visual perception (see Jack & Schyns, 2017; Mangini & Biederman, 2004; Todorov et al., 2011, 2015), the two-image forced-choice noised-based reverse correlation (hereafter 'Traditional-RC') introduced by Dotsch et al. (2008; see also Dotsch & Todorov, 2011) in social psychology is by far the most prevalent. A typical procedure involves two steps. In the first step, participants choose, across many trials, between two random variations of the same base face (i.e., noisy faces) the one that best matches their mental representation of the category of interest (e.g., 'Choose the most Moroccan-looking face'). The random instantiations consist of superimposing (by either adding or subtracting) a noise pattern (for the generation of the noise pattern, see Mangini & Biederman, 2004) onto a base image (usually a morph of several faces) as illustrated in Figure 1.

Each noisy face in a pair is maximally different from the other of the same pair because one is the mathematical opposite of the other (e.g., the luminance value of a given pixel in one image will be the mathematical opposite of the other). In essence, both faces are equidistant to the base face and any difference in classifications can only stem from the noise pattern (Dotsch & Todorov, 2011). Building the representation of a participant or a group of participants then consists of averaging all selected noise patterns to the base face. This allows obtaining the so-called *classification image* (CI), that is, the visual proxy of the mental representation of the target by a single participant (i.e., participant-level or individual CI) or by a group of participants (i.e., group-level or average CI; for the interpretation of CIs, see Brinkman et al., 2017).

In the second step, a new sample of participants (i.e., the judges) rates the CIs on the variables of interest, allowing a test of the researchers' hypotheses. In one illustrative research, Dotsch et al. (2008) captured the average facial representation of Moroccans from low, moderate and highly prejudiced individuals (as measured with an Implicit Association Task). The authors then asked a sample of independent judges to rate these representations on criminality and trustworthiness. Results revealed that the more prejudiced the individuals, the more negative their visual representation of this group.

The RC procedure offers several advantages to studying psychological effects (see Brinkman et al., 2017; Dotsch & Todorov, 2011; Jack & Schyns, 2017). First, it makes no prior assumption about what people's internal representations may look like. In other words, it allows sampling a vast set of hypotheses (that the researcher may not even have thought of or theoretically anticipated) instead of a single one and doing so agnostically. Second, participants' responses are not primed in any direction, as with pre-selected response labels which could bias their responses (e.g., 'How aggressive are Moroccans?' from 1 = not at all to 7 = totally), which can sometimes lead to different results (e.g., Axt et al., 2023; Michalak & Ackerman, 2021). The RC thus reveals 'near spontaneous use of information' (Brinkman et al., 2017, p. 336) because participants may adopt any criterion they want for the classification task (i.e., the method is said to be unconstrained). It may even allow probing representations that are ineffable for the participant (Mangini & Biederman, 2004).

FIGURE 2 Example of an average CI and some individual CIs issued from the Traditional-RC after 526 trials (from Experiment 2, where participants' task was to select the most 'Chinese-looking face').

1.2 | Current limitations of the Traditional-RC

Most studies based on the Traditional-RC have relied solely on the average CIs instead of individual CIs. The reason is that the former are usually much less noisy than the latter (see Figure 2; Cone et al., 2020; Imhoff et al., 2011, 2013). However, average CIs come with several shortcomings. First, they are valid only to the extent that participants share a common representation of the target category, which is not necessarily the case when the representation is expected to vary (Brinkman et al., 2017; Ratner et al., 2014). Also, judges may detect significant differences between group-wise CIs that may not materialize if the inter-individual variability were to be considered, that is, if judges had to rate the individual CIs. Indeed, Cone et al. (2020) showed that using average CIs in a typical two-phase reverse correlation procedure may considerably increase Type I error rates because they do not consider the inter-individual variability underlying CIs. These authors estimated that about 66% of past research might suffer from inflated false positivity rates. Thus, it is very likely that RC research suffers from replicability issues – although to date we are not aware of any such reports in the literature. Building on this result, they recommend using individual CIs to address this problem (see also the 'subgroup-level' approach; Rougier & De Houwer, 2023; Rougier et al., 2021).

A final and most critical issue from the point of view of social psychological research concerns the fact that average CIs prevent carrying out idiosyncratic analyses (e.g., analysing the relation between racial bias and visual representation bias at the participant-level; Dotsch et al., 2008). Indeed, quite a few questions that occupy social psychologists have to do with relations that may exist between contextual and individual factors. Whether researchers are focusing on the issues of when or for whom a given phenomenon emerges (i.e., the moderation question), or how it comes to materialize (i.e., the mediation question), it is often of prime importance that they are able to collect information on some individual difference of sorts. For instance, if the level of prejudice against a given stigmatized group (e.g., the extent to which people believe Muslim targets are hostile and aggressive, as measured by the faces spontaneously selected with the RC) is believed to play a moderating role on the impact of contextual information (e.g., the tightness of the job market) on discriminatory decisions (e.g., the degree of severity of immigration policies), one would hope to count on individual-level information with respect to prejudice. The same reasoning holds for the test of a mediational approach. Imagine a researcher who wants to examine how derogatory remarks from female colleagues may affect men's organizational commitment via a subjective understanding that feminists are generally less trustworthy and more dominant. Assuming the RC would be selected as a measure of choice to dig into participants' spontaneous perceptions of feminists, such a design would of course require other measures of the underlying mechanisms than the average CIs.

The above considerations emphasize the importance of relying on individual CIs (i.e., they grant access to fine-grained analyses and reduce inflated Type I error rates). Still, because individual CIs rest on substantially fewer responses than average CIs, participants would need to perform a very large number of trials to achieve an acceptable level of quality. Indeed, RC procedures usually comprise 300–1,000 trials per individual-level CI (and thus 300–1,000 trials multiplied by the number of participants within a condition for average-level CIs; Brinkman et al., 2017; Brinkman et al., 2019b; Cone et al., 2020; Dotsch & Todorov, 2011). In theory, adding more trials should enhance the quality of individual CIs. In practice, however, doing so may be detrimental because it hinders participants' motivation to mindfully complete the task (Brinkman et al., 2019b; Todorov et al., 2011; see also Lick et al., 2013). Furthermore, increasing the length of the task can be time-consuming, economically costly or simply impractical. Importantly, if the task is too short to properly cancel out the noise or if participants' responses become erratic (e.g., because of their lack of motivation), the risk is to interpret a noisy CI as being meaningful when it is not. Indeed, the RC method will always yield a CI whatever the number of trials or the meaningfulness of responses (Brinkman et al., 2019b).

To address this issue, Brinkman et al. (2019b) developed infoVal, an information value metric that assesses the degree to which an individual CI derives from signal rather than noise. The infoVal metric comes as a new standard of good practice because it addresses the issue of erroneously interpreting the CI as meaningful when it is not. More specifically, the infoVal score of an individual CI can be interpreted as a

z-score so that when the value is above the threshold of 1.96, the CI can be considered meaningful (i.e., as sufficiently different from random responding). In the same vein, recent work introduced algorithms for ad hoc detection of noisy data (i.e., participants answering randomly after a certain number of trials, participants not complying with experimenter instruction and trials with non-contrasting pairs; Kevane & Koopmann-Holm, 2021). However, although these methods allow identifying (and thus excluding) noisy data, they do not necessarily secure high-quality CIs. A different approach to improving individual CIs could be to consider the response confidence by providing multiple response alternatives (e.g., 'probably happy', 'possibly happy', 'possibly unhappy' or 'probably unhappy') and only build the CIs from trials with high-confidence responses (e.g., Brinkman et al., 2019a; Mangini & Biederman, 2004; see also Dai & Micheyl, 2010). However, this technique implies a contrast between target categories (e.g., happy vs unhappy) that is not necessarily desired as researchers may want to tap into a single target category (e.g., happy) without contrasting it with another one (Dotsch & Todorov, 2011). Furthermore, this strategy fails to shorten task length and may require even more trials if the proportion of high-confidence responses is low (Brinkman et al., 2017). Because no strategy currently allows reducing the task length while simultaneously improving the outcome quality, only a limited number of RC-based studies conducted analyses of individual CIs (e.g., Brooks et al., 2018; Degner et al., 2019; Dotsch et al., 2008, 2013; Imhoff et al., 2013).

In sum, and as of today, there is no practical and reliable solution to improve the quality of individual CIs. This not only reduces the inflated Type I error rates stemming from the use of average CIs but also prevents examining a vast number of questions of interest to social psychologists. This issue becomes even more crucial considering the replicability crisis in psychological science where strong and reliable outcomes should be encouraged (Open Science Collaboration, 2015; Simmons et al., 2011).

1.3 | The brief reverse correlation

In light of these concerns, we propose an improved version of the paradigm, the Brief-RC. The key difference between the Brief-RC and the Traditional-RC is that the former present participants with a larger number of noisy faces (i.e., 12 or 20 instead of two) to select from at each trial.

In the Brief-RC, the likelihood of finding a stimulus that carries diagnostic information of the expected signal at each trial should be higher because the set of options is larger. For instance, if we are searching for a happy-looking face, the higher the number of noisy faces at each trial, the higher the probability that one of them will seem (by chance alone) happy-looking. Conversely, the fewer the noisy faces, the more likely it is that we end up picking one purely at random because none happens to prove close enough to a happy-looking face. A greater panel of faces should therefore improve the signal-to-noise ratio of the selected noisy faces and thus accelerate the convergence towards a higher quality and more robust CI, both at the individual and at the average levels. Surely, the improvement in signal-to-noise ratio due to the increase in face stimuli per trial should reach a plateau because the visual processing capacity of the human brain is limited (e.g., participants should perform poorly with 100 faces per trial; Marois & Ivanoff, 2005; Palmer, 1990). Still, providing more faces per trial should deliver (a) clearer CIs for a given number of trials and (b) high-quality CIs with fewer trials and a shorter completion time. Said otherwise, the same quality-level CI issued from a Traditional-RC may be achieved with a Brief-RC after fewer trials and, thus, a shorter amount of time. In sum, we argue that in comparison to the Traditional-RC, the Brief-RC should improve the quality of the individual CIs without hampering participants' motivation with additional trials.

1.4 | The present research

In two experiments, we compared the quality of individual and average CIs produced from the Traditional-RC with 2 stimuli (i.e., noisy faces) per trial to the Brief-RC involving more than two stimuli: the Brief-RC12, with 12 stimuli per trial, and the Brief-RC20, with 20 stimuli per trial. We assessed the quality of the CIs by means of subjective ratings (from independent judges) and an objective metric (infoVal). To minimize the variability of both the to-be-measured facial representation and the ratings from judges, we opted for a social group – Chinese people – for which there is a relatively clear and homogeneous visual representation among our participants – that is, Americans (e.g., other-race effect; Ge et al., 2009; Kelly et al., 2007; out-group homogeneity effect; Judd & Park, 1988; Lee & Ottati, 1995).

In Experiment 1, we fixed the overall number of stimuli. Because of the improved signal-to-noise ratio, we expected the two variants of the Brief-RC to perform at the same level or better than the Traditional-RC. The decision to rely on the same number of stimuli overall can be seen as rather a conservative comparison between the methods because the number of trials in the Traditional-RC will be much larger than in the Brief-RC for the same number of stimuli presented.

In Experiment 2, we compared the RC variants across different criteria (task length, number of trials and number of stimuli). We expected the Brief-RC (Brief-RC12 and Brief-RC20) to outperform the Traditional-RC when relying on the same task length and number of trials, while we predicted at least as good a performance when holding constant the number of stimuli (as in Experiment 1). We geared our predictions towards individual CIs because they are usually noisier than average CIs. Crucially, if the Brief-RC outperforms the traditional method in enhancing individual CI quality, this would contribute to promoting the use of individual CIs and thus increase replicability in social psychology.

2 | OVERVIEW AND ANALYTICAL STRATEGY

The experiments reported in the present research followed the standard implementation of a two-phase RC paradigm, involving

two distinct samples of participants. In the first part, 'producers' completed one of the three versions of the RC task (i.e., Traditional-RC, Brief-RC12 or Brief-RC20) to capture their visual representation of a Chinese-looking face. In the second part, 'judges' evaluated the obtained CIs on how Chinese they looked. All participants took part online via Prolific Academic (www.prolific.ac) in exchange for a monetary compensation (5£/hour). Participants were from the United States, they did not participate in any of our previous similar studies, and they had an approval rate of at least 95% to ensure data quality (Peer et al., 2013). The present research complied with APA's ethical principles.

Whenever possible, we conducted the analyses using linear mixed models (LMM; Judd et al., 2012, 2017; Westfall et al., 2014). Usually more conservative, a mixed model approach also allows generalizing the results across both 'judges' (judges' unique IDs) and 'producers' (producers' unique IDs or equivalently CIs' unique IDs), our two random factors. Maximal LMM (i.e., the ones that fit the full variance–covariance structure of random effects) usually comes with a significant loss of power and may fail to converge. We report the most parsimonious LMM (i.e., the ones that maximize power while trying to minimize the Type I error rate), following Bates and colleagues' guidelines (Bates et al., 2015; see also Matuschek et al., 2017). The R scripts (available in the Open Science Framework link, see below) detail the model selection process. We report effect sizes for LMM computed with the *r2glmm* package (version 0.1.2, Jaeger, 2017). Of note, LMM effect sizes tend to be considerably smaller than those for by-judges or by-faces analyses because averaging across responses (either across judges or faces) drastically reduces the standard errors (Brysbaert & Stevens, 2018).

We relied on a set of a priori orthogonal contrast codes to compare the performance of the different RC variants. The first contrast, C_1, compares the Traditional-RC (Traditional-RC coded −2/3) to the Brief-RCs versions (Brief-RC12 and Brief-RC20 both coded +1/3), whereas the second contrast, C_2, compares the two Brief-RC variants to each other, that is, Brief-RC12 (coded −0.5) and Brief-RC20 (coded +0.5; with Traditional-RC coded 0).

On some occasions, we predicted an absence of difference between conditions. To better gauge the evidence in favour of the null hypothesis, we report the Bayes factors[1] (BF_{01}) associated with the specific predictor when the OLS or LMM did not yield a significant result (Dienes, 2014). We interpret the Bayes factor according to Jeffreys (1961) guidelines in which a BF_{01} is considered as anecdotal (1–3), substantial (3–10), strong (10–30), very strong (30–100) or decisive (> 100) evidence in support of null hypothesis (H_0) as opposed to the alternative hypothesis (H_1).

To assess the quality of the average and individual CIs, we relied on the infoVal metric developed by Brinkman et al. (2019b; see also Schmitz et al., 2020a, 2020b) that quantifies the degree to which a CI contains signal rather than noise. To improve the accuracy or the metric, we applied an oval-shaped mask to the CIs to extract the face region (e.g., Oliveira et al., 2019; Ratner et al., 2014) and computed the infoVal score on this region.

We pre-registered all experiments on the Open Science Framework (OSF; including a priori theoretical reasoning, hypotheses, power estimations, procedures and statistical analyses). We report any significant deviations from the initial pre-registrations in the core manuscript. We report all measures, manipulations and exclusions. Our pre-registrations, data, data analysis R scripts for all experiments and JavaScript scripts to run both the Traditional-RC and the Brief-RC variants are available on the following link: https://osf.io/ps9wu/?view_only=028597d60e7342bfaeb6881051cb6bca.

3 | EXPERIMENT 1

In Experiment 1, we compared the performance of two variants of the Brief-RC (Brief-RC12 and Brief-RC20) to the Traditional-RC (Traditional-RC). We kept constant the overall number of stimuli (720 noisy faces) across conditions, while the number of trials, stimuli per trial and completion time varied accordingly. We expected the Brief-RC to perform as well or possibly better than the Traditional-RC.

3.1 | Method

3.1.1 | Participants

Although there are no specific good practices regarding how to design an RC experiment (Brinkman et al., 2017), we relied on a sample size that was like a previous study (N = 28 for a three-level between-participants design; Dotsch et al., 2008, Experiment 1). In the first part, we implemented a similar design. To increase statistical power, we recruited a sample of 67 producers (and thus 67 individual CIs). In line with pre-registration, we excluded one participant (from the Traditional-RC condition) due to an abnormally high percentage (48%) of fast reaction times (< 200 ms). The final sample comprised 66 'producers' (n = 22 per condition; M_{age} = 31.86, SD_{age} = 11.52; 36 males). In the second part, we recruited 70 independent judges (M_{age} = 29.79, SD_{age} = 8.52; 33 males and 2 not identifying as either male or female). We based the judges' sample size on previous research that usually relies on 30–90 raters (e.g., Dotsch et al., 2008, 2013; Imhoff et al., 2013). A post hoc sensitivity analysis conducted on PANGEA (https://jakewestfall.shinyapps.io/pangea/) revealed that the current configuration had an 80% power to detect an effect size $\eta_p^2 \geq 0.044$ for the C_1 contrast (opposing the Traditional-RC to the Brief-RC12 and Brief-RC20).

3.1.2 | Experimental design and procedure

Part 1: Reverse correlation task. We randomly assigned producers to one of the three RC tasks (Traditional-RC, Brief-RC12 or Brief-RC20). All participants received the following instruction:

[1] Bayes factor (BF) estimations were derived from the Bayesian information criterion (BIC) with a 'unit information prior' following the guidelines from Wagenmakers (2007).

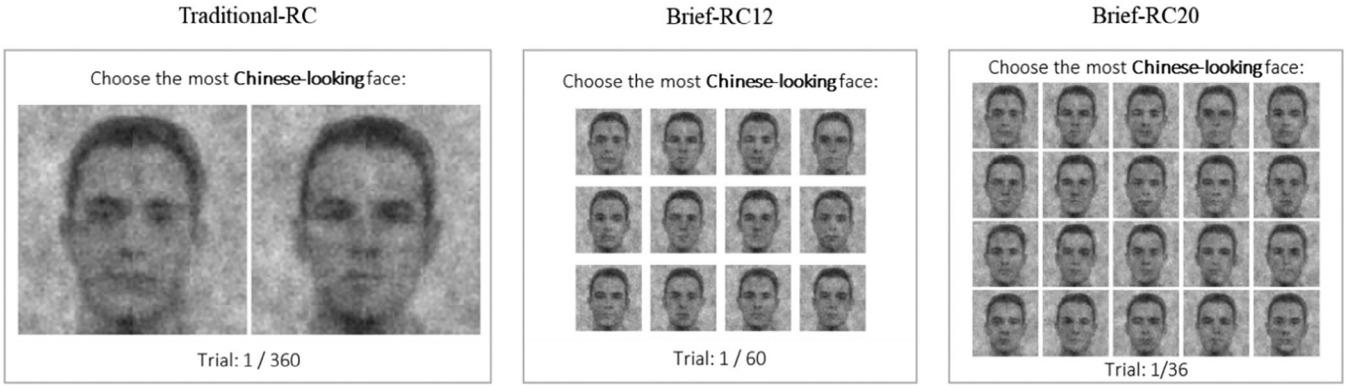

FIGURE 3 Illustration of a single trial in the Traditional-RC, Brief-RC12 and Brief-RC20 tasks. The instruction is displayed at the top of the screen and reads: 'Choose the most Chinese-looking face'. The current and total number of trials are displayed at the bottom of the screen.

> In this task, you will be presented with a series of noisy faces. Your task will be to 'Choose the most Chinese-looking face' in each trial. Use your mouse to make a choice. Note that the faces may look similar to each other, yet they are different. Try to rely on your intuition to make the best choice. Please remain fully concentrated.

The Traditional-RC task comprised 360 trials with two adjacent noisy faces (one oriented and one inverted). The Brief-RC12 comprised 60 trials with 12 noisy faces per trial (6 oriented and 6 inverted) arranged in a 3 × 4 grid. Finally, the Brief-RC20 comprised 36 trials with 20 noisy faces per trial (10 oriented and 10 inverted) arranged in a 4 × 5 grid (see Figure 3). In each condition, we randomly generated a trial order and a stimuli position within each trial and kept it constant across all the participants.

To construct the base image, we selected six Caucasian male faces from the Radboud Face Database (Langner et al., 2010), merged and blurred them (using a low spatial frequency filter) into a single black and white image using Psychomorph (Tiddeman et al., 2005). With the base image, we then generated 360 pairs of stimuli (i.e., noisy faces) using the *rcicr* R package (development version)[2] with the default settings. Pairs consisted of either adding ('oriented' image) or subtracting ('inverted' image) a sinusoidal noise pattern from the base image (see Mangini & Biederman, 2004). All stimuli initially had a 512 × 512 pixels size and were rescaled to 150 × 150 pixels size in the case of the Brief-RC12 and Brief-RC20 so they could fit into the window screen, while they kept their original size in the Traditional-RC. We built the reverse correlation tasks with a JavaScript library called jsPsych (www.jspsych.org; de Leeuw, 2014; version 6.0.3). At the end of the RC procedure, participants provided demographic information through a Qualtrics survey before being debriefed.

After collecting the RC data, we computed each participant's individual CI by averaging all the noise patterns extracted from the selected noisy faces and superimposing them on the base image.[3] The CIs were created with a constant scaling that was different as a function of the type of CI (individual vs average) and the type of condition (Traditional-RC vs Brief-RC12 vs Brief-RC20). We opted for a constant scaling because the noise range differed between conditions and levels of aggregation (average vs individual). The selected constants[4] were those minimizing the perceptual difference (based on visual inspection) in noise level in the CIs between conditions. Finally, we computed the *infoVal* metric for the individual CIs.

Part 2: CIs rating task. A new sample of participants (judges) rated the 66 individual CIs and the 3 average CIs via a Qualtrics survey. Judges were told that they were about to rate several blurred faces by means of a 7-point scale based on how Chinese-looking the faces were (from 0 = *neutral* to 6 = *extremely Chinese-looking*).[5] Before the rating task, all individual CIs were presented at once (through a matrix grid) for one minute to allow the judges to gauge the similarities and differences between them. Then, judges rated in a first block the randomly ordered individual CIs one by one. In a second block, they rated the three average CIs presented simultaneously in a matrix with one row per CI (CI position was randomized between participants). Next, judges provided demographic information, had the option to leave a comment about the study and were debriefed.

3.2 | Results

3.2.1 | Task completion time

We submitted the task completion time (in minutes) to our two contrasts of interest C_1 (opposing Traditional-RC to Brief-RC12 and

[2] We used the development version to obtain the noise matrices (for more information, see https://github.com/rdotsch/rcicr), which ease the computation of the CIs and infoVal.

[3] We adapted the functions 'generateStimuli2IFC' and 'genCI' from the *rcicr* package (Dotsch, 2017). These functions are used to compute the noisy faces (oriented and inverted pictures) and to compute the CIs, respectively. The adapted R scripts can be found on the OSF repository.

[4] The following constants were used to scale the noise patterns for the average CIs: Traditional-RC = 0.0055, Brief-RC12 = 0.0130, Brief-RC20 = 0.0150; and for the individual CIs: Traditional-RC = 0.0190, Brief-RC12 = 0.0450, Brief-RC20 = 0.0600.

[5] We slightly deviated from the pre-register scale values (-3 = *not at all Chinese-looking* to +3 = *very Chinese-looking*) to make the scale more meaningful to participants as in Brinkman et al. (2019b; 1 = 'not masculine', 9 = 'very masculine').

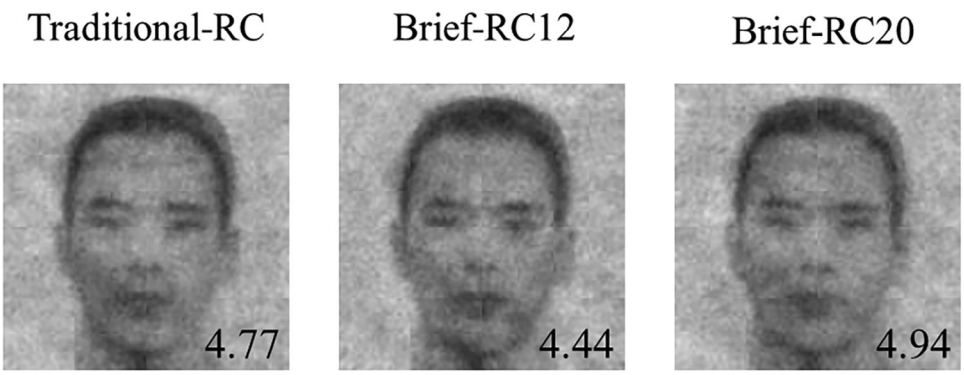

FIGURE 4 Average CI ratings by condition with the average ratings displayed in the lower-right corner of the CIs.

Brief-RC20) and C_2 (comparing Brief-RC12 to Brief-RC20) in an OLS analysis. As expected, the task completion time was significantly longer – almost twice as long – for the Traditional-RC ($M = 10.81$, $Mdn = 10.74$, $SD = 4.55$) than for the Brief-RC12 and Brief-RC20 ($M = 5.72$, $Mdn = 5.45$, $SD = 3.49$), $F(1, 67) = 25.22$, $p < .001$, $\eta_p^2 = 0.286$. Interestingly, the completion time did not significantly differ between the Brief-RC12 ($M = 6.20$, $Mdn = 5.38$, $SD = 3.70$) and Brief-RC20 ($M = 5.23$, $Mdn = 5.45$, $SD = 3.28$), $F(1, 67) = 0.69$, $p = .410$, $\eta_p^2 = 0.011$, $BF_{01} = 5.67$.

3.2.2 | Judges' ratings

We submitted judges' ratings on the Chinese-looking scale of the individual CIs to C_1 and C_2 in an LMM analysis, using judges and producers as random factors. Regardless of the condition, ratings were slightly, yet significantly, above the middle-scale point (i.e., value '3'; $M = 3.31$, $SE = 0.15$), $F(1, 192.22) = 3.91$, $p = .049$.[6] The individual CI ratings did not differ significantly between the Traditional-RC ($M = 3.13$, $SD = 1.83$) and the Brief-RC12 and Brief-RC20 ($M = 3.39$, $SD = 1.82$), $F(1, 65.81) = 1.13$, $p = .292$, $\eta_p^2 = 0.004$, $BF_{01} = 38.86$, nor between the Brief-RC12 ($M = 3.46$, $SD = 1.78$) and the Brief-RC20 ($M = 3.32$, $SD = 1.86$), $F(1, 65.81) = 0.25$, $p = .616$, $\eta_p^2 = 0.001$, $BF_{01} = 59.89$.

Next, we submitted judges' ratings of the average CIs to C_1 and C_2 in an OLS analysis. Regardless of the condition, ratings were significantly above the middle-scale point (i.e., value '3'; $M = 4.72$, $SE = 0.09$), $F(1, 67) = 338.60$, $p < .001$, $\eta_p^2 = 0.621$. The average CI ratings did not significantly differ between the Traditional-RC ($M = 4.77$, $SD = 1.08$) and the Brief-RC12 and Brief-RC20 ($M = 4.69$, $SD = 1.49$), $F(1, 67) = 0.16$, $p = .692$, $\eta_p^2 = 0.001$, $BF_{01} = 13.38$. However, the average CI ratings of the Brief-RC12 ($M = 4.44$, $SD = 1.53$) were significantly lower than those of the Brief-RC20 ($M = 4.94$, $SD = 1.41$), $F(1, 67) = 4.77$, $p = .030$, $\eta_p^2 = 0.023$ (see Figure 4).

[6] Computation of the effect size for the intercept is not available in the r2glmm R package (for mixed-models).

3.2.3 | InfoVal

We submitted the infoVal[7] scores of the individual CIs to C_1 and C_2 in an OLS analysis. There were no significant differences between the Traditional-RC ($M = 1.54$, $SD = 2.08$) and the Brief-RC12 and Brief-RC20 ($M = 1.83$, $SD = 1.80$), $F(1, 67) = 0.34$, $p = .560$, $\eta_p^2 = 0.005$, $BF_{01} = 6.79$, nor between the Brief-RC12 ($M = 1.93$, $SD = 1.74$) and the Brief-RC20 ($M = 1.74$, $SD = 1.90$), $F(1, 67) = 0.10$, $p = .748$, $\eta_p^2 = 0.002$, $BF_{01} = 7.69$. In all conditions, infoVal scores were relatively low as all of them were below the threshold of 1.96 (see Brinkman et al., 2019a).

3.3 | Discussion

In Experiment 1, we compared the visual representation of a Chinese-looking face as captured by the Brief-RC (Brief-RC12 and Brief-RC20) versus the Traditional-RC with the same number of stimuli (720 noisy faces). Neither the individual nor the average CI ratings from judges (on how Chinese they looked like) differed significantly between the various methods. The same pattern emerged when relying on a more objective metric, namely, the infoVal score of individual CIs. Overall, frequentist as well as Bayesian analyses of our data confirm that the Brief-RC and the Traditional-RC delivered similar outcomes. A crucial difference, however, is that fixing the number of stimuli as we did here resulted in the fact that the Brief-RC took half the time than the Traditional-RC. A researcher with limited resources would thus have an obvious interest in opting for the Brief-RC rather than the Traditional-RC.

One caveat of Experiment 1 concerns the selected number of stimuli (720 noisy faces). Indeed, RC-based studies generally rely on more than 360 trials. This means that one would expect to see more stimuli presented to the participants in the first part of the experiment. Moreover, one could argue that the comparison between methods should also rest on such criteria as completion time and number of trials rather

[7] We did not declare a priori hypothesis concerning infoVal for Experiment 1 in the pre-register since the reference for this metric (Brinkman et al., 2019b) was not available at that time. Nonetheless, we decided to include these analyses to better gauge the differences between the Brief-RC and Traditional-RC.

than number of stimuli alone because these factors affect participants' motivation and attention, as well as the time and economic resources from the researchers. Experiment 2 increased the length of the task and relied on completion time, number of trials and number of stimuli as comparison criteria.

4 | EXPERIMENT 2

In Experiment 2, we sought to replicate and extend the findings from Experiment 1 by comparing the performance of the Brief-RC (Brief-RC12 and Brief-RC20) versus the Traditional-RC not only after the same number of stimuli (and thus for different points in time and numbers of trials) but also after the same time (and thus for different numbers of trials and stimuli) and after the same number of trials (and thus for different points in time and numbers of stimuli). We expected the Brief-RC variants to outperform the Traditional-RC when comparing ratings after the same time and number of trials. Regarding the number of stimuli, we expected to replicate the findings from Experiment 1. To increase statistical power, we increased the sample sizes of both producers and judges.

4.1 | Method

4.1.1 | Participants

Because our design involved LMM, with producers (i.e., individual CIs) and judges as random factors, we relied on Judd et al. (2012, 2017) and Westfall et al. (2014) to select the number of producers and judges. Accordingly, we needed about 60 producers per condition and 150 judges to achieve a power of .80 to detect an effect of $d = 0.20$. To maximize power, we recruited a sample of 300 producers ($n = 100$ per condition; $M_{age} = 31.30$, $SD_{age} = 10.35$; 169 males, 4 others) and another sample of 253 independent judges. We excluded one judge who gave the same rating to all faces.[8] The final sample thus comprised 252 judges ($M_{age} = 30.50$, $SD_{age} = 10.04$; 126 females, 3 others). A post hoc sensitivity analysis revealed that the current configuration had an 80% power to detect an effect size $\eta_p^2 \geq 0.013$ for the C_1 contrast (opposing the Traditional-RC to the Brief-RC12 and Brief-RC20).

4.1.2 | Experimental design and procedure

The overall procedure was as in Experiment 1 with a few differences detailed below.

Part 1: Reverse correlation task. We estimated the number of trials for each RC task from Experiment 1 such that the median completion time length for the RC task in all three conditions would be approximately 10 minutes. We created 2000 pairs of noisy faces (stimuli) from a different base image (the same as in Dotsch et al., 2008) and a different seed than Experiment 1 for generalizability purposes. The Traditional-RC comprised 550 trials (2 noisy faces per trial; 1100 noisy faces in total), the Brief-RC12 comprised 250 trials (12 noisy faces per trial; 3000 noisy faces in total), and the Brief-RC20 comprised 200 trials (20 noisy faces per trial; 4000 in total). Because the total number of stimuli varied from one condition to another, we presented the noisy faces in the same order and kept the order of the trials constant between conditions. For instance, the two noisy faces in the first trial of the Traditional-RC were also part of the first trial of the Brief-RC12 and Brief-RC20.

Next, participants answered two questions concerning their perceived threat (2 items) and attitude (1 item) regarding Chinese people. We do not report about these measures here as they were used to inform a separate project. Producers then provided their feedback about how boring (slider from 0 = *not boring at all* to 100 = *very boring*) and difficult (slider from 0 = *very easy* to 100 = *very difficult*) they found the task. Average ratings on these two measures are reported in Supplementary Table S3. Finally, they provided the same demographic information, had the option to leave a comment about the study and were debriefed.

We computed the individual and average CIs within each task at different points according to the criteria of comparison, namely, time (approximately 5 and 10 minutes), number of trials (90 and 167) and number of stimuli presented (approximately 1050). Table 1 presents the 11 distinct points of comparison.[9] The time (5 and 10 minutes) criterion in each RC condition was applied on the trial that, on average over participants, was approximately completed after 5 or 10 minutes.

We created the individual and average CIs with the auto-scale method that matches the noise pattern to the range of pixels of the base image. Although this method is suboptimal, it is common in RC experiments because it prevents from selecting an arbitrary constant (called 'constant scaling') that may introduce a bias (see Brinkman et al., 2017; and the documentation from the *rcicr* package at https://rdrr.io/cran/rcicr/man/generateCI.html). We computed the infoVal as in Experiment 1.

Part 2: CIs rating task. Judges first saw a random sample of 55 individual CIs (5 CIs randomly sampled from each of the 11 distinct points of comparison; see Table 1) to better gauge the differences between CIs. These illustrative CIs appeared as a grid for one minute on the same page in a random fixed order. Judges then rated 88 individual CIs (8 individual CIs randomly sampled from each point of comparison). We presented individual CIs in a matrix (each row corresponding to one CI) with the row order randomized across participants. Next, judges saw the 11 average CIs corresponding to the 11 points of comparison for 20 seconds before rating each of them. In the rating task, average CIs appeared one by one and in random order across participants. Next,

[8] We pre-registered that judges with less than 5% variation on their ratings or/and with a median rating time of less than 1 second per stimulus would be excluded. However, the median rating time per stimulus was 0.5 seconds, and 94.44% of judges had rating durations shorter than 1 second/stimulus. Given that this exclusion criterion was underestimated, we omitted it.

[9] We deviated slightly from the pre-registered points of comparison to minimize the number of CIs and judgements required. To do so, we selected points that could be used to compare more than one criterion (see Table 1). These points were determined *before* computing the visual renderings.

TABLE 1 Points of comparison by time, number of trials and number of stimuli as a function of condition (Traditional-RC, Brief-RC12 and Brief-RC20).

Condition	Time	Number of trials	Number of stimuli	Points of comparison		
				By time	By trials	By stimuli
Traditional-RC	2.34	90	180		90 trials	
	4.19	167	334		167 trials	
	5.00	203	406	~5 min		
	9.99	526	1052	~10 min		~1050 stimuli
Brief-RC12	5.01	90	1080	~5 min	90 trials	~1050 stimuli
	7.73	167	2004		167 trials	
	9.99	240	2880	~10 min		
Brief-RC20	4.42	51	1020			~1050 stimuli
	5.02	62	1240	~5 min		
	6.56	90	1800		90 trials	
	10.05	167	3340	~10 min	167 trials	

judges provided demographic information, had the possibility to leave a comment about the study and were debriefed.

4.2 | Results

4.2.1 | Judges' ratings as a function of time

We submitted judges' ratings of the individual CIs to an LMM analysis with C_1 (Traditional-RC vs Brief-RC12 and Brief-RC20), C_2 (Brief-RC12 vs Brief-RC20), time (5 vs 10 minutes) and their interactions as predictors, using judges and producers as random factors. As predicted, C_1 was significant such that the individual CIs from the Traditional-RC ($M = 2.28$, $SD = 1.95$) looked less Chinese than those from the Brief-RC12 and Brief-RC20 ($M = 2.81$, $SD = 2.06$), $F(1, 319.34) = 13.71$, $p < .001$, $\eta_p^2 = 0.010$. As expected, there was a main effect of time such that the individual CIs after 5 minutes ($M = 2.56$, $SD = 1.99$) looked significantly less Chinese than those after 10 minutes ($M = 2.72$, $SD = 2.08$), $F(1, 247.84) = 19.14$, $p < .001$, $\eta_p^2 = 0.001$. Moreover, the $C_1 \times$ time interaction was significant, $F(1, 247.44) = 5.73$, $p = .017$, $\eta_p^2 < 0.001$. Specifically, individual CIs from the Traditional-RC ($M = 2.25$, $SD = 1.90$) looked less Chinese than those from the Brief-RC12 and Brief-RC20 ($M = 2.71$, $SD = 2.02$) after 5 minutes, $F(1, 341.57) = 9.27$, $p = .003$, $\eta_p^2 = 0.005$, this pattern being even more pronounced after 10 minutes ($M_{\text{traditional-RC}} = 2.31$, $SD_{\text{traditional-RC}} = 2.00$; $M_{\text{Brief-RC12\&Brief-RC20}} = 2.92$, $SD_{\text{Brief-RC12\&Brief-RC20}} = 2.10$), $F(1, 336.07) = 17.99$, $p < .001$, $\eta_p^2 = 0.010$. There was no significant effect of C_2 nor of $C_2 \times$ time (all p values $> .10$, $BF_{01} > 10$).

Next, we submitted judges' ratings of the average CIs to an OLS analysis with C_1, C_2, time and their interaction as predictors. As expected, there was a significant main effect of C_1 such that the average CIs looked less Chinese for the Traditional-RC ($M = 4.55$, $SD = 1.38$) than for the Brief-RC12 and Brief-RC20 ($M = 4.85$, $SD = 1.21$), $F(1, 246) = 19.74$, $p < .001$, $\eta_p^2 = 0.013$. There was no significant effect of C_2, time, $C_1 \times$ time nor $C_2 \times$ time (all p values $> .10$, BF_{01}'s > 10).

Figure 5(A) shows the results from judges' ratings of the individual and average ratings of CIs as a function of time. Figure 6 illustrates the average CIs as a function of time and their ratings.

4.2.2 | Judges' ratings as a function of the number of trials

We submitted judges' ratings of the individual CIs to an LMM analysis with C_1, C_2, trials (90 vs 167 trials) and their interaction as predictors, using judges and producers as random factors. As predicted, there was a main effect of C_1 such that the individual CIs from the Traditional-RC ($M = 2.15$, $SD = 1.87$) looked significantly less Chinese than those from the Brief-RC12 and Brief-RC20 ($M = 2.75$, $SD = 2.05$), $F(1, 339.58) = 20.19$, $p < .001$, $\eta_p^2 = 0.020$. The individual CIs after 90 trials ($M = 2.50$, $SD = 1.98$) looked significantly less Chinese than those after 167 trials ($M = 2.60$, $SD = 2.04$), $F(1, 246.43) = 8.12$, $p = .005$, $\eta_p^2 = 0.001$. Moreover, there was a significant $C_2 \times$ trials interaction $F(1, 245.72) = 4.67$, $p = .032$, $\eta_p^2 < 0.001$. Although the simple effects were not significant, the individual CIs from the Brief-RC12 ($M = 2.77$, $SD = 2.01$) looked descriptively more Chinese than those from the Brief-RC20 ($M = 2.66$, $SD = 2.04$) at 90 trials, $F(1, 339.22) = 0.42$, $p = .519$, $\eta_p^2 < 0.001$, whereas the individual CIs from the Brief-RC12 ($M = 2.75$, $SD = 2.01$) looked less Chinese than those from the Brief-RC20 ($M = 2.83$, $SD = 2.14$) at 167 trials, $F(1, 338.13) = 0.34$, $p = .558$, $\eta_p^2 < 0.001$. There were no significant effects of C_2 nor of $C_1 \times$ trials (all p values $> .10$, $BF_{01} > 10$).

Next, we submitted judges' ratings of the average CIs to an OLS analysis with C_1, C_2, trials and their interaction as predictors. The predicted main effect of C_1 was significant such that the average CIs looked less Chinese for Traditional-RC ($M = 4.28$, $SD = 1.44$) than for Brief-RC12 and Brief-RC20 ($M = 4.83$, $SD = 1.20$), $F(1, 246) = 60.38$,

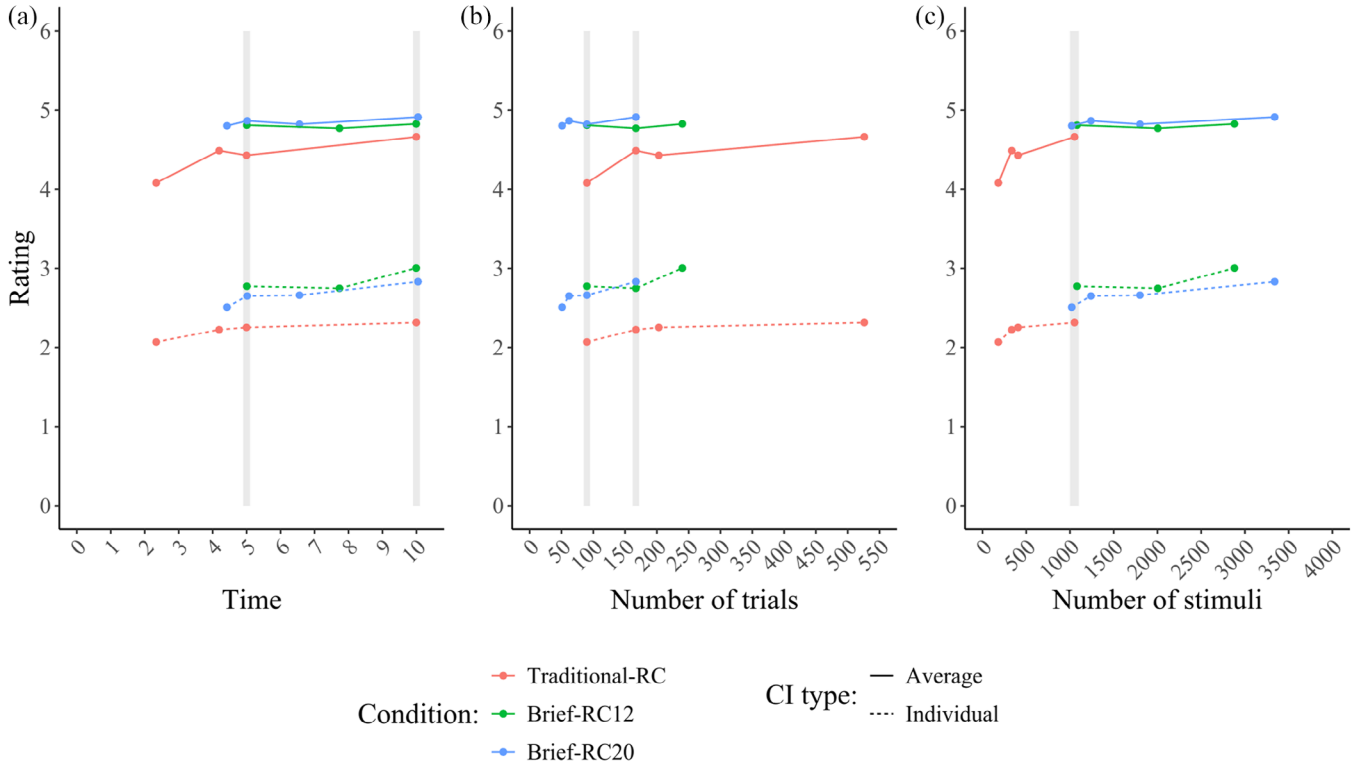

FIGURE 5 Relation between individual CIs ratings (averaged) and (A) time (median cumulative time in minutes), (B) number of trials and (C) number of stimuli, as a function of condition (different coloured lines: Traditional-RC, Brief-RC12 and Brief-RC20), and CI type (solid lines represent average CIs and the dashed lines individual CIs). The dots represent average ratings at given points of comparison (cf. Table 1). The grey frames highlight the points of comparison, that is, (A) at approximately 5 and 10 minutes, (B) at 90 and 167 trials and (C) at approximately 1050 stimuli.

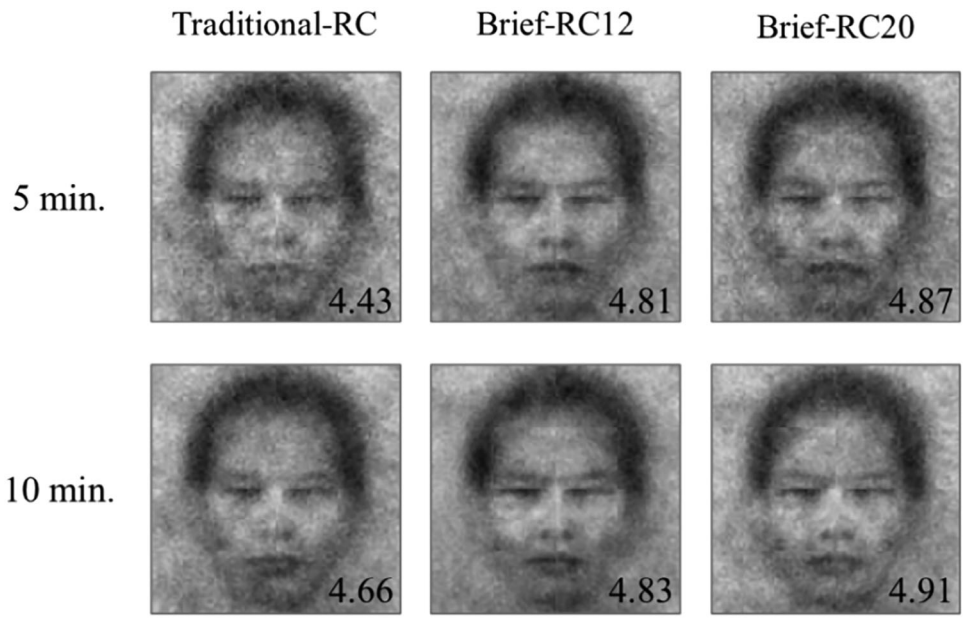

FIGURE 6 Average CIs by condition and time (5 and 10 minutes) with the average ratings displayed in the lower-right corner of the CIs.

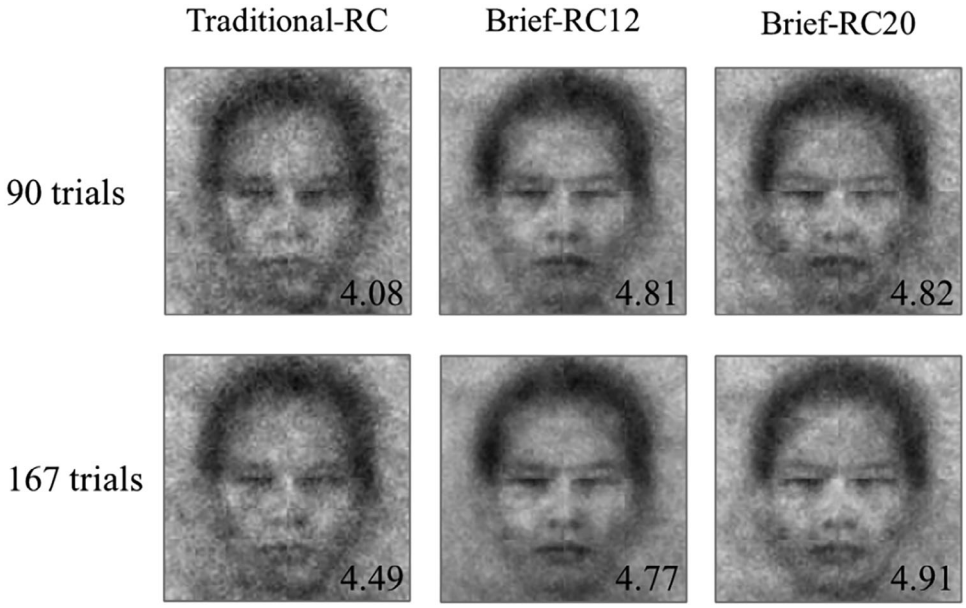

FIGURE 7 Average CIs by condition and trial (90 and 167 trials). Average ratings are displayed in the lower-right corner of the CIs.

$p < .001$, $\eta_p^2 = 0.039$. There was also a main effect of trials such that the average CIs looked significantly less Chinese after 90 trials ($M = 4.57$, $SD = 1.34$) than after 167 trials ($M = 4.72$, $SD = 1.28$), $F(1, 246) = 5.39$, $p = .020$, $\eta_p^2 = 0.004$. Moreover, there was a significant $C_1 \times$ trials interaction $F(1, 246) = 7.46$, $p = .007$, $\eta_p^2 = 0.005$. Specifically, the average CIs from the Traditional-RC ($M = 4.08$, $SD = 1.50$) looked significantly less Chinese than the Brief-RC12 and Brief-RC20 ($M = 4.82$, $SD = 1.19$) after 90 trials, $F(1, 246) = 55.15$, $p < 0.001$, $\eta_p^2 = 0.035$, this in a more pronounced manner than at 167 trials ($M_{\text{traditional-RC}} = 4.49$, $SD_{\text{traditional-RC}} = 1.36$; $M_{\text{brief-RC12\&brief-RC20}} = 4.84$, $SD_{\text{brief-RC12\&brief-RC20}} = 1.22$), $F(1, 246) = 12.69$, $p < .001$, $\eta_p^2 = 0.008$. There were no significant effects of C_2, nor $C_2 \times$ trials (all p values $> .10$, $BF_{01} > 10$).

Figure 5(B) shows the results from judges' ratings of the individual and average ratings of CIs as a function of the number of trials. Figure 7 illustrates the average CIs as a function of trials.

4.2.3 | Judges' ratings as a function of the number of stimuli

We submitted judges' ratings of the individual CIs produced after producers had been exposed to approximately 1050 stimuli (noisy faces) to an LMM analysis with C_1 and C_2 as predictors, using judges and producers as random factors. There was a main effect of C_1 such that the individual CIs from the Traditional-RC ($M = 2.31$, $SD = 2.00$) looked significantly less Chinese than those of the Brief-RC12 and Brief-RC20 ($M = 2.64$, $SD = 2.02$), $F(1, 331.36) = 4.78$, $p = .029$, $\eta_p^2 = 0.005$. There was no significant effect of C_2, $F(1, 326.76) = 2.52$, $p = .113$, $\eta_p^2 = .003$, $BF_{01} = 22.12$.

Next, we submitted judges' ratings of the average CIs to an OLS analysis with C_1 and C_2 as predictors. The significant main effect of C_1 confirmed the average CIs looked less Chinese for Traditional-RC ($M = 4.66$, $SD = 1.29$) than for the Brief-RC12 and Brief-RC20 ($M = 4.81$, $SD = 1.18$), $F(1, 249) = 59.96$, $p < .001$, $\eta_p^2 = 0.038$. There was no significant effect of C_2, $F(1, 249) = 0.91$, $p = .341$, $\eta_p^2 = 0.001$, $BF_{01} = 24.68$.

Figure 5(C) shows the results from judges' ratings of the individual and average ratings of CIs as a function of the number of stimuli. Figure 8 illustrates the average CIs and their ratings. Means and standard deviations of the ratings of individual and average CIs for each of the experimental cells are available in Supplementary Table S1. When considering all stimuli and regardless of the condition, ratings of individual CIs were slightly, but significantly, below the middle-scale point ($M = 2.53$, $SE = 0.09$), $F(1, 506.72) = 27.14$, $p < .001$. For average CIs, the average rating was significantly above the middle-scale point ($M = 4.65$, $SE = 0.03$) $F(1, 249) = 2467.00$, $p < .001$, $\eta_p^2 = 0.620$.

4.2.4 | InfoVal as a function of time

We submitted the infoVal scores of the individual CIs to an OLS analysis with C_1, C_2, time (5 vs 10 minutes) and their interactions as predictors in an OLS analysis. As a reminder, the higher infoVal, the more likely the individual CI was generated from meaningful (vs random) responses. As for C_1, there was no difference between the infoVal scores from the Traditional-RC ($M = 0.98$, $SD = 1.66$) and those from the Brief-RC12 and Brief-RC20 ($M = 1.09$, $SD = 1.79$), $F(1, 294) = 0.61$, $p = 0.434$, $\eta_p^2 = 0.001$, $BF_{01} = 17.99$. There was a significant time effect, such that the infoVal scores after 5 minutes ($M = 0.77$, $SD = 1.43$) were lower than those after 10 minutes ($M = 1.35$, $SD = 1.98$), $F(1, 294) = 16.89$, $p < .001$, $\eta_p^2 = 0.028$. There was no significant effect of C_2, $C_1 \times$ time, or $C_2 \times$ time (all $p > .10$, $BF_{01} > 10$). Figure 9(A) shows the averaged infoVal scores of individual CIs as a function of time.

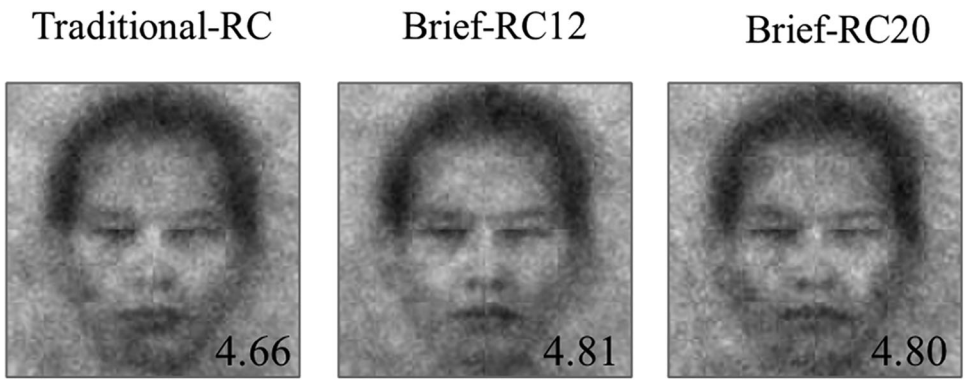

FIGURE 8 Average CIs produced after producers were exposed to approximately 1050 stimuli (noisy faces) with the average ratings displayed in the lower-right corner of the CIs.

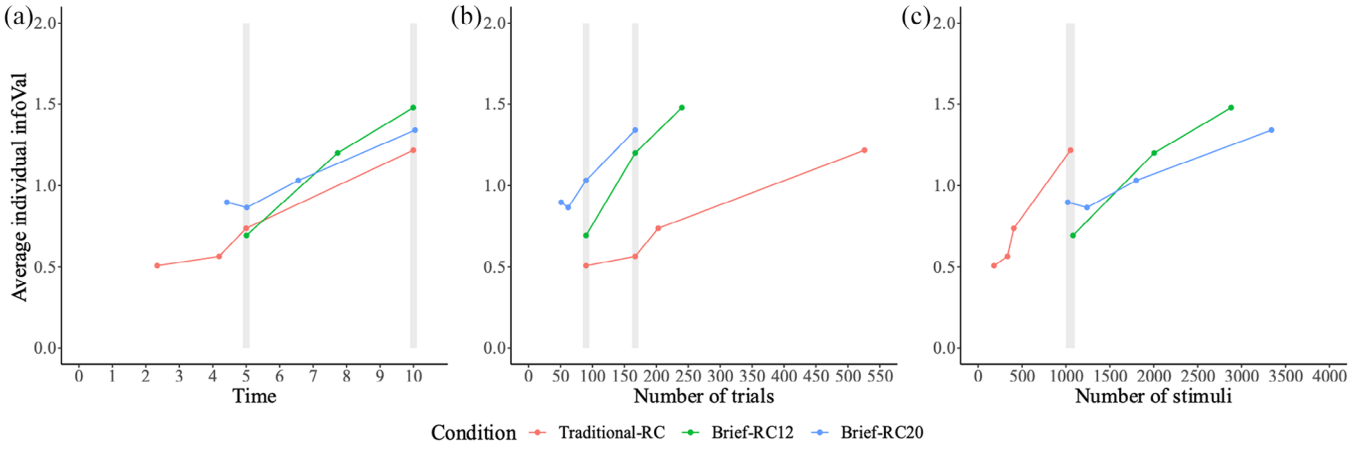

FIGURE 9 Relation between the average infoVal (information value) of individual CIs and (A) time (median cumulative time in minutes), (B) number of trials and (C) number of stimuli, as a function of condition (different coloured lines: Traditional-RC, Brief-RC12 and Brief-RC20). The dots represent the averaged infoVal at given points of comparison (cf. Table 1). The grey frames highlight the points of comparison, i.e., (A) at approximately 5 and 10 minutes, (B) at 90 and 167 trials and (C) at approximately 1050 stimuli.

4.2.5 | InfoVal as a function of the number of trials

We submitted the infoVal scores of the individual CIs to an OLS analysis with C_1, C_2, trials (90 vs 167 trials) and their interactions. C_1 was significant such that infoVal scores from the Traditional-RC ($M = 0.54$, $SD = 1.24$) were lower than those from the Brief-RC12 and Brief-RC20 ($M = 1.07$, $SD = 1.70$), $F(1, 294) = 15.60$, $p < .001$, $\eta_p^2 = 0.026$. InfoVal scores after 90 trials ($M = 0.74$, $SD = 1.47$) were significantly lower than those after 167 trials ($M = 1.04$, $SD = 1.67$), $F(1, 294) = 5.29$, $p = .022$, $\eta_p^2 = 0.009$. C_2 was not significant such that infoVal scores from the Brief-RC12 ($M = 0.95$, $SD = 1.64$) did not significantly differ from those of the Brief-RC20 ($M = 1.19$, $SD = 1.75$), $F(1, 294) = 2.38$, $p = .124$, $\eta_p^2 = 0.004$. There was no significant effect of $C_1 \times$ trials, nor $C_2 \times$ trials (all $p > .10$, $BF_{01} > 10$). Figure 9(B) shows the averaged infoVal scores of individual CIs as a function of the number of trials.

4.2.6 | InfoVal as a function of the number of stimuli

We submitted the infoVal scores of the individual CIs produced after producers had seen approximately 1050 stimuli (noisy faces) to an OLS analysis using C_1 and C_2 as predictors. C_1 was significant such that the infoVal scores from the Traditional-RC ($M = 1.22$, $SD = 1.89$) were higher than those from the Brief-RC12 and Brief-RC20 ($M = 0.80$, $SD = 1.41$), $F(1, 297) = 4.73$, $p = .030$, $\eta_p^2 = 0.016$. There were no significant differences between the infoVal scores of the Brief-RC12 ($M = 0.69$, $SD = 1.49$) and Brief-RC20 ($M = 0.90$ $SD = 1.32$), $F(1, 297) = 0.82$, $p = .364$, $\eta_p^2 = 0.003$, $BF_{01} = 11.43$.

Figure 9(C) shows the averaged infoVal scores of individual CIs as a function of the number of stimuli. Means and standard deviations of the infoVal scores of individual CIs for each of the experimental cells are available in Supplementary Table S2. When considering all stimuli, average infoVal scores observed in the Brief-RC12 ($M = 0.69$,

SD = 1.49), the Brief-RC20 (*M* = 0.90, *SD* = 1.32) and the Traditional-RC (*M* = 1.22, *SD* = 1.89) were all relatively low as all of them were below the threshold of 1.96.

4.3 | Discussion

Subjective ratings from independent judges revealed that the Brief-RC outperformed the Traditional-RC on our three criteria. With respect to time, judges evaluated individual CIs as more prototypical of the social target (i.e., a Chinese-looking face) after both 5 and 10 minutes. As for the number of trials, ratings were also higher with the Brief-RC as opposed to the Traditional-RC and, this, whether participants had completed 90 or 167 trials. In contrast to Experiment 1, Experiment 2 revealed that the Brief-RC did in fact outperform the Traditional-RC after the participant had seen an equal number of stimuli (i.e., noisy faces). One reasonable explanation for the difference between the two experiments may reside in the fact that, compared with the point of comparison in Experiment 2 (i.e., approximately 1050 stimuli), the point of comparison in Experiment 1 (i.e., 720 stimuli) may have been set too early to spot any divergence. Finally, the average CIs also came across as more prototypical on all three criteria when we relied on the Brief-RC as compared to the Traditional-RC.

We also explored the differences between methods using an objective metric. The message proves somewhat less clear than the one emerging from the subjective ratings. On the one hand, infoVal scores of individual CIs revealed that the two methods performed at a similar level after the same time, while the Brief-RC outperformed the Traditional-RC after the same number of trials. The opposite pattern emerged when holding constant the number of stimuli. We come back to this point in the General Discussion.

5 | GENERAL DISCUSSION

Producing robust and high-quality individual CIs is far from being an easy task as it requires a very large number of trials. This is not only economically costly and time-consuming but also challenging in terms of participants' motivation (Brinkman et al., 2019b; Todorov et al., 2011). For these reasons, researchers called for an improvement of the RC method to produce higher-quality outcomes, while reducing the number of trials (Todorov et al., 2011). With exactly this goal in mind, we conducted two pre-registered experiments that examine a new method, namely, the Brief-RC. This new version aims to reduce the number of trials and improve the outcome quality by increasing the number of stimuli at each trial.

In Experiment 1, we compared two variants of the Brief-RC (the Brief-RC12 with 12 stimuli per trial and Brief-RC20 with 20 stimuli per trial) to the Traditional-RC (with 2 stimuli per trial) while fixing the overall number of stimuli (i.e., noisy faces) presented to the participants across all trials. Individual CIs produced from both methods performed similarly, as assessed by judges' ratings (i.e., on how prototypically Chinese the faces looked) and by the infoVal scores – which determines the degree to which a CI stems from meaningful responses (i.e., signal) rather than random one (i.e., noise). Moreover, judges' ratings of average CIs between the two methods did not differ. In itself, this is a noteworthy achievement because the Brief-RC allows presenting the same amount of information (i.e., number of stimuli) in almost half the time and only a fraction of trials compared with the traditional method.

Experiment 2 extended these findings by also comparing the end-products of the different RC methods on additional criteria. Moreover, we increased the task length and statistical power to achieve more robust CIs. This time, subjective ratings from independent judges of the individual CIs revealed that the Brief-RC outperformed the Traditional-RC after participants had seen the same number of stimuli. Importantly, the Brief-RC variants also outperformed the Traditional-RC after the same length of time (after 5 or 10 minutes) and after the same number of trials (after 90 or 167 trials). Moreover, these results were replicated when considering the average CIs. As for infoVal, although more exploratory, the two methods performed on similar levels when considering the same length of time, whereas the Brief-RC scored better when comparing across the number of trials, but worse when comparing across the number of stimuli. This divergence between the results on the subjective and objective measures suggests that the relation between the two is not as straightforward as it may seem and needs further investigation. One possible explanation may be that the signal detected by infoVal did not (or did only partly) correspond with the one assessed by subjective ratings. For instance, it may be the case that most of the signal identified from infoVal came from the jaw, whereas subjective ratings of external judges (regarding how Chinese-looking the faces were) were mainly driven by the eye's region. In other words, some slight qualitative changes – that are not quantitatively noticeable enough for the infoVal metric – could be highly informative for judges. All of this remains very tentative at this stage and requires additional research. Finally, it is noteworthy that judges' ratings on individual CIs and infoVal scores were relatively low, regardless of the condition, signalling that increasing the number of trials may be beneficial to improve both metrics.

An interesting finding from Experiment 2 stems from the visual inspection of Figure 5. Indeed, the gap between judgements of individual CIs from the Brief-RC variants versus the Traditional-RC widens as the task progresses. Specifically, individual CIs' ratings from the Traditional-RC seem to reach a horizontal asymptote not long after the beginning of the task (around 4 minutes, or 200 trials) whereas for the Brief-RC (and particularly for the Brief-RC12), they appear to further improve, even beyond our fixed comparison points. This suggests that, on average, individual CIs may not substantially get better after a few minutes (or trials) when relying on the traditional procedure, whereas there seems to be room for improvement with our new method. At the same time, a different pattern seems to take place when considering ratings of average CIs. That is, the Brief-RC offers the advantage that it rapidly converges towards a stable average CI, whereas the Traditional-RC may take longer. In both cases, these improvements could be attributed to the enhanced signal-to-noise ratio of the Brief-RC, and perhaps also to the lack of motivation when choosing from a set of stimuli that are less likely to carry a signal (as in the Traditional-

RC). We also note that ratings of average CIs are largely superior to those of individual CIs. This is hardly surprising given that they build on thousands of trials instead of a few hundred (Cone et al., 2020).

Overall, the present findings strongly suggest that the Brief-RC offers a substantial improvement over the Traditional-RC on both individual and average CIs, at least in terms of subjective judgements. This is an important accomplishment because it should allow researchers to run RC-based studies in a much more efficient and reliable manner. Some recent research successfully adopted this new method to investigate approach/avoidance effects on social perception (Rougier et al., 2021). Replicability of RC effects should also be enhanced via the production of higher-quality CIs – but to be sure, future research could compare effects produced by the Brief-RC with those of the traditional method. Clearly, the advantage of increasing the number of stimuli per trial should also benefit other variants of the RC paradigm, as in methods that rely on three- and four-dimensional (instead of two-dimensional) stimuli (see Jack & Schyns, 2017; Walker & Keller, 2019; Walker & Vetter, 2016). A notable dividend that the Brief-RC grants access to higher-quality individual CIs paves the way for more fine-grained analyses with other variables of interest (e.g., inter-individual differences).

Differences between the two variants of the Brief-RC (i.e., Brief-RC12 and Brief-RC20) were minimal. This is an interesting outcome in and of itself as one might conjecture that an ideal observer should perform better as the number of stimuli per trial increases from 12 to 20 per trial. This was not the case here with our participants. A possible reason may stem from the limitations of the human brain when it comes to processing visual information, suggesting that adding more faces per trial should not help improve the method (Marois & Ivanoff, 2005; Palmer, 1990). Considering the present findings, the Brief-RC12 should be preferred over the Brief-RC20 as it requires fewer stimuli to achieve a similar result. However, our research was limited to two versions of the Brief-RC (Brief-RC12 and Brief-RC20) and further work should investigate the optimal configuration in terms of the number of stimuli per trial (e.g., 4, 6, 8 and 10).

Although not directly tested in the present work, another remarkable advantage of the Brief-RC is that it likely increases the external validity of the outcomes. CIs are essentially a linear combination of the stimuli (noisy faces). Increasing the number of trials and, even more so, the number of stimuli per trial substantially enlarges the unique set of stimuli (e.g., noisy faces) and thus the CI's potential face space – the universe of potential CIs that can be built from all possible combinations of stimuli from a given RC task (Jack & Schyns, 2017; Todorov et al., 2011). In particular, the visual renderings from the RC-Brief should gain external validity because they derive from an exponentially larger sample of stimuli[10] and therefore have more chances to match the underlying psychological representations. Overall, this also brings the Brief-RC a step closer to adequately sampling larger stimulus space spanned by additional dimensions (e.g., colour, depth, texture or higher dimensional noise patterns; Brinkman et al., 2017) and to capturing more complex structures unlikely to emerge with fewer stimuli (e.g., a wrinkle on the forehead when selecting older faces; Jack & Schyns, 2017). Further research should provide empirical evidence for the increased ecological validity of Brief-RC's outcomes. This could be done by testing whether the obtained individual CIs are variable (when relying on a target category that should be associated with variability in visual representations) and whether CIs correlate with meaningful individual-level measures (e.g., self-reported evaluation).

A noteworthy limitation of this work is that we only relied on the target category of Chinese-looking people. Therefore, for the sake of generalizability, any future investigations comparing the Brief-RC with Traditional-RC (or other variants) should consider varying the target category (e.g., use female-looking as the target category). In addition, future work could also examine the correlation between individual CIs and other individual-level measures to test whether the improved quality of CIs in the Brief-RC leads to larger correlations as compared to the traditional method.

In conclusion, the current work introduces an improved reverse correlation method, namely, the Brief-RC. The access to more robust and better-quality individual CIs is key to address inflated Type I error rates prevalent in traditional two-phase reverse correlation procedures. The Brief-RC technique enhances the quality of average CIs and even more so of individual CIs by improving the signal-to-noise ratio through the presentation of a greater number of stimuli at every trial. Moreover, the Brief-RC significantly reduces the overall task length, which offers clear-cut benefits in terms of maintaining the participant's motivation, but also in terms of economic resources and time constraints for the researcher.

AUTHOR CONTRIBUTION

Mathias Schmitz and Marine Rougier are equal contributors to this work and are designated as co-first authors.

ACKNOWLEDGEMENTS

This work was supported by the Fonds de la Recherche Scientifique (FNRS) grants awarded to Mathias Schmitz (number 1.A393.17) and Marine Rougier (number 1.B347.19).

CONFLICT OF INTEREST STATEMENT

We have no conflict of interest to disclose.

ETHICS STATEMENT

It received approval from the local ethics committee (institutional board).

TRANSPARENCY STATEMENT

We pre-registered all experiments on the Open Science Framework (OSF; including a priori theoretical reasoning, hypotheses, power estimations, procedures and statistical analyses). We report any significant deviations from the initial pre-registrations in the core manuscript.

[10] A reverse correlation task with m stimuli per trial and n trials can generate m^n unique individual CIs (assuming that each stimulus is unique). Therefore, a Brief-RC with m stimuli can generate $(m/2)^n - 1$ times more unique individual CIs than a Traditional-RC with the same number of trails. For instance, after $n = 3$ trials, the Brief-RC12 with $m = 12$ stimuli per trial can produce $12^3 = 1728$ unique individual CIs, whereas a Traditional-RC can only generate $2^3 = 8$. That is, the Brief-RC12 can already generate $(12/2)^3 - 1 = 215$ times more unique individual CIs after only 3 trials than the Traditional-RC ($8 + 8 \times 215 = 1728$).

We report all measures, manipulations and exclusions. Our preregistrations, data, data analysis R scripts for all experiments and JavaScript scripts to run both the Traditional-RC and the Brief-RC variants are available on the following link: https://osf.io/ps9wu/?view_only=028597d60e7342bfaeb6881051cb6bca.

ORCID

Mathias Schmitz https://orcid.org/0000-0001-9272-5874
Marine Rougier https://orcid.org/0000-0002-9467-2726
Vincent Yzerbyt https://orcid.org/0000-0003-1185-4733

REFERENCES

Ahumada, A., & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, *49*(6B), 1751–1756. https://doi.org/10.1121/1.1912577

Ahumada, A., Marken, R., & Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. *The Journal of the Acoustical Society of America*, *57*(2), 385–390. https://doi.org/10.1121/1.380453

Albohn, D. N., & Adams, R. B. (2020). Everyday beliefs about emotion perceptually derived from neutral facial appearance. *Frontiers in Psychology*, *11*, 264. https://doi.org/10.3389/fpsyg.2020.00264

Axt, J., Siemers, N., Discepola, M. N., Martinez, P., Xiao, Z., & Wehrli, E. (2023). The mind's "aye"? Investigating overlap in findings produced by reverse correlation versus self-report. *Journal of Experimental Social Psychology*, *107*, 104473. https://doi.org/10.1016/j.jesp.2023.104473

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv:1506.04967.

Brinkman, L., Dotsch, R., Zondergeld, J., Koevoets, M. G. J. C., Aarts, H., & van Haren, N. E. M. (2019b). Visualizing mental representations in schizophrenia patients: A reverse correlation approach. *Schizophrenia Research: Cognition*, *17*, 100138. https://doi.org/10.1016/j.scog.2019.100138

Brinkman, L., Goffin, S., van de Schoot, R., van Haren, N. E. M., Dotsch, R., & Aarts, H. (2019b). Quantifying the informational value of classification images. *Behavior Research Methods*, *51*(5), 2059–2073. https://doi.org/10.3758/s13428-019-01232-2

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, *28*(1), 333–361. https://doi.org/10.1080/10463283.2017.1381469

Brooks, J. A., Stolier, R. M., & Freeman, J. B. (2018). Stereotypes bias visual prototypes for sex and emotion categories. *Social Cognition*, *36*(5), 481–493. https://doi.org/10.1521/soco.2018.36.5.481

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2016). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*, *28*(1), 92–103. https://doi.org/10.1177/0956797616674999

Brown-Iannuzzi, J. L., McKee, S., & Gervais, W. M. (2018). Atheist horns and religious halos: Mental representations of atheists and theists. *Journal of Experimental Psychology: General*, *147*(2), 292–297. https://doi.org/10.1037/xge0000376

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 1–20. https://doi.org/10.5334/joc.10

Cone, J., Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2020). Type I error is inflated in the two-phase reverse correlation procedure. *Social Psychological and Personality Science*, *12*(5), 760–768. 194855062093861. https://doi.org/10.1177/1948550620938616

Dai, H., & Micheyl, C. (2010). Psychophysical reverse correlation with multiple response alternatives. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 976–993. https://doi.org/10.1037/a0017171

Degner, J., Mangels, J., & Zander, L. (2019). Visualizing gendered representations of male and female teachers using a reverse correlation paradigm. *Social Psychology*, *50*(4), 233–251. https://doi.org/10.1027/1864-9335/a000382

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dotsch, R., & Todorov, A. (2011). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571. https://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980. https://doi.org/10.1111/j.1467-9280.2008.02186.x

Dotsch, R., Wigboldus, D. H. J., & Van Knippenberg, A. (2013). Behavioral information biases the expected facial appearance of members of novel groups. *European Journal of Social Psychology*, *43*(1), 116–125. https://doi.org/10.1002/ejsp.1928

Dotsch, R. (2017). rcicr: Reverse correlation image classification toolbox (R package version 0.3.0). Retrieved from https://CRAN.R-project.org/package=rcicr

Ge, L., Zhang, H., Wang, Z., Quinn, P. C., Pascalis, O., Kelly, D., Slater, A., Tian, J., & Lee, K. (2009). Two faces of the other-race effect: Recognition and categorisation of Caucasian and Chinese faces. *Perception*, *38*(8), 1199–1210. https://doi.org/10.1068/p6136

Hinzman, L., & Maddox, K. B. (2017). Conceptual and visual representations of racial categories: Distinguishing subtypes from subgroups. *Journal of Experimental Social Psychology*, *70*, 95–109. https://doi.org/10.1016/j.jesp.2016.12.012

Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. J. (2011). Facing Europe. *Psychological Science*, *22*(12), 1583–1590. https://doi.org/10.1177/0956797611419675

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology*, *4*, 1–8. https://doi.org/10.3389/fpsyg.2013.00386

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*(1), 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Jaeger, B. (2017). r2glmm: Computes R squared for mixed (multilevel) models (R package version 0.1.2). https://CRAN.R-project.org/package=r2glmm

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788. https://doi.org/10.1037/0022-3514.54.5.778

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy. *Psychological Science*, *18*(12), 1084–1089. https://doi.org/10.1111/j.1467-9280.2007.02029.x

Kevane, M., & Koopmann-Holm, B. (2021). Improving reverse correlation analysis of faces: Diagnostics of order effects, runs, rater agreement, and image pairs. *Behavior Research Methods*, *53*(4), 1609–1647. https://doi.org/10.3758/s13428-020-01499-w

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile? *Vision Research*, *44*(13), 1493–1498. https://doi.org/10.1016/j.visres.2003.11.027

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces database. *Cognition & Emotion*, *24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Lee, Y.-T., & Ottati, V. (1995). Perceived in-group homogeneity as a function of group membership salience and stereotype threat. *Personality and Social Psychology Bulletin*, *21*(6), 610–619. https://doi.org/10.1177/0146167295216007

Lick, D. J., Carpinella, C. M., Preciado, M. A., Spunt, R. P., & Johnson, K. L. (2013). Reverse-correlating mental representations of sex-typed bodies: The effect of number of trials on image quality. *Frontiers in Psychology*, *4*, 1–9. https://doi.org/10.3389/fpsyg.2013.00476

Lloyd, E. P., Sim, M., Smalley, E., Bernstein, M. J., & Hugenberg, K. (2020). Good cop, bad cop: Race-based differences in mental representations of police. *Personality and Social Psychology Bulletin*, *46*(8), 1205–1218. https://doi.org/10.1177/0146167219898562

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209–226. https://doi.org/10.1207/s15516709cog2802_4

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, *9*(6), 296–305. https://doi.org/10.1016/j.tics.2005.04.010

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Michalak, N. M., & Ackerman, J. M. (2021). A multimethod approach to measuring mental representations of threatening others. *Journal of Experimental Psychology: General*, *150*(1), 114–134. https://doi.org/10.1037/xge0000781

Oliveira, M., Garcia-Marques, T., Dotsch, R., & Garcia-Marques, L. (2019). Dominance and competence face to face: Dissociations obtained with a reverse correlation approach. *European Journal of Social Psychology*, *49*(5), 888–902. https://doi.org/10.1002/ejsp.2569

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 332–350. https://doi.org/10.1037/0096-1523.16.2.332

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*(6), 897–911. https://doi.org/10.1037/a0036498

Rougier, M., & De Houwer, J. (2023). Unconstraining evaluative conditioning research by using the reverse correlation task. *Social Psychological and Personality Science*, *0*(0), 00–00. https://doi.org/10.1177/19485506231217526

Rougier, M., Schmitz, M., & Yzerbyt, V. (2021). When my actions shape your looks: Experience-based properties of approach/avoidance bias the visual representation of others. *Journal of Personality and Social Psychology*, *120*(5), 1146–1174. https://doi.org/10.1037/pspa0000268

Schmitz, M., Rougier, M., & Yzerbyt, V. (2020a). Comment on "Quantifying the informational value of classification images": A miscomputation of the infoVal metric. *Behavior Research Methods*, *52*(3), 1383–1386. https://doi.org/10.3758/s13428-019-01295-1

Schmitz, M., Rougier, M., Yzerbyt, V., Brinkman, L., & Dotsch, R. (2020b). Erratum to: Comment on "Quantifying the informational value of classification images": Miscomputation of infoVal metric was a minor issue and is now corrected. *Behavior Research Methods*, *52*(4), 1800–1801. https://doi.org/10.3758/s13428-020-01367-7

Schmitz, M., Vanbeneden, A., & Yzerbyt, V. (2024). The many faces of compensation: The similarities and differences between social and facial models of perception. *PloS One*, *19*(2), e0297887.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Tiddeman, B. P., Stirrat, M. R., & Perrett, D. I. (2005). Towards realism in facial image transformation: Results of a wavelet MRF method. *Computer Graphics Forum*, *24*(3), 449–456. https://doi.org/10.1111/j.1467-8659.2005.00870.x

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass*, *5*(10), 775–791. https://doi.org/10.1111/j.1751-9004.2011.00389.x

Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. *Current Opinion in Neurobiology*, *23*(3), 373–380. https://doi.org/10.1016/j.conb.2012.12.010

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Walker, M., & Keller, M. (2019). Beyond attractiveness: A multimethod approach to study enhancement in self-recognition on the big two personality dimensions. *Journal of Personality and Social Psychology*, *117*(3), 483–499. https://doi.org/10.1037/pspa0000157

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived big two and big five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*(4), 609–624. https://doi.org/10.1037/pspp0000064

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Young, A. I., Ratner, K. G., & Fazio, R. H. (2013). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, *25*(2), 503–510. https://doi.org/10.1177/0956797613510717

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schmitz, M., Rougier, M., & Yzerbyt, V. (2024). Introducing the Brief Reverse Correlation: An Improved Tool to Assess Visual Representations. *European Journal of Social Psychology*, 1–16. https://doi.org/10.1002/ejsp.3100