Introducing the Brief Reverse Correlation:

An Improved Tool to Assess Visual Representations

Mathias Schmitz[a], Marine Rougier[a], Vincent Yzerbyt[a]

[a]Institut de recherche en sciences psychologiques, Université catholique de Louvain

Authors Note

We have no conflict of interest to disclose. This work was supported by the Fonds de la Recherche Scientifique (FNRS) grants awarded to Mathias Schmitz [grant number 1.A393.17] and Marine Rougier [grant number 1.B347.19]. This work was also supported by the Louvain Social Psychology Lab.

Corresponding author at: Place du Cardinal Mercier 10, B-1348, Louvain-la-Neuve, Belgium. *E-mail address:* mathias.schmitz@uclouvain.be (M. Schmitz).

Abstract

The reverse correlation (RC) is an innovative method to capture visual mental representations (i.e., classification images, CIs) of social targets that has become increasingly popular in social psychology. Because CIs of high quality are difficult to obtain without a large number of trials, the majority of past research relied on CIs extracted from samples of participants (average CIs). This strategy, however, leads to inflated false positivity rates. Using the representation from each participant (individual CIs) offers one solution to this problem. Still, this approach requires large numbers of trials and is thus economically costly, time demanding, demotivating for the participants, or simply impractical. We introduce a new version of the reverse correlation method, namely the Brief-RC. The Brief-RC increases the quality of individual (and average) CIs and reduces the overall task length by increasing the number of stimuli (i.e., noisy faces) presented at each trial. In two experiments, assessments by external judges confirm 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 Brief-RC may thus facilitate the production of higher-quality individual CIs and alleviate the risk of false positivity rate.

Keywords: reverse correlation, Brief-RC, classification images, rcirc package, infoVal

Introducing the Brief Reverse Correlation:

An improved tool to assess visual representations

The two-image forced choice reverse correlation (Brinkman, et al., 2017; Dotsch & Todorov, 2011; Dotsch et al., 2008; Mangini & Biederman, 2004) is a method that provides visual proxies of mental representations (mostly faces). In recent years, the method has become increasingly popular, especially in social psychology (for a review, see Brinkman et al., 2017). This approach allows capturing facial representations at a group-level (i.e., from a sample of participants in a given condition) as well as at an individual-level (i.e., from a single participant). Although facial renderings produced at a group-level are of better quality than those at an individual-level, the evaluation of the former is prone to inflated Type I errors (Cone et al., 2020). Additionally, only the individual-level allows more fine-grained analyses (e.g., correlating judgments of the visual rendering with another individual-level variable). However, whereas individual-level visual outcomes are preferable, a large number of trials is usually required to achieve high-quality outcomes, which entails other issues (i.e., economically costly, time demanding, decreased participants' motivation to complete the task in a conscientious manner; Brinkman, Goffin et al., 2019; Todorov et al., 2011). Researchers have therefore noted the need to improve the method in order to "generate higher quality outcomes or reduce the number of trials" (Todorov et al., 2011, p. 787). With these concerns in mind, we propose an improved version of the method, namely 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.

The Reverse Correlation Paradigm

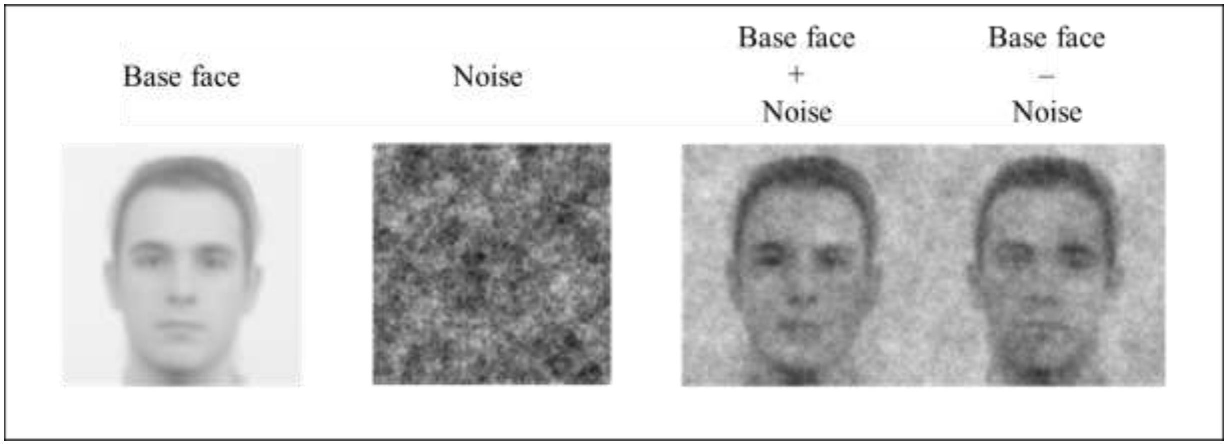
The reverse correlation (RC) is a data-driven method which provides access to the visual representation (generally facial) that people have of a given target (e.g., how a criminal person's face looks like). Over the years, the procedure has become widespread in social psychology and has proven to be particularly useful to identify the diagnostic features that drive social perception or to examine how top-down processes can bias mental images of social targets (see Brinkman et al., 2017; Jack & Schyns, 2017; Todorov et al., 2011; Todorov et al. , 2013). For instance, it contributed to uncover facial diagnostic components of social categories such as ethnicity and race (e.g., Dotsch et al., 2008; Dotsch et al., 2011; Hinzman & Maddox, 2017; Krosch & Amodio, 2014; Kunst, Dovidio et al., 2017), gender (e.g., Brooks et al., 2018; Degner et al., 2019; Gundersen & Kunst, 2018), country of origin (e.g., Imhoff & Dotsch, 2013; Imhoff et al., 2011), profession and occupation (e.g., Degner et al., 2019; Lloyd et al., 2020), age (e.g., Albohn & Adams, 2020), religion (e.g., Brown-Iannuzzi et al., 2018) but also personality traits (e.g, Dotsch & Todorov, 2011; Imhoff et al., 2013; Lin et al., 2018; Oliveira et al., 2019), and emotions (e.g., Albohn & Adams, 2020; Brooks et al., 2018). Furthermore, the RC method provides insights on how these mental templates could be distorted by a priori preferences, attitudes, beliefs, and knowledge, such as political ideology (e.g., Jackson et al., 2018; Young et al., 2013), love and attraction (e.g., Gunaydin & DeLong, 2015; Karremans et al., 2011), stereotypes and prejudice (e.g., Brown-Iannuzzi et al., 2016; Brown-Iannuzzi et al., 2018; Dotsch et al., 2008; Hinzman & Maddox, 2017), group membership (e.g., Hong & Ratner, 2020; Ratner et al., 2014), dehumanization (Kunst, Kteily et al., 2017; Petsko et al., in press), or behavioral information about novel groups (Dotsch et al., 2013).

The RC is a method rooted in signal-detection theory (Ahumada & Lovell, 1971; Ahumada et al., 1975) that aims to identify the information that underlies perception. 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) 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).

Although several implementations of this paradigm exist in the domain of visual perception (see Jack & Schyns, 2017; Todorov et al., 2011; Todorov et al., 2015), the two-image forced choice noised-based reverse correlation (hereafter Traditional-RC) developed by Dotsch et al. (2008; see also Dotsch & Todorov, 2011) is by far the most prevalent. A typical procedure (e.g., Dotsch et al., 2008) involves two steps. In the first step, participants have to 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 in superimposing (by either adding or subtracting) a noise pattern (for the generation of the noise pattern see Dotsch & Todorov, 2011; Mangini & Biederman, 2004) onto a base image (usually a morph of several faces) as illustrated in Figure 1.

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



Each noisy face in a pair is maximally different from the other of the same pair since 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 on 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). To build the representation of a participant or a group of participants, one adds the average of all selected noise patterns to the base face. This allows obtaining the so-called *classification image* (CI), i.e., 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 researchers' hypotheses. In one illustrative research, Dotsch et al. (2008) captured the average facial template 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 individual, the more negative their representation of this group.

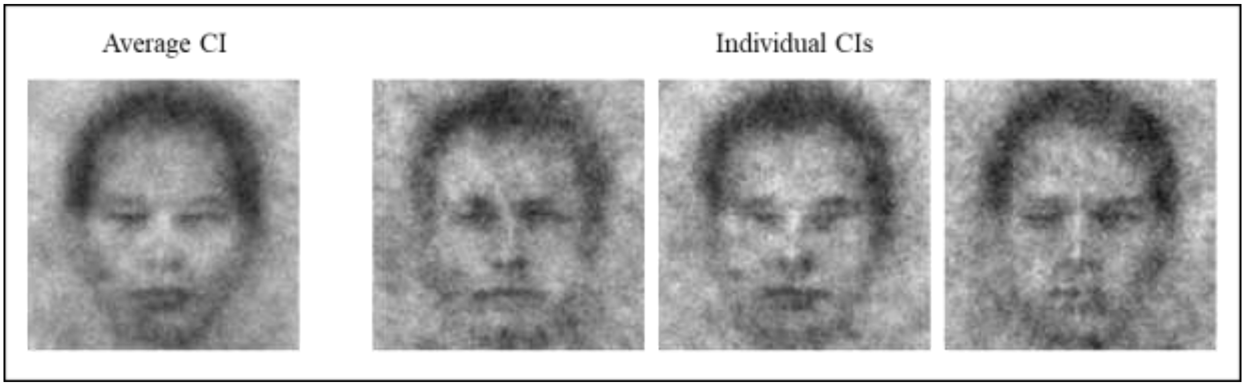
This procedure offers several advantages to study psychological effects (see Brinkman et al., 2017; Dotsch & Todorov, 2011; Jack & Schyns, 2017). First, it makes no prior assumption about what people's internal representations may look like. In other words, it allows sampling a vast set of hypotheses (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). The RC thus reveals "near spontaneous use of information" (Brinkman et al., 2017, p. 336) because participants may adopt any criteria they want for the classification task (i.e., the method is said to be unconstrained). Actually, it may even allow probing representations that are ineffable for the participant (Mangini & Biederman, 2004).

Current Limitations of the Traditional-RC

The majority of 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 less noisy than the latter (see Figure 2; Cone et al., 2020; Imhoff et al., 2011; Imhoff et al., 2013). However, average CIs present several shortcomings. They prevent carrying out idiosyncratic analyses, for instance, by analyzing the relation between racial bias and visual representation bias at the participant-level (Dotsch et al., 2008). In addition, they are only valid to the extent that participants actually share a common representation of the target category, which is not necessarily the case (Brinkman et al., 2017; Ratner et al., 2014). Also, judges may detect significant differences between group-wise CIs that may not be 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) conducted computer simulations and showed that using average CIs in typical two-phase reverse correlations procedures (the first phase corresponding to the production of the CIs and the second to their rating by external judges) may increase considerably Type I error rates because they do not take into account the inter-individual variability of the underlying CIs. These authors estimated that about 66% of past research might suffer from inflated false positivity rates. Building on this result, they recommend using individual CIs (and other methods that are currently being developed) to address this problem.

Figure 2

Example of an average CI and some individual CIs issued from the Traditional-RC after 526 trials (from Experiment 2)



Although relying on individual CIs would address the abovementioned concerns (i.e., they reflect the inter-individual variability and thus reduce inflated Type I error rates; they grant access to fine-grained analyses), they require a very large number of trials to achieve an acceptable level of quality. This is because they are composed of much fewer trials than the average CIs (Brinkman et al., 2017; Cone et al., 2020; Jack & Schyns, 2017; Ratner et al., 2014; Todorov et al., 2011). 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, Goffin et al., 2019; Cone et al., 2020; Dotsch & Todorov, 2011). Although adding more trials should in principle enhance the quality of individual CIs, in practice, however, it does not always bear fruit. In fact, doing so may even be detrimental because this hinders participants' motivation to mindfully complete the task (Brinkman, Goffin et al., 2019; 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. This would increase the chances of false positives when judging low quality individual CIs as well as the resulting average CIs. Indeed, the RC method will always yield a CI whatever the number of trials or the meaningfulness of responses (Brinkman, Goffin et al., 2019). To address this issue, Brinkman, Goffin et al. (2019) developed infoVal, an information value metric that assesses the degree to which an individual CI derives from signal rather than noise. However, although this metric allows identifying noisy individual CIs, it does not preclude obtaining poor quality renderings.

A different approach to improving individual CIs would be to take into account the response confidence by providing multiple response alternatives (e.g., "probably happy", "possibly happy", "possibly unhappy", or "probably unhappy") and only built the CIs from trials with high confidence responses (e.g., Brinkman, Dotsch, et al., 2019; 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 rate of high-confidence responses is low (Brinkman et al., 2017).

In sum, and as of today, there is no practical and reliable solution to improve the quality of individual CIs and thus reduce the inflated Type I errors rates stemming from the use of average CIs or poor-quality individual CIs. This matter becomes even more crucial in light of the replicability crisis in psychological science where strong and reliable science should be encouraged (Open Science Collaboration, 2015; Simmons et al., 2011).

The Brief Reverse Correlation

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 short, researchers have urged to "generate higher quality outcomes or reduce the number of trials" (Todorov et al., 2011, p. 787). 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 greater number of noisy faces to select from at each trial.

In the Brief-RC, the likelihood to find a stimulus that carries more diagnostic information of the expected signal, i.e., with a higher signal-to-noise ratio (SNR), 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 match (to some degree and by chance alone) our expectation of a happy-looking face. 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 our mental template of the target. A greater panel of faces should therefore improve the SNR of the selected noisy face at each trial and consequently accelerate the convergence toward a higher quality and more robust CI, both at the individual and at the average levels. To be sure, the improvement in SNR resulting from the increase of the stimulus set at each trial should reach a plateau because the visual processing capacity of the human brain is limited (Marois & Ivanoff, 2005; Palmer, 1990). Still, a reasonable increase in the selection presented to the participants should deliver clearer CIs for a given number of trials. This means that, in general, one could obtain high quality CIs with fewer trials and a shorter time of completion. Said otherwise, the same quality-level CI issued from a Traditional-RC might 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.

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 with more than two stimuli per trial. Because factors such as limitations to process visual information may constrain the benefits of the increase in the number of trials, we tested this assumption empirically with two versions of the Brief-RC—namely the Brief-RC12, with 12 stimuli per trial, and the Brief-RC20, with 20 stimuli per trial. We assessed the quality of the visual renderings by means of a subjective (ratings of the CIs from independent judges) and an objective measure (infoVal). In order 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—i.e., 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 SNR, 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 conservative because the number of trials in the Traditional-RC will be much larger than the Brief-RC for the same number of stimuli presented. In Experiment 2, we sought to replicate findings from Experiment 1. Furthermore, we increased the length of the task and compared the RC variants across different criteria (task length, number of trials, and number of stimuli). Indeed, all these features may have important practical implications in terms of improving research practices, namely, to achieve better quality outcomes, to reduce the task length (number of trials and time), to sustain participants' motivation, and to reduce the researcher's financial cost and time to carry out the task. 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 specifically geared our predictions toward individual CIs because they are usually noisier than average CIs and because enhancing individual CIs quality will contribute to a more replicable social psychology science (Cone et al., 2020; Open Science Collaboration, 2015; Simmons et al., 2011).

Overview and Analytical Strategy

The experiments reported in the present research followed the standard implementation of the reverse correlation paradigm (Dotsch et al., 2008). Each experiment comprised two parts involving two distinct samples of participants. In the first part, a sample of participants called “producers” completed one of the three version of the RC task (i.e., Traditional-RC, Brief-RC12, or Brief-RC20) in order to capture their visual representation of a Chinese-looking face. In the second part, another sample of participants called “judges” evaluated these CIs on how Chinese they look. All participants took part via the online crowdsourcing platform 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 was conducted in a manner consistent with the APA’s Ethical Principles.

Whenever possible, we conducted the analyses using linear mixed models (LMM; Judd et al., 2012; Judd et al., 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. Therefore, we only report the most parsimonious LMM (i.e., the ones that maximize power while trying to minimize the Type I error rate), following Bates et al. 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, LLM 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 created 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 favor 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 infoVal, an innovative informational value metric developed by Brinkman, Goffin et al. (2019; see also Schmitz, Rougier, & Yzerbyt, 2020; Schmitz, Rougier, Yzerbyt, Brinkman et al., 2020) that quantifies the degree to which a CI contains 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. To improve the accuracy or the metric, we applied an oval-shaped mask to the CIs in order to extract the face region (e.g., Oliveira et al., 2019; Ratner et al., 2014) and compute the infoVal score on this region.

We preregistered all experiments on 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 in these studies. Our pre-registrations, data,

[1] Bayes Factor (BF) estimations were derived from the Bayesian Information Criterion (BIC) with a “unit information prior” following the guidelines from Wagenmakers (2007).

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

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.

Method

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 similar to previous studies (e.g., 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 abnormally high percentage (48%) of fast reaction times (< 200ms). 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 others—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; Dotsch et al., 2013; Imhoff et al., 2013). Sample size was determined before any data analysis. A post-hoc sensitivity analysis 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).

Experimental design and procedure

We first asked producers to complete one of the RC variants and then invited judges to rate the produced CIs.

Part 1: Reverse correlation task. We randomly assigned producers to one of the three RC tasks (Traditional-RC, Brief-RC12, or Brief-RC20). 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 base 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 in either adding ("oriented") or subtracting ("inverted") a sinusoidal noise pattern from the base image (see Dotsch & Todorov, 2011; 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).

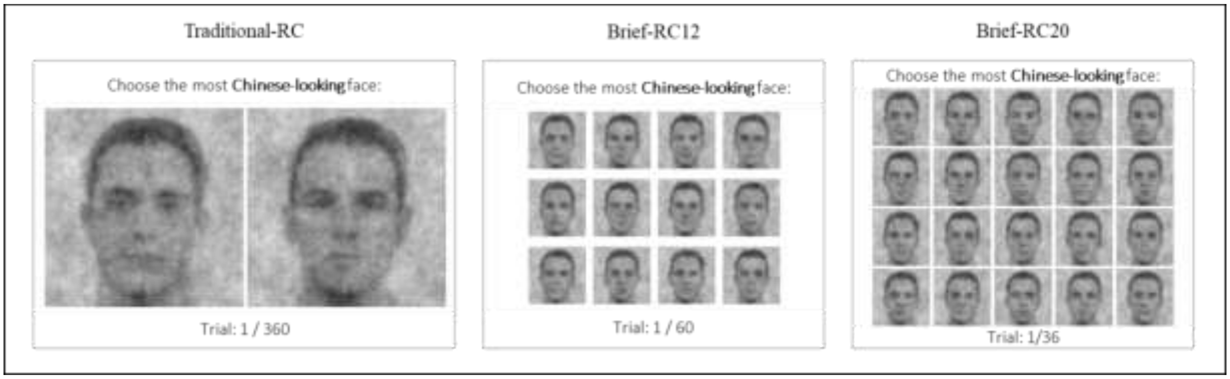
All participants received the following instruction: "*In this task, you will be presented with a series of noisy faces. Your task will be to 'Choose the most Chinese-looking face' in each trial. Use your mouse to make a choice. Note that the faces may look similar to each other, yet they are different. 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

[2] We used the development version in order to obtain the noise matrices, which ease the computation of the CIs and infoVal. The development version is available at https://github.com/rdotsch/rcicr/tree/development

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. At the end, participants provided demographic information through a Qualtrics survey before being debriefed.

Figure 3

Illustration of a single trial from 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



Traditional-RC

Choose the most **Chinese-looking** face:

Trial: 1 / 360

Brief-RC12

Choose the most **Chinese-looking** face:

Trial: 1 / 60

Brief-RC20

Choose the most **Chinese-looking** face:

Trial: 1/36

We computed each participant's individual CI by averaging all the noise patterns extracted from the selected noisy faces and superimposing it 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

[3] We adapted the function from the *rcicr* package (Dotsch, 2017) to compute the CIs in our experiments in order to better handle large noise matrices.

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 the 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 (CIs position was randomized between participants). Next, judges provided demographic information, had the option to make a comment about the study, and were debriefed.

Results

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 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* <

[4] The following constants were used to scale the noise patterns for the average CIs: Traditional-RC = .0055, Brief-RC12 = .0130, Brief-RC20 = .0150; and for the individual CIs: Traditional-RC = .0190, Brief-RC12 = .0450, Brief-RC20 = .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, Goffin et al. (2019; 1 = “not masculine”, 9 = “very masculine”).

.001, η_p^2 = .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, η_p^2 = .011, BF_{01} = 5.67.

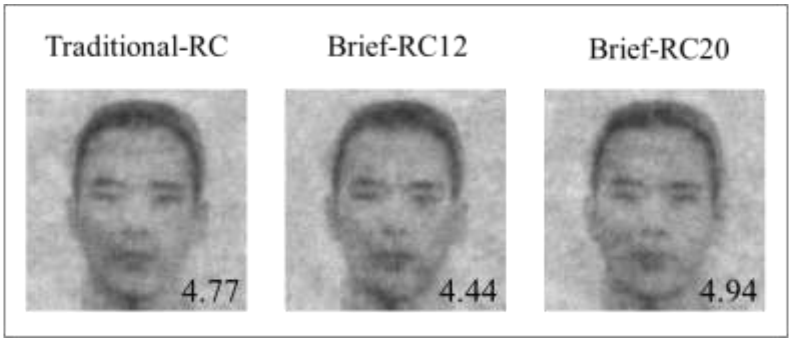
Judges' ratings

We submitted judges' ratings on the Chinese-looking scale of the individual CIs to C_1 and C_2 in a LMM analysis, using judges and producers as random factors. In line with expectations, the individual CIs 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, η_p^2 = .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, η_p^2 = .001, BF_{01} = 59.89.

Next, we submitted judges' ratings of the average CIs to C_1 and C_2 in an OLS analysis. 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, η_p^2 = .001, BF_{01} = 13.38. Interestingly, 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, η_p^2 = .023 (see Figure 4).

Figure 4

Average CIs ratings by condition. Average ratings are displayed in the lower-right corner of the CIs



InfoVal

We submitted the infoVal[6] 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), $F(1, 67) = 0.34$, $p = .560$, $\eta_p^2 = .005$, 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), $F(1, 67) = 0.10$, $p = .748$, $\eta_p^2 = .002$, BF_{01} = 7.69.

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 (Traditional-RC) after participants had been exposed to the same number of stimuli (720 noisy faces). Neither the individual nor the average CIs 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 and only

[6] We did not declare a priori hypothesis concerning infoVal for Experiment 1 in the preregister since the reference for this metric (Brinkman, Goffin et al., 2019) 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.

required as little as 60 or 36 trials. A researcher with limited resources would thus have obvious interest in opting for the Brief-RC rather than for 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 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.

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 (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, who deliver the individual CIs, and judges.

Method

Participants

Because our design involved linear mixed models, with produces (i.e., individual CIs) and judges as random factors, we relied on Judd, Westfall, and Kenny (2012; Judd et al., 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 opted for 100 produces (and thus 100 individual CIs) per condition and 250 judges. We thus recruited a sample of 300 producers (n = 100 per condition; $M_{\text{age}} = 31.30$, $SD_{\text{age}} = 10.35$; 169 males, 4 others) and another sample of 253 independent judges. We excluded one judge who gave the same rating to all faces[7]. The final sample thus comprised 252 judges ($M_{\text{age}} = 30.50$, $SD_{\text{age}} = 10.04$; 126 females, 3 others). Sample size was determined before any data analysis. 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).

Experimental design and procedure

The overall procedure was similar to 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 two thousand pairs of noisy faces (stimuli) from a different base image and a different seed than Experiment 1 for generalizability purposes. Specifically, we relied on the same base image as in Dotsch et al. (2008). The

[7] It was established in the preregister 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 second, and 94.44% of judges had rating durations shorter than 1 second/stimulus. Given that this exclusion criterion was underestimated, we omitted it.

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 belonged to a separate study. Producers then gave 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. Finally, they provided the same demographic information, had the option to make 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 (approximatively 5 and 10 minutes), number of trials (90 and 167), and number of stimuli presented (approximatively 1050). Table 1 presents the 11 distinct points of comparison[8].

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

[8] We deviated slightly from the points of comparison established in our preregister to minimize the number of CIs and judgments required. To do so, we selected points that could be used to compare more than one criterion (see Table 1). Importantly, these points were determined *before* computing the visual renderings.

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

We created the individual and average CIs with the auto-scale method that matches the noise pattern to the range of pixel 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 the same 55 randomly sampled individual CIs (5 CIs randomly sampled from each of the 11 distinct points of comparison; see Table 1) as 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; see Table 1). 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 twenty seconds before rating each of them. In the rating task, average CIs appeared one by one and in random order across participants. Next, judges provided demographic information, had the possibility to make a comment about the study, and were debriefed.

Results

Judges' ratings as a function of time

We submitted judges' ratings of the individual CIs to a LMM analysis with C_1 (Traditional-RC versus Brief-RC12 and Brief-RC20), C_2 (Brief-RC12 versus Brief-RC20), time (5 versus 10 minutes) and their interaction 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), $F(1, 319.34) = 13.71$, $p < .001$, $\eta_p^2 = .010$. 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), $F(1, 247.84) = 19.14$, $p < .001$, $\eta_p^2 = .001$. Moreover, the C_1 × time interaction was significant, $F(1, 247.44) = 5.73$, $p = .017$, $\eta_p^2 < .001$. The 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, $F(1, 341.57) = 9.27$, $p =.003$, $\eta_p^2 = .005$, this pattern being even more pronounced after 10 minutes ($M_{traditional\text{-}RC} = 2.31$, $SD_{traditional\text{-}RC} = 2.00$; $M_{Brief\text{-}RC12\&Brief\text{-}RC20} = 2.92$, $SD_{Brief\text{-}RC12\&Brief\text{-}RC20} = 2.10$), $F(1, 336.07) = 17.99$, $p < .001$, $\eta_p^2 = .010$. There was no significant effect of C_2 nor of C_2 × time (all p's > .10, BF_{01}'s > 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), $F(1, 246) = 19.74$, $p < .001$, η_p^2 =

.013. There was no significant effect of C_2, time, C_1 × time, nor C_2 × time (all *p*'s > .10, BF_{01}'s > 10).

Figure 5 (panel 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.

Figure 5

Relation between individual CIs ratings (averaged) and (A) the time (median cumulative time in minutes), (B) the number of trials, and (C) the number of stimuli, as a function of condition (different colored 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, i.e., (A) at approximately 5 and 10 minutes, (B) at 90 and 167 trials, and (C) at approximately 1050 stimuli.

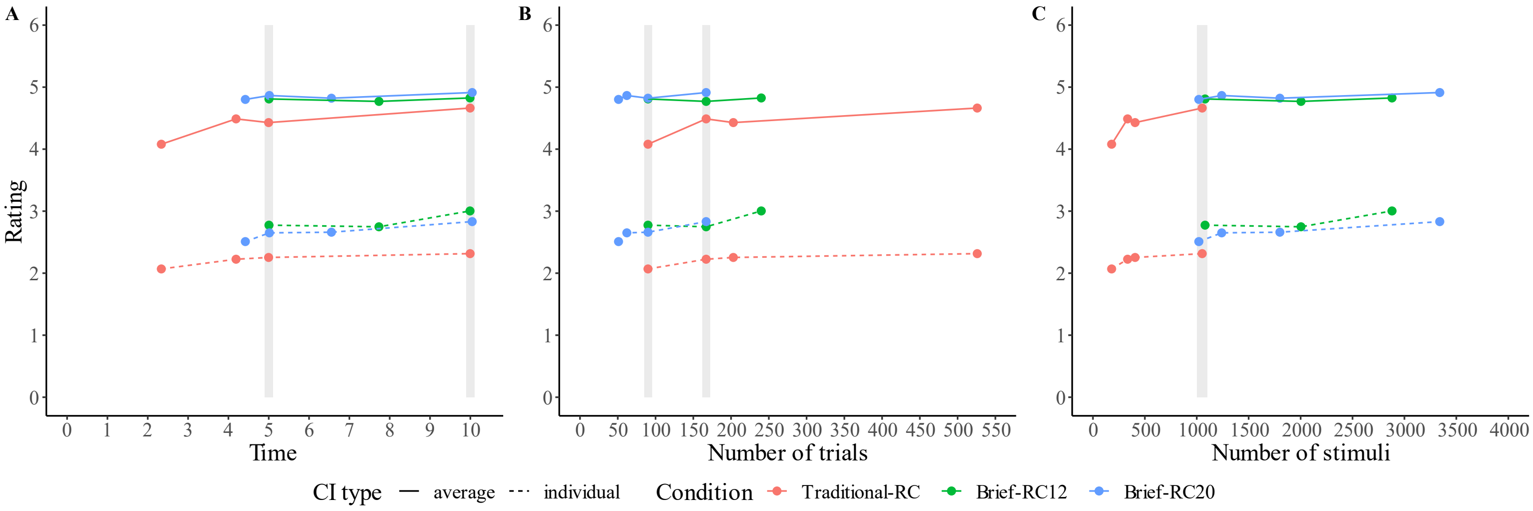
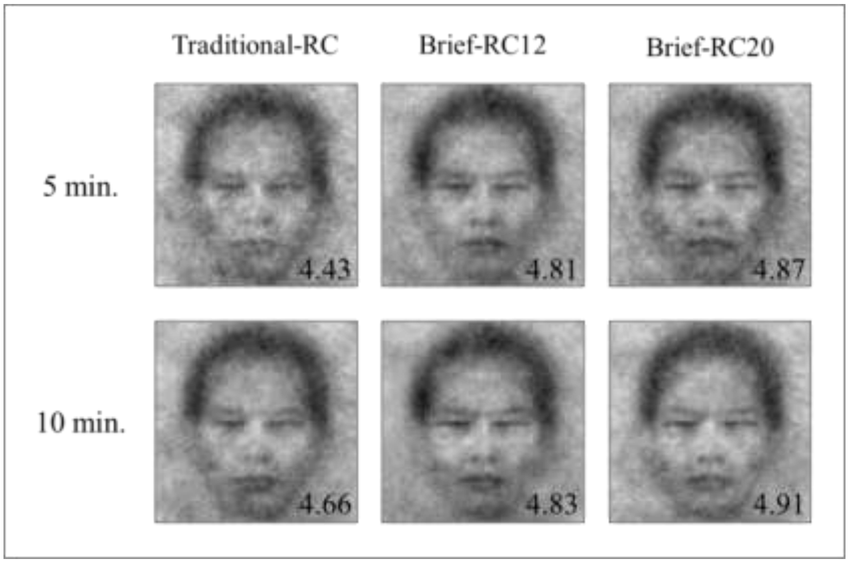


Figure 6

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



Judges' ratings as a function of number of trials

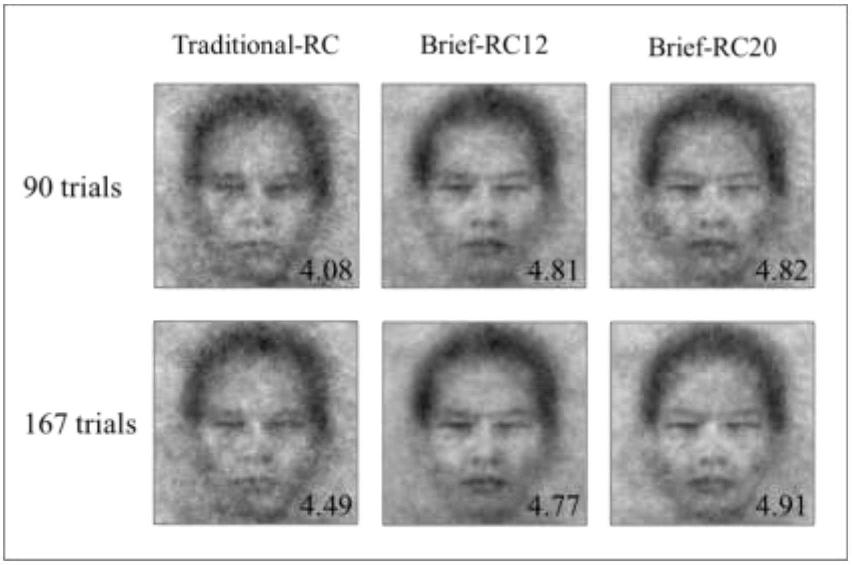
We submitted judges' ratings of the individual CIs to a LMM analysis with C_1, C_2, trials (90 versus 167 trials), and their interaction as predictors, using judges and producers as random factors. As predicted, there was a main effect of C_1 such that the individual CIs from the Traditional-RC (M = 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, η_p^2 = .020. Unsurprisingly, 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, η_p^2 = .001. Moreover, there was a significant C_2 × trials interaction F(1, 245.72) = 4.67, p = .032, η_p^2 < .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, η_p^2 < .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, η_p^2 < .001. There were no significant effects of C_2 nor of C_1 × trials (all p's > .10, BF_{01}'s > 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, $p < .001$, $\eta_p^2 = .039$. 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, $p = .020$, $\eta_p^2 = .004$. Moreover, there was a significant C_1 × trials interaction *F*(1, 246) = 7.46, $p = .007$, $\eta_p^2 = .005$. 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, $p < .001$, $\eta_p^2 = .035$, this in a more pronounced manner than at 167 trials ($M_{traditional\text{-}RC} = 4.49$, $SD_{traditional\text{-}RC} = 1.36$; $M_{brief\text{-}RC12\&Brief\text{-}RC20} = 4.84$, $SD_{brief\text{-}RC12\&Brief\text{-}RC20} = 1.22$), *F*(1, 246) = 12.69, $p < .001$, $\eta_p^2 = .008$. There were no significant effects of C_2, nor C_2 × trials (all *p*'s > .10, BF_{01}'s > 10).

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

Figure 7

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



Judges' ratings as a function of number of stimuli

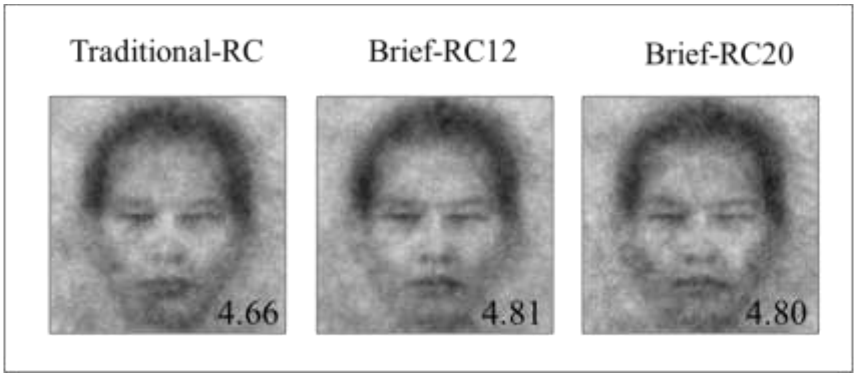
We submitted judges' ratings of the individual CIs produced after producers had been exposed to approximatively 1050 stimuli (noisy faces) to a LMM analysis with to C_1 and C_2 as predictors, using judges and producers as random factors. There was a main effect of C_1 such that the individual CIs from 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, η_p^2 = .005. There was no significant effect of C_2, *F*(1, 326.76) = 2.52, *p* = .113, η_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, η_p^2 = .038. There was no significant effect of C_2, *F*(1, 249) = 0.91, *p* = .341, η_p^2 = .001, BF_{01} = 24.68.

Figure 5 (panel C) shows the results from judges' ratings of the individual and average ratings of CIs as a function of 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 the Supplementary Materials (Table S1).

Figure 8

Average CIs produced after producers were exposed to approximatively 1050 stimuli (noisy faces). Average ratings are displayed in the lower-right corner of the CIs



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 (versus 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), $F(1, 294) = 0.61$, $p = .434$, $\eta_p^2 = .001$, $BF_{01} = 17.99$. 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), $F(1, 294) = 16.89$, $p < .001$, $\eta_p^2 = .028$. There was no significant effect of C_2, C_1 × time, or C_2 × time (all p's > .10, BF_{01}'s > 10). Figure 9 (panel A) shows the averaged infoVal scores of individual CIs as a function of time.

InfoVal as a function of number of trials

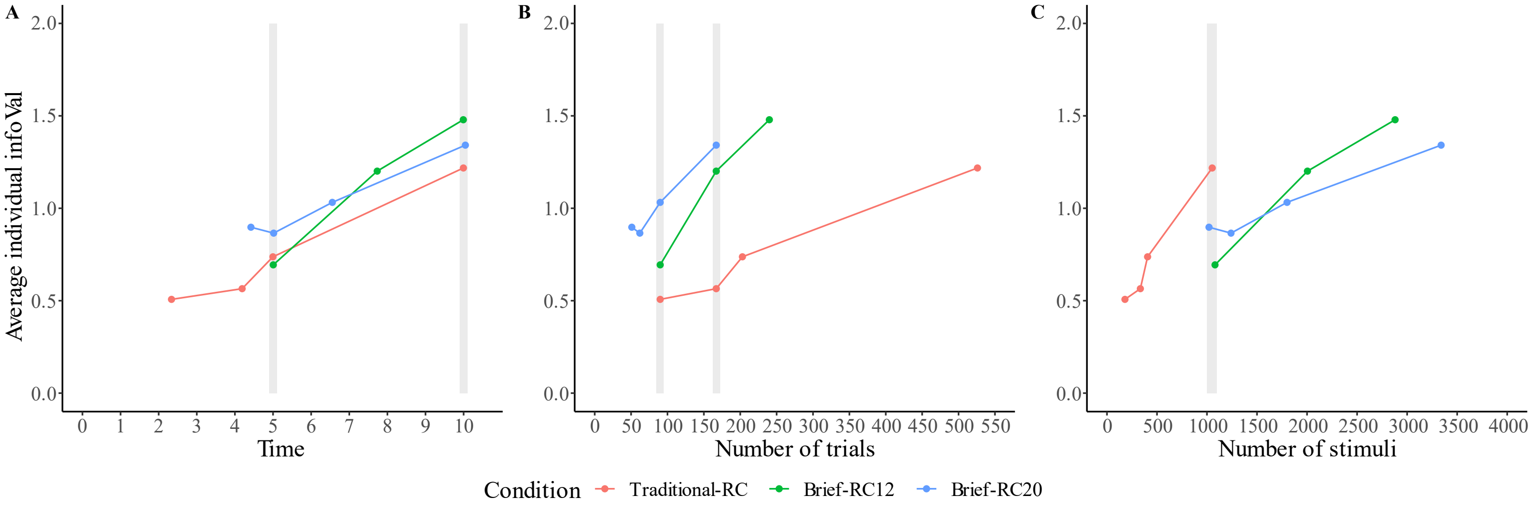
We submitted the infoVal scores of the individual CIs to an OLS analysis with C_1, C_2, trials (90 vs. 167 trials), and their interactions. C_1 was significant such that infoVal scores from the Traditional-RC (*M* = 0.54, *SD* = 1.24) were lower than those from the Brief-RC12 and Brief-RC20 (*M* = 1.07, *SD* = 1.70), *F*(1, 294) = 15.60, *p* < .001, η_p^2 = .026. We also found a trials main effect in that 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, η_p^2 = .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, η_p^2 = .004. There was no significant effect of C_1 × trials, nor C_2 × trials (all *p*'s > .10, BF_{01}'s > 10). Figure 9 (panel B) shows the averaged infoVal scores of individual CIs as a function of number of trials.

InfoVal as a function of number of stimuli

We submitted the infoVal scores of the individual CIs produced after producers had seen approximatively 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) was higher than those for the Brief-RC12 and Brief-RC20 (*M* = 0.80, *SD* = 1.41), *F*(1, 297) = 4.73, *p* = .030, η_p^2 = .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, η_p^2 = .003, BF_{01} = 11.43. Figure 9 (panel C) shows the averaged infoVal scores of individual CIs as a function of number of stimuli. Means and standard deviations of the infoVal scores of individual CIs for each of the experimental cells are available in the Supplementary Materials (Table S2).

Figure 9

Relation between the average infoVal (information value) of individual CIs and (A) the time (median cumulative time in minutes), (B) the number of trials, and (C) the number of stimuli, as a function of condition (different colored 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.



Discussion

As a reminder, Experiment 1 revealed that when participants saw the same number of stimuli, the Brief-RC performed as well as the Traditional-RC. Experiment 2 sought to replicate these findings while taking into account additional criteria that prove relevant when designing RC-based studies, namely the time of completion and the number of trials. Moreover, we increased the overall length of task and increased the statistical power to maximize the reliability of comparisons.

In line with our expectations, subjective ratings from independent judges revealed that the Brief-RC outperformed the Traditional-RC on all three criteria. With respect to time, judges evaluated individual CIs as more prototypical of the social target (i.e., a Chinese-looking face) after only 5 minutes and even more so after 10 minutes, suggesting that this trend may further improve over time. As for the number of trials, the ratings were also more satisfactory with the new as opposed to the traditional method and, this, whether participants had completed 90 or 167 trials. In contrast with Experiment 1, Experiment 2 revealed that the Brief-RC was not as good but did in fact surpass the Traditional-RC after participant had seen to the same number of stimuli (i.e., noisy faces). To be sure, the latter comparison entailed a large number of stimuli in Experiment 2 than in Experiment 1. Finally, the average CIs also came across as more prototypical on all the 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.

Taken together, these findings clearly suggest that the Brief-RC produces superior individual and average CIs at least as assessed on a subjective scale.

General Discussion

The reverse correlation (RC) is an innovative data-driven method designed to capture visual representations (or so-called classification images; CIs) of social targets. The procedure has proven very useful to uncover various important phenomena such as the top-down distortions of facial representation of social groups (e.g., Brown-Iannuzzi et al., 2016; Brown-Iannuzzi et al., 2018; Dotsch et al., 2008; Hinzman & Maddox, 2017). One issue, however, is that past research relied almost solely on ratings from average CIs, a strategy that may drastically increase Type I errors (Cone et al., 2020). This state of affairs poses a serious threat in terms of the replicability of psychological science (e.g., see the Open Science Collaboration, 2015; Simmons et al., 2011). An obvious solution is the use of individual CIs (Cone et al., 2020). This is even more tempting that individual CIs also pave the way to more fine-grained analyses (e.g., analyzing the relation between racial bias and visual representation bias at the participant-level; Dotsch et al., 2008).

However, 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, Goffin et al., 2019; Todorov et al., 2011). For these reasons, researchers called for an improvement of the 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 to 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 (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. Our findings reveal that individual CIs produced from both methods performed similarly, as assessed by judges (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) (Brinkman, Goffin et al., 2019). 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 to 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 to produce CIs that are more robust. Finally, we also increased statistical power. 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. One reasonable account for the difference between the two experiments may reside in the fact that, compared to 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. It should also be noted that Experiment 2 had more statistical power than Experiment 1.

Importantly, the Brief-RC variants also outperformed the Traditional-RC after the same time length (after 5 or 10 minutes) and after the same number of trials (after 90 or 167 trials). Moreover, these results replicated when considering the average CIs. As for infoVal, although more exploratory, the two methods performed on similar levels when considering the same time length, whereas the Brief-RC scored better when comparing across number of trials, but worse when comparing across 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 investigations. 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 were the faces where 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.

An interesting finding from Experiment 2 stems from the visual inspection of Figure 5. Indeed, the gap between judgments 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 an 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 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 hundreds (Cone et al., 2020; Imhoff et al., 2011; Imhoff et al., 2013).

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 judgments. This is an important accomplishment because it should allow researchers to run RC-based studies in a much more efficient manner. A notable dividend is that the Brief-RC may grant an access to higher quality individual CIs thus paving the way for more fine-grained analyses with other variables of interest (e.g., inter-individual differences). 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).

As anticipated, the 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. Indeed, one might conjecture that an ideal observer should perform better as the number of stimuli per trial increases from 12 to 20 per trials. 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 (Marois & Ivanoff, 2005; Palmer, 1990). In light of the present findings, the Brief-RC12 should be preferred over the Brief-RC20 as it requires less stimuli to achieve a similar result. However, our research was limited to two version of the Brief-RC (Brief-RC12 and Brief-RC20) and further work should investigate the optimal configuration in terms of number of stimuli per trial (e.g., 4, 6, 8, 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 stimulus (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 in external validity because they derive from an exponentially larger sample of stimuli[9] 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., color, depth, texture, or higher dimensional noise patterns; Brinkman et al., 2017; Jack & Schyns, 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.

In conclusion, the current work introduces an improved reverse correlation method, namely the Brief-RC. This new 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. The access to more robust and better-quality individual CIs is key to address inflated Type I errors rates prevalent in traditional two-phase reverse correlations procedures. Moreover, the Brief-RC significantly reduces the overall task length, which offers clear-cut benefits in terms of maintaining participant's motivation, but also in terms of economic resources and time constraints for the researcher.

[9] 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 ($8 + 8 \times 215 = 1728$).

Open Practices

The data, R code, and materials are publicly available on Open Science Framework (https://osf.io/ps9wu/?view_only=028597d60e7342bfaeb6881051cb6bca).

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

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

Brinkman, L., Dotsch, R., Zondergeld, J., Koevoets, M. G. J. C., Aarts, H., & van Haren, N. E. M. (2019). 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. (2019). 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*, 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. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology, 100*(6), 999–1014. https://doi.org/10.1037/a0023026

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

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

Gunaydin, G., & DeLong, J. E. (2015). Reverse Correlating Love: Highly Passionate Women Idealize Their Partner’s Facial Appearance. *PLOS ONE*, *10*(3), e0121094. https://doi.org/10.1371/journal.pone.0121094

Gundersen, A. B., & Kunst, J. R. (2018). Feminist ≠ Feminine? Feminist Women Are Visually Masculinized Whereas Feminist Men Are Feminized. *Sex Roles*, *80*(5–6), 291–309. https://doi.org/10.1007/s11199-018-0931-7

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

Hong, Y., & Ratner, K. G. (2020). Minimal but not meaningless: Seemingly arbitrary category labels can imply more than group membership. *Journal of Personality and Social Psychology*, 1–25. https://doi.org/10.1037/pspa0000255

Imhoff, R., & Dotsch, R. (2013). Do We Look Like Me or Like Us? Visual Projection as Self- or Ingroup-Projection. *Social Cognition*, *31*(6), 806–816. https://doi.org/10.1521/soco.2013.31.6.806

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

Jackson, J. C., Hester, N., & Gray, K. (2018). The faces of God in America: Revealing religious diversity across people and politics. *PLOS ONE*, *13*(6), e0198745. https://doi.org/10.1371/journal.pone.0198745

Jaeger, B. (2017). r2glmm: Computes R Squared for Mixed (Multilevel) Models (R package version 0.1.2). Retrieved from https://CRAN.R-project.org/package=r2glmm

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: 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

Karremans, J. C., Dotsch, R., & Corneille, O. (2011). Romantic relationship status biases memory of faces of attractive opposite-sex others: Evidence from a reverse-correlation paradigm. *Cognition*, *121*(3), 422–426. https://doi.org/10.1016/j.cognition.2011.07.008

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

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

Krosch, A. R., & Amodio, D. M. (2014). Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences*, *111*(25), 9079–9084. https://doi.org/10.1073/pnas.1404448111

Kunst, J. R., Dovidio, J. F., & Dotsch, R. (2017). White Look-Alikes: Mainstream Culture Adoption Makes Immigrants “Look” Phenotypically White. *Personality and Social Psychology Bulletin*, *44*(2), 265–282. https://doi.org/10.1177/0146167217739279

Kunst, J. R., Kteily, N., & Thomsen, L. (2017). “You Little Creep”: Evidence of Blatant Dehumanization of Short Groups. *Social Psychological and Personality Science*, *10*(2), 160–171. https://doi.org/10.1177/1948550617740613

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

Lin, C., Adolphs, R., & Alvarez, R. M. (2018). Inferring Whether Officials Are Corruptible From Looking at Their Faces. *Psychological Science*, *29*(11), 1807–1823. https://doi.org/10.1177/0956797618788882

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

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

Petsko, C. D., Lei, R. F., Kunst, J. R., Bruneau, E., & Kteily, N. (in press). Blatant dehumanization in the mind's eye: Prevalent even among those who explicitly reject it? *Journal of Experimental Psychology: General.* Retrieved from https://psyarxiv.com/g7w4b/

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

Schmitz, M., Rougier, M., & Yzerbyt, V. (2020). 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. (2020). 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

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

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

Supplementary Materials

Table 1S

Ratings of Individual and Average CIs as a function of Time (5 vs. 10 minutes), Number of trials (90 vs. 167 trials), and Number of stimuli (approximatively 1050 stimuli)

	Time		Number of trials		Number of stimuli
Classification images	5 minutes	10 minutes	90 trials	167 trials	~ 1050 stimuli
Individual CIs					
Traditional-RC	2.25(1.90)	2.31(2.00)^{a}	2.07(1.82)	2.23(1.92)	2.31(2.00)^{a}
Brief-RC12	2.77(2.01)^{b}	3.00(2.05)	2.77(2.01)^{b}	2.75(2.01)	2.77(2.01)^{b}
Brief-RC20	2.65(2.03)	2.83(2.14)^{c}	2.66(2.04)	2.83(2.14)^{c}	2.51(2.02)
Average CIs					
Traditional-RC	4.43(1.46)	4.66(1.29)^{d}	4.08(1.50)	4.49(1.36)	4.66(1.29)^{d}
Brief-RC12	4.81(1.12)^{e}	4.83(1.24)	4.81(1.12)^{e}	4.77(1.20)	4.81(1.12)^{e}
Brief-RC20	4.87(1.23)	4.91(1.25)^{f}	4.82(1.25)	4.91(1.25)^{f}	4.80(1.23)

Note. Standard deviations are in parenthesis. Cells with the same superscript correspond to the same average CI or set of individual CIs. Each cell of individual CIs corresponds to the ratings of 100 individual CIs which received a total of n = 2016 ratings (each individual CI was rated, on average, 20 times). Each cell of average CI corresponds to a single average CI with n = 252 ratings.

Table 2S

InfoVal scores of Individual CIs as a function of Time (5 vs. 10 minutes), Number of trials (90 vs. 167 trials), and Number of stimuli (approximatively 1050 stimuli)

	Time		Number of trials		Number of stimuli
Classification images	5 minutes	10 minutes	90 trials	167 trials	~ 1050 stimuli
Traditional-RC	0.74(1.35)	1.22(1.89)[a]	0.51(1.23)	0.56(1.24)	1.22(1.89)[a]
Brief-RC12	0.69(1.49)[b]	1.48(2.17)	0.69(1.49)[b]	1.20(1.74)	0.69(1.49)[b]
Brief-RC20	0.87(1.46)	1.34(1.87)[c]	1.03(1.63)	1.34(1.87)[c]	0.90(1.32)

Note. Standard deviations are in parenthesis. Cells with the same superscript correspond to the same set of individual CIs. Each cell corresponds to *n* = 100 infoVal scores (from 100 individual CIs).