Behav Res (2015) 47:736–743
DOI 10.3758/s13428-014-0497-4

Reliability of composite-task measurements of holistic face processing

David A. Ross · Jennifer J. Richler · Isabel Gauthier

Published online: 25 June 2014

Abstract There is growing interest in the study of individual differences in face recognition, including one of its hallmarks, holistic processing, which can be defined as a failure of selective attention to parts. These efforts demand that researchers be aware of, and try to maximize, the reliability of their measurements. Here we report on the reliability of measurements using the composite task (complete design), a measure of holistic processing that has been shown to have relatively good validity. Several studies have used the composite task to investigate individual differences, yet only one study has discussed its reliability. We investigate the reliability of composite-task measurements in eight data sets from five different samples of subjects. In general, we found reliability to be fairly low, but there was substantial variability across experiments. Researchers should keep in mind that reliability is a property of measurements, not of a task, and think about the ways in which measurements in a particular task may be improved before embarking on individual differences research.

Keywords Face · Individual differences · Reliability · Holistic processing

Faces, unlike many of the visual stimuli that we encounter each day, are processed holistically. That is, we are unable to attend to a single face part when it is presented in the context of a whole face; rather, numerous behavioral studies have revealed interference from the ostensibly unattended face parts. In a seminal article, Young, Hellawell, and Hay (1987) found that subjects were slower to identify a face half (e.g., top or bottom) when it was aligned with the complementary half of a different face than when the two halves were misaligned. In contrast, in similar paradigms but in the absence of extensive experience or training, such effects are not seen for other visual object categories: That is, subjects can selectively attend to parts of these objects without difficulty. Holistic processing may offer an insight into what is special about face recognition (McKone, Kanwisher, & Duchaine, 2007; Richler, Cheung, & Gauthier, 2011b), explain what is different about populations with particularly poor face recognition abilities (e.g., Busigny, Joubert, Felician, Ceccaldi, & Rossion, 2010; Gauthier, Klaiman, & Schultz, 2009; Palermo et al., 2011), or shed light on what is learned when we become experts with a particular category of objects (Gauthier, Williams, Tarr, & Tanaka, 1998; Richler, Wong, & Gauthier, 2011; Wong, Palmeri, & Gauthier, 2009).

One issue that has complicated progress in understanding holistic processing is that numerous different meanings and measures of holistic processing have been used in the literature (see Richler, Palmeri, & Gauthier, 2012, for a review). Perhaps the best validated of these measures is the complete design of the composite task[1] (Gauthier & Bukach, 2007). Like Young et al.'s (1987) task, described above, this task involves making decisions about one half of a composite face while ignoring the other half. Faces are presented sequentially: A study face composed of two face halves is shown, followed by a blank screen or mask, and then a test face. The task is to indicate by buttonpress whether the cued half of the test face is the same as, or different from, the study face. Both the cued and the to-be-ignored face parts can be associated with the

[1] A version of the composite task featuring half of the trials has been referred to as the "partial design" in the literature (Gauthier & Bukach, 2007); however, the partial design confounds response and congruency (all *same* trials are incongruent and all *different* trials are congruent), which has been found to be problematic, because congruency produces a response bias that cannot be separated from sensitivity without the other half of the trials (see Cheung, Richler, Palmeri, & Gauthier, 2008; Richler, Cheung, & Gauthier, 2011a; Richler & Gauthier, in press).

D. A. Ross (✉) · J. J. Richler · I. Gauthier
Department of Psychology, Vanderbilt University, PMB 407817, 2301 Vanderbilt Place, Nashville, TN 37240-7817, USA
e-mail: david.ross@vanderbilt.edu

same correct response (both parts same or both parts different; congruent trials) or with different correct responses (one part same, the other part different; incongruent trials). Holistic processing is defined by an interaction between congruency and alignment. On congruent trials, a failure to selectively attend to the cued part should facilitate performance, whereas on incongruent trials, a failure to selectively attend to the cued part should hinder performance. This difference in performance between congruent and incongruent trials is reduced on misaligned trials (e.g., Richler, Tanaka, Brown, & Gauthier, 2008). Finally, it is worth noting that, although we refer to the observed behavior as a failure of selective attention, this does not necessarily imply that the mechanism is attentional; the congruency effect observed in this task may be the result of attentional (Chua, Richler, & Gauthier, 2014) or perceptual (Rossion, 2013) mechanisms.

The complete design is a good candidate for a measure of holistic processing, since not only does it have good face validity, but it also captures an effect that is specific to faces and objects of expertise (e.g., Gauthier et al., 1998; Richler, Wong, & Gauthier, 2011; Wong et al., 2009). In addition, studies that have investigated the relationship between face recognition ability, measured with the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), and holistic processing measured with the complete design have reported a robust relationship between the two (DeGutis, Wilmer, Mercado, & Cohan, 2013; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012; Richler, Cheung, & Gauthier, 2011b), again suggesting that the task measures something that is relevant to understanding face recognition.

However, beyond the issue of validity, the conclusions in individual differences studies must be evaluated in the light of another basic concern, that of the reliability of the measurements. There are several reasons to think that the reliability of holistic processing measured using the composite task may be fairly low. First, the measure of holistic processing, Congruency × Alignment, is calculated by subtracting the difference between congruent and incongruent misaligned trials from the difference between congruent and incongruent aligned trials (e.g., Richler, Cheung, & Gauthier, 2011b). Unfortunately, difference scores are often less reliable than the scores from which they are computed (Peter, Gilbert, Churchill, & Brown, 1993)—though this depends on the reliability, and the relative variability, of the constituent measures (Malgady & Colon-Malgady, 1991; Rogosa, Brandt, & Zimowski, 1982). Second, in addition to there being only one published report of the reliability of measurements using the complete design of the composite task (DeGutis et al., 2013), the reliability calculated there was low enough to raise concerns for anyone interested in doing individual differences work (Guttman's λ_2 value of .10).[2] In that case, the authors were looking for a relationship between subjects' CFMT score and their composite task scores. Fortunately, the CFMT was highly reliable (λ_2 = .88), which together led to a theoretical upper limit for the correlation between them, computed as the geometric mean of the two reliabilities (Schmidt & Hunter, 1996), of ~.3. Such low reliability could be prohibitive if researchers wish to relate composite-task performance with anything less reliable. It would also reduce the power to find such relationships, as well as the power to find a difference between this and other relationships.

DeGutis et al. (2013) suggested one method that might improve the reliability of these measurements: Rather than using a difference of differences (i.e., aligned congruency effect – misaligned congruency effect), they suggested that it might be better to use the residuals (i.e., to regress the congruency effect for misaligned trials out of the congruency effect for aligned trials). Their reasoning, which was driven mainly by considerations of validity, was that the misaligned condition is a control condition that should not contribute as much as the aligned condition to the final estimates of holistic processing—rather, it would make more sense to partial out the variability in the misaligned condition, presumed to be irrelevant to holistic processing, which might be present in both conditions. More importantly for the present discussion, DeGutis et al. found that the reliability of the regression measure of holistic processing (λ_2 = .24) was higher than that of the difference measure.

Here, we aimed to investigate the reliability of the composite task measure of holistic processing in five experiments (eight data sets) from our lab. We also aimed to explore whether regression always provides a more reliable measure of holistic processing than do difference scores.

Method

Studies

We used the data from five different experiments, with four of these experiments including two independent measurements of holistic processing from the same sample of subjects and taken from blocks of trials separated by a few minutes (Table 1). The first three experiments came from unpublished data sets and used task parameters that were relatively standard and commensurate with the several dozens of uses of this task in the literature. In contrast, Experiments 4 and 5 used modified designs with the explicit motivation to explore whether such changes might produce better reliability. Specifically, in Experiment 4 we manipulated the contrast of the tops and bottoms of faces (see details in the Stimuli section) with the aim of getting more variability in each trial, in order to better cover the range of holistic processing ability

[2] Note that this reliability was computed for 144 trials, and we report reliabilities for 160 trials in the present work.

Table 1 Demographics and experimental design for each of the five experiments

Stimulus Set	*N*	Age (*SD*)	Sex (M/F)	Number of Faces	Fixation (ms)	Study (ms)	Mask (ms)	Test (ms)
Experiment 1								
MPI female G&R	85	21.0 (3.4)	33/52	20	200	200	2,000	200
Experiment 2								
MPI female MPI female	67	21.3 (3.3)	25/42	20	200	200	(blank) 500	200
Experiment 3								
MPI female G&R	91	22.5 (4.3)	27/39	20	200	200	(blank) 500	200
Experiment 4								
MPI female MPI male	53	24.5 (6.8)	18/35	20	200	200	1,000	150 (or less)
Experiment 5								
MPI female	43	19.3 (2.1)	6/33	5	200	200	(blank) 500	200

G&R, Goffaux and Rossion (2006)

in the population (see Fig. 1). We hypothesized that varying the contrast of the top or bottom would make it more or less difficult to selectively attend to a given part, by reducing or increasing its saliency. For example, it might be easier to attend to the top part of a face (i.e., to ignore the bottom part) if the bottom part were low in contrast—that is, it might be less distracting. Likewise, attending to the top part might be more difficult if it were low and the bottom were high in contrast. Thus, we hypothesized that this manipulation could add more variability into the trials, such that on some trials almost everybody might show some evidence of holistic processing, and on other trials only a subset of people would show evidence of holistic processing. Indeed, whereas such variability in trials is not usually a consideration for group analyses, it is central to the idea of measuring individual differences. In Experiment 5, we restricted the number of different face parts (tops and bottoms) to five, rather than 20, as in the other experiments. The rationale was that there might be variability in learning across subjects, over the course of an experiment, and that using fewer faces might cause subjects' performance to asymptote more quickly, reducing the influence of such effects. The various task parameters are listed in Table 1 and discussed below.

Stimuli

The face images in each experiment were drawn either from the Max Planck Institute Database (MPI; Troje & Bülthoff, 1996) or from the set of 20 face stimuli used by Goffaux and Rossion (2006). Both the male and female MPI faces were used, but not within every experiment (see Table 1). In the Goffaux and Rossion stimulus set, the male and female faces were intermixed within a single study block. In all cases, the stimuli were converted to grayscale and cut in half to produce a set of top and bottom halves, which were randomly recombined on every trial. The faces were also all cropped to remove hair; in the case of the MPI faces, they retained some external features (ears, jaw shape, neck), whereas the Goffaux and Rossion faces had all external features removed. Recombined faces were presented on a gray background with a white line separating the top and bottom halves. Misaligned stimuli were created by horizontally offsetting the top and bottom halves, such that the edge of one half fell approximately in the center of the other half (see Table 1 for more details about the stimuli and design in each experiment).

In addition to this standard stimulus construction, the top and bottom face parts used in Experiment 4 were adjusted to

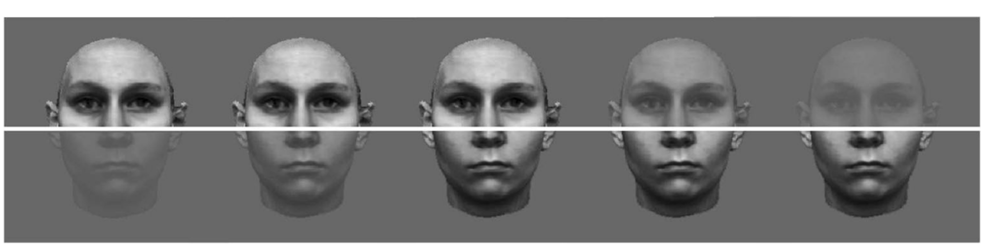

Fig. 1 Sample stimuli used in Experiment 4

create three different levels of contrast (full, medium, and low). These parts were then combined to create five possible stimulus types: top-full/bottom-low, top-full/bottom-med, top-full/bottom-full, top-med/bottom-full, and top-low/bottom-full (see Fig. 1). As we described earlier, this manipulation was designed so that some trials (when the relevant part was lower in contrast) would provide more information to discriminate subjects with a presumed low degree of holistic processing, whereas other trials (in which the relevant part was higher in contrast) might provide more information to discriminate subjects with a presumed high degree of holistic processing. In Experiment 5, all faces were shown with a single level of contrast, as in Experiments 1 to 3, but we used only five top halves and five bottom halves for the entire experiment. Reducing the set of face parts was done in the hope that this would reduce spurious variability attributable to subjects learning to identify the larger set of face parts over the course of the experiment.

Procedure

For all experiments, on each trial, a fixation cross was presented (200 ms), followed by a study face (200 ms) and then either a mask or a blank screen for 500 to 2,000 ms, depending on the experiment (see Table 1). Next, a test face was presented for 200 ms, with the exception of Experiment 5, in which the test face was presented for 150 ms (this was done to try to avoid ceiling, since the learning of individual parts could be faster in this experiment). In all cases, subjects were instructed to respond by keypress if the top part of the test face matched the top part of the study face. In Experiments 1, 2, 3, and 4, two blocks were based on different stimuli (see Table 1). The blocks were separated by a screen instructing participants that they could take a break if they wished (the order of blocks was the same for all participants). Since the number of trials influences reliability and all of the reported experiments contained at least 160 trials, we decided to analyze only the first 160 trials in each experiment (this was also close to the 144 trials used by DeGutis et al., 2013), with 40 trials in each of the experimental conditions (congruent–aligned, incongruent–aligned, congruent–misaligned, and incongruent–misaligned).

Results

We first checked each data set, removing any univariate or multivariate outliers and confirming that at the group level all five of our experiments showed the typical Congruency × Alignment interaction (see Table 2). The present effect sizes are comparable to those in prior studies (a meta-analysis of 48 experiments showed an effect size for versions of this task for upright faces of $\text{eta}^2_{\text{partial}} = .31$; Richler & Gauthier, in press). Next, individual scores for holistic processing were calculated, both by subtracting the congruency effect in misaligned d' (congruent misaligned – incongruent misaligned) from the congruency effect in aligned d' (congruent aligned – incongruent aligned), and by regressing the congruency effect for misaligned trials from the congruency effect for aligned trials. Both the subtraction (Konar, Bennett, & Sekuler, 2010; Richler, Cheung, & Gauthier, 2011b) and regression (DeGutis et al., 2013) methods have been reported in individual differences work in the literature.

We calculated the reliability of the composite task in two ways. First, we used a split-half measure of reliability, making use of the four experiments in which we had two independent samples (with the same participants) and calculating the correlations between d' scores in the two parts (these are expected to present low estimates of reliability, because the stimulus sets differ). We did this using both the difference scores and the residuals (see Table 2).

Second, following DeGutis et al. (2013), we used Guttman's λ_2, which may provide a more appropriate measure of reliability in the composite task than Cronbach's alpha, because Guttman's λ^2 is robust when measures include multiple factors (Callender & Osburn, 1979). Because failures of selective attention indicative of holistic processing are observed on aligned trials and are significantly reduced or abolished on misaligned trials, in the following calculations the congruency effect on aligned trials was considered to be the primary measure, and the congruency effect on misaligned trials was considered as the control measure.

The formula that we used to calculate the reliability of the difference score, $\rho(D)$, took the difference in variances between the primary measure, $\rho(X_1)$, and the control measure, $\rho(X_2)$, into account (see Rogosa et al., 1982):

$$\rho(D) = \frac{\sigma_{x1}{}^2\rho(X_1) + \sigma_{x2}{}^2\rho(X_2) - 2\sigma_{x1}\sigma_{x2}\rho_{x1x2}}{\sigma_{x1}{}^2 + \sigma_{x2}{}^2 - 2\sigma_{x1}\sigma_{x2}\rho_{x1x2}}, \tag{1}$$

where σ_{x1} is the standard deviation of the primary measure, σ_{x2} is the standard deviation of the control, and ρ_{x1x2} is the correlation between the primary and control conditions.

In contrast, the formula that we used to calculate the reliability of the residuals, $\rho(U)$, did not directly include any terms to describe the variances of the primary and control conditions, since they do not effect the reliability of regression (Malgady & Colon-Malgady, 1991)[3]:

$$\rho(U) = \frac{\rho(X_1) + \rho(X_2)\rho_{x1x2}{}^2 - 2\rho_{x1x2}{}^2}{1 - \rho_{x1x2}{}^2}. \tag{2}$$

[3] Note that DeGutis et al.'s (2013) reproduction of this formula included a minor error in which the primary and control conditions were switched.

Table 2 Group level analysis and split-half reliability in each of the five studies

Stimulus Set	Subjects Included	Holistic Processing ($\Delta d'$, σ)	*p* Value	Effect Size (eta$^2_{partial}$)	Split-Half Subtraction	Split-Half Regression
Experiment 1						
MPI female	76/85	0.30 (0.96)	<.05	.09	.05	.13
G&R	77/85	0.53 (0.98)	<.05	.22		
Experiment 2						
MPI female	60/67	0.53 (1.00)	<.05	.21	.04	.11
MPI female	60/67	0.90 (0.84)	<.05	.53		
Experiment 3						
MPI female	89/91	0.58 (0.86)	<.05	.31	0	.15
G&R	86/91	0.81 (0.89)	<.05	.46		
Experiment 4						
MPI female	52/53	0.18 (1.20)	>.05	.02	0	.12
MPI male	50/53	0.52 (0.81)	<.05	.29		
Experiment 5						
MPI female	40/43	0.86 (0.93)	<.05	.46		

Note that since we had only one data set from subjects in Experiment 5, we do not report split-half reliabilities. Holistic processing is calculated as the interaction between congruency and alignment. G&R, Goffaux and Rossion (2006)

Figure 2 summarizes the results of the difference-score and residual reliability calculations using both formulas, as well as the split-half measure of reliability within each data set and across the two data sets in Experiments 1, 3, and 4. It is worth noting that some of the computed reliability estimates were negative. However, since it is generally accepted that negative values are not theoretically meaningful, these values are displayed as zero (see the Discussion). The reliabilities (λ_2) for the difference scores (dark gray bars) ranged between –.22 and .50, $\mu = 22$, 95 % CI [7.63, 36.37], and the reliabilities of the residuals (light gray bars) ranged between –.54 and .67, $\mu = 19$, 95 % CI [–3.21, 41.21]. Although we have included the mean and 95 % CIs, since these may be of interest to some readers, it is important not to interpret these values as reflecting the reliability of the composite task. Fundamentally, reliability is a property of the measurement. Although the reliabilities of the two measures are highly correlated (r = .90), as would be expected because both depend at least on the variability of the aligned congruency effect, in four of the nine data sets (Exp. 1 MPI; Exp. 2 MPI female; Exp. 3 Goffaux & Rossion [2006]; and Exp. 4 MPI female), we found that the difference scores were more

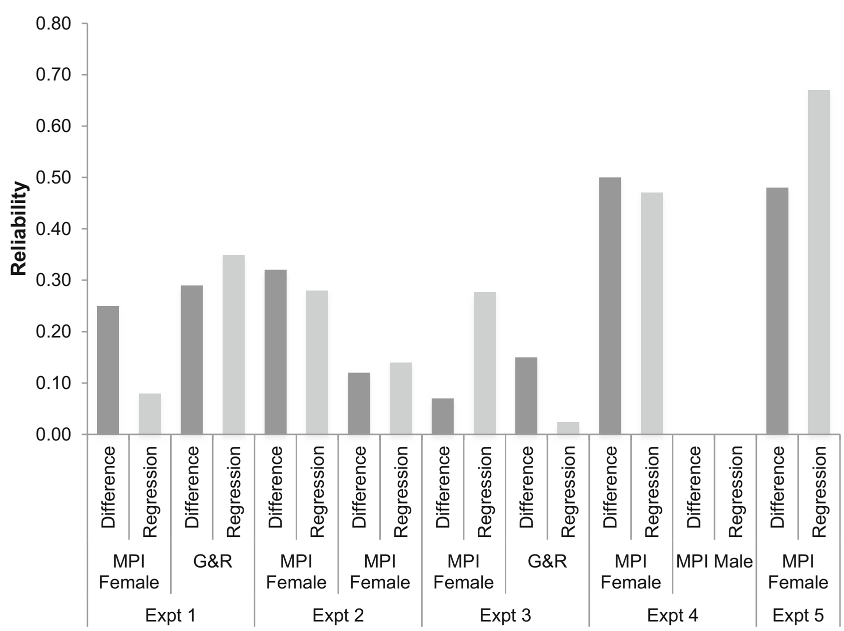

Fig. 2 Reliabilities (Guttman's λ_2) across the eight data sets calculated using difference scores and regressions. Note that negative reliability estimates have been set to zero, because values less than zero are generally assumed not to be theoretically meaningful

reliable. This is perhaps surprising, because a number of authors have suggested that residuals may be more reliable than difference scores (DeGutis et al., 2013; Peter et al., 1993).

Discussion

Here we have reported the reliabilities of the composite task in five different experiments (nine data sets). Perhaps most striking is that reliability often appears to be fairly poor, and it is also quite variable from experiment to experiment. This might be surprising to anyone tempted to consider reliability as a property of a task, but this is not a valid interpretation, since reliability is a property of measurements and is sample-dependent. However, the considerable range in the reliabilities that we obtained cannot solely be attributed to sample differences (which, we should point out, were all samples from the same undergraduate population). Although the differences in reliability between Experiments 2 and 3 with MPI female faces can only be attributed to the samples (since the measures were identical), the differences between the two measurements for each of the three samples (in Exps. 1, 3, and 4) can only be attributed to the different face sets (or to an order effect), since all other task parameters were identical. Finally, beyond differences due to the samples and the specific sets of faces used in an experiment, our efforts to increase reliability by changing aspects of the task seem to also have had an influence. Indeed, the highest reliabilities that we achieved were in Experiments 4 and 5, using the few-faces and contrast manipulation designs that were explicitly motivated as efforts to increase reliability. This success is dampened by the fact that in Experiment 4, the contrast manipulation achieved a reasonable level of reliability with the female face set, and negative reliability with the male face set (this is despite holistic processing being obtained at the group-level effect for male and not for female faces). However, it is interesting to note that of all the tasks reported here, the measure of holistic processing in Experiment 4 (using the contrast-manipulated female MPI faces) had the most between-participant variability, $\sigma = 1.20$. This goes some way toward supporting the idea that our manipulation improved reliability by adding relevant variability in the participants' measured holistic processing. It is unclear whether to attribute this to an order effect: Whereas Experiments 2 and 3 showed some reduction in reliability from one block to the other, Experiment 1 showed the reverse pattern. This is also the only one of our experiments using the MPI males stimulus set, and so we do not have sufficient bases to attribute this effect to the stimulus set. As a whole, these results suggest that although it is generally inappropriate to quote the reliability of a task from prior work (Thompson, 1994), this is particularly inadequate in the context of the composite task. As in other tasks that rely on a comparison to a baseline (e.g., the Stroop task; Eide, Kemp, Silberstein, Nathan, & Stough, 2002; Strauss, Allen, Jorgensen, & Cramer, 2005), measurements of holistic processing using the composite task often result in poor reliability. Besides sample dependence, there appears to be an influence of the faces used, although we cannot explain why some faces yield more reliable measurements than others. Our explorations of contrast manipulations or using a few faces are encouraging, suggesting that the reliability of such measurements may be improved by changing some aspects of the paradigm, although the space of parameters that may have an influence is large—including, for instance, the number and discriminability of the faces involved and the study, retention, and maximum response durations, as well as other aspects of the tasks, such as the relative-contrast manipulation we used in Experiment 4. As it stands, a combination of using a few faces and/or the contrast manipulation that we used, together with a larger number of trials, might help achieve reliabilities closer to .7, which is typically considered acceptable in the early stages of construct validation research (Nunnally & Bernstein, 1994). Whenever possible, collecting more data from each subject should also increase reliability.

Unlike DeGutis et al. (2013), who found a large difference in the reliabilities of residuals versus difference scores, we did not find much evidence to suggest that regression provides a more reliable measure of holistic processing than do difference scores. In fact, for a number of data sets the difference scores were more reliable than the residuals. Whether residuals or difference scores are more reliable depends on the precise relationships between the component measures (Chiou & Spreng, 1996; Llabre, Spitzer, Saab, Ironson, & Schneiderman, 1991; Malgady & Colon-Malgady, 1991; Zimmerman & Williams, 1998). Residuals are known to be more reliable, albeit not by very much, when the reliability of the primary condition exceeds the reliability of the control condition (Malgady & Colon-Malgady, 1991).

However, whether residuals or difference scores are preferred should not be driven entirely by concerns of reliability. Difference scores and residuals make different assumptions about what is being measured. Difference scores are fairly easy to interpret: Consider just the aligned trials in the composite task. In this case, taking the difference between the congruent and incongruent trials provides a measure of the congruency effect. Neither condition is obviously a baseline, since it is theoretically possible that congruency could facilitate and incongruency hinder performance. That is, both conditions potentially contain useful information about the construct of interest, which is the degree to which the to-be-ignored part has an influence. Regressing out one of the conditions from the other would remove all of the variance associated with that condition, and this is only appropriate if one assumes that all of the variance in the condition being regressed out is irrelevant to the construct of interest.

This decision is slightly more difficult in the case of regressing out performance on misaligned trials from that on aligned trials. On the one hand, misaligned trials are generally considered to be a baseline, with any variance in the congruency effect reflecting some general interference and not genuine holistic processing. On the other hand, if this is true, it is not clear why the misaligned condition would be necessary to measure individual differences in holistic processing (which is theoretically well defined by aligned congruent – aligned incongruent). Previous studies have shown that the misaligned condition is important for distinguishing face-like holistic processing from other kinds of interference effects in the composite task—for instance, contextual influences on spatial attention (see Richler, Bukach, & Gauthier, 2009). One particularly problematic finding for the idea of the misaligned condition as a baseline is that, for nonface objects in novices, when the misaligned condition uses images that are misaligned at study and at test (rather than only at test, as in all of the experiments in the present article), a congruency effect is obtained when both aligned and misaligned trials are randomized together (Richler, Bukach, & Gauthier, 2009). This congruency effect is not obtained for these objects in novices when the misaligned trials use an aligned study image, and it is not obtained on aligned trials when alignment is blocked, so these results indicate that the context of certain kinds of misaligned trials can influence performance on aligned trials. Until such issues are understood more fully, we advise researchers interested in measuring holistic processing not to take one specific model for granted. Indeed, it may be that, in a normal population, the congruency effect for misaligned trials is unrelated to the congruency effect for aligned trials, in which case regressing out the misaligned variance does not achieve anything (Richler & Gauthier, in press). Supporting this to some extent, in the present set of studies the Pearson's r correlations between the misaligned and aligned congruency effects were low, ranging between $r = –.19$ and .20. In this case, collecting data for twice as many aligned trials might be much more productive than collecting data from misaligned trials, only to remove the small amount of variance shared between the aligned and misaligned conditions.

Although individual differences research has a long history, it is not necessarily familiar territory for cognitive psychologists, especially in the area of high-level vision. Our results demonstrate that reliability can vary a great deal across different implementations of the composite task. We urge researchers interested in measuring individual differences in holistic processing not to rely on prior estimates of reliability, but to report the reliability of their own measurements and to consider how the said reliability constrains their claims.

Author note This work was supported by the National Science Foundation (Grant No. SBE-0542013), Vanderbilt Vision Research Center (Grant No. P30-EY008126), and National Eye Institute (Grant No. R01 EY013441-06A2). We thank Riaun Floyd and Magen Speegle for their help with data collection and stimulus construction, and Jeremy Wilmer and Joseph DeGutis for helpful discussions.

References

Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., & Rossion, B. (2010). Holistic perception of the individual face is specific and necessary: Evidence from an extensive case study of acquired prosopagnosia. *Neuropsychologia, 48,* 4057–4092. doi:10.1016/j.neuropsychologia.2010.09.017

Callender, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's Lambda-2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement, 16,* 89–99.

Cheung, O. S., Richler, J. J., Palmeri, T. J., & Gauthier, I. (2008). Revisiting the role of spatial frequencies in the holistic processing of faces. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 1327–1336. doi:10.1037/a0011752

Chiou, J., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 9,* 158–167.

Chua, K.-W., Richler, J. J., & Gauthier, I. (2014). Becoming a Lunari or Taiyo expert: Learner attention to parts drives holistic processing of faces. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1174–1182. doi:10.1037/a0035895

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition, 126,* 87–100. doi:10.1016/j.cognition.2012.09.004

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic faces. *Neuropsychologia, 44,* 576–585.

Eide, P., Kemp, A., Silberstein, R. B., Nathan, P. J., & Stough, C. (2002). Test–retest reliability of the emotional Stroop task: Examining the paradox of measurement change. *Journal of Psychology, 136,* 514–520.

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition, 103,* 322–330. doi:10.1016/j.cognition.2006.05.003

Gauthier, I., Klaiman, C., & Schultz, R. T. (2009). Face composite effects reveal abnormal face processing in autism spectrum disorders. *Vision Research, 49,* 470–478.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training "Greeble" experts: A framework for studying expert object recognition processes. *Vision Research, 38,* 2401–2428. doi:10.1016/S0042-6989(97)00442-2

Goffaux, V., & Rossion, B. (2006). Faces are "spatial"—Holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1023–1039. doi:10.1037/0096-1523.32.4.1023

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science, 21,* 38–43.

Llabre, M. M., Spitzer, S. B., Saab, P. G., Ironson, G. H., & Schneiderman, N. (1991). The reliability and specificity of delta versus residulized change as measures of cardiovascular reactivity to behavioral challenges. *Psychophysiology, 28,* 701–711.

Malgady, R. G., & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. *Educational and Psychological Measurement, 51,* 803–807.

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research, 69,* 10–22. doi:10.1016/j.visres.2012.07.014

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11,* 8–15. doi:10.1016/j.tics.2006.11.002

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw–Hill.

Palermo, R., Willis, M. L., Rivolta, D., McKone, E., Wilson, E. C., & Calder, A. J. (2011). Impaired holistic coding of facial expression and facial identity in congenital prosopagnosia. *Neuropsychologia, 49,* 1226–1235.

Peter, J., Gilbert, A., Churchill, J., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research, 19,* 662–665.

Richler, J. J., Bukach, C. M., & Gauthier, I. (2009). Context influences holistic processing of nonface objects in the composite task. *Attention, Perception, & Psychophysics, 71,* 530–540. doi:10.3758/APP.71.3.530

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011a). Beliefs alter holistic face processing . . . if response bias is not taken into account. *Journal of Vision*, *11*(13), 17:1–13. doi:10.1167/11.13.17

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011b). Holistic processing predicts face recognition. *Psychological Science, 22,* 464–471. doi:10.1177/0956797611401753

Richler, J. J., & Gauthier, I. (in press). A meta-analysis and review of holistic processing. *Psychological Bulletin*.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology, 3*(553), 1–6. doi:10.3389/fpsyg.2012.00553

Richler, J. J., Tanaka, J. W., Brown, D. D., & Gauthier, I. (2008). Why does selective attention to parts fail in face processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1356–1368.

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011d). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science, 20,* 129–134. doi:10.1177/0963721411402472

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Quantitative Methods in Psychology, 92,* 726–748.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition, 21,* 139–253.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223. doi:10.1037/1082-989X.1.2.199

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test–retest reliability of standard and emotional Stroop tasks an investigation of color–word and picture–word versions. *Assessment, 12,* 330–337.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research, 36,* 1761–1771. doi:10.1016/0042-6989(95)00230-8

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: Becoming a Ziggerin expert—but which type? *Psychological Science, 20,* 1108–1117. doi:10.1111/j.1467-9280.2009.02430.x

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception, 16,* 747–759.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology, 51,* 343–351.