The Vanderbilt Holistic Face Processing Test: A short and reliable measure of holistic face processing

Jennifer J. Richler Vanderbilt University, Nashville, TN, USA

R. Jackie Floyd Vanderbilt University, Nashville, TN, USA

Isabel Gauthier Vanderbilt University, Nashville, TN, USA

Efforts to understand individual differences in high-level vision necessitate the development of measures that have sufficient reliability, which is generally not a concern in group studies. Holistic processing is central to research on face recognition and, more recently, to the study of individual differences in this area. However, recent work has shown that the most popular measure of holistic processing, the composite task, has low reliability. This is particularly problematic for the recent surge in interest in studying individual differences in face recognition. Here, we developed and validated a new measure of holistic face processing specifically for use in individual-differences studies. It avoids some of the pitfalls of the standard composite design and capitalizes on the idea that trial variability allows for better traction on reliability. Across four experiments, we refine this test and demonstrate its reliability.

Introduction

Compared to other objects, faces are believed to be processed as wholes, or holistically, rather than as a collection of parts. Such holistic processing may facilitate discrimination between objects in a category that consist of the same features (e.g., eyes, nose, mouth) in the same configuration (e.g., eyes separated horizontally above nose, nose above mouth). We are concerned here with the construct of holistic processing whereby it is difficult to selectively attend to individual face parts (see Richler, Palmeri, & Gauthier, 2012, for a discussion of other meanings of holistic processing and the importance of not confusing them). Such failures of selective attention have been consistently demonstrated using the composite task (Young, Hellawell, & Hay, 1987). In the sequential-matching version of this task, participants are asked to judge whether one half (e.g., top) of two sequentially presented faces are the same or different while ignoring the other face half (e.g., bottom). On congruent trials, the target and irrelevant face halves are associated with the same response (e.g., both halves same or both halves different). On incongruent trials, the target and irrelevant face halves are associated with conflicting responses (e.g., one half is the same, the other half is different). Performance is better on congruent versus incongruent trials, indicating that participants could not ignore the task-irrelevant half. This congruency effect is greatly reduced or eliminated when the familiar face configuration is disrupted by misaligning the face halves (see Richler & Gauthier, 2014, for a review and meta-analysis of this effect). Importantly, the failures of selective attention in this paradigm are not found for objects in novices (e.g., Richler, Mack, Palmeri, & Gauthier, 2011) but are observed for real-world experts in their domain of expertise (Boggan, Bartlett, & Krawczyk, 2012; Bukach, Philips, & Gauthier, 2010; Gauthier, Curran, Curby, & Collins, 2003; A. C.-N. Wong et al., 2012; Y. K. Wong & Gauthier, 2010) and for novel objects following individuation training (Gauthier & Tarr, 2002; Gauthier, Williams, Tarr, & Tanaka, 1998; A. C.-N. Wong, Palmeri, & Gauthier, 2009), supporting the idea that holistic processing is a behavioral signature of face and expert object recognition.

Recently, there has been a surge in interest in understanding individual differences in holistic processing and how they relate to individual differences in face and object recognition (e.g., DeGutis, Wilmer, Mercado, & Cohan, 2013; Konar, Bennett, & Sekuler, 2010; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012; Richler, Cheung et al., 2011; Wang, Li,

Citation: Richler, J. J., Floyd, R. J., & Gauthier, I. (2014). The Vanderbilt Holistic Face Processing Test: A short and reliable measure of holistic face processing. *Journal of Vision*, *14*(11):10, 1–14, http://www.journalofvision.org/content/14/11/10, doi:10.1167/14.11.10.

doi: 10.1167/14.11.10 Received June 04, 2014; published September 16, 2014 ISSN 1534-7362 © 2014 ARVO

Fang, Tian, & Liu, 2012). However, such endeavors are complicated by the fact that, although there are many measures of face recognition and holistic processing used in group studies, few have been developed and validated for use in individual-differences contexts. Of particular interest here is that, although the composite task is a robust measure in group studies, it has poor reliability. A meta-analysis of 48 studies (Richler & Gauthier, 2014) estimated the effect size for holistic processing in the composite task as .32 95% confidence interval (CI) (.26, .38). To detect this effect with 95% power, one only needs 13 subjects, on average. But an analysis of the reliability of this measurement across six typical instantiations of this task, each using 160 trials (experiments 1–3 in Ross, Richler, & Gauthier, in press), found reliability around .2, regardless of whether holistic processing was quantified by subtracting the congruency effect for misaligned trials from that obtained in aligned trials or by regressing the misaligned congruency effect from the aligned congruency effect (DeGutis et al., 2013). Using a similar task with 144 trials, DeGutis et al. (2013) found reliabilities of .10 for the subtraction method and .24 for the regression method.

Such low reliability limits the utility of this task in detecting relationships between holistic processing and performance on other tasks because such correlations are limited by the reliability of the measures (Nunnally, 1970). Two measures cannot produce a correlation that is higher than the square root of the product of their reliabilities. Therefore, based on the reliability of the Cambridge Face Memory Task (CFMT; Duchaine & Nakayama, 2006; ~.85), a standard in the field for measuring face-recognition abilities, and the composite task (~.2), even if the two constructs were perfectly correlated, the maximum possible detectable relationship between these measures would be ~.41. Therefore, a better measure of holistic processing is needed as a complement to the extant CFMT to answer questions about whether and how holistic processing specifically (measured by the new test we develop here) is related to general face-recognition abilities (measured by the CFMT) in the normal population and whether face-recognition deficits in patient populations (e.g., prosopagnosia) are related to impaired holistic processing.

The standard composite task may have low reliability for several reasons. First, holistic processing is measured as a difference in the congruency effect between aligned and misaligned trials. Moreover, the congruency effect itself is a difference score (congruent – incongruent), and the dependent measures is d′, which is also a difference score (zHit – zFalse Alarm). Second, in the standard composite task used in group studies, parameters such as timing of the study or test face, mask duration, the homogeneity of the faces, the size of the faces, etc. are the same in all trials and are mainly chosen to avoid floor and ceiling levels in performance. However, variability among trials with regard to the construct a test aims to measure is important to the reliability of the measurements. On most tests in which all items are constructed such that the probability of success on the item is directly related to the ability measured, this means that relatively easy trials will provide more information to help discriminate among individuals who possess fairly low levels of the ability, and relatively hard trials will provide more information to help discriminate among individuals who possess fairly high levels of the ability. As an analogy, consider a group of athletes competing in the high jump. Keeping the bar at one level throughout will only tell us which athletes can jump that high and which cannot, but it does not help us determine who can jump the highest. Many cognitive tasks that have been designed to perform optimally in group studies are designed with all items at one medium level of difficulty and therefore provide the most information to discriminate individuals in a small range of ability.

Holistic processing in the composite task has been defined as a congruency effect, which means that the construct is defined by a difference between two kinds of trials: congruent and incongruent. Difficulty on these trials per se does not define holistic processing and is irrelevant: Two subjects, one who performs poorly and another who performs much better on average, would both be deemed equally able to attend selectively if both have an equivalent difference in performance between congruent and incongruent trials. This means that to optimize the reliability of this measure, it is not variability in difficulty per se that is needed but variability in average congruency effects elicited by different trials. The hope is that conditions that elicit low average levels of congruency will provide more information to discriminate people at the low end of the holistic processing continuum, and conditions that elicit high average levels of congruency will provide more information to discriminate people at the high end of the holistic processing continuum.

In this spirit, our goal here is to present a new task—The Vanderbilt Holistic Face Processing Test (VHFPT)—that measures failures of selective attention to face parts but hopefully with better reliability than the standard sequential-matching composite task. To this end, we used a three-alternative forced choice task rather than a same/different matching task. This serves two purposes. First, response bias does not affect the performance level of interest in alternative forced choice tasks, and so computation of d′ is not necessary; this eliminates one of the difference scores in the measure of holistic processing. Second, having three options reduces chance level and therefore the influence of guessing on performance. In addition, according to the principles outlined previously, we varied the proportion of the composite face that is task-irrelevant as well as the size of the entire composite face across trials to increase variability in aspects of the task that we hoped would affect holistic processing. To a large extent, this was based on intuition although large variations in face size have been found to influence holistic processing on various tasks (McKone, 2009), including the composite task (Ross & Gauthier, 2014). In addition, we conjectured that it should be more difficult to ignore a task-irrelevant segment if it makes up a greater proportion of the face and if the entire face is small such that both the target and distractor segments are foveated within a single fixation. Note that we are not interested in measuring or interpreting the effect of these manipulations; they were included solely to increase variability in the difficulty of exerting selective attention.

Finally, we included aligned and misaligned trials because in the standard sequential matching–composite task holistic processing is defined by the congruency × alignment interaction, in which congruency effects are larger on aligned versus misaligned trials. We also included trials in which task-irrelevant segments were phase-scrambled (see methods) to test the usefulness of an alternative baseline. On misaligned trials, the spatial distance between the target and distractor image segments may introduce a confound; at the extreme, it would not surprise anyone that information that is several degrees of visual angle away from task-relevant information is easier to ignore. On phase-scrambled trials, the spatial relationship between target and distractor segments is preserved. Because we scrambled the face segments used on aligned trials, apart from the phase information, all other low-level information is preserved in these images, including where the target segment is located relative to the task-irrelevant segment. We were also interested in whether this alternative baseline affords a benefit in terms of the reliability of the measurement of holistic processing.

Experiment 1: VHFPT

Methods

Participants

One hundred thirteen members of the Vanderbilt University Community (34 male; mean age = 21.5 years, range 18–38) participated in exchange for monetary compensation or course credit.

All subjects completed the CFMT (Duchaine & Nakayama, 2006), the VHFPT, and the PANAS questionnaire (Watson, Clark, & Tellegen, 1988). Here, as our goal is to evaluate the reliability of the VHFPT as a measure of holistic processing, we only discuss the results from this task. There was no significant correlation between holistic processing and CFMT performance in this data set ($r = -.17$, $p = 0.10$), but average accuracy on the VHFPT (collapsed over congruent and incongruent trials) was correlated with CFMT performance in all three conditions (aligned: $r = .38$, misaligned: $r = .48$, phase-scrambled: $r = .29$, all $ps < 0.005$). Combined with the evidence we present below that the VHFPT measures strong and reliable holistic effects, this suggests that the CFMT may depend on the ability to process face parts but not on holistic processing of faces. There were some significant but very small positive correlations with positive and negative affect, but they did not replicate in a different data set. Because this experiment included the CFMT and there is evidence that face recognition is impaired for faces from other races (e.g., Meissner & Brigham, 2001), we were only recruiting Caucasian subjects. Thus, data from two non-Caucasian individuals were excluded from the analyses. Data from an additional 14 subjects were discarded due to a computer error. Therefore, data from 97 subjects were included in the analyses.

Stimuli

Stimuli were created using 360 unfamiliar Caucasian faces (180 male, 180 female) and 81 famous Caucasian faces (45 male, 36 female) obtained from the Internet and converted to gray scale. Stimuli were divided into 40 unfamiliar and nine famous sets of six same-sex faces. One additional famous face set (female) was used to create examples for the instructions. Within each set, three faces were used to make target segments, and three faces were used to make distractor segments. There were two images of the same individual for each target face. Each set was used to create one trial for each combination of alignment (aligned, misaligned, phase-scrambled) × congruency (congruent, incongruent) for a total of six trials per set. Unfamiliar face sets were randomly assigned to target segment conditions (top two thirds, bottom two thirds, top half, bottom half, top third, bottom third, eyes, nose, mouth; four to six sets per target segment; see Figure 1). One famous face set was randomly assigned to each target segment condition. All sets were randomly assigned to size conditions (small and medium = 16 sets each; large = 17 sets). Small, medium, and large faces were 1.09, 2.09, and 3.09 in. in height, respectively. Face width was variable but held constant within a set.

An example of the six trials made from one set is shown in Figure 2. Each target face was the correct response for one alignment condition pair (congruent and incongruent trials) and served as a foil in the other alignment conditions. Study composites were created by pairing target image 1 with one of the three

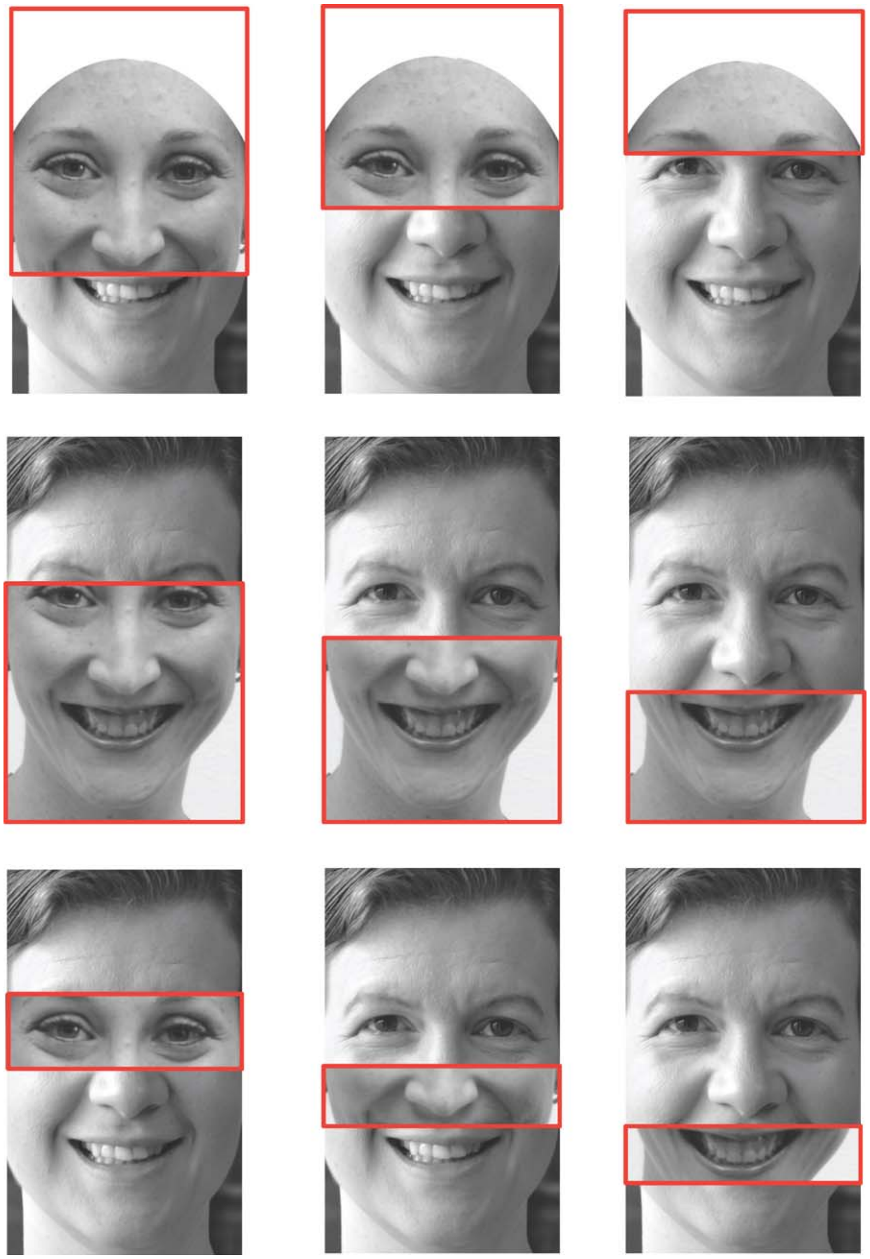

Figure 1. Example composites illustrating the nine different kinds of target segments (that were outlined in red on all VHFPT trials). In these examples, the target segment is Jenn Richler, and the distractor segment is Isabel Gauthier (these are for illustration only and were not actual stimuli in the experiment). Permissions available from the first author.

distractor segments. Target top segments in the study composite were cropped to remove hair, which is a salient nonface cue (see Figures 1 and 2). The same study composite was used for congruent and incongruent trials within each alignment condition. Test displays showed three composite faces. In congruent trials, target image 2 was paired with the same distractor segment as in the study composite. The other two faces were foil composites (foil 1, foil 2) created with the remaining two target and distractor segments. In incongruent trials, one of the foil composites (foil 1) from the congruent trial was used, and segments were crossed with the study composite such that target image 2 was paired with the foil 1 distractor segment, and the study composite distractor segment was paired with the foil 1 target segment (see Figure 2a). Thus, in incongruent trials, the target segment was paired with a different distractor than at study, and foil 2 was identical in congruent and incongruent trials within each alignment condition. The target segment in both study and test composites was outlined in red (1.5 point thick).

Example misaligned and phase-scrambled trials are shown in Figure 2b. In misaligned trials, the study

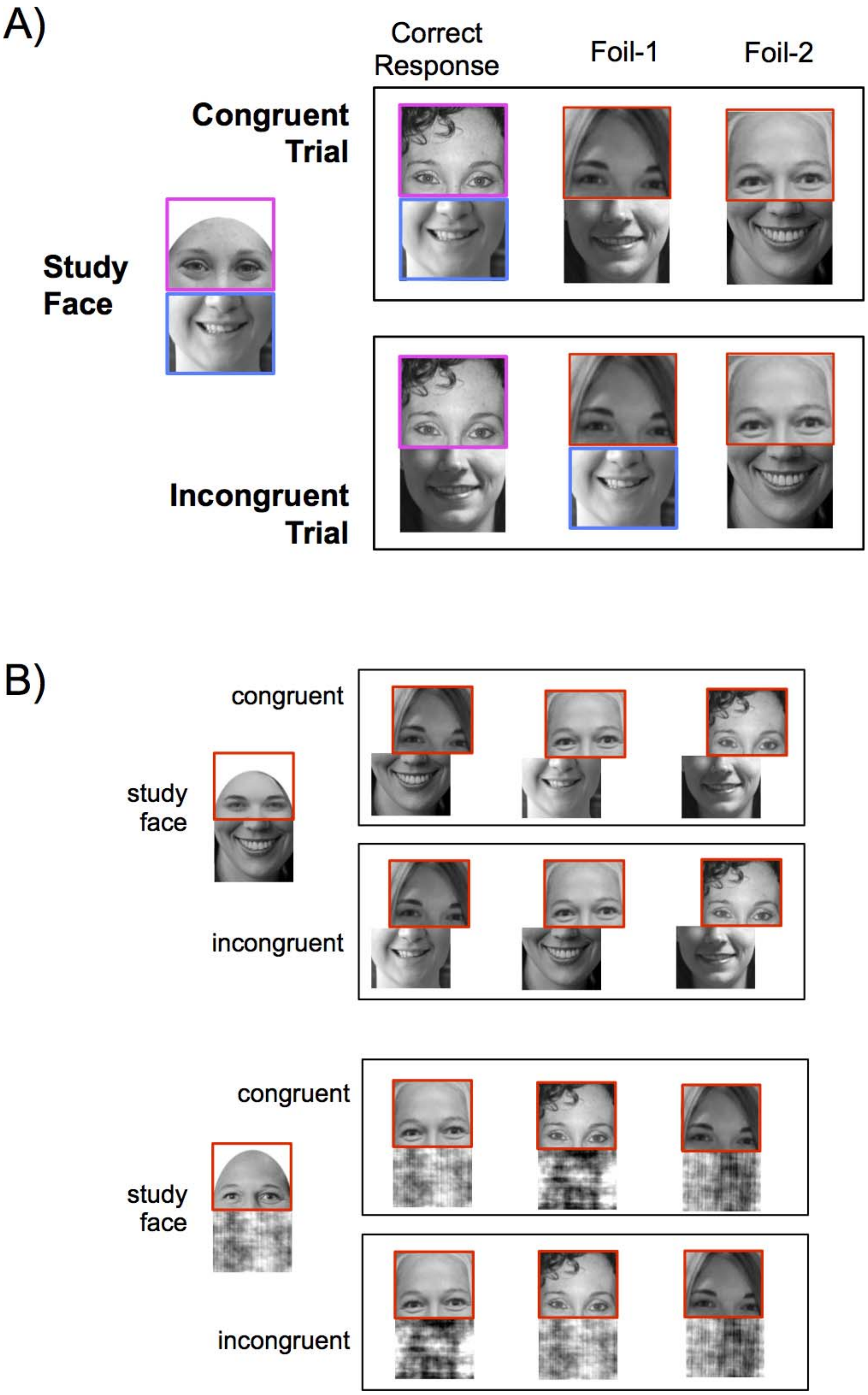

Figure 2. Example of six trials made with one illustrative face set (not included in the actual experiment). Each target face is the correct response for one alignment condition (both congruent and incongruent trials). The target segment in the study composite and the target segment in the correct response composite are different images of the same individual. (A) Example of an aligned congruent and incongruent pair. The target and distractor segment from the study composite are shown outlined in pink and blue for illustrative purposes. (B) Misaligned and phase-scrambled congruent and incongruent trials. Permissions available from the first author.

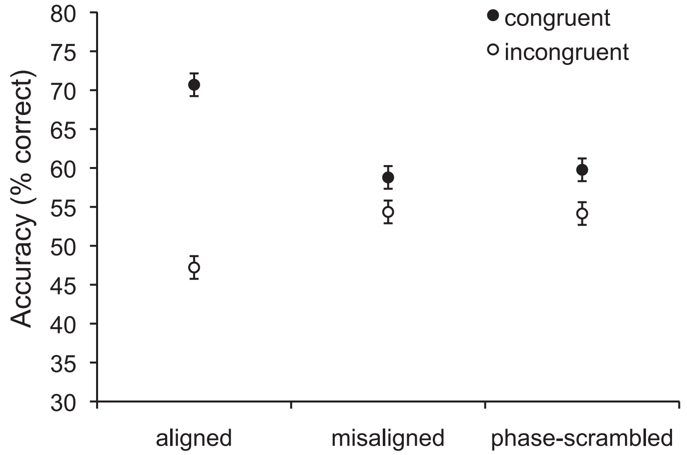

Figure 3. Accuracy (percentage correct) on congruent and incongruent trials as a function of condition. Error bars show 95% CIs of the within-subjects effect from a 3 × 2 ANOVA that included all conditions. Chance in this task is 33.33%.

composite was aligned, and target/distractor segments were misaligned in the test composites. The precise amount and direction of displacement between the segments in misaligned trials was the same within a set but varied between sets. In phase-scrambled trials, the distractor segments were degraded using the Random Image Structure Evolution algorithm (Sadr & Sinha, 2004), in which image phase components are randomized while preserving all low-level attributes and power spectrum. This degrades images with a particular level of phase coherence (65% in the images used here). Thus, in phase-scrambled trials, the distractor segments contain no face feature information, but congruency can still be defined in terms of the phase-scrambled image paired with the target segment at study and test.

Procedure

In each trial, a study composite was presented for 2 s, followed by a random pattern mask (500 ms) and then the test display. Participants were instructed to indicate which of the three composite faces contained the target segment with the same identity (outlined in red) as the study composite while ignoring the task-irrelevant distractor segments. Response keys (J, K, L) were presented below the test images. The test display was presented until participants made a response. The location of the correct response was counterbalanced within set.

One trial order was created and used for all subjects. Trial order was generally randomized with the following exceptions. Trials were blocked by target part. Within each block, the six famous face trials (one for each combination of alignment × congruency) were presented first as practice trials. These were followed by the unfamiliar face trials. At least two trials separated trials created from the same face set. There were 294 trials (54 famous, 240 unfamiliar; 49 for each alignment × congruency condition). Only unfamiliar face trials were included in the analyses.

Results and discussion

Group-level analyses

Mean accuracy in congruent and incongruent trials as a function of alignment condition (aligned, misaligned, phase-scrambled) is plotted in Figure 3.

First, we conducted a 2 (congruency) × 2 (alignment) repeated-measures ANOVA using the aligned and misaligned trials only. This is the same analysis that is used to assess holistic processing in the standard same/different composite task. There were significant main effects of alignment, $F(1, 96) = 11.66$, $MSE = 47.30$, $p = 0.001$, $\eta^2_p = .11$, and congruency, $F(1, 96) = 245.97$, $MSE = 76.81$, $p < 0.001$, $\eta^2_p = .72$. Critically, the congruency × alignment interaction indicative of holistic processing was significant, $F(1, 96) = 152.30$, $MSE = 57.76$, $p < 0.001$, $\eta^2_p = .61$.

Next, we conducted the same ANOVA using aligned and phase-scrambled trials to determine whether phase-scrambled trials can be used as an alternative baseline. The results were comparable to those with misaligned trials: There were significant main effects of condition, $F(1, 96) = 6.16$, $MSE = 62.88$, $p = 0.015$, $\eta^2_p = .06$, and congruency, $F(1, 96) = 367.75$, $MSE = 55.83$, $p < 0.001$, $\eta^2_p = .79$, and a significant congruency × condition interaction, $F(1, 96) = 140.46$, $MSE = 55.08$, $p < 0.001$, $\eta^2_p = .59$.

Finally, and as can be appreciated from Figure 3, there were no differences in performance between phase-scrambled and misaligned trials (main effect of alignment: $p = 0.64$, $\eta^2_p = .002$; congruency × alignment interaction: $p = 0.39$, $\eta^2_p = .008$).

These results suggest that phase-scrambled trials can be used as a baseline in group studies. Phase scrambling may be preferable to misalignment as a baseline because phase-scrambled trials preserve the spatial relationship between target and distractor segments. Thus, any improvements in selective attention cannot be attributed to the fact that distractor segments are simply further away from target segments.

Reliability

Following DeGutis et al. (2013) and Ross et al. (in press), we evaluated reliability with Guttmans $\lambda 2$ (Guttman, 1945), which is more robust than Cronbach's alpha when the measure includes multiple factors (Callender & Osburn, 1979). Guttman $\lambda 2$ for each individual condition in Experiments 1, 2, 3, and 4 are shown in Table 1. Because failures of selective

Experiment	Aligned C	Aligned I	Misaligned C	Misaligned I	Phase-scrambled C	Phase-scrambled I
1: VHPT	.38	.45	.43	.36	.27	.38
2: VHPT-A	.65	.71				
2: VHPT-A Retest	.61	.68				
3: VHPT Contrast	.48	.54				
4: VHPT-A 2.0	.78	.79				

Table 1. Guttman λ2 for each condition in Experiments 1, 2, 3, and 4. *Notes*: C = congruent; I = incongruent.

attention indicative of holistic processing are observed in aligned trials and are significantly reduced or abolished in misaligned or other control trials, in the following calculations the congruency effect on aligned trials is considered to be the primary measure, and the congruency effect on misaligned or phase-scrambled trials is considered to be the control measure. We therefore quantified holistic processing by regressing the congruency effect for the control condition (misaligned or phase-scrambled) from the congruency effect for aligned trials (see DeGutis et al., 2013; Ross et al., in press).

In step 1, we calculated the reliability of the congruency effect for each alignment condition as a difference score (congruent – incongruent). The formula we used to calculate the reliability of the difference score, $\rho(D)$, took the difference in variance between measures, $\rho(X_1)$ and $\rho(X_2)$, into account (see Rogosa, Brandt, & Zimowski, 1982):

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

Whereby, σ_{x1} and σ_{x2} are the standard deviations, and ρ_{x1x2} is the correlation between measures.

In the second step, we calculated the reliability of the residuals (control condition regressed from primary condition), $\rho(U)$, by the formula

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

where $\rho(X_1)$ is the primary measure, and $\rho(X_2)$ is the control measure calculated from step 1.

Reliability of holistic processing using misaligned trials as the control measure was .41, and reliability of holistic processing using phase-scrambled trials as the control measure was .43. This is generally higher than the reliability of the same/different composite task (~.2; e.g., DeGutis et al., 2013; Ross et al., in press). Importantly, this is comparable to the reliability of the congruency effect on aligned trials calculated as a difference score in step 1 (.43) without regressing out any baseline or control condition. This is consistent with the results of a recent meta-analysis, suggesting a very small correlation between the congruency effect on aligned trials and the congruency effect on misaligned trials (Richler & Gauthier, 2014); little is gained by regressing out performance in conditions that share little variance with the condition of interest in the first place. Here, the correlation between the congruency effect on aligned and misaligned trials was .14, and that between aligned and phase-scrambled trials was .007.

Experiment 2: VHFPT-A

In Experiment 1, we found that regressing a control condition from the aligned condition did not influence reliability. In fact, because there is, at best, a very small correlation between the congruency effects for aligned and baseline conditions, as is also the case for the typical composite task (Richler & Gauthier, 2014), this argues against the value of measuring and regressing the variance in these conditions in this task. Because reliability increases with the number of trials, these results suggest that when the goal is to measure individual differences, time is better spent on more aligned trials. Thus, in Experiment 2, we tested a version of the VHFPT that only includes aligned trials (VHFPT-A). We also tested a more diverse population in terms of age and race by recruiting subjects via Amazon Mechanical Turk.

Methods

Participants

One hundred nineteen subjects (54 male; mean age = 36.5, range = 21–66) were recruited via Amazon Mechanical Turk. Although we did not specifically recruit Caucasian subjects, the majority of our subjects ($n = 92$) were Caucasian. Of the remaining subjects, seven were African American, 11 were Asian, eight were Hispanic/Latino, and one subject did not disclose race.

To evaluate test–retest reliability, we contacted all subjects 18 days after they first completed the experiment and asked them to participate a second time. We ultimately obtained retest data for 65 subjects (28 male; mean age = 36.5, range = 21–66; 51 Caucasian, five African American, four Asian, and five Hispanic/Latino). Six of these subjects were excluded

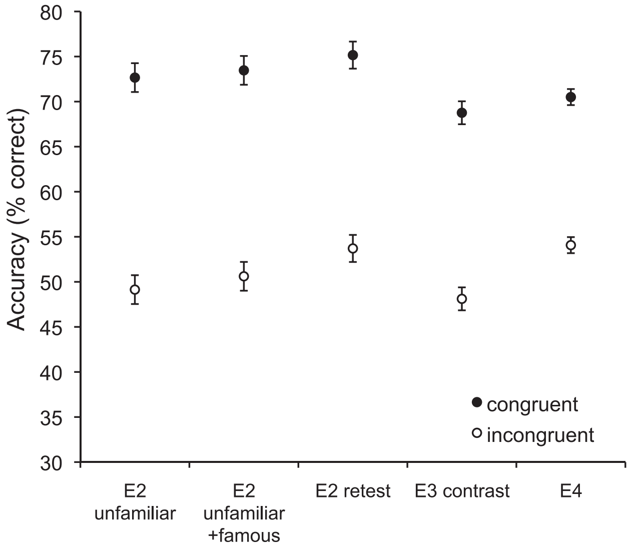

Figure 4. Accuracy on congruent and incongruent trials in Experiments 2, 3, and 4. Error bars show 95% CIs for each within-subjects congruent-incongruent comparison. Chance is 33.33%.

from the analyses of the first test session (see below) and so were not included in the test–retest analysis.

Stimuli and procedure

An additional 18 face sets (two for each target segment) were created using 162 unfamiliar Caucasian faces (81 male, 81 female). These sets were used to make additional aligned trials only. New aligned trials were added to the aligned trials from the previous experiment. A new random order of unfamiliar face trials was created with the constraint that trials created from the same set could not be consecutive. There were 134 trials (18 famous, 116 unfamiliar; 67 congruent/incongruent).

In addition, because we were running this study on Amazon Mechanical Turk, we added three introductory trials at the beginning of the experiment to ensure that participants read and understood the instructions. In the introductory trials, composites were made from Muppet faces that were presented in color. These trials were therefore very easy. There were two incongruent trials (target segment: top and bottom halves) and one congruent trial (target segment: top half). Subjects who did not get all three introductory trials correct ($n = 19$) were not included in our analyses. The trial sequence was nearly identical to the full VHFPT from Experiment 1 except a blank screen was presented for 1000 ms instead of a random pattern mask between the study face and test display. Subjects were instructed to click the face in the test display that contained the same target segment identity as the study face while ignoring the task-irrelevant face segments. Participants were instructed not to click the study face. Subjects who failed to follow this instruction on more than 10 trials ($n = 9$) were excluded from the analyses. Thus, data from 90 subjects were analyzed. Average test duration was 18 min.

Results and discussion

Mean accuracy in congruent and incongruent trials for unfamiliar faces only and all trials, including famous faces, during the first test and all trials for the retest is shown in Figure 4.

First, we only analyzed the unfamiliar face trials as in Experiment 1. At the group level, there was a large and significant congruency effect, $F(1, 89) = 426.72$, $MSE = 58.36$, $p < 0.001$, $\eta^2_p = .83$.[1] Because there is no control condition (e.g., misaligned, phase-scrambled) here, reliability was calculated based on the difference score (congruent − incongruent) as in step 1 of Experiment 1. Reliability was .50.

Next, we tested the impact of including the famous face trials in the analyses. Because variability of trial types improves reliability, it may be that these trials improve reliability because they afford a greater use of verbal labels and may be easier because they are familiar identities. At the group level, there was again a large and significant congruency effect, $F(1, 89) = 403.89$, $MSE = 58.19$, $p < 0.001$, $\eta^2_p = .82$. Although reliability was slightly higher when famous faces were included (.57), this is equivalent to the improvement in reliability that is expected when we extrapolate, using the Spearman Brown prophecy formula, the reliability for unfamiliar faces based on 58 trials to the number of trials involved when famous faces are included (67 trials). Thus, beyond increasing the number of trials, the famous face trials do not seem to provide any additional benefit in terms of reliability.

Finally, there was a large and significant congruency effect in the retest data, $F(1, 58) = 408.67$, $MSE = 33.22$, $p < 0.001$, $\eta^2_p = .88$. Reliability was lower at retest (.30), but this is close to half as many subjects as in the first test session. The correlation between the congruency effect on test 1 and test 2 (test–retest reliability) is shown in Figure 5. Test–retest reliability (Pearson's r) was .52, $p < 0.001$.

Experiment 3: VHFPT-A + contrast

In Experiment 2, we found that a version of the VHFPT with aligned trials only (VHFPT-A) had good reliability and test–retest reliability. Moreover, reliability of the congruency effect on aligned trials was higher in Experiment 2 than Experiment 1. These

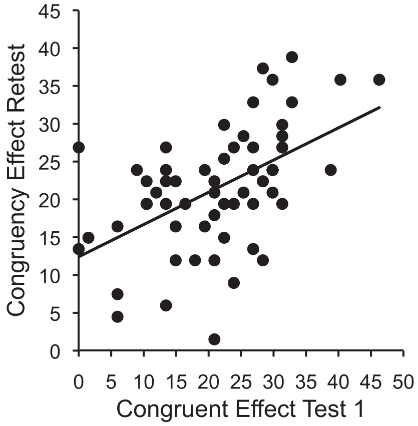

Figure 5. Correlation between the congruency effect (congruent – incongruent) on Test 1 and Retest in Experiment 2.

results support the suggestion that, because the control conditions do not share significant variance with the condition of interest (Richler & Gauthier, 2014), it is better to include more aligned trials that tap into the effect of interest to boost reliability instead of control trials.

Ross et al. (in press) found the highest reliability in the same/different composite task in an experiment in which the contrast of target and distractor segments was manipulated. The idea was that the distractor segment would be difficult to ignore if it was high contrast and the target segment was low contrast but easier to ignore if it was low contrast compared to a high-contrast target. In other words, this contrast manipulation was designed to add variability in the difficulty of selectively attending, which is directly relevant to holistic processing. In Experiment 3, we added a similar contrast manipulation to the VHFPT-A trials to test whether this would improve reliability.

Methods

Participants

One hundred twenty one subjects (58 male; mean age = 37.4 years, range = 19–66; 98 Caucasian, 10 African American, five Asian, six Hispanic/Latino, one Pacific Islander, one nondisclosed) were recruited via Amazon Mechanical Turk.

Stimuli and procedure

Face sets used in the VHPT-A were randomly assigned to low-contrast target (22 sets), low-contrast distractor (22 sets), or normal (23 sets) conditions. Low-contrast segments (target or distractor, depending on condition) were created by setting the transparency of the segment to 70% in PowerPoint. Examples are shown in Figure 6. Targets or distractors were presented in low contrast in the study image and test display for both congruent and incongruent trials in a set.

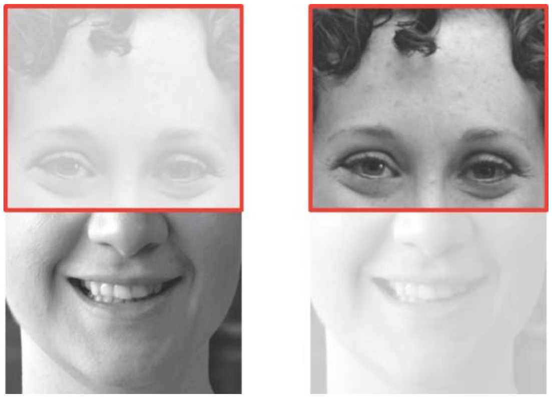

Figure 6. Example composites in which the target segment (right) and distractor segment (left) are low contrast. Permissions available from the first author.

The procedure was identical to Experiment 2, and the same trial order was used. Thus, low contrast conditions were randomized within the experiment. Average test duration was 18 min.

Results

Data from 17 subjects were discarded for failing to achieve 100% accuracy on the introductory Muppet trials. Data from 14 subjects were excluded from the analyses for clicking on the study face on more than 10 trials. Finally, data from one subject was excluded for below-chance performance in both congruent and incongruent conditions. Thus, our analyses are based on data from 89 subjects. Data from both unfamiliar and famous face trials were included in the analyses.

At the group level, there was a large and significant congruency effect, $F(1, 88) = 518.91$, $MSE = 36.56$, $p < 0.001$, $\eta^2_p = .86$ (see Figure 4). Reliability was .30. This is lower than reliability for a similar sample size without the contrast manipulation. Therefore, we found no evidence that including lower-contrast target and distractor segments improves the reliability of this measure.

Experiment 4: VHFPT-A 2.0

Based on Experiment 2, in which famous faces did not improve reliability beyond adding more trials, we added more unfamiliar face trials to the VHFPT-A and removed the famous face trials. This should afford at least comparable reliability to Experiment 2 without any potential confounds due to variability in familiarity

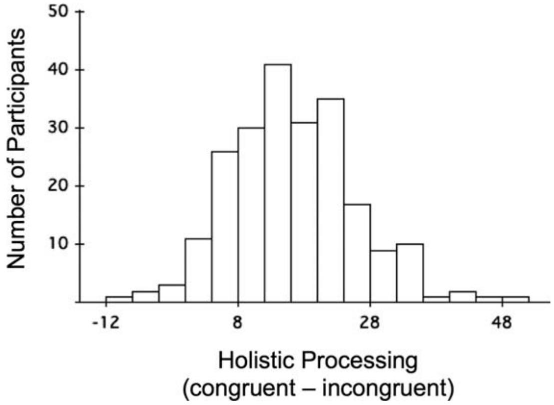

Figure 7. Histogram of holistic processing (congruent – incongruent) in Experiment 4.

with the famous faces and the fact that they could be named. In addition, because VHFPT-A 2.0 represents our current best effort at creating a reliable measure of holistic processing that will be made available for others to use, we collected data from more than double the number of subjects so that we could report age/sex norms.

Methods

Participants

Data were collected from 276 subjects (118 male; mean age = 35 years, range = 20–69; 215 Caucasian, 22 African American, 15 Asian, 16 Hispanic/Latino, three Native American/Alaskan, one Middle Eastern, five nondisclosed). Two hundred forty participants were recruited via Amazon Mechanical Turk. To increase the number of younger participants in our sample, we also went to a sample from another Amazon Turk experiment in our lab (a VET-car test as in McGugin et al., 2012) and invited all those who were born after 1983 to participate. This yielded data from another 36 subjects.

Stimuli and procedure

Only the unfamiliar trials from Experiment 2 were used. An additional 32 face sets were created from the set of computer-generated faces used in Richler, Cheung, & Gauthier (2011) for which we had two different versions of each face with different lighting conditions. The rationale for adding those kinds of faces was that they would add to the trial variability in the VHFPT, providing broader coverage of the kinds of adult faces that participants may encounter in other face-recognition tasks with which the VHFPT might be used in future studies. Only aligned trials were used in this version. Target segment and size were assigned to sets such that, when added to the aligned unfamiliar face trials from Experiment 2, there were 20 trials (10 congruent, 10 incongruent) per target segment and 60 trials (30 congruent, 30 incongruent) per size. Thus, there were a total of 180 trials in the experiment (90 congruent, 90 incongruent). A new random order of trials was created with the constraint that trials created from the same set could not be consecutive. The trial sequence and practice trials were identical to Experiment 2 with the exception that feedback was provided in the Muppet practice trials. Because of this opportunity for learning the task during the practice trials, we did not use practice trial accuracy as an exclusion criteria. Average test duration was 22 min.

Results and discussion

Data from 53 subjects were excluded from the analyses for clicking on the study face in more than 10 trials. Data from one subject were excluded for less than 10% accuracy in incongruent trials, and data from another subject were excluded for being a clear bivariate outlier in the correlation between accuracy in congruent and incongruent trials. Thus, data from 221 subjects were analyzed.

At the group level, there was a large and significant congruency effect, $F(1, 220) = 660.38$, $MSE = 45.17$, $p < 0.001$, $\eta^2_p = .75$ (see Figure 4). As can be appreciated from Figure 7, holistic processing was close to normally distributed (mean = 16.43, $SD = 9.5$, skewness = .44, kurtosis = .64, Shapiro-Wilk W = .99, $p = 0.027$). Reliability of the congruency effect was .56.[2]

Because we have a larger sample size in this study, we calculated the congruency effect in different conditions to provide a test of our conjecture that these various conditions add variance to our measurement of holistic processing. Figure 8 shows performance for each target segment. A 2 (congruency: congruent, incongruent) × 9 (target segment: bottom two thirds, top two thirds, bottom half, top half, bottom third, top third, eyes, mouth, nose) repeated-measures ANOVA revealed a significant main effect of target segment, $F(8, 1760) = 81.10$, $MSE = 246.39$, $p < 0.001$, $\eta^2_p = .27$, and a significant congruency × segment interaction, $F(8, 1760) = 40.64$, $MSE = 186.51$, $p < 0.001$, $\eta^2_p = .16$, indicating that holistic processing varied between target segments. As can be appreciated from Figure 8 and in line with our predictions, holistic processing was small when target segments were large (i.e., there was less to-be-ignored information) and increased as target size decreased (i.e., most of the face information had to be ignored). In the order the target segments are listed in Figure 8 (from large to small), the percentage of subjects with positive congruency values was 49%, 50%, 65%, 66%, 67%, 78%, 84%, 83%, and 82%. The

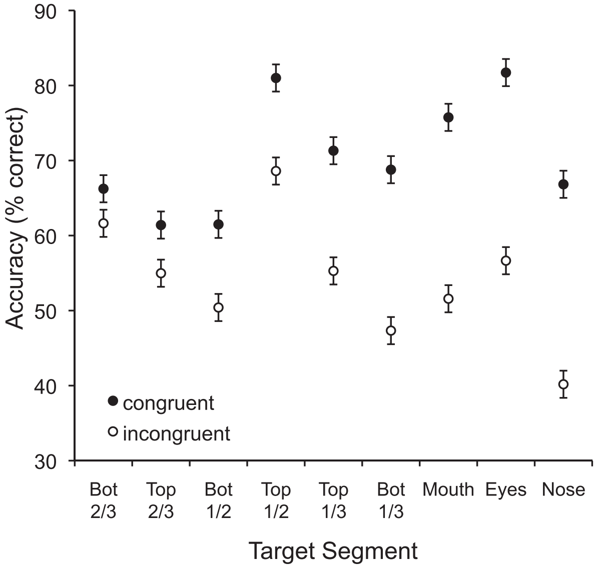

Figure 8. Accuracy in Experiment 4 on congruent and incongruent trials as a function of target segment (arranged in order from smallest to largest holistic processing effect, left to right). Error bars show 95% CIs of the within-subject effect. Chance is 33.33%.

magnitude of holistic processing was not correlated with average accuracy ($r_9 = -.08$).

Figure 9 shows performance as a function of face size. A 2 (congruency: congruent, incongruent) × 3 (size: small, medium, large) repeated-measures ANOVA revealed a main effect of size, $F(2, 440) = 10.24$, $MSE = 67.55$, $p < 0.001$, $\eta^2_p = .04$, and a significant congruency × size interaction, $F(2, 440) = 17.28$, $MSE = 54.52$, $p < 0.001$, $\eta^2_p = .07$. There was greater holistic processing when faces were small versus medium or large. Note, however, that the effect of size was quite small (maximum difference in average performance between size conditions = 2%, maximum difference in holistic processing = 5%).

Table 2 shows performance separately for male and female participants divided into five age bins. Table 3 shows the zero-order correlations between holistic processing (congruent – incongruent) and age, sex (dummy-coded male = 1, female = −1), and age × sex. Table 4 shows the results of a multiple regression with holistic processing as the dependent variable and age, sex, and age × sex entered simultaneously as predictors. All predictors were significant but only accounted for 4.1% of the variance in holistic processing. Looking at mean holistic processing, there is a striking increase in holistic processing for women in the last age group (45 and older). This was not predicted, and it is interesting in the light of the finding in large online samples that face recognition ability peaks after age 30, which is older then most other cognitive and perceptual abilities (Germine, Duchaine, & Nakayama, 2011). This effect is not associated with poorer performance overall or slower response times: The interaction between age and sex remains virtually unchanged when controlling for both overall performance and mean response time (B=−.109248, t=−2.03, p = 0.04). This interaction is small, however, and it is not present when looking at effect sizes because this group also shows considerable variability in holistic processing. It should therefore be treated with caution until it is replicated with considerably more power.

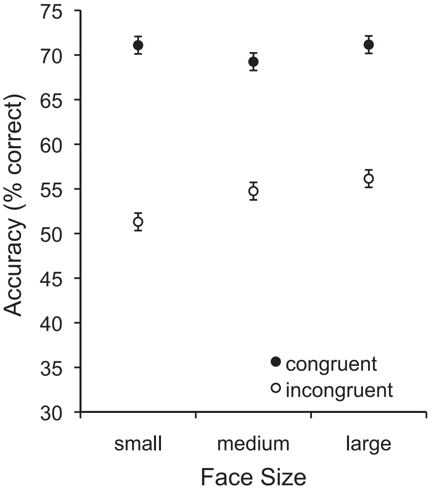

Figure 9. Accuracy in Experiment 4 on congruent and incongruent trials as a function of face size. Error bars show 95% CIs of the within-subject effect. Chance is 33.33%.

General discussion

Here we developed and refined a new test for measuring individual differences in holistic processing, the VHFPT. We tested several versions and validated its ability to detect holistic processing in the vast majority of the subjects we tested; only a very small proportion of our subjects (2%) over all the versions tested here showed a negative congruency effect for aligned trials. This is in contrast to individual differences in the more standard same/different composite task in which a larger proportion of subjects tend to show "negative" holistic processing (McGugin et al., 2012; Richler, Cheung et al., 2011), a concept that does not really have much validity. That is, although it makes sense that some subjects may show less holistic processing than others and perhaps some would show none at all, better performance in incongruent than congruent trials is difficult to interpret. Given the sequential-matching task has very low reliability, this is possibly mainly due to error.

We aimed to increase the reliability of holistic processing measurement in a number of ways. Primarily, we infused the test with trial variability that was meant

	Age bin	*n*	Mean age	Mean acc. (%) congruent	Mean acc. (%) incongruent	Mean HP	η^2_p HP
Female	<25	21	22.2 (1.4)	71.64 (10.72)	60.00 (8.85)	11.64 (7.65)	.71
	25–29	29	27.0 (1.3)	71.19 (9.30)	54.67 (10.32)	16.51 (7.77)	.82
	30–34	23	31.4 (1.4)	70.10 (8.04)	54.69 (7.07)	15.41 (10.99)	.67
	35–44	31	38.7 (3.2)	68.35 (11.27)	53.19 (10.92)	15.16 (8.88)	.75
	45+	30	55.2 (6.9)	72.56 (7.48)	51.15 (9.22)	21.41 (10.41)	.81
Male	<25	10	22.7 (1.1)	67.78 (7.72)	52.44 (7.89)	15.33 (5.85)	.88
	25–29	17	26.9 (1.6)	70.39 (0.17)	54.90 (10.91)	15.49 (7.46)	.82
	30–34	16	32.3 (1.4)	72.50 (6.96)	51.17 (13.22)	18.33 (12.68)	.69
	35–44	25	39.7 (2.9)	70.04 (9.51)	53.87 (8.31)	16.18 (6.99)	.85
	45+	19	55.3 (7.6)	69.42 (8.91)	52.22 (11.63)	17.19 (12.18)	.68

Table 2. Performance for female and male participants divided into five age bins. *Notes*: Standard deviations shown in brackets. HP = holistic processing (congruent − incongruent).

to provide better coverage of the whole range of holistic processing ability. For instance, we expected that trials in which the target segment comprised a very small part of the faces may be processed holistically by most subjects except those few who were the best at selective attention (the least holistic), and other trials in which the target segment comprised a very large part of the faces may not be processed holistically except by those few who were the worst at selective attention (the most holistic). This was supported by our results, which showed that small target segments, such as the nose, mouth, and eye areas, led to positive congruency effects in most subjects (83%) whereas larger target segments that comprised two thirds of the image led to positive congruency effects in only 50% of the subjects. Second, the size of the faces on the screen likely also contributed some variance relevant to holistic processing although those effects were smaller, and the direction of the effect was different from that observed in the standard composite task in which the size manipulation was much more important (Ross & Gauthier, 2014, see also McKone, 2009). Third, we also attempted to increase variability in the extent to which different trials would discriminate best at various levels of the holistic ability continuum by including a range of different kinds of faces: male and female with a range of different emotional expressions, makeup, and hair as well as some computer-generated faces. This should also improve the generalizability of the test, but one limitation is that we only used Caucasian faces in the VHFPT. Some recent studies (Bukach, Cottle, Ubiwa, & Miller, 2012; Harrison, Gauthier, Hayward, & Richler, 2014) have suggested that other-race faces are processed holistically, and so this limitation may not have been necessary. However, given the extent of the literature on the other-race effect (see Meissner & Brigham, 2001, for a meta-analytic review) and the fact that other-race effects are correlated with an individual's experience with the other race (Bukach et al., 2012), researchers may still prefer a test in which experience is less likely to vary across items for many applications. We repeated face segments on the test only once, each one being part of a congruent and an incongruent trial, such that the test would measure holistic processing and not face learning.

Finally, the reliability of this measure of holistic processing can be increased by including more aligned trials to replace misaligned trials that we have argued are not necessary for the quantification of this effect. This may be a sensitive issue because the misaligned (or other kinds of) baseline has played an important role in understanding the construct measured by the composite task. It was important in the development of the composite task to demonstrate that the congruency effect obtained for aligned face segments is not obtained with misaligned segments or, here, with phase-scrambled segments. In particular, there are experimental conditions under which congruency effects are observed for aligned segments, but they are not larger than for misaligned segments (Chua, Richler, & Gauthier, 2014; Gauthier, Klaiman, & Schultz, 2009; Richler, Bukach, &

	HP
Age	.19*
Sex	.01
Age × sex	−.05

Table 3. Zero-order correlations between holistic processing (congruent − incongruent) and age, sex, and age × sex with sex coded as 1 for males and −1 for females. *Notes*: * $p < 0.01$. HP = holistic processing.

Model and predictor holistic processing (R^2 − adj = 4.1%)	B	*SE*	*t*	*p*
Intercept	12.158	2.081	5.84	<0.0001
Age	0.119195	0.0538	2.22	0.0277
Sex	4.18007	2.081	2.01	0.0458
Age × sex	−0.113127	0.0538	−2.10	0.0366

Table 4. Results of multiple regression on holistic processing scores with age, sex, and age × sex entered simultaneously as predictors.

Gauthier, 2009), and in these cases, the control condition may serve to distinguish face-like holistic processing from other strategies that are not face-specific. In projects with populations or stimuli other than familiar-race faces or when testing the interaction of new manipulations with holistic processing, a control condition should be used. In fact, the phase-scrambled control condition we used here may be theoretically preferable in some cases to the misaligned condition. However, in a meta-analysis of several matching composite task studies and in the present work, we found that there is very little shared variance between the congruency effects in these two conditions. At this point, armed with the knowledge that this measure taps into face-specific holistic processing (sensitive to configuration, which is not the case for objects in novices; Richler et al., 2009) for the purpose of quantifying the effect across subjects, there is virtually nothing to be gained by devoting testing time to measuring a control condition that is not correlated with the variance of interest. Instead, assuming that in projects on individual differences several measures have to be gathered, requiring each of them to be as efficient as possible, the most compact and reliable measure will be obtained using only aligned trials.

With the final version of the VHFPT we present here, we obtained measurements of holistic processing in a sample of normal adults on Amazon Mechanical Turk that had much better reliability (.56) than the standard version of the composite task (.2, Ross et al., in press), and a previous version showed test–retest reliability of .52. In addition, a version with fewer trials showed reliability of ∼.40 for measurements in a lab sample, suggesting that the VHPT-A 2.0 should perform well in this population too. However, readers should be cautioned not to attribute reliability to a test: It is a property of measurements (Thompson, 1994); ultimately, it is the reliability of the measurements that need to be considered in interpreting their value and relationship to other measurements.

Keywords: face recognition, face perception, holistic processing, individual differences

Acknowledgments

This work was supported by the National Eye Institute (2 RO1 EY013441-06A2) and the Temporal Dynamics of Learning Center (NSF grant SBE-0542013). The authors thank Amit Khandhadia for assistance with data collection.

Commercial relationships: none.

Corresponding author: Jennifer J. Richler.
Email: jennifer.j.richler@vanderbilt.edu.
Address: Vanderbilt University, Nashville, TN, USA.

Footnotes

[1] In all our Amazon Mechanical Turk studies, we were very conservative with our inclusion criteria. However, the behavioral effect of interest did not depend on subject exclusion. For example, in Experiment 2, we also obtained a large and significant congruency effect when all subjects were included in the analyses ($p < 0.001$, $\eta^2_p = .81$).

[2] The results are similar if all subjects (except two who were discarded as outliers in behavior) are included in the analyses (congruency effect: $p < 0.001$, $\eta^2_p = .68$; reliability = .60).

References

Boggan, A. L., Bartlett, J. C., & Krawczyk, D. C. (2012). Chess masters show a hallmark of face processing with chess. *Journal of Experimental Psychology: General, 141,* 37–42.

Bukach, C. M., Cottle, J., Ubiwa, J., & Miller, J. (2012). Individuation experience predicts other-race effects in holistic processing for both Caucasian and Black participants. *Cognition, 123,* 319–324.

Bukach, C. M., Philips, W. S., & Gauthier, I. (2010). Limits of generalization between categories and implications for theories of category specificity. *Attention, Perception & Psychophysics, 72,* 1865–1874.

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.

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

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.

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 participants. *Neuropsychologia, 44,* 576–585.

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a nonmodular account of face processing. *Nature Neuroscience, 6,* 428–432.

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., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 431–446.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. W. (1998). Training "Greeble" experts: A framework for studying expert object recognition processes. *Vision Research, 38,* 2401–2428.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118*(2), 201–210.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Harrison, S. A., Gauthier, I., Hayward, W. G., & Richler, J. J. (2014). Other-race effects are quantitative not qualitative in the composite paradigm. *Visual Cognition, 22,* 843–864.

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

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.

McKone, E. (2009). Holistic processing for faces operates over a wide range of sizes but is strongest at identification rather than conversational distances. *Vision Research, 49,* 268–283.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7,* 3–35.

Nunnally, J. C., Jr. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

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.

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

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin, 140,* 1281–1302.

Richler, J. J., Mack, M. L., Palmeri, T. J., & Gauthier, I. (2011). Inverted faces are (eventually) processed holistically. *Vision Research, 51,* 333–342.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology, 3.*

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Quantitative Methods in Psychology*, *92*(3), 726–748.

Ross, D. A., & Gauthier, I. (2014). Holistic processing in the composite task depends on face size. Submitted for publication.

Ross, D. A., Richler, J. J., & Gauthier, I. (in press). Reliability of composite task measurements of holistic face processing. *Behavior Research Methods.*

Sadr, J., & Sinha, P. (2004). Object recognition and random image structure evolution. *Cognitive Science, 28,* 259–287.

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

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science, 23,* 169–177.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.

Wong, A. C.-N., Bukach, C. M., Hsiao, J., Greenspon, E., Ahern, E., Duan, Y., & Lui, K. F. (2012). Holistic processing as a hallmark of perceptual expertise for nonface categories including Chinese characters. *Journal of Vision, 12*(13):7, 1–15, http://www.journalofvision.org/content/12/13/7, doi:10.1167/12.13.7. [PubMed] [Article]

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for face-like expertise with objects: Becoming a Ziggerin expert – But which type? *Psychological Science, 20,* 1108–1117.

Wong, Y. K., & Gauthier, I. (2010). Holistic processing of musical notation: Dissociating failures of attention in experts and novices. *Journal of Cognitive and Affective Behavioral Neuroscience, 10,* 541–551.

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