Investigating the Associations between Performance Outcomes on Tasks Indexing Featural, Configural and Holistic Face Processing and Their Correlations with Face Recognition Ability

Elizabeth Nelson

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Doctorate in Philosophy in Experimental Psychology

School of Psychology
Faculty of Social Sciences
University of Ottawa

Acknowledgments

I would like to start by thanking my supervisor, Charles Collin, for his guidance and patience throughout the course of this project. It veered every way it possibly could have, and you managed it, and supported me, perfectly. My parents and brother have been a constant source of love and encouragement in my life, and they were especially so during graduate school. I love you and cannot thank you enough. To my Grandmother, thank you for the most fabulous reading weeks that we spent together in Nerja, and for encouraging me to: "get this thesis done!" Mission accomplished. Thank you as well to my fabulous group of friends, and especially Max, you all made sure that I had a life outside of school. That work-life balance was invaluable. Lastly, to the many PhD's in my life, thank you all for the special support and advice throughout this whole process.

Abstract

Many important questions remain unanswered regarding how we recognize faces. Methodological inconsistencies have contributed to confusion regarding these questions, especially those surrounding three purported face processing mechanisms—featural, configural, and holistic—and the extent to which each play a role in face recognition. The work presented here aims to 1) empirically test the assumption that several face recognition tasks index the same underlying construct(s), and 2) contribute data to a number of ongoing debates concerning the reliability and validity of various methods for assessing integrative (i.e., holistic and/or configural) aspects of face processing.

Experiment 1 tested the assumption that various tasks purporting to measure integrative face processing index the same construct(s). It is important to test this assumption because if these tasks are in fact measuring different things, then researchers should cease interpreting them as interchangeable measures. Using a within-subjects design ($N = 223$) we compared performance—as reflected by accuracy and reaction time measures, as well as two types of difference scores—across four of the most commonly used integrative face processing tasks: The *Partial Composite Face Effect Task*, the *Face Inversion Effect Task*, the *Part Whole Effect Task*, and the *Configural/Featural Difference Detection Task*.

Analyses showed that within-task correlations were much stronger than those between-tasks. This suggests that the four conditions within each task are measuring something in common; In contrast, low correlations across tasks suggest that each is measuring something unique. This in turn suggests these tasks should not be seen as assessing the same integrative face-processing construct. Exploratory factor analyses corroborated the correlation data, finding

that performance on most conditions loaded onto a single factor in unrotated solutions, but onto separate factors in direct oblimin-rotated solutions.

In Experiment 2, we investigated the question of whether integrative face processing performance is related to face recognition ability. We did this by assessing the degree to which results from four widely-used integrative face processing tasks correlate with a measure of general face recognition ability, The Cambridge Face Memory Test (CFMT). The four integrative processing tasks used in this study only partly overlapped those from in Experiment 1. They were: The *Complete Composite Face Effect Task*, the *Partial Composite Face Effect Task*, the *Part Whole Effect Task*, and the *Configural/Featural Difference Detection Task*. As with Experiment 1, we used a within-subjects design ($N = 260$) and analyzed a variety of performance variables across these tasks.

Analyses demonstrated low to moderate positive correlations between performance on the task conditions and performance on the CFMT. This suggests that the constructs the tasks reflect do contribute to face recognition ability to a modest degree. These analyses also replicated parts of Experiment 1, showing weak correlations between tasks. Also similar to Experiment 1, factor analyses generally revealed task conditions loading onto a common first factor in the unrotated factor matrix, but loading separately in the rotated factor solution.

In addition to providing evidence regarding the nature of integrative face processing tasks, the data presented here speak to a number of other questions in this domain. For instance, they contribute to the debate regarding which kinds of difference scores (subtraction-based or regression-based) are more reliable, as well as the reliability of the various tasks used to investigate integrative face processing. In addition, the data inform the debate over whether the

Complete or the *Partial* version of the *Composite Face Effect Task* is the superior measure of integrative face processing.

In summary, the studies presented here indicate that the previous literature in face recognition needs to be interpreted with care, with an eye to differences in methodology and the problems of low measurement reliability. The various methods used to investigate integrative face processing are not assessing the same thing and cannot be taken as reflecting the same underlying construct.

Statement of Candidate Contribution

Nelson solely wrote the text of the dissertation. For both experiments, Nelson collected and analyzed the data. Collin aided in a general and supervisory manner, offering feedback and assistance. Collin and Boutet conceived of the project idea. Nelson, Collin and Boutet chose the experimental design. Boutet and Watier provided additional feedback and assistance. Rainville programmed the experiments.

Table of Contents

List of Tables

Chapter 2

Chapter 3

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 1: Introduction

Recognizing someone's identity based on his or her face requires a deceptively complicated series of cognitive processes. Despite decades of research, many important questions pertaining to how the human visual system accomplishes facial recognition remain unanswered. Importantly, there are a number of methodological issues within the domain of facial recognition that complicate efforts to answer such questions. As well, there are ongoing debates regarding how best to address these methodological issues that must be resolved in order to answer the larger conceptual questions. Examples of important questions that remain to be answered include:

1) Are there several different kinds of processing applied to faces and, if so, how many? The answer to this question is widely held to involve some subset of three mechanisms, typically called featural, configural and holistic (Maurer, Le Grand, & Mondolch, 2002).

2) Assuming there are several processing mechanisms underlying face recognition, how are they related to one another?

3) What role does each of the face processing mechanisms play in general face recognition ability? Here, it has been widely assumed that integrative mechanisms, sometimes referred to as holistic and/or configural mechanisms, play a particularly important role in face recognition (although recent work has raised doubts about this).

The studies presented here provide data that contribute to a number of ongoing debates in the field regarding these and related questions. Perhaps most importantly, we empirically test a fundamental assumption pertaining to commonly-used paradigms thought to measure the integrative aspects of face processing. We stipulate that a fundamental error, namely interpreting a variety of integrative face processing tasks as interchangeable measures, has contributed to

confusion and inconsistencies in the literature regarding the nature of the association between the processing mechanisms and with the extent to which they contribute to face recognition ability in general.

The general introduction of the thesis is structured as follows. First, we provide an overview of three proposed face processing mechanisms—featural, configural, and holistic—and how they are currently measured. We then explain five face processing tasks that are thought to index some of these face processing mechanisms. We provide a rationale for the inclusion in this thesis of these five tasks, and we detail their respective methodologies. In addition, we describe the assumptions underlying how performance on each task is thought to reflect one or more of the three processing mechanisms.

Next, we provide methodological details and the rationale for including the task we chose as our measure of general face recognition ability, the Cambridge Face Memory Test. Following this, we review the literature examining patterns of performance across a variety of commonly used face processing tasks and explain why investigating this is important. We then review the literature investigating the role that integrative face processing plays in face recognition. Next, we discuss the methodological issues that exist and that have compromised the research conducted to date regarding face processing and face recognition. We summarize the ongoing debates about how to address these issues and outline how the present work serves to contribute data that will inform these debates. Lastly, we summarize the research questions and hypotheses for both manuscripts that make up this thesis.

Review of Face Processing Mechanisms

For several decades, research in face recognition has been based on the assumption that there several processing mechanisms underlying face recognition. Specifically, it was widely

thought that three mechanisms were at play: Featural, Configural, and Holistic (though the latter two are sometimes equated). Maurer et al. (2002) provided the definitions of featural, configural, and holistic processing that are most commonly accepted. Yet researchers do not use them consistently, and, as Richler, Palmeri, and Gauthier (2012) remark, more work is required in order to develop precise terminology, especially with respect to holistic processing. Notwithstanding the fact that these definitions remain to be perfected, we have elected to use these conceptualizations. Throughout the thesis, decisions made between possible alternative terms are based on which is the most commonly accepted or commonly used in the literature to date. This is done in order for our results to have the broadest applicability within the domains of facial recognition and face perception as they currently exist.

According to Maurer and colleagues (2002), *Featural processing* refers to extracting individuating information from the facial features (e.g., eyes, nose, mouth, etc.). The same authors suggest that there are two types of configural processing: First order and second order. *First-order relations* refers to the organization of facial features. All faces have the same first-order organization (two eyes placed side by side centered above a nose, which is centered above a mouth). *Second-order relations* on the other hand differ across faces and thus provide additional individuating information beyond what is extracted during featural processing. *Second-order relations* refers to the distance between features (e.g., inter-ocular distance, the distance between nose and mouth, etc.). Because configural processing involves integrating various pieces of information, it is referred to as one type of integrative processing. Another integrative processing mechanism is termed *holistic processing*. In holistic face processing people exhibit a tendency to process the face as a whole instead of by its components (i.e., people integrate facial features, and first and second order relational information into a gestalt)

(Maurer et al., 2002). One goal of the work presented here is to provide a better understanding of the associations between performance on face processing tasks thought to tap into these processing mechanisms as they are currently conceptualized.

Review of Face Processing Tasks – Experiment 1

Understanding the nature of the associations between featural, configural, and holistic processing is a necessary precondition to revising and improving the definitions of these face-processing mechanisms in the future. It is also necessary before researchers are able to disentangle the mechanisms' individual contributions to face recognition ability. To date, investigating the associations between featural, configural and holistic processing has involved studying patterns of behavioural performance across a number of face and object processing tasks. Each of these tasks is thought to primarily tap into one of the three processing types, or to contrast one of the types against another. This project, in effect, will investigate the orthogonality of featural, configural, and holistic processing. That is, how separate are these three mechanisms, and to what degree can they be measured in isolation from one another?

Researchers have used a wide variety of face processing paradigms across studies. We chose to include and investigate a subset of these tasks that are most commonly used. This decision was made in order for our results to have the broadest application to the existing literature. By doing so our results will help contextualize the largest number of previously conducted studies.

In the first study presented in this thesis, the four commonly used face-processing tasks that we included were the *Partial Composite Face Effect Task* (e.g., Young, Hellawell, & Hay, 1987), the *Face Inversion Effect Task* (e.g., Yin, 1969), the *Configural/Featural Difference Detection Task* (e.g., Freir, Lee, & Symons, 2000; Leder & Bruce, 2000), and the *Part Whole*

Effect Task (e.g., Tanaka & Farah, 1993). While many versions of each of these tasks exist, we selected the versions that have been most commonly used. This is again to give our data the broadest possible relevance to interpreting previous studies. We are not advocating for the superiority of these particular paradigms over others, and below we will discuss criticisms of these tasks and methodological limitations. Here we briefly review the tasks' methodologies and what they each purport to measure. A full description of the methodology for each of the tasks is available in the Methods Sections within the first and second manuscripts, which compose Chapters 2 and 3 of the thesis.

Partial Composite Face Effect Task. The *Partial Composite Face Effect Task* (Young et al., 1987) was designed to measure holistic processing, specifically measuring the extent to which participants are naturally prone to integrate face parts despite task instructions to do otherwise. First, we explain the task methodology and then we explain how performance on this task is thought represent holistic processing. Face stimuli in this task are called composites because each face stimulus is made up of the top half from one face (forehead to middle of nose) and the bottom half from another face (middle of nose to chin). This is a sequential matching paradigm where participants view one face, it disappears, and then they view a second face. Their task is to selectively attend to the top halves of both faces and indicate whether they are the same or different. The face halves are either presented aligned, such that they look similar to a typical face, or misaligned, such that the top half and bottom half of the face are offset horizontally, thus breaking the face into to obviously separate halves. Test and inspection faces are both presented in the same manner, that is, either both aligned or both misaligned. Likewise, test and inspection faces are either both presented upright or both inverted. In this so-called "partial" version of the *Composite Face Effect Task,* test and inspection face top halves either

match or are different, but test and inspection face *bottom* halves are always different. For a visual depiction of the trial structure in this task for "same" faces, see Figure 1.

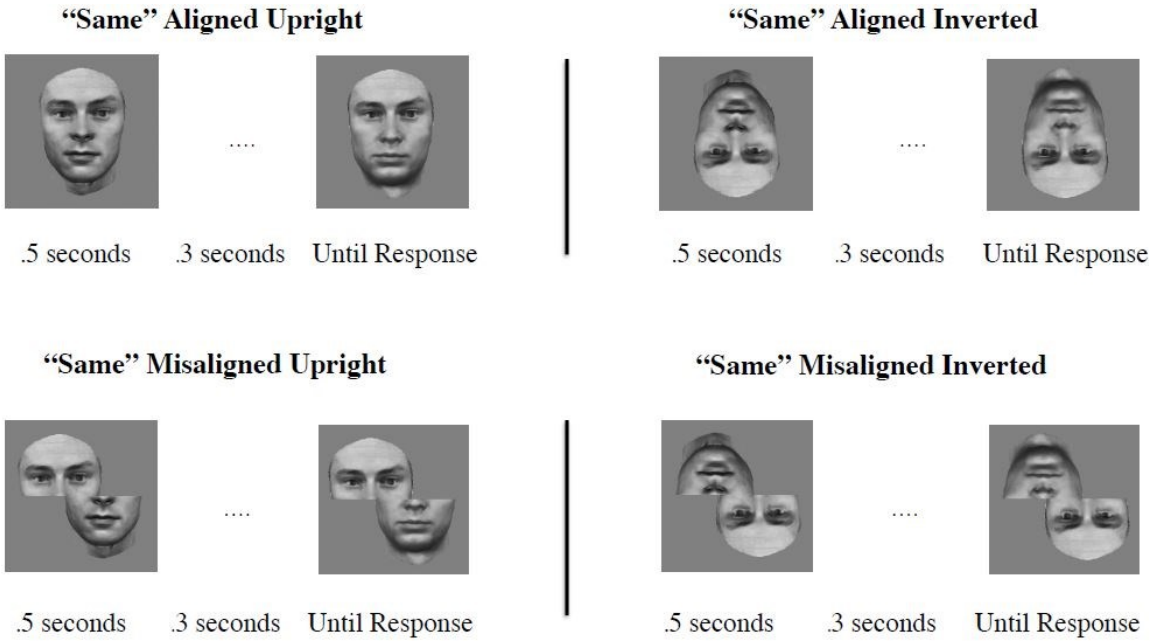

Figure 1. This figure depicts the stimuli from the first manuscript, presentation times, and the inter-stimulus interval for four types of trials in the "partial" Composite Face Effect Task. They are Aligned Upright, Aligned Inverted, Misaligned Upright, and Misaligned Inverted for "same trials" meaning the top halves of the test and inspection face match.

Recall that the participants are instructed to selectively attend to the top halves of both the test and inspection faces to determine if they are the same or if they are different. It should be noted that the term "selectively attend" in this case does not necessarily imply that an attentional mechanism is operating. While this is a possibility (e.g., Chua, Richler, & Gauthier, 2015) it has also been suggested that a perceptual mechanism is at play (Rossion, 2013). Regardless of the mechanism(s) at play, it is thought that performance patterns in this task provide strong evidence in support of holistic processing (Rossion, 2013).

Because humans have a tendency to process faces holistically (Mondolch, Pathman, Maurer, Le Grand, & de Schonen, 2007) they have difficulty being able to selectively attend to face parts (Meinhardt-Injac, Boutet, Persike, Meinhardt, & Imhof, 2017; Richler et al., 2012;

Richler & Gauthier, 2014). The task irrelevant bottom halves of the test and inspection face get integrated into a gestalt. Due to the interference from the irrelevant always-different bottom halves, participants have a tendency to mistakenly believe that two composites with the same top face halves are different. This is particularly the case in the Upright Aligned condition. In the Upright Misaligned condition, holistic processing is disrupted by misalignment. Therefore, performance tends to be better in the Upright Misaligned condition compared to the Upright Aligned condition. This is because misalignment disrupts holistic processing and makes it easier to selectively attend to the task-relevant top halves of the test and inspection faces (Gauthier & Tarr, 2002).

Holistic processing, which this task purports to measure, is thought to be impaired when faces are presented upside down (Carey & Diamond, 1977; Goffaux & Rossion, 2006; 2007;Van Belle, De Graef, Varfaillie, Rossion, & Lefèvre, 2010). Therefore, the impact of alignment is minimized when the face stimuli are presented upside-down. Nevertheless, performance is inferior on both Aligned and Misaligned Inverted trials compared to Aligned and Misaligned Upright trials, it is simply that the difference between them is reduced.

In all conditions of the partial composite task, featural information is available for processing. First and second order configural information is also available is also available in all conditions, but the processing mechanisms treating this higher order information are thought to be impaired in inverted conditions. Therefore, we expect that performance will be best during Misaligned Upright trials. We expect comparable performance between the Aligned and Misaligned Inverted conditions and the Upright Aligned condition, but for performance to be lower on these trials than on Misaligned Upright trials.

Face Inversion Effect Task. The *Face Inversion Effect Task* (Yin, 1969) examines the impact that inversion has on face recognition versus object recognition. In this task, a face or an object is presented upright or inverted. This is the inspection stimulus. Then this stimulus disappears and two of the same type of stimulus (i.e., two faces or two objects) in the same orientation (i.e., upright or inverted) are presented simultaneously. The participant must indicate which of the two stimuli matches the first stimulus they saw. In the version of the task presented in this thesis, the object class was chairs (e.g., Stein, Sterzer, & Pellen, 2012, Yovel & Kanwisher, 2005). For a visual depiction of the trial structure for this task, see Figure 2.

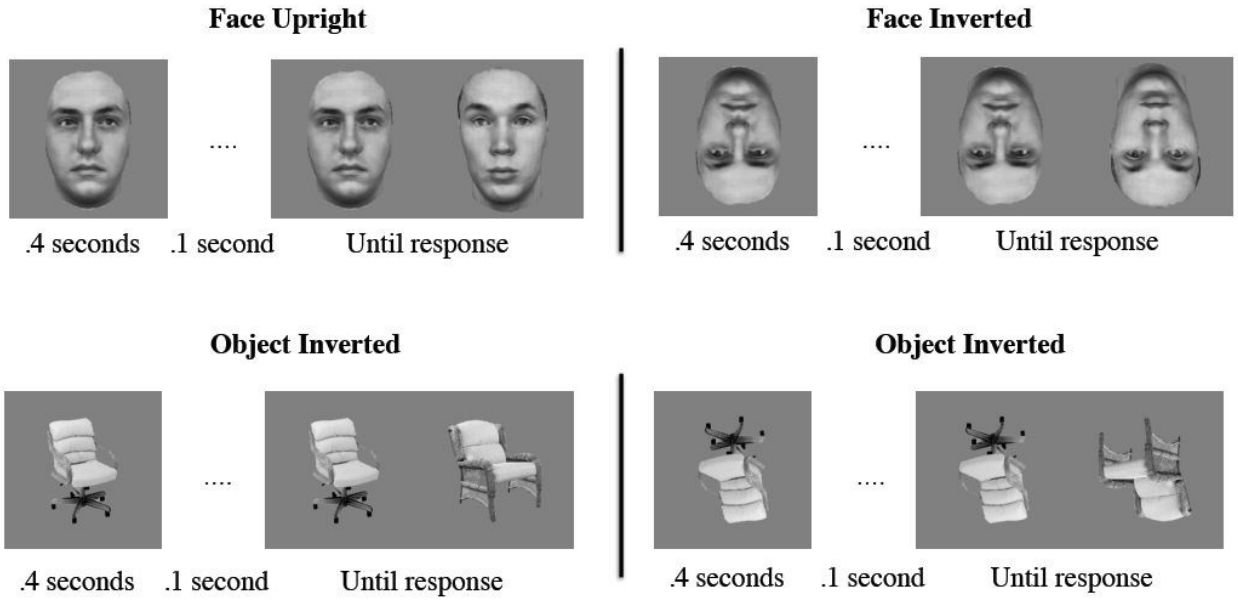

Figure 2. This figure depicts the stimuli, presentation times, and the inter-stimulus interval for upright and inverted trials in the Face Inversion Effect Task. The first row depicts the two types of face trials, upright and inverted. The second row depicts upright and inverted object trials.

It is unclear which processing mechanism(s) this task indexes. Unlike the *Composite Face Effect Task* (or the *Part Whole Effect Task* which will be discussed below), the *Face Inversion Effect Task* was not designed to be a measure of integrative processing. Rather, the difference in the effect of inversion on faces versus objects is thought to arise due to the disruption of holistic and/or configural processing, which is purportedly more important for face

recognition than object recognition. However, whether it is one, both, or neither of these mechanisms that is disrupted remains unknown.

These questions notwithstanding, the face inversion task is likely the task most commonly used to assess integrative face processing. By including it in our battery of tasks, we will be able to determine how similar or different performance on this task is compared with performance on the other three tasks included in the first manuscript. Moreover, it is the only task that includes a non-face control object and as such can provide valuable information regarding differences between the processing of faces versus other visual objects.

Part Whole Effect Task. The *Part Whole Effect Task* (Tanaka & Farah, 1993) is assumed to capture holistic processing because it compares how well people recognize facial features within the context of a whole face compared to facial features presented in isolation. In this task, participants are first presented with a face (upright or inverted). Then the face disappears and participants are then presented with two faces or two facial features. They are prompted with a word cue indicating to which feature they should attend (eyes, nose, or mouth) and must identify which feature (the one presented on the right or on the left) matches the same feature that was part of the first face they saw. Orientation (upright or inverted) is consistent throughout each trial. For a visual depiction of this tasks' trial structure see Figure 3.

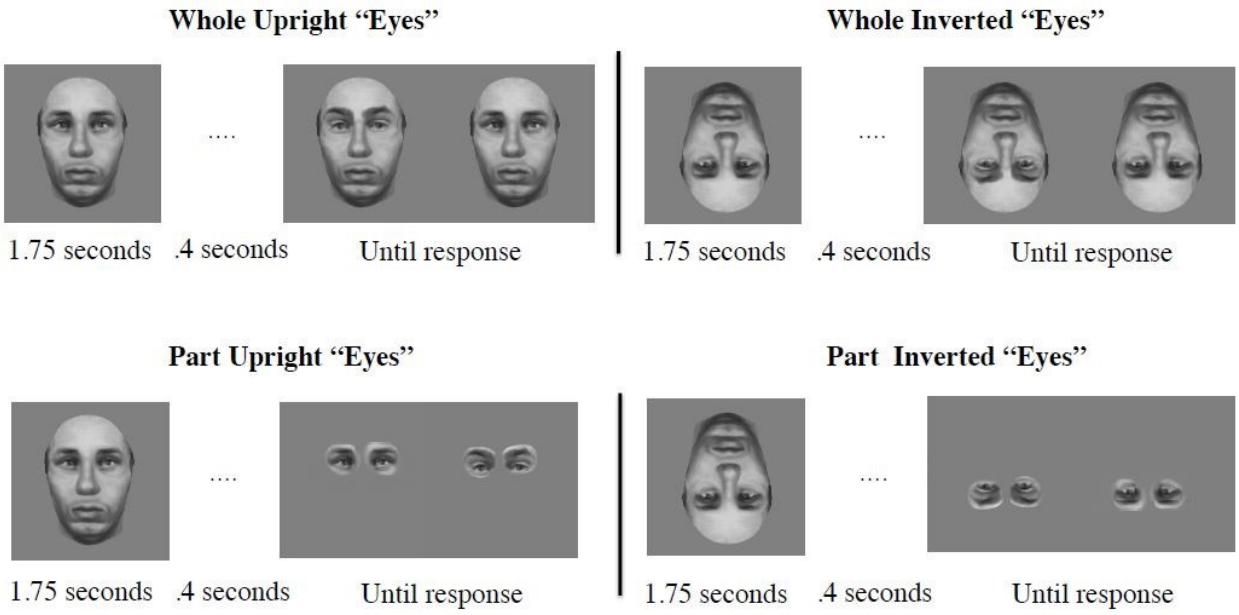

*Figure 3.*This figure depicts the stimuli from the first manuscript and trial structure for upright and inverted trials in the Part Whole Effect Task that pertain to eyes. The first row shows upright followed by inverted Whole condition trials. The second row shows upright followed by inverted Part condition trials.

Participants demonstrate a holistic advantage on this task in that they perform better in the Whole condition than in the Part condition on upright trials. On Part trials participants engage in featural processing only; this is because the only piece of information available to them in the stimuli is a single facial feature. On Whole Upright trials participants may also use featural processing; however, they also have configural and holistic information available to them. On Whole Inverted trials configural and holistic information is still available, but processing of them is thought to be impaired (Carey & Diamond, 1977; Goffaux & Rossion, 2006; 2007; Van Belle et al., 2010).

Configural/Featural Difference Detection Task. This task includes both featural and configural manipulations assumed to elicit featural and configural processing mechanisms respectively (Carbon & Leder, 2005; Friere et al., 2000). It also includes upright and inverted conditions for both types of trials in order to investigate the impact that inversion has. This is a

simultaneous discrimination task with two faces presented side by side. Both faces are either upright or inverted. Participants must decide whether the two faces are exactly the same or if they are different. Faces can differ based on a featural difference (one feature is swapped out with that same feature from another face, i.e., the two faces would be identical except for the eyes). Featural modifications also include the nose and the mouth. The other way faces can differ is based on a configural change, meaning that a facial feature was moved up or down 3 mm. This changes the *second-order* relations. Either the eyes or the nose or the mouth can be moved up or down. For a visual depiction of the trial structure for this task see Figure 4.

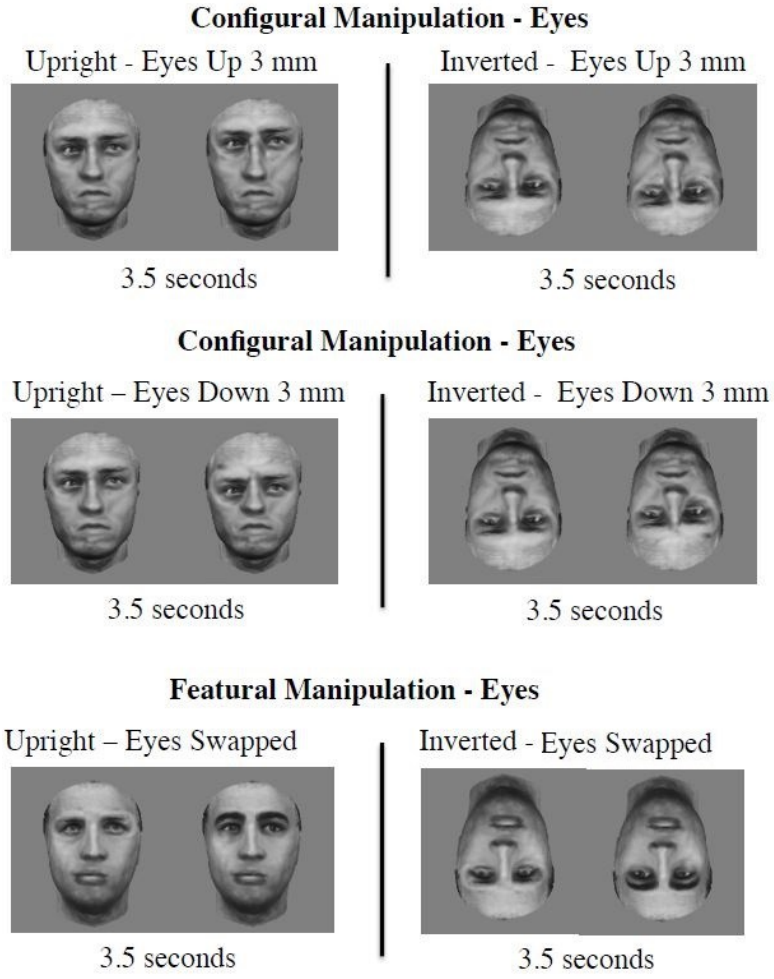

Figure 4. This figure depicts the stimuli from the first manuscript and trial structure for upright and inverted trials in the Configural/Featural Difference Detection Task that pertain to eye manipulation.

As previously stated, it is assumed that featural modifications will be detected by the featural processing mechanism and that configural modifications will be detected by a configural processing mechanism, although this remains to be determined. In the upright condition for in both types of trials participants could use a holistic strategy as well, or any combination of all three processing mechanisms. Prior research suggests that people are more sensitive to featural changes compared to second-order relational configural changes (Carbon & Leder, 2005; Freirer et al., 2000; Mercurer, Dick, & Johnson, 2008), therefore we expect that performance will be higher on featural trials compared to configural trials in the upright condition as well as in the inverted condition. In keeping with the research regarding inversion and integrative processing, it is expected that performance will be higher on upright trials compared to inverted trial for both featural and configural conditions, and importantly that performance will be impaired by inversion to a greater extent on configural trials compared to featural ones.

Review of Face Processing Tasks – Experiment 2

In the second experiment, we included three of the same tasks as in Experiment 1: *Partial Composite Face Effect Task*, *Configural/Featural Difference Detection Task*, and *Part Whole Effect Task*. We also included the *Complete* version of the *Composite Face Effect Task* (e.g., Farah, Wilson, Drain, & Tanaka, 1998; Richler et al., 2011b) and the *Cambridge Face Memory Test* (CFMT) (Duchaine & Nakayama, 2006). In this section an overview is given of the methodology and rationale behind the two tasks that were not part of Experiment 1.

Complete Composite Face Effect Task. The *Complete Composite Face Effect Task*, like the *Partial Composite Face Effect Task*, is thought to be a measure of holistic processing.

This is indexed by participants' inability to selectively attend to the target face halves (tops) resulting in interference from the task-irrelevant bottom-halves. The Complete and Partial versions of the *Composite Face Effect Task* differ in two meaningful ways. The first is that in the *Complete* version of the task stimuli are always presented in the upright orientation (vs. the Partial version, which includes both upright and inverted trials).

The second way these versions of the task differ is that instead of the task-irrelevant bottom halves of the test and inspection faces always being different (as in the *Partial* version) in the *Complete* version the bottom halves can match or be different. This manipulation is referred to as congruency. Trials are either Congruent or Incongruent. Congruent refers to the test and inspection face halves having the same degree of similarity. This means on "Same" Congruent trials, in which the top halves of the test and inspection faces match, the task-irrelevant bottom halves also match. And on "Different" Congruent trials, in which the top halves of the test and inspection faces do not match, the task-irrelevant bottom halves also do not match. In contrast, on Incongruent "Same" trials the top halves always match and the bottom halves are always different. Finally, on Incongruent "Different" trials the top halves are always different and the bottom halves always match. Richler et al. (2011b) present a diagram that may help the reader visualize the different trial types. For a depiction of the *Complete Composite Face Effect Task*'s trial structure refer to Figure 5.

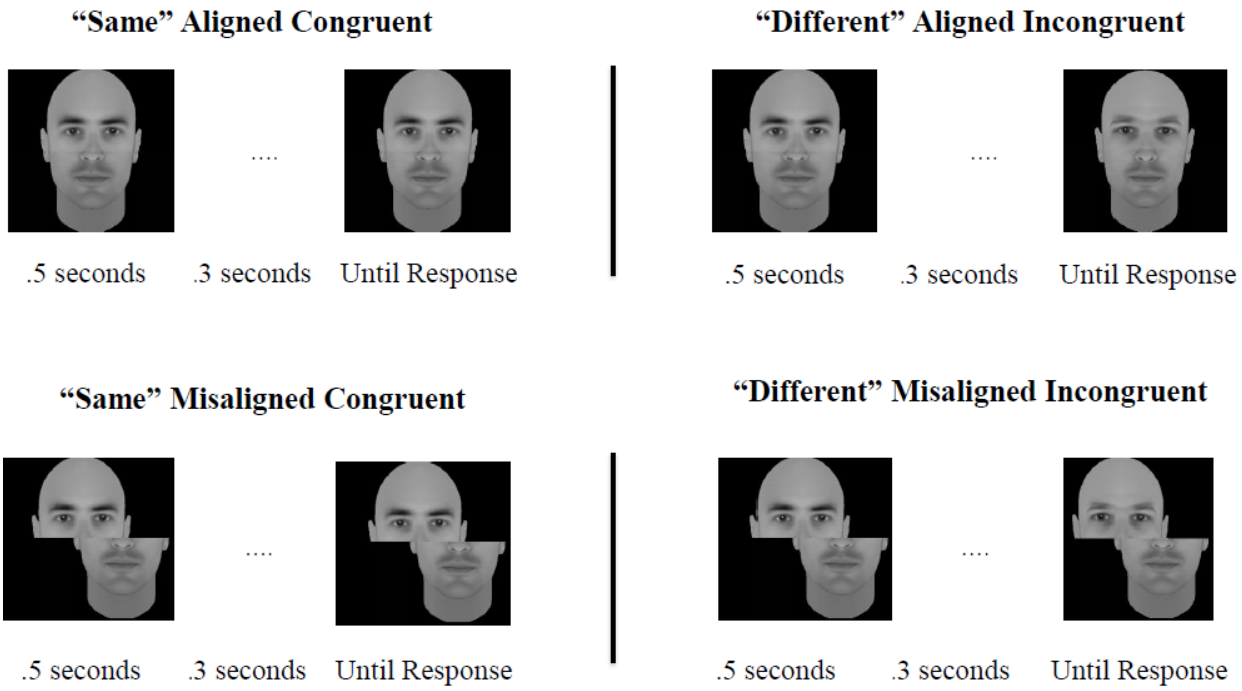

Figure 5. This figure depicts the trial structure of congruent "same" and incongruent "different" trials for both aligned and misaligned stimuli in the Complete Composite Face Effect Task. These are the "other half of the trials" that are not part of the Partial version.

What is similar between the two versions of the *Composite Face Effect Task* is that face composites are either Aligned or Misaligned, and participants are instructed to selectively attend to the top halves of the test and inspection faces in order to judge if the top halves match or are different. Again, the interaction effect is thought to index holistic processing. In the Complete Version of the *Composite Face Effect Task* the interaction effect is Congruency × Alignment. Performance is superior on congruent trials compared to incongruent trials but when composites are misaligned this reduces the magnitude of this congruency effect (Richler, Cheung, & Gauthier 2011a; Richler et al., 2011b). This is because misaligning the top and bottom face halves is thought to disrupt holistic processing (Richler, Tanaka, & Brown, 2008; Richler, Bukach, & Gauthier, 2009; Wong, Palmeri, & Gauthier, 2009). Like in the *Partial* version, featural information is available in all trials. Second-order configural and/or holistic is available on Aligned trials and to a lesser degree on Misaligned trials.

As will be discussed below in more detail, there is an ongoing debate regarding which version of the *Composite Face Effect Task* is the better measure of holistic processing (Richler & Gauthier, 2013; Richler & Gauthier, 2014); Rossion, 2013. We included both versions of this task in the second manuscript in order to: 1) Compare patterns of performance across both tasks to see, among other things, which was the most reliable, 2) Examine the associations between performance measures from each version of the composite task with those from the other face processing tasks, and 3) Examine which version was more strongly related to face recognition ability as reflected by performance on the CFMT.

Cambridge Face Memory Test (CFMT). The CFMT (Duchaine & Nakayama, 2006) was designed as a measure of general face recognition ability. In contrast to the other tasks presented to this point, it is not designed to assess any particular aspect of face processing, but rather the end result of all forms of processing working together. It is designed to minimize the usefulness of non-face-processing strategies (such as memorizing hair styles, ear shapes, blemishes, jewelry, etc.). It is the most commonly used standardized face recognition task (Cho et al., 2015). The CFMT tests people's ability to learn and subsequently recognize upright faces they have never seen before across different viewing angles as well as different levels of illumination and Gaussian noise. Each trial consists of a three alternative forced choice, wherein the participant must indicate which of the three face stimuli presents the same identity as a face that he or she learned during the study phases of the experiment. For a depiction of CFMT stimuli in various experimental blocks refer to Figure 1 in Duchaine and Nakayama (2006).

CFMT is considered to be a valid test of face memory, as it has high internal reliability, does not produce ceiling effects, and is relatively easy and quick to administer (McKone et al. 2012). Item Response Theory analysis supports the unidimensionality of the CFMT, meaning

that researchers are justified in reporting and analyzing the total score as a measure of performance (Cho et al., 2015). Moreover, Differential Item Functioning established that the test is appropriate to use for participants both above and below 20 years or age, and for both males and females (Cho et al., 2015).

To date, many researchers have used the CFMT as their measure of choice when attempting to investigate the associations(s) between face processing and face recognition ability, specifically the role that holistic processing plays in face recognition (e.g., DeGutis, Wilmer, Mercado, & Cohan, 2013; Dennette et al., 2012; McGugin et al., 2012, Richler et al., 2011b). Several pieces of evidence support the use of the CFMT for this purpose. This is because it is suggested that performance on the CFMT requires face-specific processing mechanisms. Examples of such evidence include that performance is impaired on CFMT-Inverted compared to CFMT-Upright (Duchaine & Nakayama, 2006), suggesting that CFMT taps into integrative face processing which is also impaired by inversion. Performance on the CFMT is also significantly correlated with performance on a measure of face processing called the Cambridge Face Perception Test ($r = .60$, $p < .001$) (Bowles et al., 2009). For these reasons, and because it is the measure of face recognition ability that is most commonly used, we have elected to include it as our measure of face recognition ability in Experiment 2.

Summary of Literature on Integrative Face Processing and Face Recognition

To date, results have been inconsistent across studies that have attempted to investigate the associations between face processing mechanisms as indexed by performance on face processing tasks. As previously mentioned, this may be due to myriad differences in methodologies and statistical approaches. These issues complicate attempts to investigate the role(s) that the various purported face processing mechanisms occupy in face recognition ability.

In the present section we will review the state of the face processing and face recognition literature. In the subsequent section we will review the methodological issues and summarize the debates regarding how best to address them.

First, we review results from the recent literature pertaining to the association between performance across various integrative face processing tasks. One of the studies most relevant to the present work is by Wang, Fang, Tian, and Liu (2012), who compared performance across the *Partial Composite Face Effect Task* and the *Part Whole Effect Task.* They found that performance on these two tasks was not significantly correlated. This finding is striking given that these two tasks are both thought to reflect holistic processing. DeGutis et al. (2013) investigated the patterns of performance between the *Part Whole Effect Task* and the *Complete Composite Face Effect Task*. In contrast to Wang et al. (2012), they found that performance on these two measures is significantly correlated ($r = .44$, $p < .01$). More recently, Richler and Gauthier (2014) conducted a meta-analysis investigating the effect sizes for the *Partial* and the *Complete* versions of the *Composite Face Effect Task* across studies and determined that the effect sizes from the two tasks were not significantly correlated. This suggests that assumptions about these tasks measuring the same construct are incorrect. The most recent study on this issue is by Rezlescu, Susilo, Wilmer, & Caramazza, (2017). They found a significant correlation between the *Face Inversion Effect Task* and the *Part Whole Effect Task* ($r = .28$, $p < .001$). However, correlations were not significant between performance on the *Face Inversion Effect Task* and *Partial Composite Face Effect Task*, nor between the P*artial Composite Face Effect Task* and the *Part Whole Effect Task*. Taken together, these results tend to support the statement that these tasks are not measuring the same construct. However, the picture is mixed, as there are

some significant correlations that emerge. Clearly, more research needs to be done to determine which integrative tasks correlate with one another.

Another way to investigate these integrative face-processing tasks is to examine how performance on each is related to facial recognition performance. Konar, Bennett, and Sekuler (2010) determined that performance was not significantly correlated between the *Partial Composite Face Effect Task* and their own test of upright face identification ability. However, when the CFMT was used as the measure of general face recognition ability, Richler et al. (2011b) determined that there was a significant correlation, with a medium magnitude ($r = .40$, $p = .01$), between face recognition ability and the Congruency × Alignment interaction in the *Complete Composite Face Effect Task* for *d*' as well as for reaction time ($r = .33$, $p = .04$). Results from DeGutis et al. (2013) indicate that there was a significant correlation between CFMT performance and performance on the *Complete Composite Face Effect Task* ($r = .36$, $p < .05$) as well as the *Part Whole Effect Task* ($r = .46$, $p < .01$). Rezlescu et al. (2017) determined that performance on the *Face Inversion Effect Task* was significantly correlated with their chosen measure of face recognition, the *Cambridge Face Perception Task* ($r = .42$, $p < .001$), as was performance on the *Part Whole Effect Task*, although with a smaller magnitude ($r = .25$, $p < .001$). Verhallen et al. (2017) compared performance on the CFMT and on the *Partial Composite Face Effect Task* and found that there was no significant correlation, nor was there are any between the *Partial Composite Face Effect Task* and the *Mooney Face Test* or the *Glasgow Face Matching Test*. These results are inconsistent and it remains uncertain to what extent the three face processing mechanisms are related to face recognition ability. This is perhaps not surprising because the extent to which the various integrative face-processing tasks are related, and what they are actually measuring, remains to be soundly established.

Methodological Limitations and Debates About How to Address Them

Methodological limitations are at least partially responsible for the state of confusion that exists with respect to the orthogonality of face processing mechanisms and their respective roles in face recognition. In this section we review a number of the key methodological limitations in the literature and summarize the ongoing debates regarding how best to address them. We then outline the approaches we have taken in this thesis to mitigate these limitations.

To date, researchers investigating face processing and face recognition have reported results based on accuracy, reaction time, and *d*' data. However, because the nature of featural, configural, and holistic processing mechanisms is not fully understood, it remains to be determined which performance measure(s) best reflects each mechanism. This issue is complicated further by the fact that evidence suggests that the tasks purporting to index these processing mechanisms may not be doing so in comparable ways. This could mean that in one task thought to index holistic processing (e.g., the *Part Whole Effect Task*) the holistic effect is best reflected in accuracy data while in another (e.g., the *Partial Composite Face Effect Task*) the holistic effect is best reflected by reaction time or *d*'. The way we address this issue is by analyzing both accuracy and reaction time. We report on accuracy within the main manuscripts and include reaction time analyses in the Appendices. This way, if there is a meaningful difference in the group-level performance data between accuracy and reaction time, we can bring this to the attention of the reader. That stated, in the present two studies, accuracy and reaction time results largely mirror each other.

In addition to group-level data regarding accuracy, reaction time and d', some researchers report individual difference scores. This again may be contributing to apparently inconsistent results across studies because individual difference data captures performance in a different

manner than group-level analysis does. For example, group-level data averages performance across participants in order to investigate differences in performance between conditions. This approach considers differences in performance between participants to be noise. However, this is not necessarily the best approach, as such differences could hold important information. Bosten, Mollon, Peterzell, and Webster (2017) note that precisely this type of information derived from an individual difference approach could "reflect differences in the mechanisms underlying perception and performance" (p. 1). Yovel, Wilmer, and Duchaine (2014) note that using an individual differences approach compliments and extend findings based on group-level data. Therefore, we chose to present both group-level data as well as individual difference data in the present work.

Because each task typically reveals a particular effect via a significant interaction, one could argue that it is not the performance on each of the conditions within the task that reveals integrative processing, but rather the performance differences between pairs of conditions. For example, in the *Face Inversion Effect Task*, it is the finding that inversion disproportionately impairs the recognition of faces as compared to non-face objects that is taken as evidence of integrative face processing. To probe this possibility, we investigated individual difference scores using two methods: subtraction and regression (DeGutis et al., 2013).

Analyzing individual difference scores allows us to investigate the effect that orientation has within a pair of conditions in each task. For example, in the *Face Inversion Effect Task* the pair of conditions that make up the "Face" individual difference scores would be: Face Upright and Face Inverted. A similar logic applies for the calculation of "Object" individual difference scores. The pair of task conditions would be: Object Upright and Object Inverted. Within this manuscript we present both subtraction-based and regression-based (a.k.a., d-residuals; Degutis

et al., 2013) individual difference scores with respect to orientation. This is because orientation is widely held to result in impairments in facial recognition due to disruption of integrative processing, while there is less consensus regarding what the various other task manipulations (e.g., alignment, part vs. whole, face vs. object) reflect.

To illustrate one of the debates in the literature regarding analysis methods, let us use two studies that we have previously mentioned as an example. Konar et al. (2010) and Richler et al. (2011b) used the subtraction method of calculating difference scores to investigate the patterns of performance on facial recognition tasks and their results were inconsistent with one another. DeGutis et al. (2013) proposed an explanation for why there are inconsistent results amongst these studies. That is, these authors argued that the subtraction method is not the appropriate way to calculate difference scores.

The rationale behind the subtraction method of calculating individual difference scores is as follows. Within each task, the researcher identifies the control condition and the condition of interest. For example, in the *Part Whole Effect Task*, the "Whole Upright" condition is thought to be the measure of holistic processing and is thus the condition of interest. Therefore, the performance score on the "Whole Inverted" condition is subtracted from the performance score on the "Whole Upright" condition to produce the "Whole" individual difference score. According to DeGutis et al. (2013), the primary concern that arises from using the subtraction method is that variability from the control condition confounds the results. Variability from the control condition and the condition of interest both impact the calculation of subtraction-based difference scores. It would be possible to obtain a low subtraction score for multiple reasons: if the participant does very well on the "Whole Inverted" condition of the *Part Whole Effect Task*,

or if the participant does very poorly on the "Whole Upright" condition, or because of a combination of both outcomes.

Conversely, Degutis et al. (2013) point out, this issue does not arise when using the regression method to calculate individual difference scores, termed d-residuals, because the way the regression-based difference scores are calculated is such that they only contain the variability from the condition of interest. It is for this reason that DeGutis et al. (2013) promote the use of regression to generate residual scores, termed d-residuals, over the use of subtraction-based difference scores. D-residuals are calculated as follows: Performance in the condition of interest (e.g., the "Whole Upright" condition in the *Part Whole Effect Task*) is regressed against performance in the control condition (e.g., "Whole Inverted" condition), and the regression line of best fit is obtained. Then one uses the equation of the line of best fit to calculate each participant's expected score on the condition of interest, given their performance in the control condition. The values of the control condition (e.g., "Whole Inverted") are entered into the equation of the regression line of best fit to generate expected performance scores in the condition of interest (e.g., "Whole Upright"). The expected scores are then subtracted from the actual performance scores on the "Whole Upright" condition to obtain the "Whole" d-residual value for each participant. Thus, the d-residual indicates the degree to which a given participant's performance is above or below the expected value. A high residual, in the example of the *Part Whole Effect Task*, indicates that an individual is performing better on the "Whole Upright" condition trials than would be expected based on their performance on the "Whole Inverted" condition trials. This may suggest that this person is exhibiting an especially holistic processing style.

There is disagreement in the literature concerning which type of individual difference score, subtraction-based or regression-based, is the more reliable. Reliability is an important issue when examining the correlation between measures, because outcomes from tasks with low reliability will necessarily correlate poorly with one another, regardless of whether the tasks are measuring the same construct or not. DeGutis et al. (2013) found that d-residuals had higher reliability scores than did subtraction scores using the *Part Whole Effect Task* and the *Complete Composite Face Effect Task*. However, Ross, Richler, and Gauthier (2015) compared the reliability of subtraction scores and d-residuals on the *Complete Composite Face Effect Task* using data from multiple data sets and determined that regression residual scores were not more reliable than subtraction scores. Because these results contradict those of DeGutis et al. (2013), and because researchers have only investigated a subset of face processing tasks in terms of comparing performance and reliability of both subtraction scores and d-residuals, we elected to analyze and report on both sets of individual difference scores. Calculating and analyzing both types of individual difference scores makes it possible to compare Guttman's λ_2 measure of reliability across the two methods.

Reliability is one aspect that should be considered when choosing which type of individual difference score to use; however, researchers should also consider the underlying assumptions regarding the conditions of interest and the control conditions within each task (Ross et al., 2015). To date, the problem remains that it has not been empirically established what each task condition is measuring. Therefore, it remains unclear which, if any, task conditions should be conceptualized as purely "control conditions" and thus candidates for regression-based individual difference calculations. This is because regression removes the variability from the "control condition" so it does not influence the calculation of this type of

individual difference score. It is also unclear which, if any, task conditions should be conceptualized as impure "control conditions" (meaning that the "control condition" does contain elements of the construct of interest and that therefore these individual difference scores should be calculated using the subtraction-based method which includes the variability from both pairs of scores). As a result, we elected to use both methods to calculate of individual difference scores.

As we have mentioned throughout this section, it has not yet been established what the various integrative face processing tasks are actually tapping into. Any given task may be indexing one or more of the processing mechanisms, or some form of higher-order processing, or something entirely different. Fully addressing this methodological issue is beyond the scope of the current project. However, what our data can do is answer an important preliminary question; That is, are these commonly-used face processing tasks interchangeable measures in integrative processing? We accomplish this by comparing patterns of performance across multiple performance measures (i.e., group and individual difference scores for both accuracy and reaction time) on a number of commonly used face processing tasks. We compare performance using correlational analysis as well as factor analysis.

One important point of comparison is between performance on the *Complete* and the *Partial* versions of the *Composite Face Effect Task* in Experiment 2. We bring this to the reader's attention here because there is an ongoing debate concerning which is the better measure of holistic processing. A detailed account of the debate can be found in Rossion (2013), who advocates for the *Partial* version, and Richer and Gauthier (2013), who advocate for the *Complete* version. Other publications, such as Richer and Gauthier (2014)'s meta-analysis and review of holistic processing, also discuss the two versions of the *Composite Face Effect Task* at

length. In brief, Rossion (2013) makes over 160 arguments against the use of the Composite version, most of which have been refuted (Richler & Gauthier 2013). Richler and Gauthier argue that the *Complete* version is the superior paradigm because this version of the task has the other half of the trial types that the *Partial* design lacks (i.e., Congruent Same and Incongruent Different), and as a result its results are not subject to response biases the way results from the *Partial* design are. By including both versions we can compare reliability scores and patterns of performance between them in order to see how similarly or differently they are related to the other processing tasks and the CFMT. Thus, we are able to contribute to this ongoing debate.

To summarize, the state of affairs within the domain of face processing and face recognition is such that conceptualizations of processing mechanisms are inconsistent and incomplete. Moreover, tasks designed to capture one or more of these processing mechanisms may not in fact do so; Researchers have yet to establish empirically what construct(s) each task indexes. Furthermore, there is limited research investigating patterns of performance between the face processing tasks. In addition, and perhaps most importantly, all of these issues call into question the body of research concerned with determining the extent to which face recognition depends on one or more of these processing mechanisms.

Research Questions and Hypotheses

Before presenting the two manuscripts that make up the central portion of this thesis, we here present the main questions and hypotheses that motivated them. We present relatively general hypotheses here, and provide more detail in each of the manuscripts.

In the first manuscript the central question was: **What is the nature of the associations between performance (accuracy and reaction time) across a variety of integrative face-processing tasks?** This is an important first step that needs to be taken in order to test (and

potentially correct) the widely held assumption that these tasks are measuring the same aspects of integrative processing in face recognition. Our research design allows us to investigate how performance is correlated across tasks and how performance outcomes load into factors. If performance across tasks were strongly correlated and loaded onto a common factor, then this would suggest that they are essentially interchangeable measures of integrative processing. If correlations were weak across tasks and the tasks load onto separate factors, then this would suggest that the tasks are not interchangeable measures, which could account, at least partially, for the discrepancies in results across studies in the past literature.

In the second Experiment we endeavored to answer three main questions: **1) Are there distinct holistic and configural processing mechanisms in face recognition? 2) If so, to what degree are they predictive of face recognition ability? and 3) Which tasks are best suited to evaluate integrative processing?** As with Experiment 1, we used correlational analysis and factor analysis to address these questions. While it remains to be determined what exactly each face-processing task is measuring, our data will at least show how performance across them is related. Factor analyses in particular will help to establish if the tasks (and their conditions) load in such a way that supports the possibility that they are indexing distinct integrative mechanisms. Regarding the second research question, our data provide evidence regarding which tasks (and conditions thereof) are related to face recognition ability. This was examined by determining if any significant correlations exist between performance on the CFMT and on four widely-used integrative face processing tasks. We also examined this question in terms of factor loading patterns, with the hypothesis that outcomes from integrative face processing tasks would load onto the same factor as the CFMT. With respect to the third question, this is a more challenging endeavor because what constitutes a good (or best) measure of integrative processing is open to

debate. However, by comparing patterns of performance across behavioural measures on a variety of face processing tasks at the group-level and individual difference level, their respective reliabilities, and with the role they play in face recognition as measured by performance on the CFMT, our data will provide some information regarding the psychometric characteristics of the various measures. More generally, we have aimed to provide data that will inform where the fields of integrative face processing and face recognition currently stand, and to provide an empirically sound base from which researchers can build towards answering the numerous more complex questions that remain unanswered, such as what each of these tasks is actually measuring.

References

Bosten, J. M., Mollon, J. D., Peterzell, D. H., & Webster, M. A. (2017). Individual differences as a window into the structure and function of the visual system. *Vision Research*, *141*, 1-3.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., ... & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423-455.

Carbon, C. C., & Leder, H. (2005). When feature information comes first! Early processing of inverted faces. *Perception*, *34*(9), 1117-1134.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, *195*(4275), 312-314.

Cho, S. J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., ... & Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, *27*(2), 552-256.

Chua, K. W., Richler, J. J., & Gauthier, I. (2015). Holistic processing from learned attention to parts. *Journal of Experimental Psychology: General*, *144*(4), 723 – 729.

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*(1), 87-100.

Dennett, H. W., McKone, E., Edwards, M., & Susilo, T. (2012). Face aftereffects predict individual differences in face recognition ability. *Psychological Science*, *23*(11), 1279-1287.

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*(4), 576-585.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception? *Psychological Review*, *105*(3), 482 – 498.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception*, *29*(2), 159-170.

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*(2), 431-446.

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*(4), 1023-1039.

Goffaux, V., & Rossion, B. (2007). Face inversion disproportionately impairs the perception of vertical but not horizontal relations between features. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 995 – 1001.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science*, *21*(1), 38-43.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(2), 513-536.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255-260.

McKone, E., Davies, A. A., Darke, H., Crookes, K., Wickramariyaratne, T., Zappia, S., ... Fernando, D. (2013). Importance of the inverted control in measuring holistic face processing with the composite effect and part-whole effect. *Frontiers in Psychology*, *4*, 1-33.

McKone, E., Stokes, S., Liu, J., Cohan, S., Fiorentini, C., Pidcock, M., ... & Pelleg, M. (2012). A robust method of measuring other-race and other-ethnicity effects: The Cambridge Face Memory Test format. *PLoS One*, *7*(10), e47956.

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences*, *109*(42), 17063-17068.

Meinhardt-Injac, B., Boutet, I., Persike, M., Meinhardt, G., & Imhof, M. (2017). From development to aging: Holistic face perception in children, younger and older adults. *Cognition*, *158*, 134-146.

Mercure, E., Dick, F., & Johnson, M. H. (2008). Featural and configural face processing differentially modulate ERP components. *Brain Research*, *1239*, 162-170.

Mondloch, C. J., Pathman, T., Maurer, D., Le Grand, R., & de Schonen, S. (2007). The composite face effect in six-year-old children: Evidence of adult-like holistic face processing. *Visual Cognition*, *15*(5), 564-577.

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1961-1973.

Richler, J. J., Bukach, C. M., & Gauthier, I. (2009). Context influences holistic processing of nonface objects in the composite task. *Attention, Perception, & Psychophysics*, *71*(3), 530-540.

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), 1- 13.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011b). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464-471.

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254-276.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281–1302.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology*, *3*, 1-6.

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*(6), 1356 - 1368.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*, *47*(3), 736–743.

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

Stein, T., Sterzer, P., & Peelen, M. V. (2012). Privileged detection of conspecifics: Evidence from inversion effects during continuous flash suppression. *Cognition*, *125*(1), 64-79.

Sunday, M. A., Richler, J. J., & Gauthier, I. (2017). Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Attention, Perception, & Psychophysics*, *79*(5), 1453-1465.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*(2), 225-245.

Van Belle, G., De Graef, P., Verfaillie, K., Rossion, B., & Lefèvre, P. (2010). Face inversion impairs holistic perception: Evidence from gaze-contingent stimulation. *Journal of Vision*, *10*(5), 1-13.

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision research*, *141*, 217-227.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169-177.

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*(9), 1108-1117.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141-145.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747-759.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256-2262.

Yovel, G., Wilmer, J. B., & Duchaine, B. (2014). What can individual differences reveal about face processing? *Frontiers in Human Neuroscience*, (*8)*, 1-9.

Chapter 2: Investigating the Orthogonality of Performance on Four Commonly Used Integrative Facial Processing Tasks

This study has been submitted to the Journal of Experimental Psychology: Human Perception and Performance and is formatted in accordance with the requirements of that journal.

We note that because this manuscript is formatted as a stand-alone publication, much of the material in the introduction will repeat that which presented in the general introduction of the thesis.

Abstract

Researchers have used a wide variety of tasks to investigate integrative aspects of face recognition (i.e., holistic and configural processing). This study examined patterns of performance across four of the most commonly used tasks in this area: The Face Inversion Effect Task, the Part Whole Task, the "partial" Composite Face Effect Task, and the Configural/Featural Difference Detection Task. We employed a within-subjects design (N = 223) and used correlation and exploratory factor analyses to investigate patterns of performance (accuracy, reaction time, and individual difference scores) across task conditions. Performance on most task conditions was modestly correlated; however, the strongest correlations were between conditions within each task. Exploratory factor analysis corroborated these results: The unrotated factor solutions showed task conditions loading onto a common first factor while the rotated factor solutions showed each task loading onto its own respective factor. There was no evidence that the tasks grouped into separate configural vs. holistic sets. In summary, we find that all four tasks measure the same underlying construct to a modest degree, but that they are each primarily measuring something unique. Therefore, these tasks should not be taken as interchangeable measures of the integrative aspects of face processing.

Keywords: face recognition, holistic processing, featural processing, configural processing, individual difference scores

Public Significance Statement

This study demonstrates that while four common face-processing tasks all tap into the same underlying construct, they do so to only a modest degree. Moreover, these tasks also each would appear to measure independent components of face processing. Therefore, we advise researchers against using and interpreting these tasks interchangeably. Because of the fact that these tasks have been used interchangeably to date, we suggest that researchers and consumers of research should only compare patterns of performance on facial recognition tasks if the same task methodology has been used across studies.

Introduction

The concepts of configural and holistic processing have driven a great deal of research in the field of face recognition for many decades. Throughout this time, there has been debate about how to define these concepts and how to best measure them (e.g., Carey & Diamond, 1977; Maurer, Le Grand, and Mondloch, 2002; Rakover, 2002; Tanaka & Sengco, 1997; Wang, Guo, and Fu, 2016; Burton, Schwienberger, Jenkins, & Kaufmann, 2015; Gold, Mundy, Tjan, 2012; Konar, Bennett, & Sekuler, 2010; Richler, Cheung, & Gauthier, 2011). In order for the study of facial recognition to progress, there must be comprehensive and generally accepted operational definitions of the terms *featural*, *configural*, and *holistic* within the context of visual processing mechanisms (Tanaka & Farah, 1993). This is a necessary precondition before we can have confidence regarding the experimental tasks designed to study them. For example, numerous commonly-used tasks purport to measure various aspects of featural, configural, and holistic information, but it has yet to be established if and how such tasks measure this information, and how this information, as measured by these tasks, relates to the processing mechanisms that underlie facial recognition ability.

The purpose of the present study is to compare performance across four of the most commonly used integrative face processing tasks in order to determine if and how performance on said tasks is related. This will indicate if these tasks are measuring the same or different construct(s) (i.e., processing mechanism) or component(s) thereof. We examined this question via both correlational analyses and exploratory factor analyses. Using correlational analyses we examined the associations between performance on the 16 task conditions tested across the four tasks measured. That is, each of the four tasks comprises four task conditions. Using exploratory factor analysis, we examined whether the different tasks tap into a unitary

mechanism or separable mechanisms. We analyzed group level performance using data from the 16 task conditions, as well as individual difference scores, which were calculated via both subtraction and regression (see Individual Difference Scores section below). The second purpose of our study was to examine how the four integrative face processing tasks load onto underlying factors, and to what degree this is compatible with the idea that they measure separate configural and holistic aspects of face processing.

In their 2002 paper, Maurer and colleagues noted that a lack of consensus remained among researchers regarding the nature of featural, configural, and holistic processing. To help address this issue they provided the following definitions: *Featural* processing can be thought of as piecemeal processing, whereby characteristic information about the appearance of an individual's isolated facial features is processed. The authors used findings from cognitive psychology, developmental psychology, and neuroscience to substantiate the existence of three types of *configural* face processing: 1) sensitivity to first-order relations (recognizing that faces consist of two eyes, above a nose, above a mouth), 2) sensitivity to second-order relations (attending to the spacing or distance between features), and 3) *holistic processing* (which combines the facial features and the first and second order configurations into a gestalt, or a whole). These definitions are currently the most commonly used in the facial recognition literature, although they are not used with consistency (Richler, Palmeri, & Gauthier, 2012). Richler et al. (2012) note that much work remains to be done with respect to developing more precise terminology, especially with regard to holistic processing.

In addition to the lack of consistency regarding the use of face-processing terminology, there are also conflicting perspectives regarding the existence and the importance of holistic processing within the domain of facial recognition. For instance, two studies investigating the

patterns of performance across tasks thought to measure holistic processing and measures of general face recognition ability have generated incompatible results (Konar et al., 2010; Richler et al., 2011). Konar et al. (2010) found no correlation (r = .05, p = .66) between holistic processing (as measured by the *d'* on the "partial" *Composite Face Effect* task[1]) and upright face identification ability; whereas Richler et al. (2011) found a medium strength correlation (r = .40, p = .01) between performance (as measured by *d'* on the *Cambridge Face Memory Test* (CFMT; (Duchaine & Nakayama, 2006) and the "complete" *Composite Face Effect* task (also see Richler, Floyd, & Gauthier, 2015).[2] Given the lack of consensus regarding how to operationally define holistic processing, and because different tasks are being used that may not be interchangeable measures, such disagreements between studies are unsurprising. Consistency will only be possible once we understand more precisely what the various tasks used to index integrative aspects of face processing are actually measuring. This will provide the necessary information to ensure that comparable tasks are being used across studies.

In addition to attempting to determine the association between integrative processing and face recognition ability, researchers have also investigated correlations between performance on different face processing tasks that are thought to measure integrative processing. Specifically,

---

[1] This refers to the original paradigm in which the bottom halves of faces are always "different". There are two versions of the C*omposite Effect Task*, both of which are sequential matching paradigms. Refer to Richler et al., (2011) for a description of the differences between what they refer to as the "complete" version and the "partial" version of this task. The so-called "partial" version of the task has also been referred to as the "standard" version (Rossion, 2013), while the "complete" version has also been referred to as the "congruency effect" (McKone et al., 2013). In this manuscript we implemented the "standard" or "partial" version, and we will simply refer to the task as the *Composite Face Effect Task*. When referring to other studies that used the *Composite Face Effect Task* we will differentiate between the versions by using "partial" or "complete" before naming the task.

[2] Their findings suggest that the magnitude of thecorrelation between performance data on the "complete" *Composite Face Effect* task and the CFMT depends on the extent of stimulus repetition.

researchers have examined performance on the "partial" *Composite Effect Task* and the *Part Whole Effect Task* (Wang, Fang, Tian & Liu, 2012). Results from this study suggest that there is no correlation ($r = .00$, $p = .55$) between these measures of integrative processing, even though they are purported to measure the same construct. Moreover, Richler and Gauthier (2014) conducted a meta-analysis of studies that have used the "partial" and the "complete" versions of the *Composite Face Effect* task and determined that there were no significant correlations between effect sizes. These findings suggest that there are distinct abilities required to complete these tasks successfully. Consequently, it may be that each of these distinct capabilities makes a unique contribution to one's face recognition ability.

Researchers have recently suggested that the lack of significant correlations between integrative processing tasks (as well as their lack of consistency regarding their correlations with face recognition ability (e.g., performance on the CFMT)) could be due to the fact that these tasks were designed for group level analysis. Task conditions often lack adequate reliability required to analyze individual difference scores (Sunday, Richler, & Gauthier, 2017). Individual difference scores are a measure of performance between a pair of task conditions, and are typically used to index configural or holistic processing.

Other researchers have proposed that the lack of significant correlations between integrative processing tasks (e.g., Konar et al., 2010; Richler et al., 2011, Wang et al., 2012) arises due to the method by which individual difference scores are calculated (DeGutis, Wilmer, Mercado, & Cohan, 2013). Using what they argue is a superior method for calculating individual difference scores, DeGutis et al. (2013) investigated the correlations between performance on two tasks purporting to measure integrative processing (*Part Whole Effect Task*, and the "complete" *Composite Effect Task*), as well as their association with face recognition

memory as measured by the CFMT. They demonstrated that the two measures of holistic processing do significantly correlate with each other ($r = .44$, $p < .01$). Furthermore, the "complete" *Composite Face Effect* task and the *Part Whole Effect* task both significantly correlated with performance on the CFMT ("complete" *Composite Effect* task: $r = .36$, $p < .05$; *Part Whole Effect* task: $r = .46$, $p < .01$). Importantly, this was the case when individual difference scores were calculated using the *regression method* rather than the *subtraction method,* the latter being the method used in previous studies. These methods for calculating individual difference scores, and their relative advantages and disadvantages, are further discussed in a later subsection.

In order for research in the field of facial recognition to advance systematically, researchers require as complete an understanding as possible regarding the types of face processing mechanisms that are involved in encoding the identity of familiar and unfamiliar faces. An important aspect of this understanding is that we must be able to empirically differentiate between the mechanisms that underlie face processing. The most common method researchers have used to explore the putative types of face processing is by investigating the interaction effects from a variety of face recognition paradigms (e.g., the tasks we have chosen to include in this manuscript).

For instance, in the *Part-Whole Task* (Tanaka & Farah, 1993) participants are shown a face either right side up or upside down. This face disappears and then participants are presented with two faces or two features (eyes, noses, or mouths) in the same orientation as the study face. Participants are cued to which feature they must base their "match" answer indicating which feature was part of the study face. Participants find it easier to correctly match a face part (eyes or nose or mouth) in the context of a whole face compared to when that same feature is presented

in isolation, but this effect is diminished when the stimuli are inverted. This interaction effect is thought to reflect holistic processing.

A similar logic applies to the *Composite Face Task* (Young, Hellawell, & Hay, 1987). In this task, the face stimuli are composites, meaning they are composed of a top half and a bottom half from different faces. These face halves are either aligned to create an intact face or the halves are offset horizontally, separating them. These aligned or misaligned faces are presented upright or inverted. In this sequential matching task participants are asked to indicate if the test and inspection top face halves match or are different. An interaction between the alignment state of face halves and the orientation of the stimuli is thought to also reflect holistic face processing. In this case, participants respond more accurately (or quickly) when faces are misaligned upright compared to aligned upright, but this advantage is diminished when faces are inverted.

Importantly, it is not clear if the interactions seen in these different tasks reflect the same underlying mechanism. It is intuitively appealing to believe that the task outcomes reflect a similar construct, because the tasks both involve comparing parts of faces to whole faces, and because the interaction effects are similar in structure. Based on this, it is widely assumed in the literature that these and a number of other commonly employed facial recognition paradigms index the same construct. However, as described above, some studies have cast doubt on this assumption (e.g. Richler et al., 2011; Richler & Gauthier, 2014; Wang et al., 2012).

One way to investigate if a set of facial recognition tasks is measuring the same construct, or component thereof, is to use statistical approaches such as exploratory factor analysis. Factor analysis can be employed to compare performance across tasks, as well as performance within tasks, with the results indicating how performance on each of the conditions within each task loads onto a set of factors. In this case a factor might represent a face processing mechanism, or

a component thereof. Results from the exploratory factor analyses would provide evidence regarding whether commonly used face recognition tasks are measuring the same or different constructs. If the pattern that emerges suggests the tasks are measuring different constructs, then this could serve to provide a possible explanation as to why there are discrepant results across studies that are ostensibly trying to measure the same construct.

Performance Measures

A question that arises when comparing performance across tasks concerns which behavioural measures to compare. In this study, we analyzed and reported both accuracy and reaction time. Analyzing one or the other of these measures would not provide an adequate reflection of the associations between performance scores on each of the tasks. This is because it is not known if face processing is better captured by accuracy or by reaction time, and this may differ depending on the task. Within the manuscript, we report accuracy analyses. Reaction time analyses are presented in Appendix A. This approach was chosen because the patterns of results between the two performance measures are very similar. Where there are substantive differences between the accuracy and reaction time results, this has been noted in the manuscript.

The four facial recognition tasks that are assessed in the present study are: the *Face Inversion Effect Task* (Yin, 1969), the *Part-Whole Effect Task* (Tanaka & Farah, 1993) the *Composite Face Effect Task* (Young et al., 1987), and the *Configural/Featural Difference Detection Task* (Freire, Lee, & Symons, 2000; Leder & Bruce, 2000). These tasks were selected because of their prevalence in the literature. It is presumed, based on previous literature, that each task primarily indexes one of two integrative face-processing types (i.e., second-order configural vs. holistic). In some instances the tasks are thought to contrast these higher-order-face-processing types against featural processing. For instance, the *Part Whole Effect Task* is

thought to primarily tap into holistic processing. This is because it directly compares the effect of providing a holistic face context vs. not doing so (i.e., when presenting facial features in isolation). The *Composite Effect Task* is likewise thought to mainly examine holistic processing, as it examines the recognition of whole faces vs. distinguishable face halves. Conversely, the *Configural/Featural Task* is thought to primarily index configural processing, as it directly examines the effects of changing the position (and therefore the configuration) of features vs. the appearance of features themselves through a featural swap (featural processing). The case of the *Face Inversion Effect Task* is less clear. Here we are simply examining the effect of inversion on the recognition of faces vs. objects (in this case, chairs), so this task may index both configural and holistic processing at once.

We note that within our experiment we have implemented versions of each task that are similar to the original authors' implementations. Some authors have argued in favour of alternative versions of some of these paradigms (e.g., the "complete" version of the *Composite Effect Task*; Gauthier & Bukach, 2007; Richler et al., 2011; Richler & Gauthier, 2013, Richler & Gauthier, 2014). Others have critiqued the methodology of some of the tasks. For instance, it has been argued that featural manipulations within the *Configural/Featural Difference Detection Task* are not purely featural manipulations, but instead are necessarily a combination of featural and configural changes (Riesenhuber & Wolff, 2009). Moreover, it has been suggested that configural manipulations in this task are too extreme compared to naturally occurring configurations of features and that therefore performance on the task is not generalizable to face recognition in the real world (Taschereau-Dumouchel, Rossion, Schyns, & Gosselin, 2010). However, it is not our intent to defend or promote any particular version of the tasks. Instead, we employed procedures that we judged to be most widely used, so that our data regarding the

associations between performance across tasks would have the widest possible applicability to the literature.

In addition to group-level performance data, we also investigated patterns across individual difference data (both subtraction-based and regression-based scores). Because integrative processing is thought to be evidenced by interaction effects within each of the tasks, it is possible that performance indexing integrative processing is better captured by difference scores than by performance on the individual task conditions. It is for this reason that we chose to include individual difference scores as additional behavioural measures. They allow us, for example, to investigate the extent to which the orientation of stimuli impacts a pair of task conditions. We can investigate this pattern across tasks to determine if inversion impacts performance on them in comparable ways. This is an interesting question to ask because inversion is thought to disrupt integrative face processing to a greater extent than featural processing (e.g., Murray, Yong, & Rhodes, 2000; Tanaka & Farah, 1993), and this pattern may provide clues as to which of these two processing types these tasks are indexing.

In the following sections, we describe the four face recognition tasks that we examine, noting how they purportedly capture featural, configural, and/or holistic processing. We also note expected results within each task. Hypotheses regarding the expected associations between tasks are given below in their own subsection.

Face Inversion Effect Task. A variety of methods have been used to examine the *Face Inversion Effect*. In one typical paradigm, a sequential matching task is used in which one face or one object is first presented either upright or inverted. Subsequently, the viewer must identify which of two simultaneously-presented faces or objects matches the stimulus he or she saw initially (See Figure 1). In neurotypical humans, inversion disrupts recognition ability for faces

to a greater degree than it does for other visual objects (e.g., Haxby et al., 1999; Yin, 1969). When a face is presented upside-down it is difficult for humans to encode it (Freire et al., 2000) and subsequently to recognize it (Diamond & Carey, 1986; Valentine, 1988; Yin, 1969). This finding is known as the Face Inversion Effect, and the common contention is that it arises because inversion disrupts configural and/or holistic processing (e.g., Murray et al., 2000; Tanaka & Farah, 1993). Inversion thus leaves the viewer to rely on 1) featural information, which is less impacted by inversion than configural and/or holistic processing is, as well as 2) impaired configural and/or holistic processing (Richler, Mack, Palmeri, & Gauthier, 2011). It is unclear if performance on the *Face Inversion Effect Task* indexes primarily configural processing, holistic processing, or both.

We expect to see the standard face inversion effect in our data. That is, we expect a greater effect of inversion on performance on face trials than on object trials (see Figure 1 for examples of trial structures in the Face Inversion Effect Task). In this case the control objects are chairs, a commonly used class of control object (e.g. Stein, Sterzer & Peelen, 2012; Yovel & Kanwisher, 2005).

Part Whole Effect Task. In this sequential matching task, participants first see a face that is either presented upright or inverted. Once the face disappears two faces or two face "parts," i.e. features, are presented simultaneously. A cue word indicates to which feature the viewer must attend: eyes, nose, or mouth. His or her task is to identify which of the features, the one of the left or right, matches the feature that was part of the initially presented face. For a visual depiction of task trial structures see Figure 2.

The whole advantage is characterized by better recognition of individual features when the feature is presented within the context of a face ("Whole" trials) rather than presented in

isolation ("Part" trials) in the upright orientation condition (Tanaka and Farah, 1993). Performance in the "Part" conditions reflects featural processing and general visual perception. Performance in the "Whole" conditions is thought to reflect featural processing, holistic processing, as well as general visual perception (DeGutis et al., 2013).

Inversion is believed to impair performance more profoundly during the "Whole" trials compared to "Part" trials because of its disproportionate impact on configural and holistic processing compared to featural processing (Carey & Diamond, 1977; Goffaux & Rossion, 2006; 2007; Van Belle, De Graef, Verfaillie, Rossion, & Lefèvre, 2010). Therefore, we expect the performance difference between upright and inverted trials to be greater in the "Whole" condition compared to the "Part" condition.

Composite Face Effect Task. In this sequential matching task participants first view one face. The top half of the face (meaning middle of the nose to the forehead) is either aligned with the bottom half or misaligned (offset horizontally). The face is either presented upright or inverted. The participants then view a second face in the same orientation and alignment manipulation condition as the initial face. Their task is to indicate whether the top half of both faces is the same or if it is different. The bottom halves of the test and target faces are always different. Figure 3 depicts trial structures within the *Composite Face Effect Task*.

When face halves are aligned and upright, it is difficult to selectively attend to face parts (e.g. the top half of faces in this case) (Meinhardt-Injac, Boutet, Persike, Meinhardt & Imhof, 2017; Richler et al., 2012; Richler & Gauthier 2014) because of the tendency for faces to be processed holistically (Mondloch, Pathman, Maurer, Le Grand, & de Schonen, 2007; but also see Rossion, 2013). As such, information from the task-irrelevant bottom half of the face image interferes even though one is trying to selectively attend to the top half of the face. Because of

the tendency for faces to be processed holistically, this interference can influence recognition judgment towards mistakenly judging the *same* top half of the faces to be *different* because of the "irrelevant" *different* bottom half. Misalignment of the top and bottom face halves disrupts holistic processing and makes it easier to selectively attend to the top half of the faces (Gauthier & Tarr, 2002).

In all trials featural information is available. In Upright Aligned task condition participants may also use first and second order configural information and/or holistic processing (to combine facial identity information into a gestalt) in order to detect differences or similarities between the two sequentially presented faces. Participants may also do so in the Inverted Aligned task condition, but with possible impairment.

As with previous experiments using this paradigm, we expect that performance will be poorer when faces are aligned upright compared to when they are misaligned upright. As previously stated, inversion is thought to disrupt holistic and/or configural processing. Because of this, we do not anticipate an additional recognition advantage on inverted misaligned trials compared to inverted aligned trials. We do expect performance to be lower across alignment conditions on inverted trials compared to upright trials.

Configural vs. Featural Difference Detection Task. Because human faces all have the same first-order relations, recognition ability is thought to depend on information about the appearance of specific features, and the distances between these features (Maurer et al., 2002). As the name implies, the *Configural/Featural Difference Detection Task* is thought to be a measure of configural vs. featural processing. Participants are shown a pair of faces that are either upright or inverted, and they are asked to make a same/different judgment. Of trials in which the faces differ, half involve configural changes and half involve featural changes. In

configural modification trials a facial feature (eyes, nose or mouth) is moved up or down, which modifies the second order relational information. In featural modification trials one feature is swapped with a feature from a different face (Carbon & Leder, 2005; Freire et al., 2000, Mercure, Dick, & Johnson, 2008; Renzi et al., 2013). The participant decides if the two faces are the same or if they are different (See Figure 4). Detecting configural changes should engage configural processing more than featural, and vice-versa. In either type of trial, participants may also use holistic processing to combine this information into a gestalt in order to detect differences or similarities between the two simultaneously presented faces.

Many studies have established that humans are better able to detect featural differences compared to second-order relational differences (Carbon & Leder, 2005; Freire et al., 2000, Mercure et al., 2008). And inversion impairs performance on configural modification trials more so than on featural trials (Leder & Bruce, 2000). Therefore we expect that performance will be better on featural difference detection trials compared to configural difference detection trials for both upright and inverted trials. We expect that performance will be better on upright trails across stimulus conditions compared to inverted trials. Most importantly, we expect to see bigger effects of inversion on configural trials than on featural trials.

The Importance of the Inverted Condition within Each Task

In this study, we have included inverted conditions within all four tasks, not just the *Face Inversion Effect Task*. According to McKone et al. (2013), the inverted control condition should always be included if one wishes to assess holistic or configural processing. She argues that any holistic effect that is present in the upright conditions should only be interpreted as a pure measure of holistic processing if the effect is reduced in the inverted conditions.

Including the orientation manipulation in all our tasks allows us to assess how face inversion impacts performance across paradigms. We expect to find the greatest performance deficits in the inverted conditions of tasks that are thought to measure configural and/or holistic processing.[3] While identifying the task conditions that are thought to be configural and/or holistic, and therefore most susceptible to performance deficits as a result of inversion, remains a speculative process based on intuition we believe said conditions are: 1) the "Face" condition of the *Face Inversion Effect Task*, 2) the "Whole" condition of the *Part Whole Effect Task*, 3) the "Aligned" condition of the *Composite Face Effect Task* and 4) the "Configural" condition of the *Configural/Featural Difference Detection Task*.

Inter-Task Analyses and Hypotheses

As previously mentioned, classifying face processing tasks according to which mechanism they primarily index—featural, configural, and holistic—has largely relied on intuition. We believe that it is important to establish empirically how performance is related between these tasks in order to determine if they are in fact measuring the same construct(s). After this has been accomplished, the next step would be then to empirically determine which construct(s) each of these tasks is measuring. This second step is beyond the scope of this manuscript although the present work will help provide the knowledge base necessary to do so.

In order to begin to address these gaps in the facial recognition literature, we conducted correlational analyses and factor analyses to establish how performance across the tasks is related. This allowed us to examine whether the tasks correlate strongly with each other, and if they load on to the same factor(s). Among other things, this course of analysis will allow us to

[3] This is a reasonable expectation because the amount of time participants had to look at the faces in the present tasks was limited. According to Richler, Mack et al. (2011) given enough time looking at an inverted face, humans do engage in integrative processing. As will be described in the methods section, presentation time was matter of seconds, which should not allow our participants to engage in unimpaired holistic processing on inverted trials.

determine the degree to which the various tasks can be taken as equivalent and interchangeable indices of the underlying construct(s). If they are not interchangeable measures, then this may, in part, explain a variety of discrepancies in the literature.

Based on past data and theory we anticipated a number of potential patterns in the performance data. For example, if all the tasks primarily measure a single construct (e.g., they all tap into a combined holistic/configural mechanism) then we would expect performance to be strongly correlated between all four tasks, and for performance on all four tasks to load onto a single factor. Alternatively, the four tasks may be measuring the same construct(s) but to different degrees. If this were the case there would be a wide range of correlation coefficients between conditions across tasks. Moreover, we would expect to see item cross loadings in the factor analysis, indicating that the different tasks partially tap into similar underlying constructs. We expect patterns for individual difference scores to mirror those of the accuracy and reaction time performance data.

If the tasks were tapping into several distinct processing mechanisms then we would expect task performance to load onto two or three factors, which could possibly be conceptualized as featural, configural, and/or holistic. Performance on each task condition would be strongly correlated with other task conditions that are measuring the same construct, and weakly correlated with task conditions measuring a different construct, loading onto factors accordingly. For instance, if configural and holistic processing are aspects of a common construct, and the Upright Whole condition of the *Part Whole Effect Task*, the Aligned Upright condition of the *Composite Face Effect Task,* the Upright Face condition of the *Face Inversion Effect Task,* and the Upright Configural conditions of the *Configural Featural Difference Detection Task* do indeed measure configural/holistic processing in comparable ways, then we

would expect these four task conditions to load onto a single factor. If, however, configural processing and holistic processing were each separate constructs, and these tasks measure these separate constructs, then we would expect these task conditions to load onto one of two factors in our factor analysis. In the event that each task (and the conditions that compose it) measures its own individual construct, or component thereof, and loads onto one of four separate factors, then this would provide a strong challenge to the widely accepted view that the four tasks tap into one or more common face processing mechanisms. Again, we expect patterns for individual difference scores to mirror those of the accuracy and reaction time performance data.

While there are examples of studies that have compared performance across some of the commonly used facial recognition tasks (e.g., Konar et al., 2010; Richler et al., 2011; Richler & Gauthier, 2014), to the best of our knowledge this is the first study to compare performance patterns via correlation analyses and factor analyses across four commonly used facial recognition tasks, and to have included both subtraction and regression-based individual difference scores in the analyses. Furthermore, we investigated the reliability of both types of individual difference scores in an attempt to reconcile discrepancies in the literature pertaining to individual difference scores. Because there are discrepancies in results across studies that are comparing performance on some of these tasks, we have reason to suspect that these commonly used tasks are, in fact, not comparable measures of face processing and should not be used interchangeably.

Method

The participant and materials information is identical for the four tasks examined in this study. As such, this information will be provided first, followed by details regarding stimuli and experimental procedure for each task. Task order was randomized across participants, as was the

order of conditions within each task. Participants took approximately one hour and thirty minutes to complete the four tasks.

Participants

Undergraduate students from the University of Ottawa ($N = 223$) participated in this study, which was approved by the University of Ottawa Social Sciences and Humanities Research Ethics Board. Participants were awarded course credit for participation. The initial sample size was $N = 240$; however, due to the within-subjects design of this study, we deleted participants listwise who quit the study before completing all four tasks, or who did not correctly follow instructions and pushed invalid keyboard keys.

We used G-Power software (Faul, Erdfelder, Lang, & Buchner, 2007) to ensure that our sample size would provide us with adequate statistical power. In order to establish that the typical interaction effects were present in each of the four tasks we ran four 2 × 2 repeated measures ANOVAs with task manipulation and stimulus orientation as the two factors. In terms of the statistical test we selected the repeated measures within factors F test. We conducted a post hoc power analysis to determine the output parameter for power ($1 - \beta$). We implemented the following input parameters: effect size $f = 0.1$, $\alpha = 0.05$, total sample size $N = 223$, two groups, four measurements, a correlation among repeated measures of .56 (based on the average correlation coefficient as calculated in the correlational analysis section below), and nonsphericity correction of 1 because our data met the assumption of sphericity (as discussed in the results section below). Based on these parameters our level of statistical power is .98.

Regarding our correlation analyses we used G-Power software to determine if our sample size was adequate based on the following assumptions: $\alpha = 0.05$, a power of 0.85, and effect size ($\rho = 0.2$). The required sample size was $N = 217$, and our obtained sample size was $N = 223$.

Costello and Osborne (2005) outlined the best practices when conducting an exploratory factor analysis. With regard to sample size adequacy, they acknowledge that there are no "strict" rules, and "more is better". The rule of thumb most researchers use to determine a priori sample size for an exploratory factor analysis is a participant to item ratio of 10:1. Our participant to item ratio is 223:16 which equates to roughly 14:1.

Materials

Across all tasks, the face stimuli were male and the images were grey-scaled. Face stimuli were taken from the Max Planck Institute for Biological Cybernetics database (Troje, & Bülthoff, 1996) and were scaled to 9 × 9 cm (9° × 9° at 57 cm). All faces were presented in the frontal (0°) view. Using Adobe Photoshop CS4 (Adobe Systems Inc., San Jose, CA; Adobe.com) and Matlab 2010a (Mathworks Inc., Natick, MA; http://www.mathworks.com) psychophysics toolbox, the background was removed and was replaced with a uniform grey field. Ears and hair were also removed. All stimuli were grey-scaled using the Matlab built in rgb2gray function and were equated for mean luminance and RMS contrast. The experiments were conducted on a series of identical Dell Optiplex 9010 LCD computers with an i5 Intel core, running Windows 7 on 17-inch monitors with a screen resolution of 1920 × 1080 and refresh rate of 60 hertz.

Task 1 – Face Inversion Effect Task

Stimuli. In the *Face Inversion Effect Task* (Yin, 1969), faces and objects are presented either upright or inverted. The class of objects used in this experiment was chairs. We chose to use this class of objects because chairs are typically seen upright, as faces are, and because chairs have been used often before in the literature (e.g. Boutet, Collin, & Faubert, 2003; Stein et al., 2012; Yovel & Kanwisher, 2005). The stimuli consisted of 80 faces and 80 chairs. The authors

created the chair images by taking pictures of arbitrarily chosen chairs. We used the Matlab rgb2gray function to grey-scale the images, and inserted a uniform grey background for all chair images so that the object images would only differ from the face images in terms of the stimulus type.

Procedure. Following instructions, the task began with four practice trials (one for each experimental condition). Before the experimental phase began, the instructions were presented to the participant again, at which point he or she had the opportunity to ask for clarification. Participants were told verbally to respond as quickly as possible. This was done because this task is relatively easy and accuracy levels are subject to ceiling effects.

Our *Face Inversion Effect Task* involved a 2 Alternative Forced Choice Sequential Matching paradigm with a 2 × 2 design. The factors were Stimulus Manipulation[4] (face or object), and Orientation (upright or inverted). Face and Object trials were presented in random order; half of the trials presented object stimuli, and the other half presented face stimuli. Orientation was blocked in batches of 40 trials. There were two inverted blocks and two upright blocks. The trial breakdown was as follows: 40 upright face trials, 40 inverted face trials, 40 upright object trials, 40 inverted object trials. This task consisted of 160 trials and took approximately 15 minutes to complete.

Figure 1 illustrates the trial structure in the "upright face" condition and in the "inverted face" condition in the *Face Inversion Effect* task. First, a single inspection image (face or object) was presented for .4 seconds. This was followed by an inter-stimulus interval of 1 second, during which a blank screen was presented. After this, two test images (two faces or two chairs, depending on the task manipulation condition) were presented. Depending on the orientation

[4] We use the term "Stimulus Manipulation" here to refer to the way in which the stimuli differ across trials other than by orientation. Because the nature of this manipulation changes across task, we elected to use the generic term Stimulus Manipulation.

block, images in a trial (i.e., both inspection and test images) were either all upright or all inverted. The participants' task was to identify which of the two simultaneously presented test images matched the inspection image. They did so by pushing one of two side by side keyboard keys, one corresponding to the face on the left and the other corresponding to the face on the right. Participants could use whichever finger(s) they preferred to enter their responses. The test images remained on the screen until the participant responded. Both accuracy and reaction time were recorded automatically.

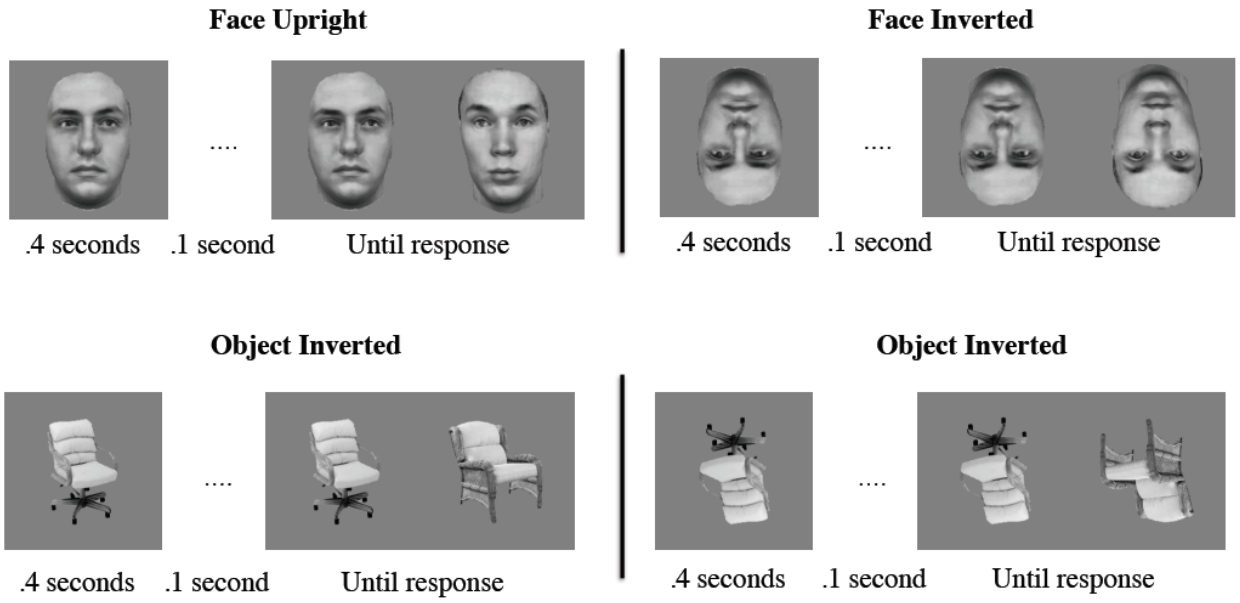

Figure 1. This figure depicts the stimuli, presentation times, and the inter-stimulus interval for upright and inverted trials in the Face Inversion Effect Task.

Task 2 – Part Whole Effect Task

Stimuli. In the *Part Whole Effect Task* (Tanaka & Farah, 1993) stimuli are either whole faces or isolated face parts (features: eyes, nose, or mouth). To make the stimuli for this task we began by creating 10 *base faces*. *Base faces* are faces whose inner features (eye, nose and mouth) were each replaced by the same feature from different faces. The faces that were the source of the replacement features were not otherwise used in this experiment. This featural

swap was done so that the base faces would not seem odder or modified than the modified faces that make up the remainder of the stimuli in this task. From the 10 *base faces* we created three modified versions of each by swapping out either the eyes, or the nose, or the mouth, with that feature from yet another different face. For example, Base Face A was used to create version 1 with the eyes from Face B, version 2 with the nose from Face C, and version 3 with the mouth from Face D.

The *base faces* and three modified versions are the face stimuli that appear in the upright and inverted "Whole" conditions of this task. Along with the facial features that were used to form the 10 *base faces*, the features that were used to create the modified faces are the featural stimuli that appear in the upright and inverted "Part" conditions of this task. In total we created 10 *base faces*, 30 modified faces with a swapped feature from a different face, 30 base parts, and 30 parts from different faces.

Procedure. The task began with instructions, followed by four practice trials. Before the experimental phase began, the instructions were presented to the participant again, at which point he or she had the opportunity to ask for clarification. The *Part Whole Effect Task* involved a 2 Alternative Forced Choice Sequential Matching paradigm with a 2 × 2 design. The factors were Stimulus Manipulation (part vs. whole) and Orientation (upright vs. inverted). An equal number of "Part" and "Whole" trials were presented in random order. Orientation was blocked into batches of 60 trials. There were 120 trials in total, and this task took participants approximately 10 minutes to complete.

The procedure for each trial was as follows (and is presented visually in Figure 2). First, one of 10 base faces was presented as the inspection face for 1.75 seconds followed by an inter-stimulus interval of .4 seconds, at which point two test faces (or test parts, both of the same

feature) were presented simultaneously side-by-side. Above these two test faces (or parts), the word "Eyes" or "Nose" or "Mouth" appeared, cueing the participants to make their recognition judgment regarding that feature. This cueing is necessary in the "Whole" conditions so the participant knows which feature to compare between test and inspection faces. However, the cue also appeared in the "Part" conditions, despite it being obvious to which feature the participant must attend and about which he or she must subsequently respond. By presenting the cue in both "Whole" and "Part" conditions it equates for any distracting effects the cue might have. On the "Whole" trials, the participant's task is to identify which of the two test faces contains the specified feature that matches the feature from the inspection face; On the "Part" trials, the participant's task is to identify which of the two test parts (features) was contained within the inspection face. Participants responded with one of two keyboard keys to indicate their response. The test faces or features remained on the screen until the participant responded. Both accuracy and reaction time were recorded automatically.

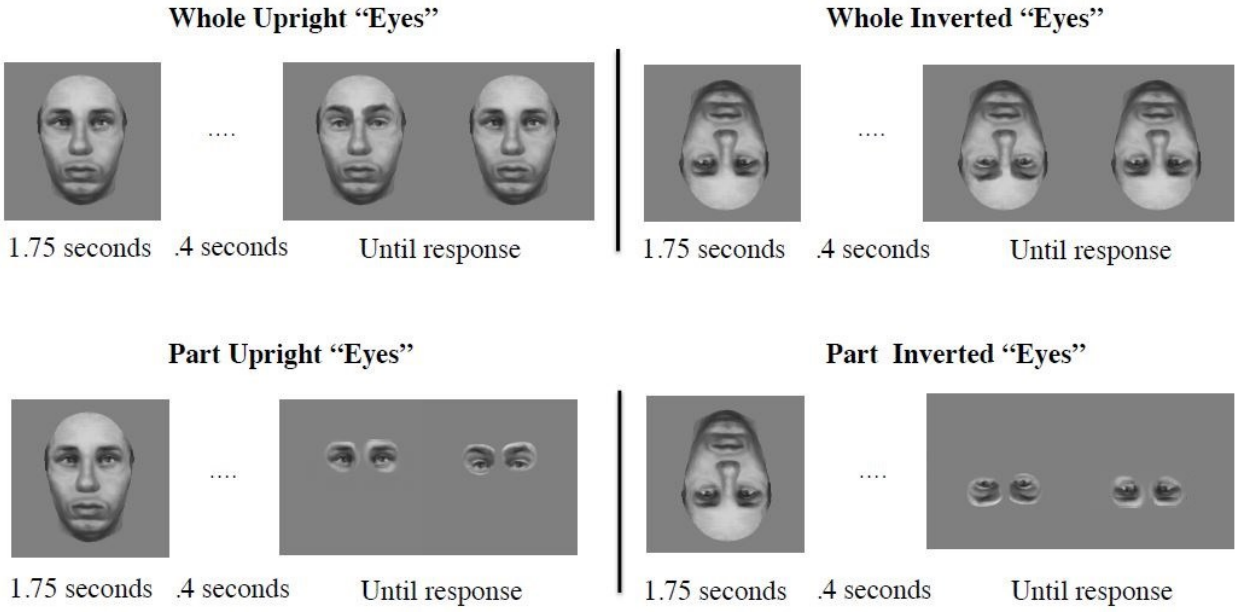

Figure 2. This figure depicts the trial structure for upright and inverted trials in the Part Whole Effect Task that pertain to eyes. Test images remained on the screen until the participant entered his or her response.

Task 3 – Composite Face Effect Task

Stimuli. In the *Composite Face Effect Task* (Young et al., 1987) stimuli consisted of 240 composite face stimuli. We created composite face stimuli by combining the top half of one face and the bottom half of a different face. That is, all faces presented in this task were a hybrid of two faces. Composite faces were either presented with the two halves aligned such that they combine to look like a whole face, or with the top and bottom halves offset, creating a misaligned face. In the case of the misaligned face, the right edge of the bottom half of the face lines up with the middle of the nose from the top half of the face. This creates a face that is broken into two obviously separate halves.

Procedure. The task began with instructions, followed by eight practice trials. Before the experimental phase began, the instructions were presented to the participant again, at which point he or she had the opportunity to ask for clarification. The researcher verbally emphasized to participants that they should selectively attend the top half of faces (middle of nose to forehead) in both the upright and inverted conditions.

The Composite Face Effect task is a 2 Alternative Forced Choice Sequential Matching paradigm with a 2 × 2 design. The factors were Stimulus Manipulation (aligned vs. misaligned) and Orientation (upright vs. inverted). An equal number of "Aligned" and "Misaligned" trials were presented in random order. Orientation was blocked in batches of 30 trials. In all there were 240 trials, and participants took approximately 20 minutes to complete this task.

Each trial began with one composite face (inspection face) presented for .5 seconds, followed by an inter-stimulus interval of .3 seconds, after which a second composite face (test face) was presented until response. Both faces were either upright or inverted, depending on the orientation condition, and both were either aligned or misaligned, depending on the alignment

condition. The top halves of the inspection and test faces either matched ("same" trials) or did not match ("different" trials), and the bottom halves never matched (i.e. were always different). The participants' task was to judge if the top halves of the inspection and test faces (i.e., the part from the nose to the forehead, regardless of orientation) were the same or if they were different. Participants indicated their response by pressing one of two keyboard keys. Both accuracy and reaction time were recorded. For examples of "same" aligned and misaligned trial structures see Figure 3.

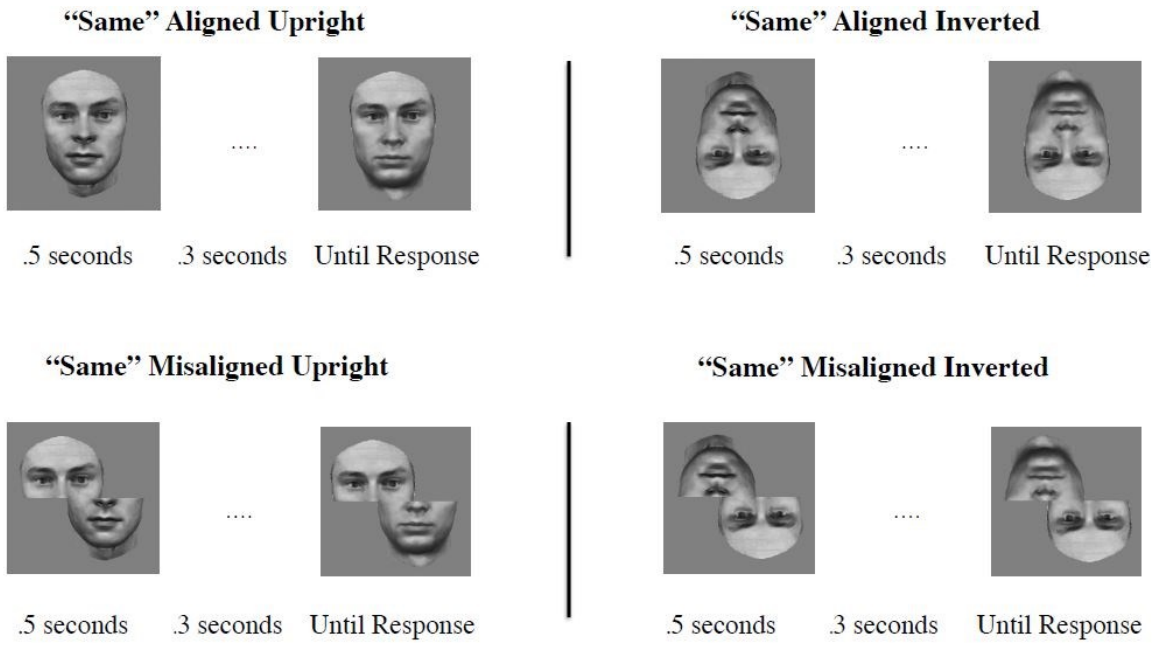

Figure 3. This figure depicts the stimuli, presentation times, and the inter-stimulus interval for upright and inverted "same" trials in the Composite Face Effect Task.

Task 4 – Configural/Featural Difference Detection Task

Stimuli. In the *Configural/Featural Difference Detection Task* (Freire et al., 2000) face stimuli were either unmodified, or modified. There were two ways in which the face could be modified: configurally or featurally. Configural modification means moving a face part (eyes, nose, or mouth), either up or down .30º, which is enough to change the second-order relational

information without overly distorting the face. Featural modification means that a face part (eyes, nose, or mouth) was replaced with that same feature from a different face that was not used elsewhere in the experiment. Four modified base faces were created so that all face stimuli would look equally modified. These modified base faces were presented, along with the nine variations on each base face, (i.e., both upward and downward configural manipulations for eyes, nose, and mouth, as well as the featural manipulation for each facial feature) making a total of 36 manipulated faces, which served as the stimuli in 96 trials.

Procedure. The task began with instructions, followed by eight practice trials (2 for each condition) that contained faces different from those used in the experimental conditions. Before the experimental phase began, the instructions were presented to the participant again, at which point he or she had the opportunity to ask for clarification. The *Configural/Featural Difference Detection Task* involved a Simultaneous Matching paradigm with a 2 × 2 design. The factors were Stimulus Manipulation (configural vs. featural) and Orientation (upright vs. inverted). Trials were presented in random order. Half of the trails presented two of the same face and half presented two different faces. Half of the "different" trials presented a featural manipulation and the other half presented a configural manipulation. Orientation was blocked in two batches of 48 trials. There were therefore 96 trials in this experiment, and participants required approximately 10 minutes to complete the task.

Each trial began with two faces presented simultaneously for 3.5 seconds. Participants were tasked with indicating if the two faces were the same or if they were different by pushing one of two keyboard keys to register their response. After 3.5 seconds both faces were masked and participants were required to make a judgment if they had not indicated their response yet (see Figure 4). Both accuracy and reaction time were recorded.

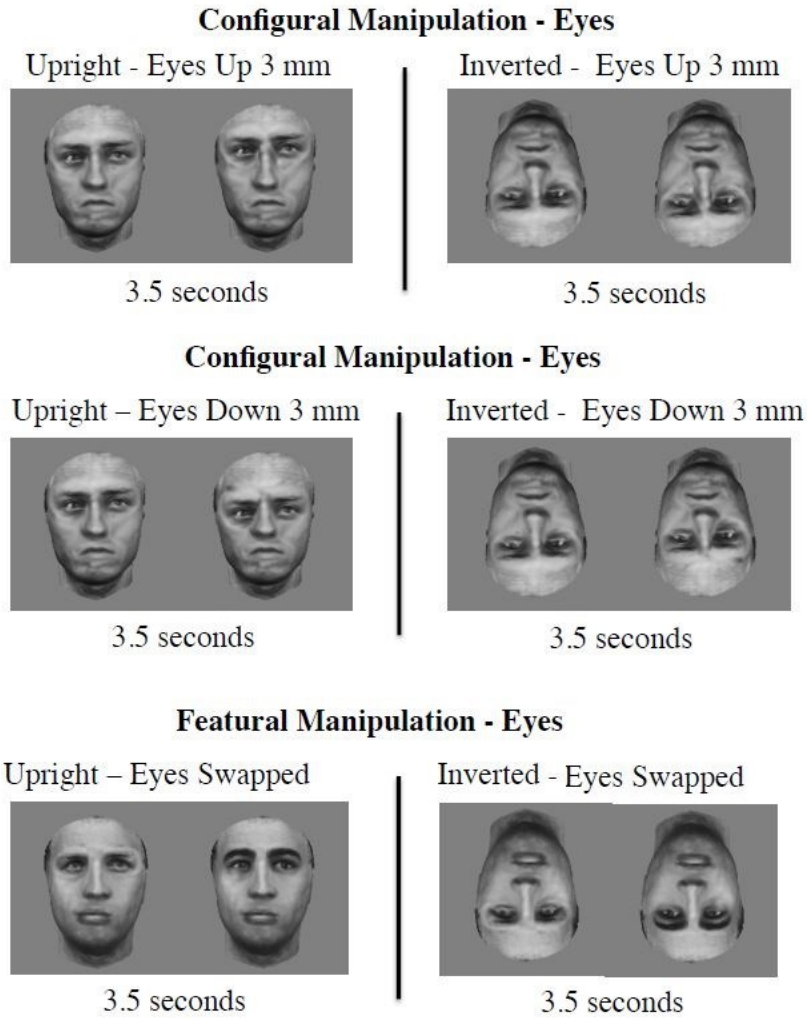

Figure 4. This figure depicts the stimuli, presentation times, and the inter-stimulus interval for upright and inverted trials in the Configural/Featural Difference Detection Task that pertain to eye manipulation.

Results

Dependent Variables

As previously stated, accuracy results are reported in this section with parallel analyses for reaction time available in Appendix A. Where there are meaningful differences between accuracy and reaction time this is noted in the text. After addressing outliers (see below for details), we calculated accuracy and reaction time difference scores using both the subtraction method and the regression method. Table 3 identifies the conditions from each task that we

chose to be the control condition and condition of interest in our analyses of accuracy and reaction time individual difference scores. .

Data Cleaning

For all four tasks, we began by analyzing accuracy and reaction time data. Data cleaning involved first calculating z-scores in all 16 task conditions (i.e., the four conditions in each of the four tasks) to identify univariate outliers. We chose the z-score cut off value of +/- 2.54, which corresponds to an α of .01. Values 2.54 standard deviations above or below the mean score were considered to be outliers, and were "tucked in". This technique, which is similar to winsorizing (Huber, 1981; Hogg, 1979), entails replacing outlier values with the closest score that is within z = +/- 2.54 standard deviations of the mean. Less than 2% of all scores were tucked in.

We used Mahalanobis distances to identify multivariate outliers in the accuracy and reaction time data. This identified 3 and 10 participants out of 223 as outliers for accuracy and RT, respectively. However, because we are interested in differences in performance across 16 task conditions, and it is possible that these participants are valid examples demonstrating such differences, we opted not to remove those participants who were multivariate outliers.

Data Normality

We investigated skewness and kurtosis values in order to determine if the accuracy scores were normally distributed within each task condition. Data were considered skewed if the estimated skew was more than twice the standard error of the skew, with the same logic applied for identifying kurtosis. In addition, we visually inspected the data using histograms with fitted normal curves and QQ-plots. Converging evidence suggested that accuracy scores from all conditions within the *Face Inversion Effect Task* demonstrated negative (left) skew and were leptokurtic. Whole conditions within the *Part Whole Effect Task* were negatively (left) skewed.

All conditions within the *Composite Effect Task* were negatively (left) skewed, and the misaligned upright condition was also leptokurtic. Lastly, within the *Configural Featural Difference Detection Task* the configural upright and featural upright task conditions demonstrated negative (left) skew and the configural inverted condition was positively (right) skewed.

In light of recent findings that log transformation has negative implications on subsequent data analyses (Feng, Wang, Lu, et al., 2014), and in order to facilitate interpretation of our results, we elected to leave our data untransformed and use non-parametric tests when appropriate. Moreover, the various violations of normality would have required different transformations, making interpretation very challenging.

We assessed multivariate normality by plotting the Mahalanobis distance values against estimated chi-square quantiles in a Q-Q plot in SPSS. Based on visual inspection, and given that we elected to leave in the 13 multivariate outliers, we determined that accuracy scores did not meet the assumption of multivariate normality.

Analysis of Variance

Before conducting our ANOVAs we inspected each of the task conditions' standardized residuals and determined that they were approximately normally distributed. We determined this by examining histograms with a superimposed normal distribution curve and by examining Normal Q-Q plots for residual scores from each of the task conditions. In order to ascertain whether we replicated the expected interaction effects for each task, we conducted a series of 2 × 2 repeated measures ANOVAs on the data from each of the four tasks. The independent variables were Orientation (upright and inverted) and Stimulus Manipulation (which varied by task). The pattern of the interactions in each of the ANOVAs is consistent with the typical

findings for each task. All interactions and main effects were significant. Because these findings are simply replications of much previous work, and because they are not central to our hypotheses, the detailed results from the ANOVAs are presented in the supplementary materials.[5]

Task Reliability

We measured task reliability of all four tasks using Guttman's λ_2. We elected to use this measure of internal reliability because it is robust in instances when the measures are composed of multiple factors (Callender & Osburn, 1979). This is an important consideration because our tasks each contain four types of trials and are thought to measure various combinations of configural/holistic processing as well as featural processing. We calculated the reliability within each task using the four task conditions. For example, to calculate the reliability of the *Part Whole Effect Task* we included: part upright, part inverted, whole upright and whole inverted conditions. All four tasks show good reliability. Guttman's λ_2 reliability scores per task are as follows: *Face Inversion Effect Task* (.92), *Part Whole Effect Task* (.79), *Composite Effect Task* (.90), and *Configural/Featural Difference Detection Task* (.80).

Correlation Analysis

In order to investigate how performance measures are related between each of the four tasks, we first conducted a correlational analysis using the accuracy data for the 16 conditions tested across the four tasks. Recall that our data did show skewness in 13 of the 16 conditions and kurtosis in five. We used variance statistics to assess homoscedasticity, comparing the ratio of the largest and smallest variance statistics from the 16 task conditions. Accuracy scores were

[5] Because these effects are often interpreted in terms of interactions, we also investigated the "individual interaction strength" data (or difference in differences) in a number of ways; However, variability was very low on these measures and therefore the results were not meaningful.

heteroscedastic as the ratio of largest to smallest variance statistic was greater than 1.5. Because our data violates the assumption of homoscedasticity, we calculated the non-parametric Spearman's Rank-Order Correlation coefficients (r_s) also known as rho. With the exception of the *Face Inversion Effect Task* (which showed ceiling effects) the data for the task conditions met the assumption of a monotonic relationship as evidenced by examination of Q-Q plots. Figure 5 depicts the correlational analysis of accuracy scores.

		Face Upright	Face Inverted	Object Upright	Object Inverted	Whole Upright	Whole Inverted	Part Upright	Part Inverted	Aligned Upright	Aligned Inverted	Misaligned Upright	Misaligned Inverted	Configural Upright	Configural Inverted	Featural Upright	Featural Inverted
Face Inversion Effect Task	Face Upright		.58	.45	.39	.26	.19	.15	.14	.32	.35	.41	.33	.26	.18	.24	.22
	Face Inverted	.58		.44	.40	.34	.36	.24	.15	.33	.30	.42	.31	.27	.29	.37	.30
	Object Upright	.45	.44		.38	.16	.13	.14	.07	.27	.20	.22	.15	.02	.16	.08	.22
	Object Inverted	.39	.40	.38		.14	.09	.17	.14	.21	.28	.23	.28	.05	.01	.16	.17
Part/Whole Task	Whole Upright	.26	.34	.16	.14		.57	.58	.43	.41	.30	.40	.26	.35	.20	.34	.27
	Whole Inverted	.19	.36	.13	.09	.57		.54	.61	.41	.26	.35	.30	.31	.39	.43	.42
	Part Upright	.15	.24	.14	.17	.58	.54		.63	.33	.30	.30	.24	.31	.31	.34	.32
	Part Inverted	.14	.15	.07	.14	.43	.61	.63		.32	.18	.32	.21	.20	.30	.31	.30
Composite Face Effect Task	Aligned Upright	.32	.33	.27	.21	.41	.41	.33	.32		.68	.71	.58	.26	.35	.30	.36
	Aligned Inverted	.35	.30	.20	.28	.30	.26	.30	.18	.68		.67	.83	.29	.27	.44	.44
	Misaligned Upright	.41	.42	.22	.23	.40	.35	.30	.32	.71	.67		.63	.35	.28	.47	.44
	Misaligned Inverted	.33	.31	.15	.28	.26	.30	.24	.21	.58	.83	.63		.34	.23	.48	.45
Configural Featural Task	Configural Upright	.26	.27	.02	.05	.35	.31	.31	.20	.26	.29	.35	.34		.53	.62	.44
	Configural Inverted	.18	.29	.16	0.01	.20	.39	.31	.30	.35	.27	.28	.23	.53		.61	.55
	Featural Upright	.24	.37	.08	.16	.34	.43	.34	.31	.30	.44	.47	.48	.62	.61		.80
	Featural Inverted	.22	.30	.22	.17	.27	.42	.32	.30	.36	.44	.44	.45	.44	.55	.80	
		Face Upright	Face Inverted	Object Upright	Object Inverted	Whole Upright	Whole Inverted	Part Upright	Part Inverted	Aligned Upright	Aligned Inverted	Misaligned Upright	Misaligned Inverted	Configural Upright	Configural Inverted	Featural Upright	Featural Inverted
		Face Inversion Effect Task				Part/Whole Task				Composite Face Effect Task				Configural Featural Task			

(Figure caption appears on next page)

Figure 5. Spearman's Rho Correlations ($N = 223$) of Accuracy Data Between Task Conditions. r_s values between .132 - .172 correspond to $p < .05$, two-tailed. r_s values between .173 - .187 correspond to $p < .01$, two-tailed. Part Whole Effect Task (Part/Whole Task), Configural/Featural Difference Detection Task (Configural Featural Task).

Two patterns of interest emerged from the correlational data. First, the average within-task correlation coefficient was r_s = .48, 95% CI [.37, .56] with a range between r_s = .33, and r_s = .64. Interestingly, reaction time data demonstrated a much stronger average within task correlation coefficient across the four tasks which was r_s = .80, 95% CI [.75, .84]. The reaction time data make a compelling case for the possibility that within each task its respective task conditions are measuring something similar. In comparison, the accuracy data's moderate within-task correlations suggest that conditions within each of these four tasks could be measuring something similar while also measuring different constructs (or components thereof). This is the first example in the present study of how reporting different behavioural measures can lead to different interpretations.

The magnitude of the average between-task correlation for accuracy data was weak (r_s = .23, 95% CI [.10, .35] with a range between r_s = -.01, and r_s = .41). The same analysis for reaction time also yielded a weak magnitude (mean r_s = .32, 95% CI [.20, .43]. Overall it seems there is something about each task that distinguishes it from the others (evidenced by strong within task correlations), and that performance across task conditions is weak. This suggests that these tasks measure something in common but primarily are measuring different constructs (i.e. processing mechanisms) (or components thereof) and therefore should not be used interchangeably. It remains to be determined what each of the four tasks is measuring and how each might relate to one or more processing mechanisms.

Schmidt and Hunter (1996) underscored the importance of correcting for measurement error in one's data. One way to account for measurement error in our correlational data is to use the reliability scores to calculate theoretical upper bound limits on the correlations and then use these limits to "scale-up" the obtained correlations to produce "maximum possible correlations".

Two examples of relevant manuscripts that used this methodology are Susilo, Rezlescu, and Duchaine (2013), and DeGutis et al. (2013). We calculated the "maximum possible correlations" and the analyses are presented in Appendix A.[6] The scaled up correlational analysis revealed the same pattern of performance as the correlation analysis reported above.

Exploratory Factor Analysis

To further assess the underlying associations between performance on the four tasks, we conducted an exploratory factor analysis. This analysis allows us to determine how many latent variables these four tasks are measuring. First, we determined that our data showed multicolinearity by checking the variance inflation factor in an iterative fashion such that each variable was regressed against the remaining 15 using linear regression. Multicolinearity was expected because performance is strongly correlated across conditions within each task, as it should be if they are measuring the same underlying factor. Next, we determined that principle axis factoring was the appropriate factor extraction method to use because accuracy data did not meet the assumption of multivariate normality (Costello & Osborne, 2005) as explained previously.

[6] The "theoretical upper bound limit" is obtained as follows: multiply the two reliability scores together and then take the square root of the product. The obtained correlation can then be "scaled-up" by dividing it by the theoretical upper bound limit value to produce the "maximum possible correlation". This technique helps to put the obtained correlation coefficients into context. For example, rarely is it the case that a correlation coefficient of r = 1.0 is possible between two tasks. Yet when researchers conduct correlational analysis, and assess the magnitude of the correlation coefficient, the assessment is based on where the correlation coefficient value falls within a range of r = .00 to r = 1.0. What researchers could do instead (or could do in addition) is use the reliability scores of their tasks to calculate the highest correlation that is possible between pairs of their tasks, then scale up the obtained correlations to calculate what the "maximum possible correlation" coefficient would be given the theoretical upper limit. Consider the following example: If the theoretical upper bound limit on the correlation between two tasks were r = .50, then an obtained r value of .20 would be considered moderately strong rather than weak (as it would be labeled under the assumption that a correlation of r = 1.0 is possible) because scaling up the obtained correlation of r = .20 results in a "maximum possible correlation" of r = .40.

In terms of selecting the number of factors to retain we followed the recommendation presented in Achim (2017)'s paper detailing the Next Eigenvalue Sufficiency Test (NEST), which is a conservative approach to factor extraction. Ashim (2017) tested eight factor extraction approaches (including 4 NEST variant models) across a range of critical eigenvalues and determined that the NESTip variant was the model that most accurately identified the correct number of factors that should be extracted when all retained eigenvalues are above 1.0. Among others, NESTip outperformed the commonly used Revised Parallel Analysis and Comparison Data approaches when critical eigenvalues were 1.0 or higher. Its primary advantage is the probability that NESTip would overestimate the number of factors that should be retained is significantly less than 5%.

We chose to use a direct oblimin rotation, which is an oblique rotation, because our data are moderately to highly correlated within each task, and more weakly correlated between tasks (Osborne & Costello, 2009). Compared to the unrotated factor structure, rotation serves to clean the factor structure, providing a simpler factor solution by eliminating cross-loadings. In accordance with Tabachnick and Fidell (2001), we suppressed factor loadings below .32 in the tables shown below. Both rotated and unrotated factor solutions are reported. Investigating both rotated and un-rotated factor structures provides us with the most complete picture of if and how performance on these tasks is related.

Results from the conservative NESTip factor extraction approach and the Scree test approach converged, indicating that accuracy data is composed of four factors. All retained factors comprised 3 or more items, which indicates that the factors were stable (Osborne & Costello, 2009). The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was high, at .85, and Bartlett's test of sphericity was significant (χ^2 (120) = 2079.40, $p < .001$). The four-factor

solution explained 62.23% of the total variance explained. Factor 1 explained 33.98% of the variance, Factor 2, 15.10%, Factor 3, 7.58% and Factor 4, 5.57%. Item communalities for the most part were within the low to moderate range of .40 - .70, and increased after extraction. Because unrotated and rotated factor solutions suggest different interpretations, we have elected to report both matrices. Tables 1 and 2 display the unrotated and rotated factor structures, respectively.

Table 1

Unrotated Four-Factor Solution Factor Analysis for Accuracy

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
FIE	Face Up	.61	-.62			.73	.77
	Face Inv	.65	-.53			.69	.72
	Object Up	.55	-.75			.81	.87
	Object Inv	.55	-.70			.76	.79
PW	Whole Up	.55				.42	.53
	Whole Inv	.53		.39		.42	.50
	Part Up	.50		.40		.41	.50
	Part Inv	.44		.44		.40	.51
CFE	Aligned Up	.69				.64	.63
	Aligned Inv	.72		-.47		.73	.80
	Misaligned Up	.72				.65	.67
	Misaligned Inv	.66		-.42		.68	.69
C/F	Configural Up	.43	.32			.40	.39
	Configural Inv	.46			.34	.41	.41
	Featural Up	.60	.42		.46	.60	.76
	Featural Inv	.55	.35			.50	.51

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F), Upright (Up), Inverted (Inv).

Table 2

Pattern Matrix, Rotated Four-Factor Structure of Accuracy

Task		Factor			
		1	2	3	4
FIE	Face Up		-.86		
	Face Inv		-.80		
	Object Up		-.95		
	Object Inv		-.89		
PW	Whole Up			.57	
	Whole Inv			.66	
	Part Up			.70	
	Part Inv			.75	
CFE	Aligned Up	.73			
	Aligned Inv	.91			
	Misaligned Up	.74			
	Misaligned Inv	.83			
C/F	Configural Up				.60
	Configural Inv				.65
	Featural Up				.88
	Featural Inv				.63

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F), Upright (Up), Inverted (Inv).

The unrotated solution demonstrates that all the task conditions load onto a single first factor, suggesting that it reflects a general processing mechanism for faces and objects. With the exception of Object Upright and Object Inverted task conditions all other items that load onto multiple factors loaded more strongly on the first common factor. The second factor includes the *Face Inversion Effect Task* conditions, as well as *Configural Featural Task* conditions with the exception of Configural Inverted. The *Part Whole Effect Task* conditions (less Whole Upright) load onto factor 3 along with inverted conditions from the *Composite Face Effect Task*. The fourth factor comprises two conditions from *Configural Featural Task*; Configural Inverted and Featural Upright.

Attempting to label the higher factors (i.e., beyond the first) from the unrotated solution is difficult, as there is no clear pattern that fits any of the predictions. For instance, there is no

suggestion of separate holistic and configural factors, nor any obvious factor that might index featural processing.

As previously mentioned, rotation cleans up the factor structure by minimizing cross loadings. The oblimin-rotated solution suggests four factors, with each task loading onto its own respective factor, and thus indexing its own particular aspect of face recognition. In our data, we see stronger correlations within tasks compared to between tasks, and the rotated factor matrix provides evidence for four separate factors with each task loading onto its own factor. This is contrast to the unrotated factor structure, which demonstrates a general first factor with all tasks loading onto it. One interpretation of our results is that performance on each of these four commonly used tasks reflects both a distinct component of higher-order face processing and a common unitary visual processing mechanism.

Individual Difference Scores

Because each task typically reveals a particular effect via a significant interaction, it could be argued that it is the difference in performance between pairs of conditions that reveals featural, configural, and/or holistic processing and not the performance on each of the task conditions. To investigate this possibility, we calculated difference scores for each task using two methods: subtraction and regression. In all cases we subtracted/regressed across task orientation conditions for each stimulus manipulation category. The control condition is regressed against or subtracted from the condition of interest. Table 3 identifies the control condition and condition of interest that were used to calculate the individual difference scores.

Table 3

Task Conditions Used to Calculate Difference Subtraction Scores and D-Residuals

Task	Calculation of Difference Scores Condition of Interest - Control Condition	Difference Score Label
FIE	Face Upright – Face Inverted	Face
FIE	Object Upright – Object Inverted	Object
PW	Whole Upright – Whole Inverted	Whole
PW	Part Upright – Part Inverted	Part
CFE	Aligned Upright – Aligned Inverted	Aligned
CFE	Misaligned Upright – Misaligned Inverted	Misaligned
C/F	Configural Upright – Configural Inverted	Configural
C/F	Featural Upright – Featural Inverted	Featural

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Task Reliability of Individual Difference Scores

Next we calculated the reliability of the subtraction difference scores and regression d-residuals for accuracy using Guttman's λ_2 (Table 4). Researchers should be aware of the proper way to calculate reliability scores of individual difference scores. It involves multiple steps, which are outlined in detail in the supplemental materials of DeGutis et al. (2013). Generally speaking, one first calculates the reliability score for each task condition using raw trial data (not summary statistics). Next, one calculates the correlation between the two task conditions that related to the individual difference score one wishes to calculate. The researcher then decides which condition is the condition of interest and which is the control condition and chooses whether subtraction or regression based difference scores should be used. Lastly, the researcher inserts these variables into the equation to calculate the reliability of subtraction-based or regression-based individual difference scores as outlined in DeGutis et al. (2013).

Table 4

Task Reliabilities of Subtraction Difference Scores and Regression Difference D-residuals for Accuracy Data

Task	Individual Difference Scores	Guttman's λ_2 Subtraction	Guttman's λ_2 Regression
FIE	Face	.69	.70
FIE	Object	.89	.90
PW	Whole	.32	.37
PW	Part	.08	.18
CFE	Aligned	.50	.56
CFE	Misaligned	.58	.63
C/F	Configural	.16	.15
C/F	Featural	.00	.10

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

DeGutis et al. (2013) noted that the reliabilities of both sets of individual difference scores should be expected to be lower than the reliabilities for performance data. This was the case in our study. In contradiction with findings from DeGutis et al. (2013), the reliability of d-residuals for accuracy data was not consistently superior to the subtraction-based reliability scores. Indeed, the values are generally very similar. The reliability of d-residuals was superior for half of our difference scores: Whole and Part and Aligned and Misaligned difference scores; however, reliability was comparable for Face, Object, and Configural difference scores. Featural d-residual difference scores were more reliability than were subtraction-based scores but both are extremely low. Our reliability data for the *Part Whole Face Effect Task* and the *Configural/Featural Difference Detection Task* corroborate Sunday et al. (2017) who note that these tasks were not designed to study individual differences and often lack adequate reliability to do so.

A parallel reliability analysis was carried out for reaction time individual difference scores (Appendix A). Unlike with the accuracy data, Guttman's λ_2 reliability scores for both

subtraction and regression-based individual difference scores were adequately high for all difference scores ranging from .48 to .83 for subtraction-based scores and .62 to .85 for d-residuals. The greatest differences in reliability scores between accuracy and reaction time were in the Configural, Featural, Whole, and Part difference scores both for subtraction and regression-based scores. Interestingly, Sunday et al. (2017)'s statement that these face processing tasks were not designed specifically to investigate individual differences which accounts for low reliability scores does not apply to the reaction time data in this experiment as it does to the accuracy data. This is another reason why we believe it is great importance for researchers to report both behavioural measures.

Correlation Analysis for Individual Difference Scores

Data met the criteria for non-parametric correlational analyses. Tables 5 and 6 depict the correlational analysis of accuracy subtraction difference scores and accuracy d-residuals respectively. The corresponding scaled-up correlation tables are available in Appendix A.

Table 5

Non-parametric Spearman Rho Correlations for Accuracy Subtraction Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.14*	-						
PW	3. Whole	.13	-.04	-					
	4. Part	-.02	-.04	.07	-				
CFE	5. Aligned	-.09	.10	-.02	-.11	-			
	6. Misaligned	-.11	.05	.07	-.07	.28**	-		
C/F	7. Configural	.08	-.07	.21**	.02	-.06	.05	-	
	8. Featural	-.05	-.14*	.06	.02	-.02	.04	.12	-

Note. *$p < .05$, two-tailed. **$p < .01$, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table 6

Non-parametric Spearman Rho Correlations for Accuracy Regression D-Residual Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.14*	-						
PW	3. Whole	.10	.00	-					
	4. Part	.01	-.01	.23**	-				
CFE	5. Aligned	-.02	.13	.14*	. 07	-			
	6. Misaligned	-.00	.07	.20**	.06	.40**	-		
C/F	7. Configural	.00	-.13	.20**	.12	.01	.12	-	
	8. Featural	-.05	-.14*	.10	.10	-.00	.11	.27**	-

Note. **p* < .05, two-tailed. ***p* < .01, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

The pattern of interest that emerged from the correlational data was that both subtraction scores and d-residuals correlate more strongly within tasks than between tasks. For subtraction scores the average within-task correlation coefficient was r_s = .15, 95% CI [.02, .28] (correlations range between r_s =. 07 and r = .28) while the average between-tasks correlation coefficient was r_s = .07, 95% CI [-.06, .20] (correlations range between r_s = .02 and r_s = .21). For regression d-residual scores the average within-task correlation coefficient was r_s = .26, 95% CI [.13, .38] (correlations range between r_s = .14 and r_s = .40) while the average between-tasks correlation coefficient for was r_s = .08, 95% CI [-.05, .21] (correlations range between r_s = .01 and r_s = .20).

There are only two significant between task correlations for subtraction-based individual difference scores. This is likely due to the low reliability scores, which is why, as previously mentioned, some authors advocate for reporting scaled-up correlations. The two notable correlations are between the "Configural" difference scores and "Whole" difference scores (r_s =

.21, 95% CI [.08, .33]) suggesting that they may have a common mechanism, or overlapping mechanisms required to accurately recognize faces within these conditions. Because it is thought that holistic processing includes aspects of configural processing this correlation between these two difference scores is not surprising. The second between-task significant correlation is between the Featural difference scores and Object difference scores (r_s = -.14, 95% CI [.01, .27]. This correlation magnitude makes sense given that non-face objects are thought to be processed in a more feature-based manner (Biederman, 1987); however, the direction of the association is surprising. It should be noted that the reliability of the "Featural" difference score is very low and therefore its results should be interpreted with caution.

Within the d-residual correlational data, the "Whole" scores of the *Part Whole Effect Task* showed significant correlations with all scores except for "Object" within the *Face Inversion Effect Task* and "Featural" scores within the *Configural/Featural Difference Detection Task*. Given that the reliability of Featural regression-based difference scores is so low (λ_2 = .10), this affects the extent to which performance on this difference score correlates with others. This finding that "Whole" d-residual scores correlate with most other individual difference scores (less "Object" and "Featural" suggests that conditions where faces or facial features are displayed share a common mechanism.

Accuracy d-residuals (and subtraction-based difference scores, albeit to a lesser extent due to their lower reliability) replicated the pattern observed in the accuracy group-level performance data correlation analysis reported (see Figure 5). Not surprisingly, due to lower reliability the effect sizes are smaller in the individual difference correlation matrices compared to those from the accuracy correlation analysis.

Parallel analyses of reaction time data are presented in Appendix A. Due to better reliability on the task conditions, as well as for the individual difference scores (both subtraction and regression-based) the magnitudes of the within and between task correlations are higher than those of accuracy. The same pattern emerges as with the accuracy data: Individual difference scores are more strongly correlated within task than they are between tasks suggesting that these tasks are measuring are measuring something unique and to a lesser extent something in common.

Exploratory Factor Analysis of Individual Difference Scores

Subtraction Scores. To further assess the associations between performance levels across the tasks, we conducted exploratory factor analyses on the accuracy subtraction scores and regression d-residuals. As with our analysis of accuracy data on the 16-task conditions, we selected principle axis factoring with a direct oblimin rotation. Again we present both the rotated and unrotated factor structures for comparison purposes.

First, we report the exploratory factor analysis for subtraction scores. Recall that according to Ashim (2017), NESTip is the most accurate method in terms of selecting the correct number of factors to retain when the smallest eigenvalues of the factors retained are 1.0 or above. In this case, eigenvalues were all above 1. As a result, we began by conducting the NESTip analysis, which suggested retaining one factor. Parallel Analysis suggested retaining four factors, and Catell's Scree test suggested that four factors be retained.[7] The NESTip proposed one-factor solution is not in line with the number of factors retained for accuracy performance data (four). We judged, based on the recommendations of Costello and Osborne (2005) that the

[7] We bring this to the reader's attention to highlight the impact that statistical decisions can have on the results and to encourage other researchers to explicitly state the decisions they made along with their rationale.

evidence converged in support of the four-factor solution. Specifically, we determined visually that four factors should be retained based on the "break point" on the scree plot graph, and furthermore we ran factor analyses with a forced number of factors (one, two, three, and four) and compared their respective patterns following rotation as well as the amount of variance explained. The four factor solution was clean, meaning both rotated and unrotated factor solutions had few cross loading items (only 1 in the rotated solution).

Below is the factor analysis for accuracy subtraction difference scores within task manipulation across orientation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was adequately high, .51. Because the sampling adequacy was at the cut off value of .50 we also investigated the anti-image correlation matrix. The diagonal values of the anti-image correlation matrix were mostly > .50, with three approaching .50 ("Face", "Whole" and "Part") suggesting there was adequate colinearity among the variables. Bartlett's test of sphericity was significant (χ^2 (28) = 85.15, $p < .001$). The four-factor solution explained 28.41% of the total variance. Factor 1 explained 11.63% of the variance, Factor 2, 7.38%, Factor 3, 6.42%, and Factor 4, 2.98%. Table 7 and 8 display the unrotated and rotated factor structures, respectively.

Table 7

Unrotated Factor Matrix for Accuracy Subtraction Difference Scores Within Task Manipulation Across Orientation

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
FIE	Face			.55		.07	.39
	Object			.34		.07	.20
PW	Whole		.58			.09	.38
	Part				.37	.04	.20
CFE	Aligned	.71				.19	.52
	Misaligned	.51				.17	.33
C/F	Configural					.07	.16
	Featural					.03	.10

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table 8

Rotated Factor Matrix for Accuracy Subtraction Difference Scores Within Task Manipulation Across Orientation

Task		Factor			
		1	2	3	4
FIE	Face			.62	
	Object		-.35		
PW	Whole		.34	.33	
	Part				.45
CFE	Aligned	.69			
	Misaligned	.58			
C/F	Configural		.36		
	Featural		.32		

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

In both the unrotated and rotated factor structures there is again little evidence to suggest that these tasks can be grouped into configural vs. holistic sets. In the unrotated factor matrix we see a pattern emerge whereby individual differences from three of the four tasks loaded across the four factors. Configural/Featural difference scores communalities were low even after extraction and as a result they failed to load above the cut off value of .32. The two difference scores from the *Composite Face Effect Task* loaded together onto the first factor, and the two difference scores from the *Face Inversion Face Effect Task* loaded together on the third factor. Difference scores from the *Part Whole effect Task* loaded separately with "Whole" on the second factor and "Part" on the fourth.

In the rotated factor solution, four difference scores loaded together: both "Configural" and "Featural", as well was the "Whole", and "Object" (which loaded negatively). This negative loading indicates that the higher participants' accuracy difference was for *Face Inversion Effect Task* "Object" trials, the lower their accuracy difference is on the construct that factor three represents. "Whole" cross-loaded and also appeared with "Face" on the third factor. "Aligned" and "Misaligned" difference scores again loaded together onto their respective factor, and "Part"

loaded alone on the fourth factor again. From subtraction-based individual difference scores the evidence is not as compelling as it was in the accuracy group-level performance data factor analyses for the tasks loading onto four separate factors. However, individual difference scores did not load together either. The low reliability scores may be contributing to these difficult to interpret patterns.

Subtraction-based reaction time individual difference scores on the other hand demonstrated a clear picture with difference scores from each task loading together onto their respective factor in the rotated solution (see Appendix A). Furthermore, five of the eight difference scores loaded onto a common first factor in the unrotated solution. Reaction time subtraction individual difference scores mirrored the results from the group-level accuracy performance data, and supports the interpretation that there is something unique to each of these face-processing tasks acknowledging that they are also measuring something in common.

Regression scores. The factor analyses for d-residuals are reported below. Results from the NESTip indicated that two factors should be extracted. The Scree Test and Parallel analysis both suggested a three-factor solution. We also ran factor analyses with a forced number of factors (two, three, and four) and compared their respective patterns following rotation as well as the amount of variance explained. We elected to choose a three-factor solution, which retained eigenvalues above 1, and provided the cleanest solution with few cross loadings (only one in the unrotated solution).

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was adequately high, .57. Bartlett's test of sphericity was significant (χ^2 (28) = 138.30, $p < .001$). The three-factor solution explained 29.45% of the total variance. Factor 1 explained 14.23% of the variance, Factor 2,

9.79%, Factor 3, 5.43%. Table 9 and 10 display the unrotated and rotated factor structures respectively.

Table 9

Unrotated Factor Matrix for Accuracy Regression D-Residual Difference Scores Within Task Manipulation Across Orientation

Task		Factor			Communality	
		1	2	3	Initial	Extraction
FIE	Face			.45	.07	.23
	Object		-.36		.10	.23
PW	Whole	.47			.17	.33
	Part				.07	.09
CFE	Aligned	.60	-.39		.21	.56
	Misaligned	.59			.28	.39
C/F	Configural		.47		.15	.32
	Featural		.40		.11	.21

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table 10

Rotated Factor Matrix for Accuracy Regression D-Residual Difference Scores Within Task Manipulation Across Orientation

Task		Factor		
		1	2	3
FIE	Face			.49
	Object			.41
PW	Whole		.51	
	Part			
CFE	Aligned	.73		
	Misaligned	.78		
C/F	Configural		.55	
	Featural		.42	

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

The rotated solution demonstrates that individual difference scores from each task tend to load together onto respective factors. The exception is "Whole" which loads with the difference scores from the *Configural Featural Difference Detection Task*, and "Part" did not meet the

loading suppression cutoff of .32. These findings provide additional support for the findings from group-level accuracy performance data. These results are also consistent with the unrotated and rotated accuracy subtraction scores factor analyses (which demonstrated this pattern albeit less clearly).

Turning to the unrotated d-residual matrix, we see difference scores loading onto two factors with the exception of "Face" which loads alone onto the third factor. "Aligned" and "Misaligned" load together with "Whole" on the first factor. The second factor is composed of "Object" (loading negatively), "Aligned" (cross loaded with factor 1 and loading negatively), along with "Configural and Featural" which loaded positively. These groupings suggest that difference scores are tapping into common constructs (or components thereof). Importantly the mix of positive and negative loadings on factor two suggest that the more accurately participants are on "Object" and "Aligned" trials the worse they perform on tasks requiring the construct (or component thereof) that Factor 2 represents. The difference between the rotated and unrotated solutions concerning how the difference scores load underscores the importance of reporting both types of analyses because there are discrepancies in the pattern they each present. However, both unrotated and rotated solutions fail to provide convincing evidence to suggest that these conditions cannot be grouped into separable configural vs. holistic sets.

Overall, this pattern of results suggests that one cannot readily equate task conditions with a specific underlying mechanism such as holistic or featural processing. We therefore suggest that researchers avoid treating these tasks as interchangeable measures of the mechanism(s) or constructs implicated in facial recognition. Factor analysis results from the reaction time subtraction and regression-based individual difference scores further supports this

contention. In both cases the rotated factor solution clearly demonstrates that difference scores from each task load together on one of four factors (see Appendix A).

Discussion

There has been a limited amount of research investigating performance patterns across the different paradigms that purport to index integrative aspects of face processing (Richler et al., 2014). An important area of research within this domain concerns determining whether researchers are justified in interpreting the results from a number of commonly used facial recognition tasks as if they are interchangeable. The evidence to date suggested that various tasks are not tapping into the same underlying construct, or perhaps that they are not doing so in a comparable way (e.g., DeGutis et al., 2013; Richler et al., 2012; Richler & Gauthier, 2014; Wang et al., 2012). However, a limited selection of tasks has been included in this line of inquiry. In the present study we aimed to shed light on the associations between four tasks that have frequently been used to examine face processing: the *Face Inversion Effect Task*, the *Part Whole Effect Task*, the "partial" *Composite Face Effect Task*, and the *Configural/Featural Difference Detection Task*. Our central questions were: 1) To what degree are these four paradigms related and therefore measuring the same underlying construct? and 2) To what degree can they be said to measure separate configural and holistic aspects of face processing?

To investigate these questions, we assessed patterns of performance across the different conditions of each task using correlation analysis and factor analysis using three different performance measures: accuracy scores for each of the 16 conditions tested across the four tasks, as well as two types of individual difference scores. We calculated individual difference scores using two methods that have been implemented in previous studies: subtraction-based scores and regression-based scores. Doing so allowed us to compare the patterns of performance across the

two methods. This was done with the goal of providing evidence to the debate regarding how best to calculate individual difference scores with these tasks.

Mean accuracy scores, subtraction scores, and d-residual scores yielded a comparable pattern of results. The four tasks appear to be related to one another, but only weakly. The shared variance in ranking explained in the accuracy score data across all 16-task conditions is approximately 7.8%, whereas the variance in ranking explained for the conditions within each task is 23%. This indicates that each of these tasks primarily measure its own specific construct. Correlational analysis of the accuracy data demonstrated weak associations between tasks from different paradigms, and correlation coefficients at, or approaching, $r_s = .00$ for individual difference scores. However, in all cases the strongest correlations were between conditions *within* each task, suggesting that each task is measuring something unique to itself. For the subtraction scores the shared variance in ranking explained is approximately .64% across tasks, and 2.3% within tasks. For d-residual scores the shared variance in ranking explained across tasks is .1.2%, and 6.7% within tasks.

One possible explanation for the relatively low correlations between tasks is that they are a consequence of low reliability scores. Reliability is a reflection of measurements, not of the tasks per se (Ross et al., 2015). Previous research has used various reliability measures and has produced internal reliability scores for some face recognition tasks (DeGutis et al., 2013; Ross, Richler, & Gauthier, 2015; Richler et al., 2015; Wang et al., 2012). A primary concern is that if measures have low reliability scores, then this will impact subsequent analyses. For example, tasks that are not internally reliable cannot be expected to correlate strongly with one another. In the present study we found relatively good reliability scores for group-level accuracy (and reaction time data); however, reliability for individual differences was low for some of the

difference scores. With respect to accuracy, for subtraction-based and regression-based scores reliability was problematic for the "Configural", "Featural" and "Part" difference scores. Specifically, we found Guttman's λ_2 values ranging between .79 and .92 for accuracy scores, between .00 and .89 for subtraction scores, and between .10 to .90 for d-residual scores.

Even compensating for the task reliability by scaling up the obtained correlations (DeGutis et al., 2013) we find that the amount of shared variance in ranking explained between tasks is modest, at 27.21% for accuracy group-level data from the 16-task conditions (scaled-up from 7.86%), and within-task shared variance in ranking explained is 55.57% (scaled-up from 23.23%). For subtraction scores, shared variance in ranking explained across tasks after scaling up correlations is 17.14% (up from 0.64%), and within-tasks the shared variance in ranking explained is 37.85% (up from 2.35%). For d-residual scores, shared variance in ranking explained across tasks is 30.03% (scaled-up from 1.23%), and within-tasks the shared variance in ranking explained is 64.84% (scaled-up from 6.72%). Please note this magnitude is driven largely by a scaled-up correlation that was well above 1.0 (2.25) between configural and featural difference scores.

The factor analyses provided corroborating evidence for our correlational analyses. That is, these showed that the majority of task conditions loaded onto a single factor in the unrotated solution, but that each task loads onto its own respective factor within the rotated solution. This was most clearly demonstrated in the group-level accuracy data, followed by d-residuals, then subtraction-based individual difference scores. For reaction time group-level and both sets of individual difference clearly demonstrated this pattern (see Appendix A). Across performance measures, including individual difference scores, there was no evidence in either the rotated or

unrotated solutions that the conditions within each task group together into separate featural, configural or holistic mechanisms.

We conclude that while the four tasks measure something in common about facial processing, they also measure separate constructs and therefore should not be used interchangeably. This applies whether analyzing performance on task conditions, or individual difference scores measured as subtraction or d-residuals. These findings add to a growing body of evidence showing that the various tasks used to index configural or holistic processing are not fungible (e.g., Richler & Gauthier, 2014), and may help to explain some discrepant findings in the literature (e.g., DeGutis et al., 2013, Konar et al., 2010, Richler et al., 2011; Wang et al., 2012).

Readers should recall that rotation does not improve the analysis in the sense of expanding the amount of variance explained by each of the factors; however, what rotation does do is clarify and simplify the obtained results (Osborne & Costello, 2009). We have found different patterns in our data with the unrotated factor structure and the rotated factor structure, suggesting that future research should include both rotated and unrotated factor structures, as together they provide a more complete report about how performance on these facial recognition tasks is related.

Regarding the question of whether subtraction or regression is the superior method for calculating individual difference scores, our obtained reliability scores for accuracy supported DeGutis et al. (2013) such that regression d-residual scores were more reliable than scores obtained via subtraction for most difference scores, or were at least as reliable (e.g., Accuracy: the *Face Inversion Effect Task*'s difference scores). For reaction time data individual difference

scores' reliability was consistent between subtraction and regression scores for all but the "Part" difference score for which the regression produced a more reliable score (λ_2 =.48 to λ_2 = .62).

While inherent differences in the tasks used to index holistic and configural processing are likely the greatest source of discrepancies in the literature, other sources of variation exist. One of these concerns the fact that researchers differ in the performance results they report. Most researchers report only accuracy or sensitivity data, some include reaction time, and a few report some kind of conglomerate score such as efficiency or inverse efficiency (Goffaux et al., 2005; Shore et al., 2006). Efficiency scores do report all the data, albeit in a conglomerated format. These sorts of scores can also account for speed accuracy trade-offs, and provide a comprehensive account of performance because, for instance, holistic processing might be demonstrated by accuracy differences in the *Composite Face Effect Task*, but by reaction time differences in the *Face Inversion Effect Task*. However, there exists no consensus on the best way to calculate efficiency, and its psychometric properties have not been established. In particular, there is no established method for calculating the reliability of efficiency data. As a result, in order to provide an analysis of the psychometric properties of performance across these tasks we elected to report both accuracy and reaction time findings.

One limitation to the current study is that our findings likely pertain more to processing unfamiliar faces than familiar ones. This is because we have focused on integrative processing, which may be less important when recognizing well-known faces (Burton et al., 2015). Future research should investigate performance patterns across these tasks using familiar faces and compare task reliabilities, correlational analyses, and factor analyses to the results obtained here. For example, Megreya and Burton (2006) demonstrated that there is a correlation between familiar and unfamiliar face matching. Face matching is required in the *Configural/Featural*

Difference Detection Task as it is a sequential matching task. It would be valuable to determine if face familiarity impacts performance patterns on sequential matching tasks in a similar way.

Taken together, our results suggest that there is primarily something unique that each of these tasks is measuring, and there may also be a common underlying processing mechanism that all four tasks are tapping into. Researchers should therefore not generalize results across studies that use these different face-processing tasks. This is likely a contributing factor to the discrepancies in results across past studies in this domain. Questions that our study leaves open are how, and to what extent, each task is different from the others. That is, if each task is measuring a separate construct, or component thereof, what is that construct or component? And moreover, how are the tasks related to underlying processing mechanisms? Richler et al. (2012) have offered a framework for understanding the various integrative aspects of holistic face processing. This framework presents a variety of what they call "hypothesized mechanisms" of holistic processing, "alternative meanings" of holistic processing onto which holistic tasks either fall exclusively or overlap, and lastly the "relation to face-selective markers." Future research may wish to investigate whether the four tasks used in the present study differ in ways that are compatible with this framework. Knowing which constructs each task indexes would allow researchers to more effectively select the research paradigm that best fits their research question. Until this next step is taken we advocate for researchers to err on the side of caution and to refrain from generalizing results across studies that have used different tasks.

Another important research question that the current project does not address is the degree to which the various tasks relate to general face recognition abilities, as measured by tasks like the CFMT or the Vanderbilt Holistic Face Processing Test (VHPT-F) (Richler et al., 2014). We believe it is important to first determine empirically if the tasks are measuring the

same construct(s), or measuring the same construct(s) in comparable ways, before investigating the associations between these tasks and a general measure of facial recognition ability about which there is considerable debate as well (for a discussion of these two measures of general face recognition ability see Richler et al., 2015). Building upon Richler and Gauthier (2014), and the results presented within this manuscript, a future study may wish to include the "complete" Composite Face Effect Task into the design to see how performance on that task relates to performance on the others, and include a measure of general face recognition ability.

Conclusion

In conclusion, our performance data and our individual difference scores data suggested that four commonly used face recognition tasks primarily measure four independent aspects of face processing. The quantitative and/or qualitative nature of this task independence remains to be established, and we therefore caution researchers not to use these tasks interchangeably. It is our hope that future research will be able to empirically establish what each task is measuring which will allow researchers to select the task(s) that is most appropriate given their research question.

We found that all four tasks are reasonably reliable for both accuracy and reaction time group-level data, as were all reaction time individual difference scores. Reliably was low for three accuracy individual difference scores at least as implemented here. We urge researchers to report reliability data in their studies because internal reliability has obvious implications for the interpretation of findings. Finally, we advocate the analysis and reporting of both accuracy and reaction time data because it has yet to be determined how performance is best captured in each task.

Establishing if these tasks are measuring the same construct(s) in the same way is a necessary precondition in order for research within the domain of facial recognition to progress. Generally speaking, because our results suggest these tasks are measuring things differently, or are measuring different things, this calls into question previous research that investigated the correlations between two or more of these tasks, and it also has implications for research that investigated the associations between integrative processing and general facial recognition ability in the sense that findings should not be generalized beyond the tasks that were used.

References

Achim, A. E. (2017). Testing the number of required dimensions in exploratory factor analysis. *The Quantitative Methods for Psychology*, *13*(1), 64-74.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115-147.

Boutet, I., Collin, C., & Faubert, J. (2003). Configural face encoding and spatial frequency information. *Attention, Perception, & Psychophysics*, *65*(7), 1078-1093.

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science*, *10*(4), 482-496.

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*(2), 89-99.

Carbon, C. C., & Leder, H. (2005). When feature information comes first! Early processing of inverted faces. *Perception*, *34*(9), 1117-1134.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, *195*(4275), 312-314.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1-9.

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*(1), 87-100.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107-117.

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*(4), 576-585.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, *26*(2), 105-109.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception*, *29*(2), 159-170.

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition*, *103*(2), 322-330.

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*(2), 431.

Goffaux, V., Hault, B., Michel, C., Vuong, Q. C., & Rossion, B. (2005). The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception*, *34*(1), 77-86.

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*(4), 1023-1039.

Goffaux, V., & Rossion, B. (2007). Face inversion disproportionately impairs the perception of vertical but not horizontal relations between features. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 995 – 1001.

Gold, J. M., Mundy, P. J., & Tjan, B. S. (2012). The perception of a face is no more than the sum of its parts. *Psychological Science*, *23*(4), 427-434.

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*(1), 189-199.

Hogg, R. V. (1979). An introduction to robust estimation. *Robustness in Statistics*, 1-17.

Huber, P.J. (1981). *Robust Statistics*. New York, NY: John Wiley and Sons.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science*, *21*(1), 38-43.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(2), 513-536.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255-260.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876.

Meinhardt-Injac, B., Boutet, I., Persike, M., Meinhardt, G., & Imhof, M. (2017). From development to aging: Holistic face perception in children, younger and older adults. *Cognition*, *158*, 134-146.

McKone, E., Davies, A. A., Darke, H., Crookes, K., Wickramariyaratne, T., Zappia, S., ... Fernando, D. (2013). Importance of the inverted control in measuring holistic face processing with the composite effect and part-whole effect. *Frontiers in Psychology*, *4*, 1-33.

Mercure, E., Dick, F., & Johnson, M. H. (2008). Featural and configural face processing differentially modulate ERP components. *Brain Research*, *1239*, 162-170.

Mondloch, C. J., Pathman, T., Maurer, D., Le Grand, R., & de Schonen, S. (2007). The composite face effect in six-year-old children: Evidence of adult-like holistic face processing. *Visual Cognition*, *15*(5), 564-577.

Murray, J. E., Yong, E., & Rhodes, G. (2000). Revisiting the perception of upside-down faces. *Psychological Science, 11,* 492–496.

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology*, *10*(1), 56-67.

Osborne, J. W., & Costello, A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review*, *12*(2), 131-146.

Rakover, S. S. (2002). Featural vs. configurational information in faces: A conceptual and empirical analysis. *British Journal of Psychology*, *93*(1), 1-30.

Renzi, C., Schiavi, S., Carbon, C. C., Vecchi, T., Silvanto, J., & Cattaneo, Z. (2013). Processing of featural and configural aspects of faces is lateralized in dorsolateral prefrontal cortex: A TMS study. *Neuroimage*, *74*, 45-51.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464-471.

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.

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition ability and holistic processing. *Journal of Vision*, *15*(9), 1-12.

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254-276.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281–1302.

Richler, J. J., Mack, M. L., Palmeri, T. J., & Gauthier, I. (2011). Inverted faces are (eventually) processed holistically. *Vision Research*, *51*(3), 333-342.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology*, *3*, 1-6.

Riesenhuber, M., & Wolff, B. S. (2009). Task effects, performance levels, features, configurations, and holistic face processing: A reply to Rossion. *Acta Psychologica*, *132*(3), 286-292.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*, *47*(3), 736–743.

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*(2), 199-223.

Shore, D. I., Barnes, M. E., & Spence, C. (2006). Temporal aspects of the visuotactile congruency effect. *Neuroscience Letters*, *392*(1), 96-100.

Stein, T., Sterzer, P., & Peelen, M. V. (2012). Privileged detection of conspecifics: Evidence from inversion effects during continuous flash suppression. *Cognition*, *125*(1), 64-79.

Sunday, M. A., Richler, J. J., & Gauthier, I. (2017). Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Attention, Perception, & Psychophysics*, *79*(5), 1453-1465.

Susilo, T., Rezlescu, C., & Duchaine, B. (2013). The composite effect for inverted faces is reliable at large sample sizes and requires the basic face configuration. *Journal of Vision*, *13*(13), 14-14.

Tabachnick, B. G. & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th Ed.). Needham Heights, MA: Allyn & Bacon.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Fournal of Experimental Psychology*, *46*(2), 225-245.

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, *25*(5), 583-592.

Taschereau-Dumouchel, V., Rossion, B., Schyns, P. G., & Gosselin, F. (2010). Interattribute distances do not represent the identity of real world faces. *Frontiers in Psychology*, *1*, 159.

Troje, N. and H. H. Bülthoff: Face recognition under varying poses: The role of texture and shape. Vision Research 36, 1761-1771 (1996).

Valentine, T. (1988). Upside- down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*(4), 471-491.

Van Belle, G., De Graef, P., Verfaillie, K., Rossion, B., & Lefèvre, P. (2010). Face inversion impairs holistic perception: Evidence from gaze-contingent stimulation. *Journal of Vision*, *10*(5), 1-13.

Wang, H., Guo, S. and Fu, S. (2016), Double dissociation of configural and featural face processing on P1 and P2 components as a function of spatial attention. Psychophysiology, *53*(8), 1165–1173.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169-177.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141-145.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747-759.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256-2262.

Appendix A

Supplementary Materials: Experiment 1

The supplementary materials section contains additional data analyses that were not presented within the manuscript. The analyses presented here, in conjunction with the content in the manuscript, provide a thorough overview of performance data i.e., accuracy and reaction time data from the four integrative face recognition tasks. We present the ANOVAs for accuracy and reaction time data to demonstrate that we replicated the expected interaction patterns thus justifying the inclusion of these tasks in our experiment. We then provide reliability analyses including upper bound limits on correlation coefficients, and scaled-up (or *disattenuated*) correlations, as well as correlations for reaction time group-level performance data. Next, we present the factor solutions, both unrotated and rotated, for reaction time group-level data. We also present these same analyses on reaction time individual difference scores (across orientation within task condition) calculated using both subtraction and regression methods.

Analysis of Variance of Performance Data

Effect sizes are reported as η_p^2 along with their 95% confidence intervals. We used the same method as Richler and Gauthier (2014) for calculating 95% confidence intervals around η_p^2. This process involved 1) converting η_p^2 values into rho values by taking the square root, 2) entering the rho value and sample size ($N = 223$) into an online statistical calculator (http://vassarstats.net/rho.html) which generated the 95% confidence interval endpoints around rho, and then 3) squaring the end point values to convert them into 95% confidence intervals around η_p^2. We would like to note that the error bars shown in Figures A1 to A8 of the supplemental materials are correlation-adjusted 95% confidence intervals of the means, appropriate for within-subject designs (O'Brien & Cousineau, 2014).

Accuracy

Face Inversion Effect Task. In the *Face Inversion Effect Task* the stimulus manipulation was Face or Object, and the orientation manipulation was either upright or inverted. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,222) = 206.66$ $p < .001$, $\eta_p^2 = .48$ (95% CI: .38, .57), a main effect of Orientation, $F(1,222) = 110.43$, $p < .001$, $\eta_p^2 = .33$ (95% CI: .22, .43), and the interaction term was significant, $F(1,222) = 116.41$, $p < .001$, $\eta_p^2 = .34$ (95% CI: .24, .44). Figure A1 depicts these results.

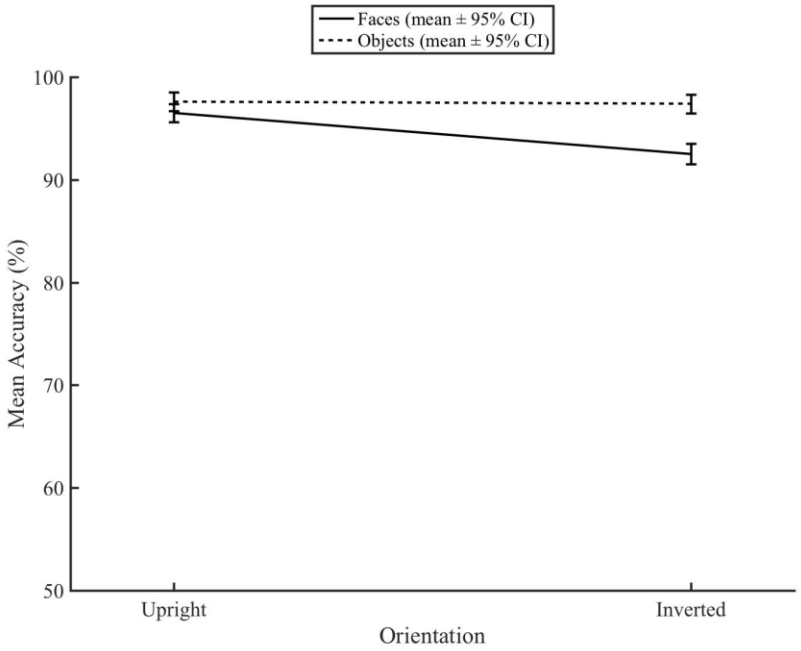

Figure A1. The main effect of Stimulus, of Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. As expected, participants performed significantly more accurately when recognizing an upright face compared to inverted face, as integrative processing is thought to be disrupted by inversion. The effect of inversion was more pronounced in the face task manipulation than the object manipulation.

Part Whole Effect Task. In the *Part Whole Effect Task* the stimulus manipulation was Part or Whole face stimuli, and the orientation manipulation was either upright or inverted

stimuli. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,222) = 152.56$, $p < .001$, $\eta_p^2 = .41$ (95% CI: .31, .51), a main effect of Orientation, $F(1,222) = 191.34$, $p < .001$, $\eta_p^2 = .46$ (95% CI: .36, .55), and the interaction term was significant, $F(1,222) = 30.65$, $p < .001$, $\eta_p^2 = .12$ (95% CI: .05, .21). Figure A2 depicts these results.

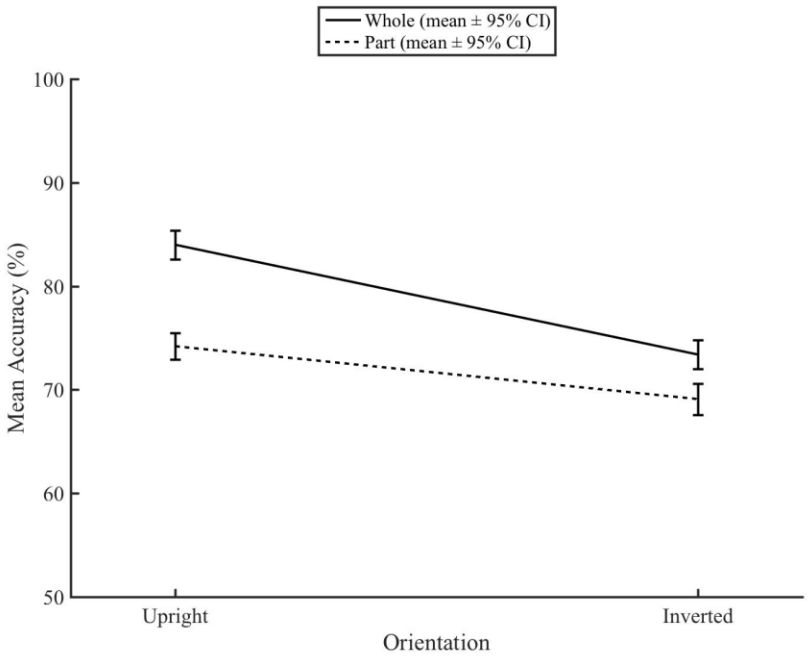

Figure A2. The main effect of Stimulus, of Orientation, and the interaction are all significant. The graph presents within-subject 95% confidence intervals around the mean values. As expected, face parts are recognized more accurately in the context of a whole face than they are in isolation, and this effect is reduced by inversion. Holistic processing is thought to be more affected by inversion than is featural processing, and our results conform to this expected pattern, with accuracy in the whole condition impaired to a greater degree by inversion than accuracy in the part condition. Inversion did significantly impair recognition accuracy on Part trials.

Composite Face Effect Task. In the *Composite Face Effect Task* the stimulus manipulation was aligned or misaligned (referring to the alignment of top and bottom face halves) and the orientation manipulation was upright or inverted. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,222) =$

15.46, $p < .001$, $\eta_p^2 = .07$ (.95% CI: .02, .14), a main effect of Orientation, $F(1,222) = 101.75$, $p < .001$, $\eta_p^2 = .31$ (.95% CI: .21, .41), and the interaction term was significant, $F(1,222) = 80.65$, $p < .001$, $\eta_p^2 = .27$ (95% CI: .17, .37). Figure A3 depicts these results.

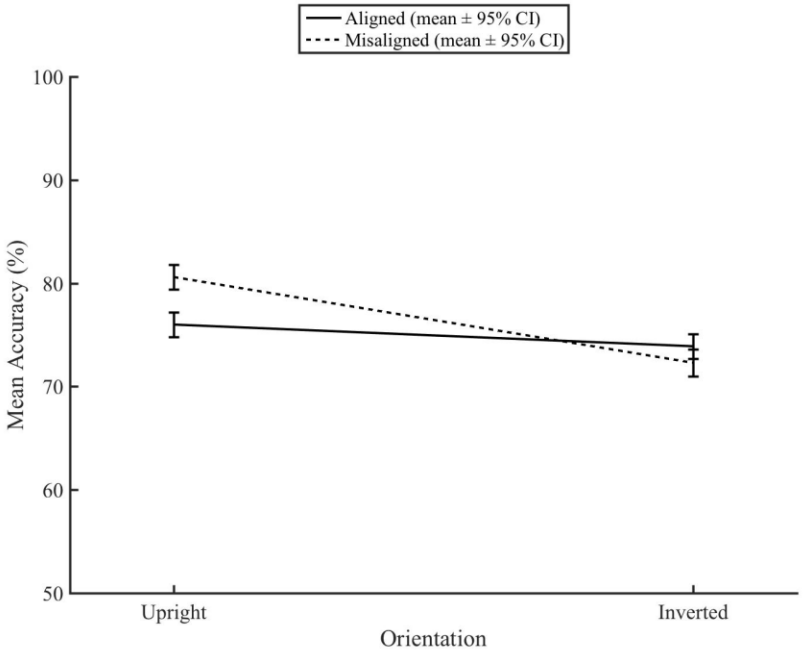

Figure A3. The main effect of Stimulus, of Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. As expected, alignment and inversion each have a significant effect on accuracy performance. Performance is more accurate on misaligned upright trials compared to aligned upright trials suggesting that participants integrated the irrelevant bottom half of the face into their judgments due to the tendency to process faces holistically despite instructions not to do so. Inversion disrupts holistic processing, as does misalignment, therefore, as expected, accuracy performance across stimulus types in the inverted condition is similar (although participants performed more accurately on inverted aligned trials than inverted misaligned trials). Moreover, inversion disrupts performance accuracy on misaligned trials to a greater degree than it does aligned trials.

Configural Featural Difference Detection Task. In the *Configural/Featural Difference Detection Task* the stimulus manipulation was a featural swap or shifting a feature up or down 3 mm. The orientation manipulation was upright or inverted stimuli presentation. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,222) = 321.02$, $p < .001$, $\eta_p^2 = .59$ (95% CI: .51, .67) a main effect of Orientation, $F(1,222) =$

418.42, $p < .001$, $\eta_p^2 = .65$ (95% CI: .58, .72) and the interaction term was significant, $F(1,222) = 137.35$, $p < .001$, $\eta_p^2 = .38$ (95% CI: .28, .48). Figure A4 depicts these results.

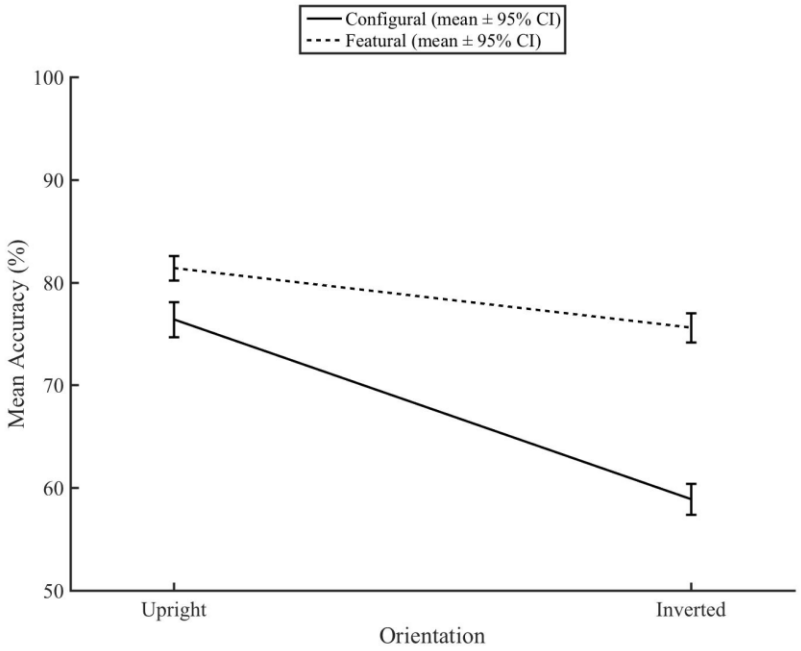

Figure A4. The main effect of Stimulus, of Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. As expected, inversion disrupted recognition accuracy in both featural and configural trials; however, inversion had a greater impact on configural trials suggesting they require holistic processing to a greater extent than featural trials do. Across orientation people were more accurate in their recognition judgments on featural trials than on configural trials.

Reaction Time

Face Inversion Effect. The task manipulation in the *Face Inversion Effect Task* was Face or Object, and orientation manipulation was upright or inverted stimuli presentation. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimuls Manipulation, $F(1,222) = 686.07$ $p < .001$, $\eta_p^2 = .76$ (95% CI: .70, .81), a main effect of Orientation, $F(1,222) = 79.72$, $p < .001$, $\eta_p^2 = .26$ (95% CI: .16, .36), and the interaction term was significant, $F(1,222) = 63.64$, $p < .001$, $\eta_p^2 = .22$ (95% CI: .16, .35). Figure A5 depicts these results.

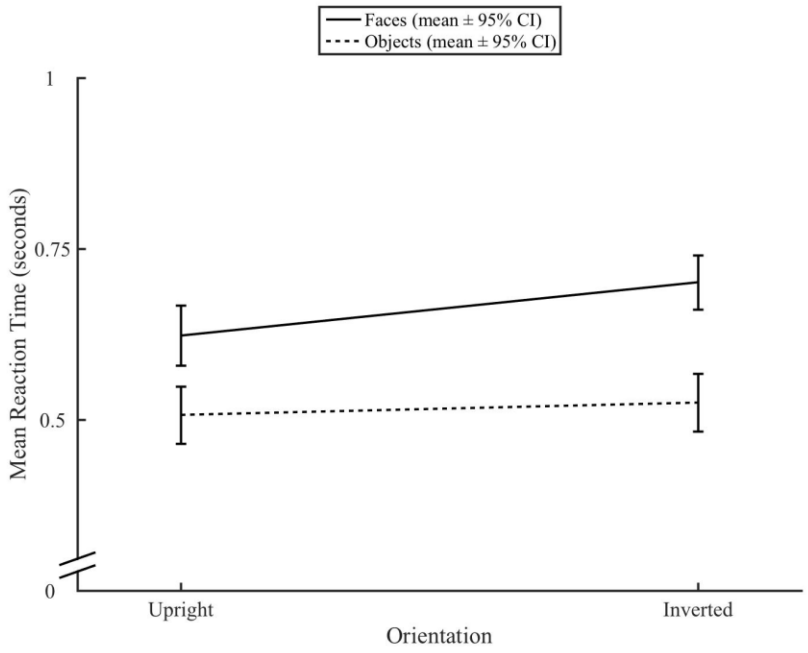

Figure A5. The main effect of Stimulus Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Faces and objects are both recognized more slowly when the stimuli are presented upside down compared to upright. The impact of inversion on reaction time is more pronounced when the faces are inverted compared to when objects are inverted. These results are expected given that presenting faces upside down makes holistic processing more difficult. Overall faces are recognized more slowly than objects are across orientation conditions.

Part Whole Effect Task. In the *Part Whole Effect Task* the stimulus manipulation was Part or Whole and the orientation manipulation was upright or inverted. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,222) = 222.19$, $p < .001$, $\eta_p^2 = .50$ (95% CI: .41, .59), a main effect of Orientation, $F(1,222) = 5.82$, $p < .05$, $\eta_p^2 = .03$ (95% CI: .00, .09), and the interaction term was significant, $F(1,222) = 48.14$, $p < .001$, $\eta_p^2 = .18$ (95% CI: .09, .27). Figure A6 depicts these results.

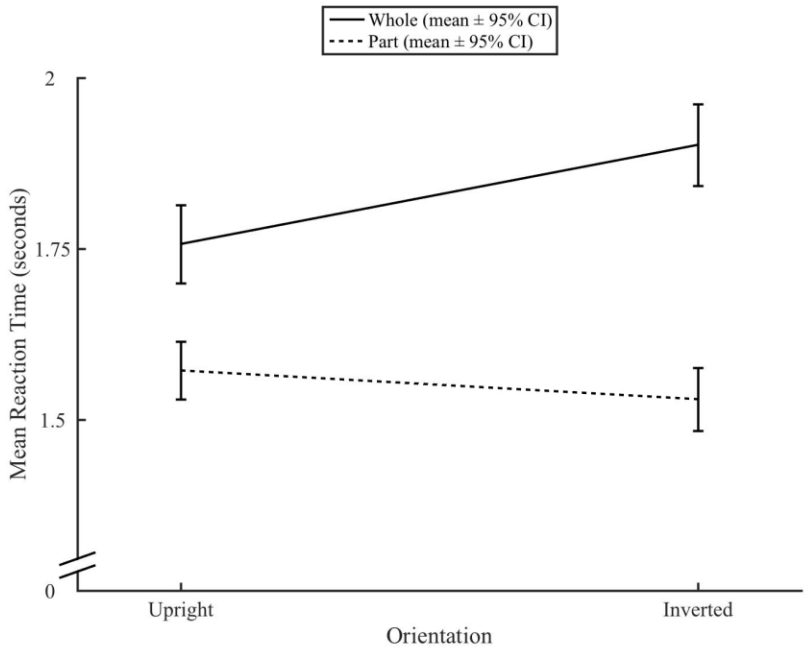

Figure A6. The main effect of Stimulus, Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Reaction time patterns show that in both the upright and inverted orientation, participants take longer to make a recognition judgment when facial features are presented within the context of a face compared to when they are presented in isolation. Participants took significantly more time to respond during inverted whole trials compared to upright whole trials. Moreover, the impact of inversion on reaction time is more pronounced in the whole condition than it is in the part condition, which is consistent with the belief that inversion disrupts holistic processing to a greater extent than it disrupts featural processing.

Composite Face Effect Task. The Stimulus Manipulation in the *Composite Face Effect Task* was aligned or misaligned face halves. The other manipulation was Orientation with stimuli presented upright or inverted. A 2 × 2 within-subjects analysis of variance showed that there was a main effect of task manipulation, $F(1,222) = 29.78$, $p < .001$, $\eta_p^2 = .12$ (95% CI: .05, .21), a main effect of orientation, $F(1,222) = 14.91$, $p < .001$, $\eta_p^2 = .06$, (95% CI: .01, .13) and the interaction term was significant, $F(1,222) = 27.44$, $p < .001$, $\eta_p^2 = .11$ (95% CI: .04, .20). Figure A7 depicts these results.

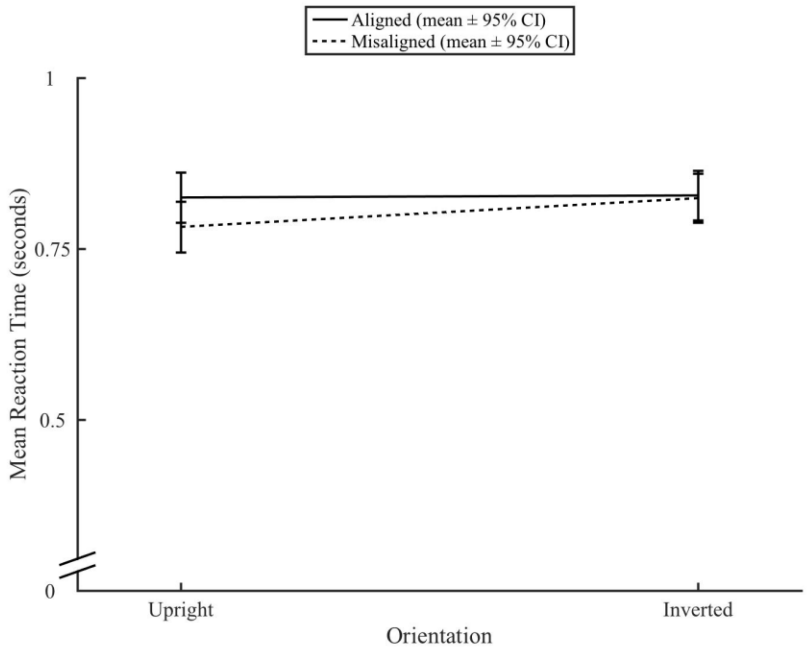

Figure A7. The main effect of Stimulus Manipulation, of Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Reaction time data shows that in the upright condition aligned faces are recognized more slowly than misaligned faces. This provides support for the alignment effect which means task difficulty is higher when face halves are aligned due to the tendency to process faces holistically, and task difficulty is lower when face halves are horizontally offset, which disrupts holistic processing. Inverted faces in the misaligned condition are recognized significantly more slowly than are misaligned upright faces. Participants do not differ in the amount of time they looked at aligned faces across orientation conditions. Notably, when faces are presented upside down, alignment does not contribute additional task difficulty beyond that of face inversion. This is evidenced by similarity in reaction time across alignment conditions within the inverted orientation.

Configural Featural Difference Detection Task. The stimulus manipulation in the Configural Featural Difference Detection Task was a featural swap or shifting a feature up or down 3 mm. The stimuli were either presented upright or inverted. An analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,222) = 68.62$, $p < .001$, $\eta_p^2 = .24$ (95% CI: .15, .34), a main effect of Orientation, $F(1,222) = 14.83$, $p < .001$, $\eta_p^2 = .06$ (95% CI: .01, .14), and the interaction term was significant, although there was a very small effect size

for this interaction, $F(1,222) = 5.93$, $p < .001$, $\eta_p^2 = .03$. (95% CI: .00, .09). Figure A8 depicts these results.

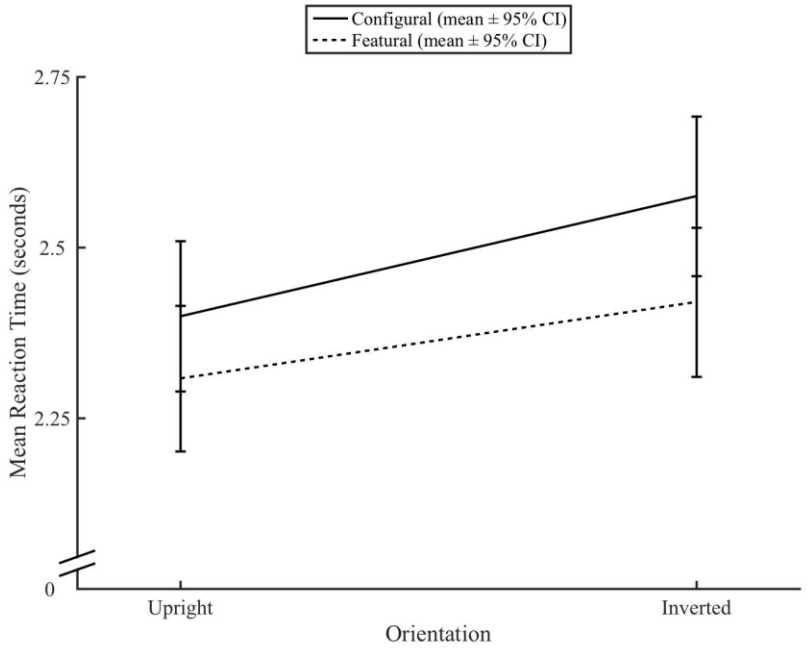

Figure A8. The main effect of Stimulus, of Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Participants took longer when responding on configural manipulation trials than they did to featural manipulation trials across orientation conditions. And participants took longer to respond when faces were presented upside down compared to upright on both configural and featural trials. This impairment due to inversion was more pronounced on configural trials than on featural trials as is expected because inversion is thought to disrupt holistic/configural processing to a greater extent than featural processing.

Reliability Analysis and Upper Bound Limit on Correlation of Performance Data

Accuracy

Guttman's λ_2 reliability scores per task for accuracy data were as follows: *Face Inversion Effect Task* (.92), *Part Whole Effect Task* (.79), *Composite Effect Task* (.90), and *Configural/Featural Difference Detection Task* (.80). Table A1 shows the accuracy theoretical upper bound correlation limits. Once the reliabilities were calculated for group-level performance data for each task we used them to obtain what is called the theoretical upper bound

limit on a correlation. If the true correlation between the two measures were 1.0, this is the highest correlation that one could expect given the reliability of the scores on the two tasks after accounting for measurement error. The theoretical upper limit is calculated as follows: the reliabilities of the two tasks are multiplied together, and then one takes the square root of the product. This is a useful tool because researchers can scale up the obtained correlations by multiplying them by 1/Upper Bound limit (Schmidt & Hunter, 1996). Table S1 contains the theoretical upper bound correlation limits for accuracy scores.

Table A1

Accuracy Data Theoretical Upper Bound Limit on Correlation

Task	FIE	PW	CFE	C/F
FIE	.92			
PW	.85	.79		
CFE	.91	.84	.90	
C/F	.86	.79	.85	.80

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Reaction Time

We measured task reliability of reaction time scores for all four tasks using Guttman's λ_2. All four tasks showed good reliability. Guttman's λ_2 reliability scores per task for reaction time data were as follows: *Face Inversion Effect Task* (.95), *Part Whole Effect Task* (.87), *Composite Effect Task* (.97), and *Configural/Featural Difference Detection Task* (.95). Table A2 shows the reaction time upper bound correlations.

Table A2

Reaction Time Data Theoretical Upper Bound Limit on Correlation

Task	FIE	PW	CFE	C/F
FIE	.95			
PW	.91	.87		
CFE	.96	.92	.97	
C/F	.95	.91	.96	.95

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Scaled-Up Correlation Analysis - Accuracy

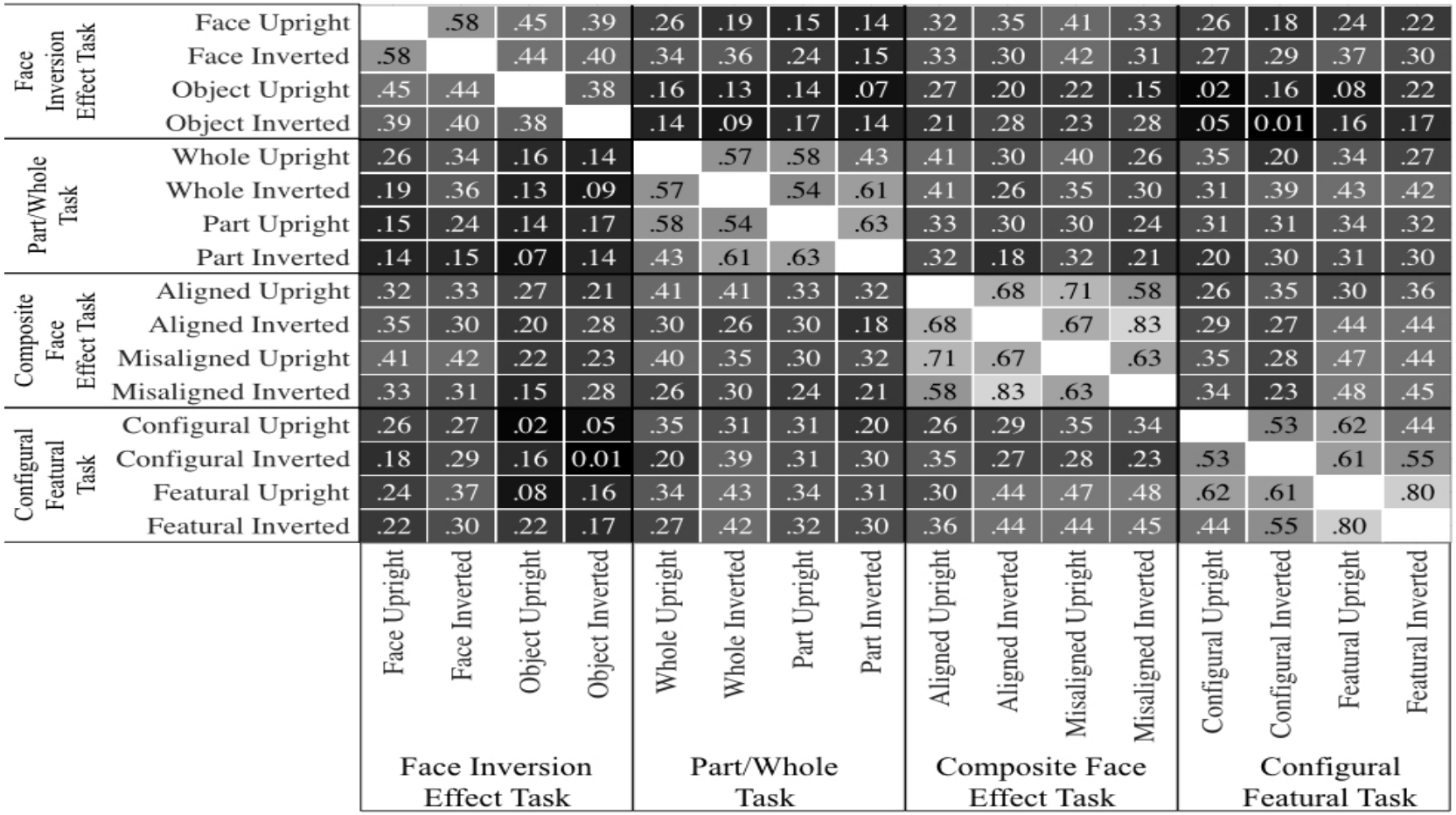

Figure A9. Scaled-Up Spearman's Rho Correlations ($N = 223$) of Accuracy Data Between Task Conditions. r_s values between .132 - .172 correspond to $p < .05$, two-tailed. r_s values between .173 - .187 correspond to $p < .01$, two-tailed. Part Whole Effect Task (Part/Whole Task), Configural/Featural Difference Detection Task (Configural Featural Task).

Reaction Time

		Face Upright	Face Inverted	Object Upright	Object Inverted	Whole Upright	Whole Inverted	Part Upright	Part Inverted	Aligned Upright	Aligned Inverted	Misaligned Upright	Misaligned Inverted	Configural Upright	Configural Inverted	Featural Upright	Featural Inverted
Face Inversion Effect Task	Face Upright		.85	.87	.83	.38	.23	.37	.28	.44	.41	.46	.43	.26	.12	.20	.14
	Face Inverted	.85		.77	.85	.39	.27	.43	.36	.42	.38	.40	.41	.25	.16	.21	.17
	Object Upright	.87	.77		.86	.31	.16	.30	.20	.44	.39	.44	.41	.16	.06	.13	.07
	Object Inverted	.83	.85	.86		.31	.21	.33	.30	.42	.39	.43	.42	.16	.07	.12	.08
Part/Whole Task	Whole Upright	.38	.39	.31	.31		.62	.70	.50	.45	.43	.47	.43	.36	.27	.36	.28
	Whole Inverted	.23	.27	.16	.21	.62		.64	.74	.49	.49	.44	.47	.32	.43	.30	.41
	Part Upright	.37	.43	.30	.33	.70	.64		.68	.50	.51	.48	.47	.33	.36	.31	.37
	Part Inverted	.28	.36	.20	.30	.50	.74	.68		.51	.53	.45	.50	.25	.37	.22	.35
Composite Face Effect Task	Aligned Upright	.44	.42	.44	.42	.45	.49	.50	.51		.84	.89	.84	.19	.23	.19	.23
	Aligned Inverted	.41	.38	.39	.39	.43	.49	.51	.53	.84		.86	.92	.27	.35	.29	.35
	Misaligned Upright	.46	.40	.44	.43	.47	.44	.48	.45	.89	.86		.87	.22	.24	.23	.25
	Misaligned Inverted	.43	.41	.41	.42	.43	.47	.47	.50	.84	.92	.87		.23	.28	.25	.28
Configural Featural Task	Configural Upright	.26	.25	.16	.16	.36	.32	.33	.25	.19	.27	.22	.23		.74	.93	.75
	Configural Inverted	.12	.16	.06	.07	.27	.43	.36	.37	.23	.35	.24	.28	.74		.75	.95
	Featural Upright	.20	.21	.13	.12	.36	.30	.31	.22	.19	.29	.23	.25	.93	.75		.77
	Featural Inverted	.14	.17	.07	.08	.28	.41	.37	.35	.23	.35	.25	.28	.75	.95	.77	
		Face Upright	Face Inverted	Object Upright	Object Inverted	Whole Upright	Whole Inverted	Part Upright	Part Inverted	Aligned Upright	Aligned Inverted	Misaligned Upright	Misaligned Inverted	Configural Upright	Configural Inverted	Featural Upright	Featural Inverted
		Face Inversion Effect Task				Part/Whole Task				Composite Face Effect Task				Configural Featural Task			

Figure A10. Spearman's Rho Correlations ($N = 223$) of Reaction Time Data Between Task Conditions. r_s values between .132 - .172 correspond to $p < .05$, two-tailed. r_s values between .173 - .187 correspond to $p < .01$, two-tailed. Part Whole Effect Task (Part/Whole Task), Configural/Featural Difference Detection Task (Configural Featural Task).

Scaled-Up Correlation Analysis – Reaction Time

		Face Upright	Face Inverted	Object Upright	Object Inverted	Whole Upright	Whole Inverted	Part Upright	Part Inverted	Aligned Upright	Aligned Inverted	Misaligned Upright	Misaligned Inverted	Configural Upright	Configural Inverted	Featural Upright	Featural Inverted
Face Inversion Effect Task	Face Upright		.89	.91	.87	.41	.25	.40	.30	.45	.42	.48	.45	.27	.13	.21	.14
	Face Inverted	.89		.81	.89	.43	.29	.47	.40	.42	.39	.42	.41	.27	.17	.22	.18
	Object Upright	.91	.81		.90	.35	.17	.33	.22	.46	.40	.46	.43	.17	.07	.14	.07
	Object Inverted	.87	.89	.90		.34	.23	.36	.33	.44	.41	.44	.44	.17	.08	.12	.08
Part/Whole Task	Whole Upright	.41	.43	.35	.34		.71	.80	.58	.48	.47	.51	.47	.40	.29	.39	.31
	Whole Inverted	.25	.29	.17	.23	.71		.73	.85	.53	.54	.48	.51	.35	.47	.33	.45
	Part Upright	.40	.47	.33	.36	.80	.73		.78	.54	.56	.53	.51	.36	.39	.34	.41
	Part Inverted	.30	.40	.22	.33	.58	.85	.78		.56	.57	.49	.54	.27	.40	.24	.39
Composite Face Effect Task	Aligned Upright	.45	.42	.46	.44	.48	.53	.54	.56		.87	.92	.86	.20	.23	.20	.24
	Aligned Inverted	.42	.39	.40	.41	.47	.54	.56	.57	.87		.89	.94	.28	.37	.30	.36
	Misaligned Upright	.48	.42	.46	.44	.51	.48	.53	.49	.92	.89		.89	.23	.25	.24	.26
	Misaligned Inverted	.45	.41	.43	.44	.47	.51	.51	.54	.86	.94	.89		.24	.30	.26	.29
Configural Featural Task	Configural Upright	.27	.27	.17	.17	.40	.35	.36	.27	.20	.28	.23	.24		.78	.98	.79
	Configural Inverted	.13	.17	.07	.08	.29	.47	.39	.40	.23	.37	.25	.30	.78		.79	.99
	Featural Upright	.21	.22	.14	.12	.39	.33	.34	.24	.20	.30	.24	.26	.98	.79		.81
	Featural Inverted	.14	.18	.07	.08	.31	.45	.41	.39	.24	.36	.26	.29	.79	.99	.81	
		Face Upright	Face Inverted	Object Upright	Object Inverted	Whole Upright	Whole Inverted	Part Upright	Part Inverted	Aligned Upright	Aligned Inverted	Misaligned Upright	Misaligned Inverted	Configural Upright	Configural Inverted	Featural Upright	Featural Inverted
		Face Inversion Effect Task				Part/Whole Task				Composite Face Effect Task				Configural Featural Task			

Figure A11. Scaled-Up Spearman's Rho Correlations ($N = 223$) of Reaction Time Data Between Task Conditions. r_s values between .132 - .172 correspond to $p < .05$, two-tailed. r_s values between .173 - .187 correspond to $p < .01$, two-tailed. Part Whole Effect Task (Part/Whole Task), Configural/Featural Difference Detection Task (Configural Featural Task).

Factor Analysis of Performance Data

Reaction Time

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was high, .84. Bartlett's test of sphericity was significant (χ^2 (120) = 4292.86, $p < .001$). The same decisions about rotation and item suppression were taken as were described in the manuscript text pertaining to accuracy scores. With regards to factor extraction, both the scree test approach and the conservative NESTip approach suggested retaining four factors. The four-factor solution explained 81.24% of the total variance. Factor 1 explained 42.44% of the variance, Factor 2, 20.43%, Factor 3, 11.80% and Factor 4, 6.57%. Table A3 and A4 display the unrotated and unrotated factor structures respectively.

Table A3

Unrotated Factor Matrix for Reaction Time Data

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
FIE	Face Up	.60	-.58	.45		.91	.91
	Face Inv	.69	-.41	.33		.77	.76
	Object Up	.56	-.63	.46		.92	.92
	Object Inv	.58	-.61	.38		.85	.85
PW	Whole Up	.63			.34	.63	.54
	Whole Inv	.59			.45	.68	.75
	Part Up	.71			.40	.67	.71
	Part Inv	.62			.30	.67	.66
CFE	Aligned Up	.78		-.38		.84	.83
	Aligned Inv	.82		-.39		.90	.91
	Misaligned Up	.82		-.33		.90	.91
	Misaligned Inv	.82		-.37		.92	.93
C/F	Configural Up	.53	.59	.44		.90	.84
	Configural Inv	.48	.71			.89	.80
	Featural Up	.51	.61	.42		.91	.85
	Featural Inv	.50	.70			.89	.52

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F), Upright (Up), Inverted (Inv).

Table A4

Pattern Matrix, Rotated Four-Factor Structure of Reaction Time Data

Task		Factor			
		1	2	3	4
FIE	Face Up			.96	
	Face Inv			.79	
	Object Up			.97	
	Object Inv			.90	
PW	Whole Up				.66
	Whole Inv				.88
	Part Up				.77
	Part Inv				.81
CFE	Aligned Up	.88			
	Aligned Inv	.96			
	Misaligned Up	.94			
	Misaligned Inv	.97			
C/F	Configural Up		.92		
	Configural Inv		.85		
	Featural Up		.95		
	Featural Inv		.87		

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F), Upright (Up), Inverted (Inv).

Reliability Analysis and Upper Bound Limit on Correlation: Individual Difference Scores Within Task Across Orientation

Accuracy

Table A5

Accuracy Subtraction Difference Scores Upper Bound Limit on Correlation Within Task Manipulation Across Orientation

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	.69							
	2. Object	.78	.89						
PW	3. Whole	.22	.53	.32					
	4. Part	.23	.26	.16	.08				
CFE	5. Aligned	.58	.67	.40	.20	.50			
	6. Misaligned	.63	.72	.43	.22	.54	.58		
C/F	7. Configural	.33	.38	.23	.11	.28	.30	.16	
	8. Featural	.00	.00	.00	.00	.00	.00	.00	.00

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table A6

Accuracy Regression D-Residual Difference Scores Upper Bound Limit on Correlation Within Task Manipulation Across Orientation

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	.70							
	2. Object	.79	.90						
PW	3. Whole	.51	.58	.37					
	4. Part	.44	.50	.32	.28				
CFE	5. Aligned	.62	.71	.46	.40	.56			
	6. Misaligned	.66	.75	.48	.42	.60	.63		
C/F	7. Configural	.32	.37	.24	.20	.29	.31	.15	
	8. Featural	.26	.30	.19	.17	.24	.25	.12	.10

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Reaction Time

Table A7

Reaction Time Task Reliabilities of Subtraction Difference Scores and Regression Difference D-residuals Within Task Manipulation Across Orientation

Individual Difference Scores	**Guttman's λ_2 Subtraction**	**Guttman's λ_2 Regression**
Face	.73	.74
Object	.71	.72
Aligned	.76	.80
Misaligned	.48	.62
Whole	.67	.69
Part	.69	.71
Configural	.83	.85
Featural	.72	.75

Table S8

Reaction Time Subtraction Difference Scores Upper Bound Limit on Correlation Within Task Manipulation Across Orientation

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	.73							
	2. Object	.72	.71						
PW	3. Whole	.75	.73	.76					
	4. Part	.59	.58	.60	.48				
CFE	5. Aligned	.70	.69	.71	.57	.67			
	6. Misaligned	.71	.70	.72	.58	.68	.69		
C/F	7. Configural	.78	.77	.79	.63	.74	.76	.83	
	8. Featural	.72	.71	.74	.58	.70	.70	.77	.72

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table A9

Reaction Time Regression D-Residual Difference Scores Upper Bound Limit on Correlation Within Task Manipulation Across Orientation

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	.74							
	2. Object	.73	.72						
PW	3. Whole	.76	.76	.80					
	4. Part	.68	.67	.70	.62				
CFE	5. Aligned	.71	.70	.74	.65	.69			
	6. Misaligned	.72	.71	.75	.66	.70	.71		
C/F	7. Configural	.79	.78	.82	.73	.77	.78	.85	
	8. Featural	.74	.73	.77	.68	.72	.73	.80	.75

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Correlation Analysis: Individual Difference Scores Within Task Across Orientation

Accuracy

Scaled Up Correlation Analysis.

Table A10

Accuracy Scaled Up Spearman's Rho Correlation of Subtraction Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.18**	-						
PW	3. Whole	.59**	-.07	-					
	4. Part	-.09	-.15*	.44**	-				
CFE	5. Aligned	-.16*	.15*	-.05	-.55**	-			
	6. Misaligned	-.17**	.07	.16**	-.32**	.52**	-		
C/F	7. Configural	.24**	-.18**	.91**	.18**	-.21**	.17**	-	
	8. Featural	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-

Note. *$p < .05$, two-tailed. ** $p < .01$, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F). N/A indicates we cannot calculate a scaled-up correlation because the reliability of that condition is 0 and therefore so too is the theoretical upper limit of correlation.

Table A11

Accuracy Scaled Up Spearman's Rho Correlation of Regression D-Residual Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.18**	-						
PW	3. Whole	.20**	N/A	-					
	4. Part	.02	-.02	.72**	-				
CFE	5. Aligned	-.03	.19**	.30**	.18**	-			
	6. Misaligned	N/A	.09	.42**	.14*	.67**	-		
C/F	7. Configural	N/A	-.35**	.83**	.60**	.03	.39**	-	
	8. Featural	-.19**	-.47**	.52**	.59**	N/A	.44**	2.25**	-

Note. *$p < .05$, two-tailed. **$p < .01$, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F). N/A indicates we could not calculate a scaled-up correlation because the correlation between the two conditions was 0. Note that it is possible to obtain scaled-up correlations with a magnitude greater than 1.0.

Reaction Time

Table A12

Non-parametric Spearman Rho Correlations for Reaction Time Subtraction Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.40**	-						
PW	3. Whole	.09	.15*	-					
	4. Part	.06	.20**	.41**	-				
CFE	5. Aligned	-.04	.10	.01	-.05	-			
	6. Misaligned	.11	.05	.15*	.15*	.32**	-		
C/F	7. Configural	.13	-.07	.26**	.18**	.15*	.06	-	
	8. Featural	.10	-.14*	.28**	.15*	.08	.10	.70**	-

Note. *$p < .05$, two-tailed. **$p < .01$, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table A13

Non-parametric Spearman Rho Correlations for Reaction Time Regression D-Residual Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.16*	-						
PW	3. Whole	.25**	.19**	-					
	4. Part	.25**	.21**	.49**	-				
CFE	5. Aligned	.14*	.11	.01	.01	-			
	6. Misaligned	.11	.08	.17*	.18**	.34**	-		
C/F	7. Configural	.01	.20**	.28**	.21**	.08	.05	-	
	8. Featural	.06	.21**	.24**	.14*	.04	.07	.74**	-

Note. **p* < .05, two-tailed. ***p* < .01, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Scaled Up Correlation Analysis

Table A14

Scaled Up Non-parametric Spearman Rho Correlations for Reaction Time Subtraction Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.56**	-						
PW	3. Whole	.12	.21**	-					
	4. Part	.17**	.34**	.68**	-				
CFE	5. Aligned	.06	.14*	.02	.08	-			
	6. Misaligned	.15*	.07	.13*	.26**	.47**	-		
C/F	7. Configural	.67**	-.09	.33**	.29**	.20**	.08	-	
	8. Featural	.39**	-.19**	.38**	.26**	.11	.14*	.91**	-

Note. **p* < .05, two-tailed. ***p* < .01, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table A15

Scaled Up Non-parametric Spearman Rho Correlations for Reaction Time Regression D-Residual Difference Scores Within Task Manipulation Across Orientation (N = 223)

Task	Variables	1	2	3	4	5	6	7	8
FIE	1. Face	-							
	2. Object	.22**	-						
PW	3. Whole	.33**	.25**	-					
	4. Part	.37**	.31**	.70**	-				
CFE	5. Aligned	.20**	.16**	.01	.02	-			
	6. Misaligned	.15*	.11	.23**	.27**	.49**	-		
C/F	7. Configural	.01	.25**	.34**	.29**	.10	.06	-	
	8. Featural	.08	.29**	.31**	.21**	.06	.10	.93**	-

Note. *$p < .05$, two-tailed. **$p < .01$, two tailed. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Factor Analysis: Individual Difference Scores Within Task Across Orientation

Reaction Time

Subtraction Difference Scores. Below is the factor analysis for reaction time subtraction difference scores within task manipulation across orientation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was adequately high, .60. Bartlett's test of sphericity was significant (χ^2 (28) = 419.13, $p < .001$). We determined the number of factors to retain using the conservative NESTip approach, which suggested retaining four factors. The scree test and parallel analysis corroborated this result. The four-factor solution explained 54.13% of the total variance. Factor 1 explained 26.58% of the variance, Factor 2, 10.47%, Factor 3, 8.82%, and Factor 4, 8.26%. Table A16 and A17 display the unrotated and rotated factor structures respectively.

Table A16

Unrotated Factor Matrix for Reaction Time Subtraction Difference Scores Within Task Manipulation Across Orientation

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
FIE	Face	.37				.23	.40
	Object			-.36	.38	.21	.47
PW	Whole	.55		.44		.32	.55
	Part	.40		.46		.25	.44
CFE	Aligned		.43		-.38	.20	.40
	Misaligned		.52		-.32	.21	.49
C/F	Configural	.82				.64	.78
	Featural	.70				.63	.79

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table A17

Rotated Factor Matrix for Reaction Time Subtraction Difference Scores Within Task Manipulation Across Orientation

Task		Factor			
		1	2	3	4
FIE	Face				.61
	Object				.70
PW	Whole			.69	
	Part			.67	
CFE	Aligned		.62		
	Misaligned		.68		
C/F	Configural	.85			
	Featural	.90			

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Regression D-Residual Scores. Below is the factor analysis for reaction time regression difference d-residual scores within task manipulation across orientation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was adequately high, .62. Bartlett's test of sphericity was significant (χ^2 (28) = 440.92, $p < .001$). We determined the number of factors to retain using the

conservative NESTip approach, which suggested retaining four factors. The scree test and parallel analysis approaches also corroborated this result. The four-factor solution explained 53.00% of the total variance. Factor 1 explained 27.65% of the variance, Factor 2, 12.06%, Factor 3, 8.12%, and Factor 4, 5.17%. Table A18 and A19 display the unrotated and rotated factor structures respectively.

Table A18

Unrotated Factor Matrix for Reaction Time Regression Difference D- Residual Scores Within Task Manipulation Across Orientation

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
FIE	Face	.38			.34	.17	.32
	Object				.47	.12	.33
PW	Whole	.61		-.35		.36	.62
	Part	.47	.33	.47		.30	.45
CFE	Aligned		.38	.46		.19	.42
	Misaligned		.47	.34		.21	.46
C/F	Configural	.79	-.36			.67	.77
	Featural	.78	-.48			.66	.87

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Table A19

Rotated Factor Matrix for Reaction Time Regression Difference D-Residual Scores Across Orientation Within Task Manipulation

Task		Factor			
		1	2	3	4
FIE	Face				.48
	Object				.58
PW	Whole			-.75	
	Part			-.66	
CFE	Aligned		.64		
	Misaligned		.66		
C/F	Configural	.83			
	Featural	.95			

Note. Face Inversion Effect Task (FIE), Part Whole Effect Task (PW), Composite Face Effect Task (CFE), Configural/Featural Difference Detection Task (C/F).

Chapter 3: Investigating the Associations between Face Recognition Ability and Four Tasks Thought to Index Integrative Aspects of Face Processing

This manuscript has been prepared with the intent of submitting it to the Journal of Experimental Psychology: Human Perception & Performance. It is therefore formatted in accordance with the requirements of that journal.

We note that because this paper was prepared as a stand-alone manuscript, much of the information in the introduction repeats what is found in the general introduction of the thesis.

Investigating the Associations between Face Recognition Ability and Four Tasks Thought to Index Integrative Aspects of Face Processing

Elizabeth A. Nelson[1], Isabelle Boutet[1], Charles A. Collin[1]

University of Ottawa[1]

Abstract

A number of studies have investigated whether various tasks thought to measure integrative aspects of face processing (i.e., holistic and/or configural processing) yield outcomes that are related to face recognition ability. A related question is whether and how the various integrative processing tasks are themselves related to one another. However, only a small subset of tasks has been investigated, and approaches to analysis have been highly heterogeneous. This study employed a within-subjects design ($N = 260$) in order to examine 1) patterns of performance across four commonly used face processing tasks: *Complete Composite Face Effect Task*, *Partial Composite Face Effect Task*, *Part Whole Effect Task*, and the *Configural/Featural Difference Detection Task*, and 2) the nature of the association that performance on each face processing task has with face recognition ability as measured by performance on the *Cambridge Face Memory Test* (CFMT). Performance data included accuracy and reaction time measures, as well as accuracy and reaction time difference scores (both subtraction-based and regression-based). Correlational analyses and factor analyses for group-level data suggested that the four integrative processing tasks are each primarily measuring unique constructs. Individual difference scores provide similar evidence, although with low psychometric reliability. Regarding how integrative face-processing performance is related to performance on the CFMT, group-level data showed a low to moderate positive correlations across most task conditions. Individual difference scores' association with CFMT was less consistent.

Introduction

Face recognition is thought to rely critically on integrative processing (e.g., Richler & Gauthier, 2014; Rossion, 2013). This involves levels of processing above that of analyzing individual face features, which is termed *featural processing*. Two broad classes of integrative processing are widely discussed in the literature: Configural and holistic. Configural processing refers to using the relations between stimulus features to recognize and distinguish between stimuli. Maurer, Le Grand, and Mondolch (2002) suggest that there are two kinds of configural relations that can be used in this kind of analysis. One is *First-order relations,* which refers to the qualitative organization of features on a face. Human faces all have the same first order configuration: two eyes are placed side by side and are centered above a nose, which is centered above a mouth. *Second-order relations* refers to the metric spacing between these features, and includes such things as inter-ocular distance and the distance between mouth and nose. A second form of integrative processing, called holistic processing, refers to humans' tendency to process faces as intact wholes rather than by their distinguishable parts (i.e. featural aspects) (Maurer et al., 2002).

The association between configural and holistic processing is debated, with some seeing them as distinct, and others as inter-related in various ways. For instance, some see configural processing as preliminary to holistic processing, while others see them as operating in parallel. Both types of integrative processing are widely held to be important aspects of face processing (e.g., Rezlescu, Susilo, Wilmer, & Caramazza, 2017; Tanaka & Farah, 1993) although this is not universally accepted (e.g., Donnelly, Cornes, & Menneer, 2012; Fitousi, 2015).

One reason that integrative processing is thought to be important to face recognition is that humans are quite adept at correctly individuating faces despite similarities in general face

shape, distances between facial features, and the appearance of individual features themselves. Integrative processing mechanisms are thought to incorporate and elaborate on such information, allowing for successful individuation of faces despite their homogeneity. Richler and Gauthier (2014) and Rossion (2013) have both published reviews supporting the position that holistic processing is important for face recognition. Corroborating evidence for this position includes the work from a number of groups (e.g., DeGutis, Wilmer, Mercado, & Cohan, 2013; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012; Richler, Cheung, & Gauthier, 2011b) all of whom used correlational analyses to demonstrate that there is a small but significant correlations between indexes of holistic processing and face recognition ($r = 0.2 – 0.4$). Moreover, findings from research investigating holistic processing and face recognition in people with prosopagnosia (Busigny, Joubert, Felician, Ceccaldi, & Rossion, 2010; DeGutis, Cohan, Mercado, Wilmer, & Nakayama, 2012; Palermo et al., 2011), as well as a parallel line of inquiry pertaining to those with autism spectrum disorder (Hobson, Ouston, & Lee, 1988; Joseph & Tanaka, 2003), suggest that when holistic processing is impaired so too is face recognition ability.

However, results from a number of recent studies (Richler, Floyd, & Gauthier, 2014, 2015; Konar, Bennett, & Sekuler, 2010; Rezlescu et al., 2017) pose a direct challenge to the assertion that holistic processing ability is important to face recognition. For instance, Richler et al., suggest that the degree of stimulus repetition drives the association between integrative processing and face recognition ability, such that with little stimulus repetition the correlations between performance on holistic processing tasks and face recognition ability becomes non-significant (Richler et al., 2014; Richler et al., 2015). Having stimulus repetition allows for participants to learn the stimuli (in this case, learn the face halves that are included in the

Composite Face Effect Task). When the same five face halves were repeatedly used as stimuli, performance was significantly correlated with accuracy on the Cambridge Face Memory Test (CFMT)[8] (Duchaine & Nakayama, 2006). However, when 95 different faces halves were presented in the same task with little repetition, performance was not significantly correlated with performance on the CFMT (Richler et al., 2015).

Additional challenges to the importance of holistic processing in face recognition come from Konar et al. (2010) and Rezlescu et al. (2017). Konar et al. (2010)'s results showed no significant correlation between holistic processing, as measured by the *partial Composite Face Effect Task*, and upright face identification ability ($r = .05$, $p = .66$). Similarly, Rezlescu et al. (2017) found no significant correlations between performance on the *partial Composite Face Effect Task* and face recognition ability, as measured by the *Cambridge Face Perception Test* (CFPT) (Duchaine, Germine, & Nakayama, 2007) ($r = .04$).

Because there are inconsistencies in results pertaining to the association between holistic processing and face recognition ability, it is important to consider the tasks being used. Different tasks may not measure holistic processing, or general face recognition ability, in comparable ways. For example, Konar et al. (2010) and Rezlescu et al. (2017) both used the so-called partial version of the *Composite Face Effect Task* (Young, Hellawell, & Hay, 1987)[9] in their respective

[8] The CFMT is a commonly used task designed to measure face recognition ability. Participants first study, then must recognize, a variety of faces presented at a variety of viewing angles and levels of luminance. This task is described in detail in its respective section in the introduction as well as the methods section.

[9] The term "partial" composite effect task refers to the original paradigm in which the bottom halves of faces are always different from one another. In contrast, the "complete" composite effect task also includes trials in which the bottom halves are the same. Refer to Richler, Cheung, and Gauthier (2011) for a fuller description of the differences between what they refer to as the "complete" and "partial" versions of this task. This distinction is also explained in the relevant section within this manuscript. The so-called "partial" version of the task has also been referred

studies. While some authors advocate for the use of this version of the task (e.g., Michel, Rossion, Han, Chung, & Caldara, 2006; Robbins & McKone, 2007; Rossion, 2013; Susilo, Rezlescu, & Duchaine, 2013), others challenge it and prefer the so-called complete *Composite Face Effect Task* version (see Richler & Gauthier, 2014 for a review). Indeed, Richler et al. (2015) have suggested that Konar et al.'s (2010) finding of no significant correlation between integrative processing and face recognition ability may be due to the fact that they elected to use the "partial" version of the *Composite Face Effect Task* as their measure of holistic face processing.[10]

This same critique applies to Rezlescu et al. (2017), who also found no significant correlation between performance on the *Partial* Composite Face Effect Task and face recognition ability, as measured by the Cambridge Face Perception Test (CFPT) (Duchaine, Germine, & Nakayama, 2007). Importantly though, Rezlescu et al. (2017) did find significant correlations between face recognition ability (CFPT) and performance on two additional integrative face processing tasks: The *Face Inversion Effect Task* (Yin, 1969) (r = .42) and the *Part Whole Effect Task* (Tanaka & Farah, 1993) (r = .25). This makes it clear that, as researchers endeavour to better understand the nature of the association between integrative processing and face recognition ability, we must pay particular attention to the tasks that are used.

---

to as the "standard" version (Rossion, 2013), while the "complete" version has also been referred to as the "congruency effect" (McKone et al., 2013). In this manuscript we implemented both versions, which we label as the "partial" or "complete" *Composite Face Effect Task*.

[10] Richler & Gauthier (2014) provide a review of what they believe to be the inherent problem within the "partial" version of the *Composite Face Effect Task,* which is one of response bias that confounds the measure of holistic processing. The authors contend that this response bias is due to the fact that bottom face halves are always different and as a result 1) all the "same" trials (in which the top face halves match) are incongruent trials and 2) all the "different" trials (in which the top face halves do not match) are congruent trials.

As the literature review to this point indicates, there is a great deal of discrepancy regarding findings that speak to some of the most basic questions in face recognition research. These questions include: 1) Are there distinct holistic and configural processing mechanisms in face recognition? 2) If so, to what degree are they predictive of face recognition ability? 3) Which experimental tasks are best suited to evaluate integrative processing? To answer these questions it is important to know how to interpret the previous literature investigating integrative face processing. One concern that arises in this context is the degree to which the variety of tasks used in the past measure the same construct. The present study focuses on investigating the associations between four commonly-used integrative processing tasks, as well as each of these tasks' association with general face recognition ability (as measured by performance on the CFMT). Specifically, we examine the *Partial* and *Complete* versions of the *Composite Face Effect Task*, the *Configural Featural Difference Detection Task*, and the *Part Whole Effect Task*. If these tasks do not measure the same construct, then discrepancies between them in the previous literature are easily interpreted, though it will force a re-evaluation of past work. Conversely, if these tasks are found to measure the same construct, then discrepancies will have to be evaluated as arising from methodological errors, lack of power, or some other confounding factors.

There is some evidence to support the notion that the various face-processing tasks used in the past literature on integrative face processing index different mechanisms. For example, Wang, Li, Fang, Tian, and Liu (2012) found a non-significant correlation between performance on the *Part Whole Effect Task* and the *Partial Composite Face Effect Task*. And Richler and Gauthier (2014) investigated the relationship between effect sizes in the *Partial* and *Complete Composite Face Effect Tasks* (Farah, Wilson, Drain, & Tanaka, 1999; Gauthier & Bukach, 2007)

by conducting a meta-analysis and found no significant associations between these two versions of the task. Furthermore, Rezlescu et al. (2017) found only weak correlations between the *Partial Composite Face Effect Task*, the *Part Whole Effect Task*, and the *Face Inversion Effect Task* that varied from r = -.03 to r = .28. Taken together, these findings suggest that these tasks are not all measuring the same construct, or that they are doing so with very weak reliability. Furthermore, with the inconsistencies found across studies, the variety of tasks used, and the possible impact of stimulus repetition that is often not accounted for (Ross, Richler, & Gauthier, 2015), it seems inadvisable to make simple and direct comparisons between previous experiments. As a result, the body of work thus far pertaining to face recognition and face processing leaves us with more questions than it does answers. One way to begin to tease apart what these processing tasks are measuring (and subsequently what their role is in face recognition), is to investigate patterns of performance across a number of tasks using a within-subjects design.

Nelson, Boutet, Watier, and Collin (in preparation) investigated within-subject patterns of performance across commonly used face-processing tasks. This was done in order to determine if these tasks could be interpreted as interchangeable measures of integrative processing. The tasks investigated were the "Partial" *Composite Face Effect Task*, the *Configural/Featural Difference Detection Task* (Freire, Lee, & Symons, 2000; Leder & Bruce, 2000), the *Face Inversion Effect Task*, and the *Part Whole Effect Task*. Results showed that performance (as measured by accuracy and reaction time, as well as difference scores derived from accuracy and reaction time data) was more strongly correlated within each task than it was across tasks, suggesting that these tasks are measuring unique constructs. Factor analyses confirmed these results, showing that performance on each task loaded onto its own respective factor.

The present study was conducted as a follow-up to Nelson et al. (in preparation), with the goal of examining the degree to which various face recognition tasks correlate with face recognition ability. As well, we were motivated to re-examine the question of whether different integrative face recognition tasks measure the same construct. The role that integrative face processing plays in facial recognition ability requires further investigation, and more needs to be understood regarding the associations between commonly-used face processing tasks. For instance, it may be that differences in the construct being indexed by the various integrative face processing tasks contributes to discrepancies in results regarding the link between integrative face processing and face recognition ability (e.g., Konar et al., 2010; Rezlescu et al., 2017). To address these possibilities, the present article 1) examines the associations that exist between performance on a variety of face processing tasks ("*Complete*" *Composite Face Effect Task*, "*Partial*" *Composite Face Effect Task*, *Configural/Featural Difference Detection Task*, and the *Part Whole Effect Task*), and 2) investigates if, and to what extent, performance on these four face processing tasks is related to face recognition ability as measured by performance on the *Cambridge Face Memory Test* (CFMT) (Duchaine & Nakayama, 2006).

As with Nelson et al. (in preparation), the selection of tasks to be examined was based primarily on the prevalence of these tasks in the literature, rather than theoretical concerns. Thus, while recent research has made compelling arguments for the use of alternate tasks, such as the Vanderbilt Holistic Face Processing Test VHPT-F (Richler et al., 2014, Richler et al., 2015) as a measure of holistic processing, we did not include them in the current study due to the relatively low number of studies to use these methods. That is, we selected commonly used tasks, and particular versions of these tasks, in order for our data to have the broadest applicability to existing data.

We wish to draw attention to two differences in the integrative face processing tasks that are included in the present experiment relative to those examined in Nelson et al. (in preparation). The first is that we elected to not include the *Face Inversion Effect Task* in this experiment. This was for two reasons. First, during initial data collection we experienced ceiling effects in performance on this task and were concerned this would not allow us to analyze its association with the other tasks. The second reason we excluded the *Face Inversion Effect Task* was a practical one: We wished to keep the experiment length under two hours in order to obtain high quality data. As will be explained below, we have included inverted conditions within three of the four tasks, and felt as a result that including the *Face Inversion Effect Task* was not crucial.

The second difference in the integrative tasks used between the present experiment and Nelson et al. (in preparation) is that we included two versions of the *Composite Face Effect Task*, the so-called "partial design" and "complete design" in this manuscript. There is debate in the literature as to which is the better measure of holistic processing, with Rossion (2013) advocating for the use of the *Partial* design and Gauthier and colleagues advocating for the use of the *Complete* design (e.g. Cheung, Richler, Palmeri, & Gauthier, 2008; Richler, Cheung, & Gauthier, 2011a, b; Richler, Mack, Palmeri, & Gauthier 2011, Richler & Gauthier, 2013). By including both versions of the *Composite Face Effect Task* we were able to compare patterns of performance between them regarding their respective association with other integrative processing tasks and with performance on the CFMT.

Behavioural Measures of Performance: Accuracy, RT, and Difference Scores

The behavioural measures we elected to analyze in this manuscript are accuracy and reaction time. Because there is considerable redundancy between the two measures' results,

accuracy data are presented within the manuscript and reaction time data are presented in Appendix B. Where there are meaningful discrepancies between accuracy and reaction time data, this is noted in the text. We thought it prudent to report both measures in order to capture performance more completely. Also, it remains to be determined whether integrative processing is best reflected in accuracy data or reaction time data, and whether the best measure might vary from one task to another.

In addition to analyzing accuracy and reaction time data, we also investigated patterns of performance using an individual differences approach. When using a group-means approach, the variation in performance data occurring across participants is considered to be noise. Individual difference scores, on the other hand, represent a measure of performance between pairs of task conditions. This approach treats variations in performance across participants as important information. Individual difference scores are often used to index configural or holistic processing because these mechanisms are thought to be reflected more by certain task conditions than others. According to Yovel, Wilmer, and Duchaine (2014) investigating face processing using an individual difference approach is important because it can complement or extend findings based on the performance data.

There are two commonly-used methods of calculating individual difference scores, subtraction and regression, and there is debate in the literature regarding which yields the better measure. Before summarizing this debate, we first provide a brief explanation of how these scores are calculated. Each processing task in the present study contains four task conditions. Using the *Part Whole Effect Task* as an example, there are "Whole Upright", "Whole Inverted", "Part Upright", and "Part Inverted" task conditions. When calculating individual difference scores, regardless of the calculation method used, the researcher must identify which are the

"control" conditions and which are the "conditions of interest". This process depends on the research question at hand. For the sake of an example, if attempting to measure holistic processing one could consider the "Whole Upright" condition to be the condition of interest. As for choosing the control condition, the researcher could select "Whole Inverted" because inversion is thought to disrupt holistic processing. Or the researcher could select "Part Upright" because holistic processing is not possible when parts are presented in isolation. Where the debate takes place is with respect to how much the variability from the "control" condition should be included in the calculation of the individual difference score.

DeGutis et al. (2013) argue that regression is the superior method because it isolates the variability from the "control" condition during the calculation of difference scores. They showed that reliability scores were higher for their regression-based difference scores than their subtraction-based ones. However, Ross et al. (2015) did not find this to be the case in their study. Rather, they noted that in a number of their datasets subtraction-based difference scores were more reliable for *Complete Composite Face Effect Task* than were the reliability scores for regression-based individual difference data. Additionally, they argue that the "control" conditions are not purely control conditions. For instance, in the "Inverted Whole" condition of the *Part Whole Effect Task*, some degree of holistic processing likely remains. As such, they argue that the control conditions contain elements of the construct of interest and that their variability should therefore remain in the equation when calculating difference scores (i.e., via subtraction). In order to present the most complete picture regarding if and how performance is related across these four face-processing tasks and the CFMT, we present data for both subtraction-based and regression-based difference scores. Doing so also contributes to the debate regarding which, if either, is the superior method.

We calculated individual difference scores across orientation/congruency and within task manipulation, and it is these scores that are presented within the manuscript. For instance, with the *Part Whole Effect Task*, we compared Whole Upright to Whole Inverted, and Part Upright to Part Inverted. Calculating difference scores this way permits us to explore the effect that orientation has on the pair of stimulus manipulation conditions within each task (e.g., aligned and misaligned face halves in the *partial Composite Face Effect Task*). How the individual difference scores, and their reliability scores, are calculated was explained in detail in Nelson et al. (in preparation) and the methods are described briefly below in the Results section.

Integrative Face Processing Tasks

The four integrative face-processing tasks that were investigated in this experiment were the *Partial Composite Face Effect Task*, the *Complete Composite Face Effect Task*, the *Configural/Featural Difference Detection Task*, and the *Part Whole Effect Task*. It is widely held that each of these tasks compares one of two higher-order face-processing types (i.e., either second-order configural or holistic) to featural processing. To date, researchers have used interaction effects to explore the different face processing mechanisms. For example, in the *Part-Whole Task* participants are better and faster when identifying a face part that is within the context of a whole face compared to how they make the same judgment about a face part that is presented in isolation. Importantly, this effect is reduced when the stimuli are presented upside down. This interaction effect is thought to reflect holistic processing. As will be described below in the Results Section, we investigated interaction effects to ensure they were replicated the main effects and interactions found in the literature before conducting other analyses.

Next, for each of the four integrative face-processing tasks we describe the way they purportedly capture featural, configural, and holistic processing, and outline the expected results

per task. Following this, we describe the measure of general face recognition we are using, the CFMT. A full description of each task is presented in the Methods Section. Our hypotheses pertaining to the associations between the four integrative tasks, as well as those pertaining to their association with the CFMT, are given later in their own subsection.

Partial Composite Face Effect Task

This task was first presented in Young et al. (1987), and was developed as a measure of the extent to which participants integrate face parts (i.e., top and bottom face halves). Each face stimulus in this task is a composite face, meaning that the top half and the bottom half of each stimulus were originally part of two different faces. Face halves are either aligned to create the appearance of an intact face, or they are misaligned such that the edge of the top half of the face aligns with the center of the bottom half. This breaks up the face into two obvious halves. For a visual depiction of task trial structures see Figure 2. We employed a sequential matching version of the task. The fact that this is the "partial" version of the task (compare the "complete" version below) means that the bottom halves of both the test and inspection faces are always different, and the top halves can either be the same (match) or be different. The task requires that participants selectively attend to the top halves of both test and inspection faces and determine if they match or if they are different.

Rossion, in his 2013 review paper, explains that this task is thought to provide strong evidence of holistic processing because, despite task instructions to the contrary, participants show performance patterns that indicate a processing strategy that mandatorily integrates both face halves. That is, rather than selectively attending to the top half of each face, which would be advantageous in this task, participants seem to be forced to attend and/or process the entire face as an integrated whole, at least to some degree. It is in part because of this holistic tendency

that humans make recognition errors on this task. The fact that the (irrelevant) bottom face halves differ fools participants into reporting that the (same) top halves are in fact different.

In terms of the face processing that this task requires, featural information is available to be processed in all trials. In both Upright and Inverted Aligned trials it is also possible for participants to use first and second order configural information and/or holistic processing. Importantly it is difficult to selectively attend to face parts (i.e., the top half of faces in this case) when face halves are aligned and upright (Meinhardt-Injac, Boutet, Persike, Meinhardt & Imhof, 2017; Richler, Palmeri, & Gauthier, 2012; Richler & Gauthier 2014). This is thought to be because people have a tendency to process faces holistically (Mondloch, Pathman, Maurer, Le Grand, & de Schonen, 2007; but also see Rossion, 2013). As a consequence, the task-irrelevant information from the bottom half of the face image interferes with the recognition judgment. Offsetting the alignment of the top and bottom face halves makes it easier to selectively attend to the top half of the faces because holistic processing is disrupted in the misaligned condition and therefore the irrelevant bottom halves interfere to a lesser extent (Gauthier & Tarr, 2002).

In line with results from previous experiments that have used this paradigm, we expect that performance will be worse when faces are aligned upright compared to when faces are misaligned upright. Because inversion is thought to disrupt integrative processing (Carey & Diamond, 1977; Goffaux & Rossion, 2006; 2007; Van Belle, De Graef, Verfaillie, Rossion, & Lefèvre, 2010) we expect performance to be similar on inverted aligned and misaligned trials. Lastly, we expect performance to be lower on inverted trials (for both aligned and misaligned trials) compared to performance on aligned and misaligned upright trials.

Complete Composite Face Effect Task

This version of the task, first presented in Farah et al. (1998) is a more "complete" than the "partial" version of the task in terms of the types of trials that are presented. Recall that in the *Partial* version the bottom halves of faces were always different. In this *Complete* version the bottom halves of test and inspection faces can either be the same or be different (refer to Richler et al., 2011b for diagram of trial types). As with the *Partial Composite Face Effect Task*, face halves are either aligned or misaligned and a sequential matching paradigm is employed. In this case, however, all faces are presented upright only. For a visual depiction of task trial structures see Figure1. Participants were still instructed to selectively attend to the top half of both faces when making their judgments.[11]

The reason that this version of the task was created was to address the congruency response bias inherent in the *Partial* version of the task. Ross et al., (2015) note that in the *Partial* design the *same* trials (when top halves match) are all incongruent, in the sense that the top halves are always matched and the bottom halves mismatched; Conversely, the *different* trials (when top halves do not match) are all congruent, in the sense that the top halves are mismatches and so are the bottom halves.

In the *Complete* version of the task half of the *same* trials are congruent, such that the bottom halves of the faces also match one another, and half are incongruent, such that the bottom halves differ while the top halves are the same. Likewise, half of the *different* trials are congruent, whereby the bottom halves differ as the top halves do, and half are incongruent, where the bottom halves are the same while the top halves are different. Having both same and

[11] The term "selectively attend" should not be taken to mean that an attentional mechanism is necessarily at play (although this is a possibility e.g. Chua, Richler, & Gauthier, 2015). Research has suggested that this congruency effect could be the result of perceptual mechanisms (Rossion, 2013).

different trials being both congruent and incongruent in the *Complete* design allows researchers to separate and address the congruency response bias (see Cheung et al., 2008; Richler, Cheung, & Gauthier, 2011a, Richler & Gauthier, 2014).

Holistic processing is understood to be reflected in this task if there is an interaction between the factors of alignment and congruency. The congruent condition consists of "same" and "different" trials. On *same congruent* trials both test and inspection top and bottom face halves match; On *different congruent* trials test and inspection top and bottom face halves are different. Therefore, if participants fail to selectively attend to the top half of the faces, which is they integrate the face halves, then performance could be enhanced on congruent trials. However, on *same incongruent* trials the top halves match and the bottom halves differ and on *different incongruent* trials the top halves differ and the bottom halves match. On these trials adopting a holistic strategy should lead to performance impairment because irrelevant misleading information will interfere with recognition judgments. Performance is typically greater on congruent trials compared to incongruent trials (Cheung et al., 2008; Farah et al., 1998; Gauthier, Curran, Curby, & Collins, 2003; Richler et al., 2009; Richler, Tanaka, Brown, & Gauthier, 2008). Importantly, these performance differences are reduced when face halves are offset (misaligned) compared to when the face halves are aligned (Richler et al., 2008; Richler, Bukach, and Gauthier, 2009; Wong, Palmeri, & Gauthier, 2009). This is thought to arise because offsetting the face halves disrupts holistic processing, such that the congruency of the top and bottom halves of the stimuli becomes less effective.

In this task, featural information is thought to be equally available on all trials, whereas second order configural and/or holistic processing is more effective on aligned trials. Because it is challenging to selectively attend to the top half of aligned faces (i.e., overriding the tendency

to engage in holistic processing) when these faces are presented upright (Meinhardt-Injac et al., 2017; Richler et al., 2012; Richler & Gauthier 2014) the bottom halves of the test and inspection faces can interfere with the recognition judgment regarding the top halves of the two faces despite the task instruction and participants' best efforts.

As in previous literature using this paradigm, we expect performance to be better on misaligned incongruent trials than on aligned incongruent trials. We do not expect alignment to impact performance on congruent trials because both top and bottom faces halves of both faces are either the same or they are different; Therefore, using a holistic processing strategy that factors in the "irrelevant" bottom half will not adversely affect the accuracy of the recognition judgment.

Configural/Featural Difference Detection Task

As the name implies, this task is thought to be a relatively direct measure of configural vs. featural processing (Friere et al., 2000). Each trial in the task begins with a pair of faces displayed simultaneously. Participants are asked to judge whether the two faces are the same or if they differ in any way. Same trials present two identical faces, and constitute half the trials. Different trials present two face stimuli that differ either configurally or featurally. Trials that contain a configural modification have a difference pertaining to the location of a facial feature (eyes, nose or mouth). One of these three features is moved up or down by a short distance; This change is enough to modify the second order relational information. Trials that contain a featural modification have a difference pertaining to one facial feature (eyes, nose, or mouth) such that one of these features is swapped with that same feature from a different face (Carbon & Leder, 2005; Freire et al., 2000, Mercure, Dick, & Johnson, 2008; Renzi et al., 2013). For a visual depiction of task trial structures see Figure 3.

Configural trials require the ability to detect configural changes and thus should engage the configural aspect of integrative processing to a greater extent than featural processing, and vice-versa. Participants may also use featural information and/or use a holistic processing strategy on both kinds of trials.

In terms of expected results, a number of studies have found that humans are better at detecting featural differences than they are at detecting second-order relational (configural) differences (Carbon & Leder, 2005; Freire et al., 2000, Mercure et al., 2008). And the impact of face inversion is more pronounced on configural trials than featural trials (Leder & Bruce, 2000). Therefore, we expect that participants will do better on featural difference detection trials in both the upright and inverted conditions (compared to configural difference detection trials). We expect that participants will do better on upright trials than on inverted trials for both featural and configural difference detection trials. Most importantly, we anticipate seeing a greater impact of inversion on performance on configural trials than on featural trials. It is this interaction effect that is typically held to reflect configural processing in this task.

Part Whole Effect Task

The *Part Whole Effect Task* (Tanaka & Farah, 1993) is a sequential matching task wherein participants are presented with a face that is either right side up or upside down. After the face disappears, either two faces or two face "parts" (i.e., features: eyes, nose, or mouth) are presented simultaneously. The participant then must identify which of the features is the same as the feature that was part of the initially presented face. For a visual depiction of task trial structures see Figure 4.

This task is thought to be a measure of holistic processing. The so-called "whole advantage" refers to one's ability to better recognize individual features when they are presented

as part of a whole face ("whole" trials) than when they are presented in isolation ("part" trials) (Tanaka & Farah, 1993). Only featural processing is assumed to take place during the "part" trials. Featural processing is also thought to take place during the "whole" trials too, alongside holistic processing.

With regard to expected results, inversion is thought to disrupt performance more profoundly during "whole" trials compared to "part" trials. This is because inversion has a disproportionate impact on configural and holistic processing compared to featural processing (Carey & Diamond, 1977; Goffaux & Rossion, 2006; 2007; Van Belle et al., 2010). Therefore, we expect that the performance deficit as a result of inversion will be more pronounced on the "whole" trials than on the "part" trials. We expect that performance will be better on upright whole trials compared to inverted whole trials. And we expect that the inversion effect will be smaller on upright and inverted part trials and so performance will be similar on part trials with slightly more of a performance deficit on inverted part trials compared to upright part trials.

Cambridge Face Memory Test (CFMT)

The CFMT (Duchaine & Nakayama, 2006) is a standardized test of face recognition ability and it has been used broadly within the face recognition literature (Cho et al., 2015). It tests one's ability to learn and subsequently recognize faces. Researchers have used this paradigm to investigate the extent to which holistic processing is important for facial recognition (e.g., DeGutis et al., 2013; Dennett, McKone, Edwards, & Susilo, 2012; McGugin et al., 2012; Richler et al., 2011b). There are several pieces of evidence suggesting that CFMT taps into face-specific processing mechanisms and is therefore an appropriate task to use to address this research question. For example, in neurotypical participants, accuracy performance drops significantly on the CFMT-inverted version of the task (58.4%) compared to the CFMT upright

version (80.4%) (Duchaine & Nakayama, 2006) as does performance on other tasks thought to measure integrative face processing. Performance on the CFMT is significantly correlated with performance on a long-term memory face test Before They Were Famous ($r = .70$, $p < .001$) (Russell, Duchaine, & Nakayama, 2009) and Famous Faces Test ($r = .51$, $p < .001$) (Wilmer et al., 2010). In addition there is some evidence that performance on the CFMT is related to face processing (e.g., to performance on the Cambridge Face Perception Test ($r = .60$, $p < .001$) (Bowles et al., 2009). Because CFMT is a widely used measure of face recognition we have elected to use it in this experiment in order for our results to have the broadest application. In terms of anticipated results, we expect to replicate the pattern from the literature, which is to say that we expect to find low to moderate positive correlation coefficients between integrative face processing and face recognition ability.

The task methodology for CFMT is described in detail in the Methods section. Briefly described here, the CFMT requires participants to study a number of faces at three viewing angles: straight on, 1/3 left profile, and 1/3 right profile. Trials require participants to identify which of three simultaneously presented faces has the same identity as one of the faces they learned during the study phase. Task manipulations include changes to viewing angle, illumination, and level of Gaussian noise.

The Importance of the Inverted Task Conditions

Because we are endeavoring to measure holistic processing and/or configural processing, we followed McKone et al.'s (2013) recommendations regarding the importance of inverted conditions. Specifically, we included inverted conditions in three of the four integrative face-processing tasks (i.e., all but the *Complete Composite Face Effect Task*). McKone et al. (2013)'s argument is that a holistic effect in the upright condition should only be interpreted as such if this

same effect is reduced in the inverted condition. By including inverted conditions in three of the four tasks we are able to investigate how face inversion impacts performance across these paradigms. Inversion is not commonly included in the *Complete Composite Face Effect Task* paradigm, where congruency variations take its place. We therefore did not include such a manipulation in this task.

Based on past literature, we expect that the conditions most susceptible to disruption of integrative processing via inversion will be: 1) the "Aligned" condition of the "partial" *Composite Face Effect Task, 2)* the "Configural" condition of the *Configural/Featural Difference Detection Task,* and 3) the "Whole" condition of the *Part Whole Effect Task*.

A Special Consideration: Reliability Scores

In the present study we include the most commonly used face processing tasks, which have traditionally been used to investigate group level effects. However, we also calculate and analyze individual difference scores from performance on these tasks. According to Sunday, Richler, and Gauthier (2017) only the *complete Composite Face Effect Task* has been adapted for the study of individual differences (meaning that task condition reliability is adequately high). Because we endeavor to analyze individual differences on tasks that were not designed to do so specifically, we must pay particular attention to reliability scores.

When conducting analyses on group level effects reliability is an important consideration. In general if performance within a given task is not internally consistent or reliable, then we cannot expect its performance to be significantly correlated with another task. However, reliability is an even more important consideration in the case of individual difference research because the reliability of an individual difference score depends on the reliability of each of the two individual task conditions that compose it, as well as the correlation between the two

conditions. If one or both of the task conditions that make up an individual difference score does not have adequate reliability this will lower the reliability of the difference score in and of itself as well as effect the correlation between the two scores. Both of these aspects will result in the reliability for the difference score being low as well. This will then impact all subsequent analyses.

In the case of these tasks, which are designed to measure group level effects, the reliability of the individual task conditions may not be adequately high for individual difference research. We are aware that the tasks we have used here may generate unreliable difference scores, which could lead to a lack of significant correlations and contribute to low factor loadings or failures to load. It is for this reason that we also analyze group level effects. Some authors (e.g. DeGutis et al., 2013) address this low reliability issue by reporting scaled-up correlation coefficients. Nelson et al. (in preparation) reported such correlations. While this approach increases the magnitude of correlation coefficients, the same patterns of correlation magnitudes between tasks are maintained. Moreover, it is possible to obtain correlation coefficients above 1.0 using this technique. For these reasons we have elected to report the unmodified (or "attenuated") correlations only in this manuscript.

Hypotheses and Inter-task Analyses

As previously stated, the purpose of this experiment is to provide data that will help to address three fundamental questions in face recognition research: 1) Are there distinct holistic and configural processing mechanisms in face recognition? 2) If so, to what degree are they predictive of face recognition ability? 3) Which tasks are best suited to evaluate integrative processing?

If the construct indexed by any or all of the four integrative face-processing tasks were related to face recognition ability (as measured by performance on the CFMT) then we would expect to see significant correlations between the outputs of the integrative tasks and accuracy on the CFMT. In addition, we would expect that a factor analysis to show the integrative tasks loading onto the same factor as the CFMT. Similarly, if the tasks are measuring the same construct, then performance values arising from them should load onto the same factor in a factor analysis. Conversely, if the tasks are measuring different constructs, then we would expect behavioural measures from the tasks to load onto separate factor(s).

Simply put, if instruments (i.e. face processing tasks) are measuring the same thing, then we expect their outputs to correlate strongly. By that logic, if the various face-processing tasks we are examining here are all measuring integrative face processing, then we expect their outputs to correlate strongly. Conversely, to the degree that they are measuring different aspects of integrative processing we would expect performance to be only weakly correlated between tasks, with stronger correlations across conditions within each task.

The two statistical methods we employ to investigate these questions are correlational analysis and factor analysis. With regard to the correlations between performance on these integrative face-processing tasks and performance on the CFMT, we expect to see small to moderate magnitudes in our correlational analyses. From the correlational analyses we can also compare the magnitudes of within-task and between-task correlation coefficients. This will indicate if performance is more strongly related on conditions within each task than it is between tasks, which would suggest that the tasks are measuring something unique. Based on previous research (Nelson et al., in preparation) we hypothesize that we will replicate the same pattern of performance that they observed. Specifically, there will be stronger within-task correlation

magnitudes than between task magnitudes. We anticipate individual difference scores to replicate these patterns both between integrative face-processing tasks and in terms of their respective association with performance on the CFMT as well.

Results from the factor analyses will indicate how performance levels on task conditions, and individual difference scores, load onto a set of factors. If many of the tasks load onto a single factor, this would suggest that there is a single integrative face processing mechanism. Alternatively, if the tasks load onto two factors, this may indicate separate configural and holistic mechanisms at play. However, because we expect to see strong within-task correlations and weaker between-task correlations we also expect to see task conditions load onto four factors, one for each task. This finding would replicate the findings from Nelson et al. (in preparation).

The way in which general face recognition, as indexed by the CFMT, will relate to the other factors is difficult to predict. Previous research has suggested that there is only a weak correlation between integrative face processing measures and CFMT results (e.g., Richler et al., 2014; 2015; Konar et al., 2010; Rezlescu et al., 2017), so it is possible that we will fail to see CFMT load onto any factors. Another possibility is that it will load with all of the integrative face processing tasks under a common single factor. Yet another outcome that might be anticipated based on the literature is that we will see CFMT cross-load onto separate configural and holistic factors because integrative processing is thought to play a role in face recognition (Richler & Gauthier 2014; 2015; Rossion, 2013) and configural and holistic processing may be separate mechanisms. Finally, because the CFMT shows upright faces we might expect to see stronger correlations with upright task conditions. And because the "partial" and "complete" *Composite Face Effect Tasks*, as well as the *Part-Whole Effect Task,* are thought to be measures of holistic processing, we expect to see stronger correlations on the upright trials of these tasks

compared to the upright conditions of the *Configural Featural Difference Detection Task* with performance on the CFMT.

There is, however, a range of possible results. For example, if all four of the face-processing tasks primarily measure a common construct (e.g., they all tap into a combined holistic/configural mechanism) then we would expect to see strong correlations between all task conditions, or pairs of conditions (i.e., difference scores), across all four tasks. Furthermore, we would expect to see performance on all task conditions, and difference scores, loading onto a single common factor. Alternatively, these four tasks may each be measuring the same construct(s) but to varying degrees. If this were the case then we would expect to see a wide range of correlation coefficients between conditions, and difference scores, across tasks in terms of magnitude and also direction. Consequently, in the factor analysis within this scenario we would expect to see item cross loadings with task conditions, or pairs of conditions loading onto multiple factors to varying degrees.

If however, these four face-processing tasks were tapping into various processing mechanisms then we would expect task performance to load onto two or three factors. These factors could be conceptualized as featural, configural, and/or holistic processing. Performance on each task condition (and difference score) would be strongly correlated with other task conditions (and difference scores) that are measuring the same construct, and weakly correlated with task conditions (and difference scores) measuring a different construct. And these task conditions (and difference scores) would load onto separate factors accordingly. If the patterns that emerge from our data analyses suggest that these four face-processing tasks are measuring different constructs, then this may account for the discrepancies in results across studies regarding the correlations between performance on the various face processing tasks.

Finally, both statistical analyses, correlation and factor analysis, will help to elucidate the association between face processing and face recognition ability by demonstrating if and how performance on these processing tasks is associated with performance on the CFMT. If these tasks were measuring similar aspects of face processing then we would expect these tasks to have similar associations with face recognition ability (as evidenced by comparable correlations). In terms of factor analysis results, we would expect all related conditions to load together onto a common factor. If these tasks however are measuring different aspects of face processing, and these aspects are differentially related to face recognition ability, then we would expect to see significant correlations between task conditions (and difference scores) and the CFMT accordingly. With respect to the factor analysis results in this case we would expect the CFMT to load with the task conditions (and difference scores) to which performance is most closely related in the rotated factor solution, and perhaps CFMT doing the same, or cross loading in the unrotated factor solution. As previously stated, regardless of the obtained results, our data will be informative to various debates that currently exist in the literature.

Method

The information regarding the participants is the same for each of the five tasks that are included in the study. The information pertaining to the materials is the same for four of the five tasks. Therefore, we present this common information first, followed by the materials information for the CFMT task. In the section that follows, we provide the details about task stimuli and experimental procedure for each task. The task order was randomized across participants. So too was the order in which the conditions within each task were presented. It took participants approximately one hour and forty-five minutes to complete the five tasks.

Participants

Undergraduate students from the University of Ottawa ($N = 260$) participated in this study, which was approved by the University of Ottawa Social Sciences and Humanities Research Ethics Board. Participants received course credit as compensation for their participation. The initial sample size was $N = 280$; however, because the experiment used a within-subjects design we deleted participants listwise. Reasons for listwise deletion included failing to complete one or more of the five tasks, and/or using an incorrect keyboard key to enter responses on one or more of the tasks.

In order to ensure that our statistical analyses had adequate power we used G-Power software (Faul, Erdfelder, Lang, & Buchner, 2007) and ran post hoc tests for the analyses we conducted. For the ANOVA analyses, we chose the following parameters: effect size $f = .10$, $\alpha = .05$, one group, sample size was $N = 260$, number of measures was four and the correlation among repeated measures was .28. Because our data met the assumption of sphericity (this is discussed in the results section below), the nonsphericity correction of 1 was applied. Our obtained level of power ($1 - \beta$ error probability) was .90.

We also used G-POWER software to determine that we obtained an adequate level of power in our correlation analyses for performance scores (accuracy and reaction time) as well as for both sets of difference scores (subtraction and regression). We acknowledge, as will be described below, that our data were not normally distributed, and as such, we elected to use Spearman's non-parametric correlation. Unfortunately, G-POWER lacks non-parametric options to assess power level adequacy so we elected to use the next best option available, which was Correlation: Bivariate normal model. We used this model based on the following assumptions: two tailed test, a correlation ρ for H1 of .28 (which was our obtained average correlation

coefficient across task conditions), $\alpha = .05$, a sample size of $N = 260$, the correlation ρ for H 0 of .50 (representing that performance on these tasks are moderately to strongly correlated across tasks). From this calculation we determined that our obtained level of power was .98. For our subtraction and regression-based difference scores we used the same parameters except for the correlation ρ for H1, which were .01 and .18 respectively, corresponding to an obtained level of power of 1.0 and .99 respectively.

Regarding sample size adequacy for exploratory factor analysis, we followed the best practices outlined in Costello and Osborne (2005). The rule of thumb is a subject to item ratio of 10:1. Our subject to item ratio is 260:17 which equates to roughly 15:1.

Materials

The testing sessions took place at the University of Ottawa on a series of identical Dell Optiplex 9010 computers with an i5 Intel core, running Windows 7 on 17-inch monitors with a screen resolution of 1920 × 1080 and refresh rate of 60 hertz. The face stimuli for the *Complete Composite Effect Task*, the *Partial Composite Effect Task*, the *Configural/Featural Difference Detection Task*, and the *Part Whole Effect Task* were from the Matheson-McMullen Face Database (Matheson & McMullen, 2011). Using Photoshop and Matlab 2010a (Mathworks Inc., Natick, MA; http://www.mathworks.com) psychophysics toolbox, all stimuli were equated for mean luminance and RMS contrast, and then converted to grey-scale using Matlab's rgb2gray function for continuity between tasks and experiments. All face stimuli were male, and were scaled to 9 × 9 cm (9° × 9° at a viewing distance of 57 cm). In all tasks, stimuli appeared in the frontal (0°) view.

Face stimuli for the CFMT task were those included in the standard version of the task (Duchaine & Nakayama, 2006). They are faces of males in their 20s and early 30s. In this set of

stimuli, the original authors removed hair as well as any blemishes to avoid facilitating recognition judgments based on non-face components of the images. Stimuli were six target faces and there were 46 distracter faces (with many faces repeated throughout the task). The six males selected to be target face images were photographed in 12 poses consisting of various viewing angles and lighting conditions, and always with a neutral expression on his face. The viewing angles and lighting conditions were held constant across the target and distractor face stimuli within each experimental condition. Stimulus dimensions were not available in the published literature; however, by measuring them as they appeared on the testing computer monitors, we determined that they were approximately 5 cm in height and 3 cm in width. Participants sitting at a viewing distance of approximately 57 cm corresponds to a viewing angle of 5.0° × 3.0° for the CFMT stimuli.

Task 1– Complete Composite Face Effect Task

Stimuli. In the Complete Composite Face Effect task (Farah et al., 1998; Gauthier et al., 1998; Richler, et al., 2011b), we used 20 base faces to create 380 different composite face stimuli. We created the composite faces by taking the top half of one face and combining it with the bottom half of a different face. That is, all faces presented in this task were a hybrid of two faces. Top halves of the test and inspection faces either matched, or were different, and (unlike in the partial version of the task) the bottom halves of each face either matched or did not match.

Procedure. The task began with instructions, followed by eight practice trials, two for each condition. Before the experimental phase began, the instructions appeared on the screen again, at which point the participant had the opportunity to ask for clarification. Orally, we reminded participants to selectively attend to the top half of faces (middle of nose to forehead).

The *Complete Composite Effect Task* involved a 2-Alternative Forced Choice Sequential Matching paradigm with a 2 × 2 design. The factors were Alignment (aligned or misaligned), and Congruency (congruent or incongruent). Note that we did not examine the effects of orientation in this task; This is because we wished to use a task method similar to that which is commonly used in the literature (e.g., Richler et al., 2011b; Richler et al., 2015; Richler, Mack, Gauthier, Palmeri, 2009). Aligned and Misaligned refer to the alignment of the face halves. In the aligned condition, the face halves are lined up so that they appear to combine into a single face; In the misaligned condition, the top and bottom halves are offset horizontally, thus breaking the face into two obvious pieces, top and bottom. Congruency refers to whether the state of equivalence is the same for the top and bottom halves of the test and inspection face stimuli. In congruent trials both top and bottom halves of the test face either 1) match those of the inspection face (same) or 2) both halves of the test face do not match those of the inspection face (different). In incongruent trials, either 1) only the top half of the inspection face and test face match (same), or 2) only the task irrelevant bottom halves match (different). On "same" trials the top halves of inspection and test faces match, and on "different" trials the top halves do not match. For a diagram explaining the face stimuli in each of the experimental conditions please refer to Figure 1 in Richler et al., 2011b.

Trials from the four different conditions were presented in random order. The experiment consisted of an equal number of trials in each of the four task conditions. There were 512 trials, and the task took approximately 35 minutes to complete.

The experimental phase began with one composite face (inspection face) presented for .5 seconds. This was followed by an inter-stimulus interval of .3 seconds, after which a second composite face (test face) was presented. Participants were required to judge if the top halves of

the inspection and test faces were the same or if they were different. Participants entered their response using one of two keyboard keys. The test face remained on the screen until the subject responded. Both accuracy and reaction time were recorded. For examples of aligned and misaligned congruent and incongruent trial structures that are unique to the *Complete* version of this task, see Figure 1.

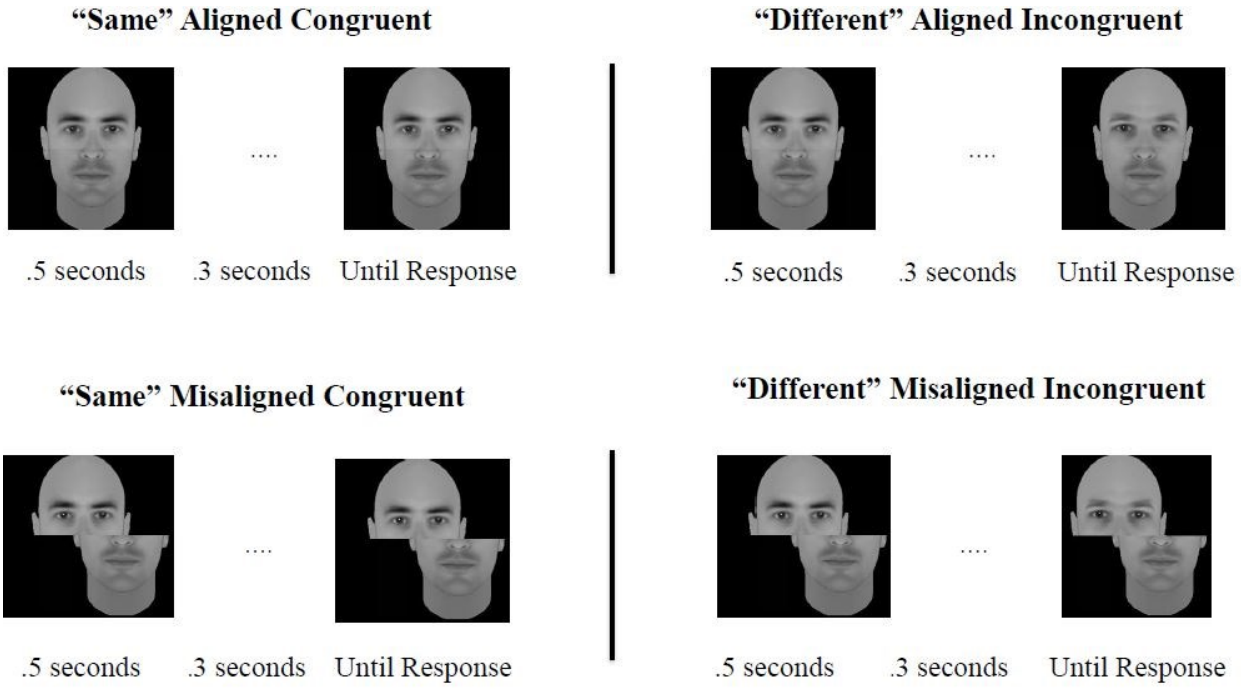

Figure 1. This figure depicts the trial structure of congruent "same" and incongruent "different" trials for both aligned and misaligned stimuli in the Complete Composite Face Effect Task. These are the "other half of the trials" that are not part of the Partial version.

Task 2 – Partial Composite Face Effect Task

Stimuli. In the *Partial Composite Face Effect Task* (Young et al., 1987), there were 120 different composite face stimuli that were presented over the course of 240 trials (60 different test faces and 60 different inspection faces each repeated four times). Like the stimuli in the *Complete Composite Face Effect Task*, the hybrid face stimuli were made from the top half of

one face and the bottom half of a different face. This way all face stimuli looked equally "composite". The face halves were either aligned to create a whole face or were misaligned horizontally to create a face that was obviously broken into two separate halves, top and bottom. This was accomplished by shifting the top halves of the faces so that the right edge of the bottom face halves and the middle of the nose from the top face halves align. Face stimuli were either presented upright or inverted. Unlike in the *Complete Composite Face Effect Task*, the bottom halves of the face stimuli in this experiment were always different. That is, there were only congruent different trials and incongruent same trials in the *Partial Composite Face Effect Task*.

Procedure. The task began with instructions, followed by eight practice trials. Before the experimental phase began participants viewed the task instructions again at which point he or she could ask for clarification. The researcher reiterated, orally, that the participant was to only pay attention to the top half of the faces (middle of the nose to the forehead regardless of orientation) when making his or her judgments. When the participant was ready the experimental phase began.

The *Partial Composite Face Effect Task* is a 2 Alternative Forced Choice Sequential Matching paradigm with a 2 × 2 design. One factor was Alignment (aligned or misaligned), and the other was Orientation (upright or inverted). Alignment refers to the alignment of the face halves as was previously discussed. Orientation refers to whether the faces were both presented upright or inverted. The experiment consisted of 120 aligned trials and 120 misaligned trials presented in a random order. Upright and inverted trials were blocked in batches of 30 trials. In total there were 240 trials in this experimental task and it took approximately 20 minutes to complete.

The experimental trial structure was as follows: the inspection face was presented and remained on the screen for .5 seconds, next there was an inter-stimulus interval of .3 seconds, followed by the test face, which remained on the screen until the participant responded. Depending on the orientation condition, face stimuli were either all upright or all inverted. Top halves of the inspection and test faces either matched or were different, and the bottom halves of the faces were always different. Participants were required to indicate whether they thought the top halves of the faces (middle of the nose to forehead regardless of orientation) matched or if they were different. They indicated their judgment by pressing one of two keyboard keys. Both accuracy and reaction time were recorded. For examples of aligned and misaligned trial structure with the same top halves (and different bottom halves), see Figure 2.

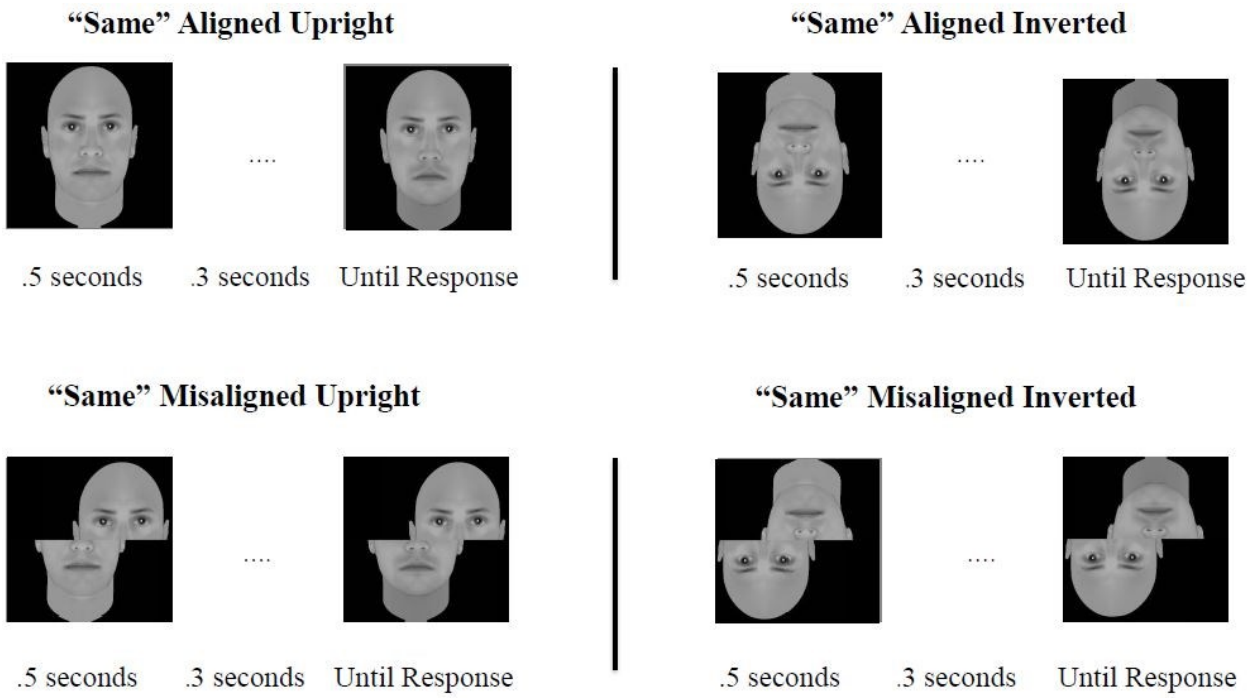

Figure 2. This figure depicts the stimuli, presentation times, and the interstimulus interval for upright and inverted trials in the Composite Face Effect Task.

Task 3 – Configural/Featural Difference Detection Task

Stimuli. First, we created four base faces so that all face stimuli in this task would look equally modified. We created a base face by taking one face stimulus and removing its eyes, nose, and mouth. Then we replaced each feature with the eyes from Face A, the nose from Face B, and the Mouth from Face C and inserting them onto the base face to create Base Face 1. The face stimuli from which we extracted facial features were not use elsewhere in this experiment. Half the face stimuli had a featural manipulation. These stimuli were modified by swapping out a face part with the same face part (feature) from a different face. This was done three ways, to the eyes, to the nose, and to the mouth. These swapped face features were new features, meaning they were not presented elsewhere in the experimental conditions either alone, or as part of a face. Configural manipulation meant moving face features (eyes, nose, or mouth) up or down three millimeters within the face. This distance is enough to modify the second-order relational information in a distinguishable way without overly distorting the appearance of the face. There were nine modified versions of four base faces, plus the four base faces, totaling 40 face stimuli in this task.

Procedure. Participants began the task with a set of task instructions followed by eight practice trials (2 practice trials for each task condition). After the practice session, and before the experiment began, participants viewed the task instructions again, at which point they could ask for clarification. The task was a Simultaneous Matching paradigm with a 2 × 2 design. One factor was Task Manipulation (configural or featural), and the other was Orientation (upright or inverted). Half the trials involved a configural manipulation and the other half involved a featural manipulation, and the presentation order was randomized. Orientation was blocked into batches of 48 trials. There were 96 experimental trials in total, and the task took approximately 10 minutes to complete.

The experimental phase consisted of two faces presented simultaneously for 3.5 seconds, at which point the faces were replaced with two masking squares. Participants indicated, by pushing one of two keyboard keys, if the two faces were the same or if they were different. Both accuracy and reaction time were recorded. For an example of trial structure for configural and featural trials please refer to Figure 3.

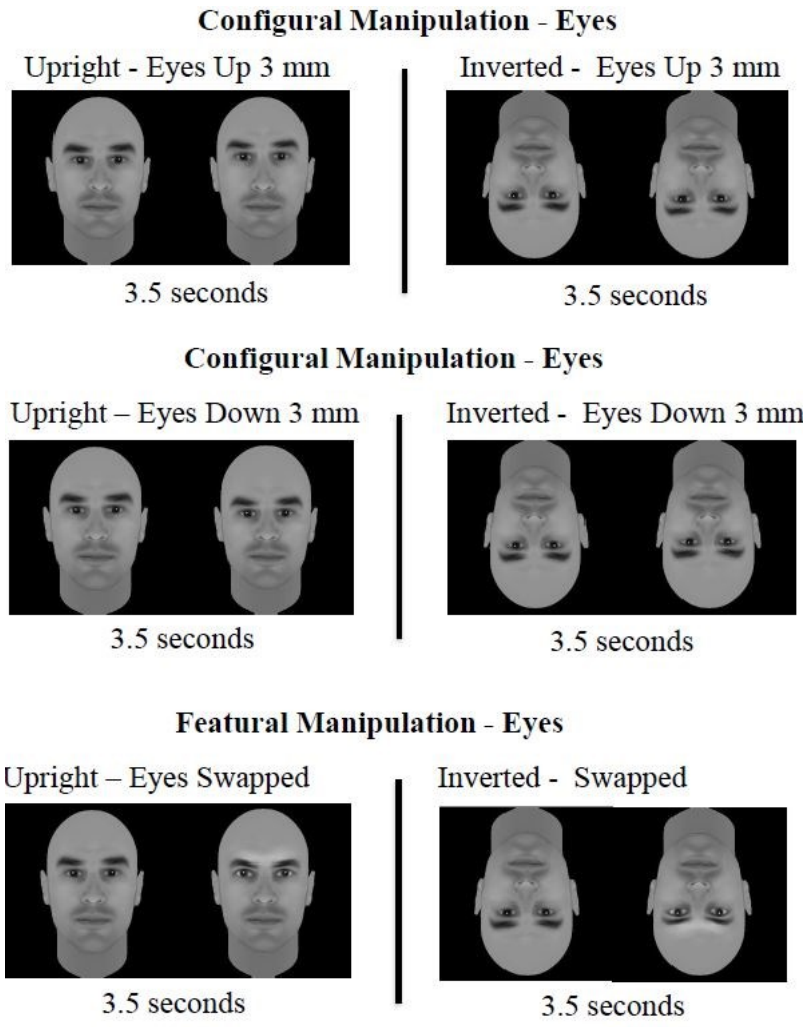

Figure 3. This figure depicts the stimuli, presentation times, and the interstimulus interval for upright and inverted trials in the Configural/Featural Difference Detection Task that pertain to eye manipulations.

Task 4 – Part Whole Effect

Stimuli. In the *Part Whole Effect Task* (Tanaka & Farah, 1993) stimuli are whole faces or isolated facial features (eyes, nose, or mouth). When creating the stimuli for this task we first made 10 base faces that all look equally modified. To do so, we took a face image then replaced the eyes with eyes from Face A, the nose with the nose from Face B, and the mouth with the mouth from Face C. Faces A, B, and C were not used elsewhere in the experiment. Next we modified these 10 base faces to create the rest of the stimuli. We then repeated the process described above, but swapping out only one feature at a time (eyes, nose, or mouth) with that feature from yet another face. Thus, three modified version of each base face were created: Base Face 1 with swapped eyes, Base Face 1 with swapped nose, and Base Face 1 with swapped mouth.

The stimuli in the Whole condition of the task were the base faces and the modified versions of the base faces. The Part condition stimuli consisted of the features that were used to create the base faces as well as the features that were used in the face modifications described above. In all, the stimuli in this task included 10 base faces, 30 modified faces, 30 parts used to make the base faces, and 30 parts from different faces to create the modified faces.

Procedure. Participants were first shown task instructions then they completed four practice trials. The task instructions were then presented a second time at which point the participants could ask questions if they were unclear about the procedure. Next, the experimental phase began. The *Part Whole Effect Task* consisted of a 2 Alternative Forced Choice Sequential Matching paradigm with a 2 × 2 design. The factors were Condition (part or whole) and Orientation (upright or inverted). Half of the trials were Part trials, and the other Whole trials. These were presented in random order. Orientation however was blocked into

batches of 60. The order of the orientation blocks was counterbalanced across participants. In all, there were 120 trials in this task and participants took approximately 10 minutes to complete it.

The trial structure for the experimental phase was as follows (and can be seen visually below in Figure 4). An inspection face (of which there are 10) appears and remains on the screen for 1.75 seconds. Next, an inter-stimulus interval lasts for .4 seconds. Following that, two test faces (or two test parts e.g. two pairs or eyes, two noses, or two mouths) appear side-by-side simultaneously. Feature labels appear above the two test faces or two test parts to cue the participant toward which feature he or she should attend and subsequently on which he or she should base the "match" decision. While it is obvious in the part condition to which feature one must attend, the label appears in both task conditions to compensate for any possible distracting effects it may have.

During Whole trials, participants must identify which one of the two test faces has the eyes (or nose, or mouth depending on the cue label) that match the eyes that were part of the inspection face they saw previously; In the Part trials, the participants identify which feature is the one that matches that same feature that was part of the inspection face. The test faces or face parts remained visible until a response was given. Again, participants entered their responses by pressing one of two keyboard keys. Both accuracy and reaction time were recorded automatically.

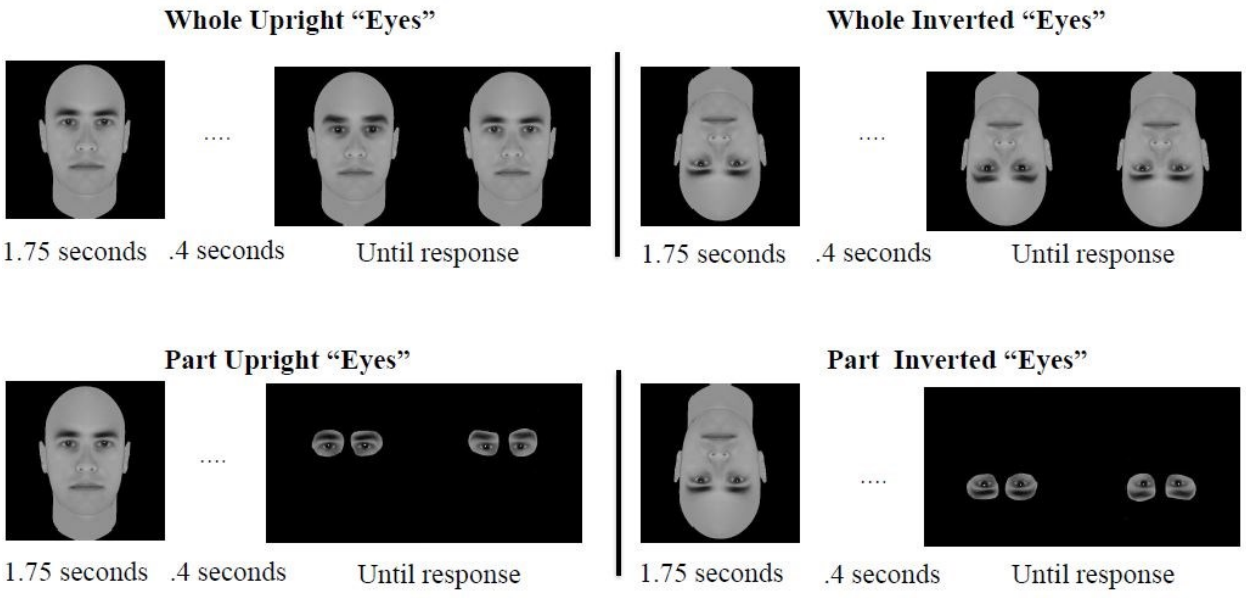

Figure 4. This figure depicts the stimuli, presentation times, and the interstimulus interval for upright and inverted trials in the Part Whole Effect Task.

Task 5 – Cambridge Face Memory Test

Procedure. Practice trials presented cartoon images in what is called the "same image stage", wherein the inspection images and the test images are in the same angular position (either left 1/3 profile, front view, or right 1/3 profile view). Bart Simpson is shown in these three views, each for three seconds, followed by a line-up of three images in the same angled position, one of which is Bart Simpson and two of which are distracter images of other Simpson's characters. The participants' task was to push a keyboard key (1, 2, or 3) to indicate which of the three test images matched the inspection image (i.e., Bart).

After the practice phase came the "same image stage" of the experiment, which consisted of 18 trials. This experimental condition has a similar structure and parameters to the practice trials, except that the inspection stimuli are six real human male target images. The procedure was as described above, and repeated until all six target faces had been presented. Once this

phase was completed, participants were given a refresher, wherein the frontal view of each of the six target faces was presented to them again for 20 seconds.

Following the refresher phase, the second experimental condition began, “novel images”, which consisted of 30 3-alternative forced choice trials. Each of the six target faces was presented five times. The stimuli were presented in a fixed random order. Each test item contained one target face and two distracter face images. However, the “novel images” condition means that the lighting, or the angle of the pose, or both, can be varied between the inspection and the test images. Participants were tasked with recognizing which of the three test faces has the same identity as one of the six target faces.

After this phase participants were given another refresher wherein the six target faces were presented in frontal view for 20 seconds. Lastly, the “novel images with noise” condition began. This condition consisted of 24 3-alternative forced choice trials during which six target images were presented four times each, again in a fixed random order. In addition to the images being novel in terms of pose and/or lighting, as described above in the previous condition, in this condition there were also different levels of Gaussian luminance noise added to the images. Within each of the three test images within a given trial the level of Gaussian noise was the same. This task took between 10-15 minutes to complete. To view experimental stimuli from each of the task conditions please refer to Duchaine and Nakayama (2006) p. 578.

Results

Dependent Variables

As previously mentioned, accuracy data are presented within the manuscript. The same analyses were conducted on reaction time data and these results are presented in Appendix B. In addition to group-level performance data we also reported individual difference scores (both

subtraction and regression-based). Individual difference scores for accuracy are presented within the manuscript and the parallel reaction time analyses are presented in Appendix B. Individual difference scores were calculated across orientation as well as across task manipulation. The "across orientation" difference data are reported within the manuscript. This is because inversion is thought to impact holistic/configural processing; but it remains to be determined what each task condition is actually measuring. The "across task manipulation" difference data are presented in Appendix B. For the most part, patterns were comparable across reaction time and accuracy data. In the case of meaningful differences between accuracy and reaction time this is noted in the manuscript.

Data Cleaning

We analyzed accuracy and reaction time data ($N = 260$) from all five tasks and used z-scores to identify outliers. A score was considered to be an outlier if it was 2.54 standard deviations above or below the mean. These cut off z-score values correspond to an α of .01. Scores that were identified as outliers according to this method were "tucked in" also referred to as winsorizing (Huber, 1981; Hogg, 1979). In this procedure the closest score within $z = +/-$ 2.54 standard deviations of the mean is used to replace the outlying values. Outlier analysis of accuracy scores revealed 101 out of 4437 performance scores to be outliers while reaction time outlier analysis revealed 114 outlying scores. This equates to 215 outliers out of 8874 scores, which is to say 2.4% of all performance scores were tucked in.

To identify multivariate outliers in accuracy and reaction time data we calculated Mahalanobis distances and compared these values to the critical value from the chi square distribution with 17 degrees of freedom (because there are 17 task conditions) at $p < .001$. Scores are considered to be multivariate outliers if the mahalanobis distance score is above the

critical chi square value, which in this case is 40.79. Two participants were multivariate outliers in accuracy data and nine were multivariate outliers in the reaction time data. Rather than transforming our data, which can lead to difficulties in interpretation (Feng, Wang, Lu et al., 2014) we opted to use non-parametric statistical procedures when necessary, and used principle axis factoring in our factor analyses.

Data Normality

In order to determine if the accuracy data were normally distributed we looked at the skewness and kurtosis values and their respective standard error values for each of the task conditions. Data in each task condition were considered to be skewed if the skewness value was greater than twice the standard error of the skew. The same logic applied to kurtosis values. Additionally, we visually inspected the accuracy data within each task condition via histograms with fitted normal curves and also investigated the corresponding QQ-plots.

There was converging evidence that data in all four conditions of the *Complete Composite Face Effect Task* and *Partial Composite Face Effect Task* were left (negatively) skewed. Within the *Part Whole Face Effect Task*, Whole Upright data were negatively (left) skewed. Regarding kurtosis, data from the Aligned Congruent, Misaligned Congruent, and Misaligned Incongruent conditions of the *Complete Composite Face Effect Task* were platykurtic. The Misaligned Upright condition of the *Partial Composite Face Effect Task* was also platykurtic. Because not all task conditions showed severe univariate normality violations, and because data transformations can complicate data interpretations we instead decided to use non-parametric statistical procedures and make other appropriate decisions throughout our analyses.

Next we used visual inspection to investigate multivariate normality. We did so by using SPSS to generate a Q-Q plot that plotted the Mahalanobis distance values against estimated chi-square quintiles. Upon visual inspection we determined that accuracy data met the assumption of multivariate normality while reaction time data did not.

Analysis of Variance

For all tasks except for the CFMT we conducted 2 × 2 repeated measures ANOVAs to demonstrate that we have replicated the expected interaction effects thought to reflect integrative processing. Before doing so, first, we determined that the standardized residuals for each of the task conditions were normally distributed (or at least were near-normally distributed). We did this by visually inspecting a histogram with a superimposed normal distribution curve as well as by inspecting Normal Q-Q plots for residual scores from each of the task conditions. Because there are only two levels of repeated measures, sphericity is not a concern. For the *Partial Composite Face Effect Task*, the *Configural/Featural Difference Detection Task*, and the *Part Whole Effect Task* the independent variables were Orientation (Upright or Inverted) and Task Manipulation (which varied by task). For the *Complete Composite Face Effect Task* the independent variables were Congruency (Congruent or Incongruent) and Alignment (Aligned and Misaligned). Interactions were all significant and in the expected directions. The results and interpretations of these repeated measures ANOVAs can be found in Appendix B.[12] With

[12] The error bars presented in Figures B1 to B8 of Appendix B are correlation-adjusted 95% confidence intervals around the mean values in within-subjects designs (O'Brien & Cousineau, 2014). Effect sizes for the main effects and interaction terms were reported as η_p^2 along with their 95% confidence intervals. These 95% confidence intervals around η_p^2 were calculated following the procedure outlined in Richler & Gauthier (2014).

respect to reaction time data, the interaction in the *Configural/Featural Difference Detection Task* was not significant, but all others were.

Task Reliability

We used Guttman's λ_2 to measure task reliability for all five tasks' group-level performance data. We did this because it is an appropriate measure to use when tasks are made up of multiple factors (Callender & Osburn, 1979) as we suspect these tasks may be because each task contains more than one type of trial. Importantly, it has been shown that Guttman's λ_2 and Chronbach's α reliability measures are comparable for the CFMT (DeGutis et al., 2013). For four of the tasks (CFMT was the exception) we entered their four task conditions into the task reliability calculation. For example, to calculate the reliability of the *Complete Composite Face Effect Task* we included: Congruent Aligned, Congruent Misaligned, Incongruent Aligned, and Incongruent Misaligned trials. For the CFMT we entered the summary scores from the three task conditions into the calculation: same image, novel image, novel image with noise. All five tasks showed good reliability. Guttman's λ_2 reliability scores per task were as follows: *Complete Composite Face Effect Task* (.96), *Partial Composite Face Effect Task* (.87), *Configural/Featural Difference Detection Task* (.77), *Part Whole Effect Task* (.74), and CFMT was (.88)

Correlation Analysis

We conducted a correlational analysis because we are interested in how accuracy scores are related across the five tasks. Recall that our accuracy data showed skewness and kurtosis in some of the task conditions. To check homoscedasticity we obtained variance statistics for each task condition and calculated the ratio of the largest and the smallest statistics. Accuracy scores were heteroscedastic because the obtained variance ratio was greater than 1.5. Therefore, after

establishing that the data met the assumption for a monotonic relationship by visually inspecting Q-Q plots, we elected to conduct the non-parametric Spearman's Rank Order Correlational analysis. Figure 5 depicts the correlational analysis of accuracy scores.

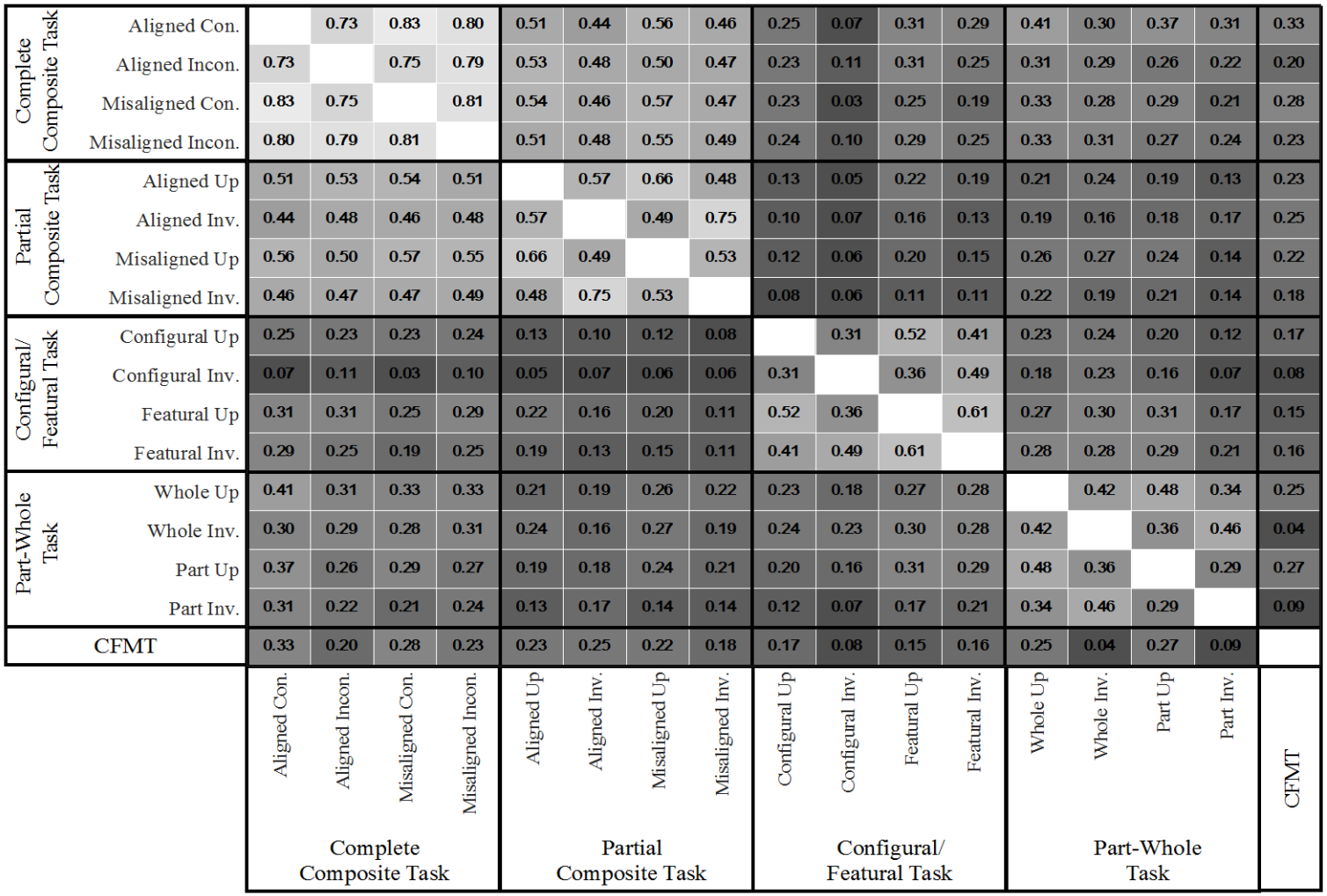

Figure Caption on Next Page

Figure 5. Non-parametric correlations of accuracy scores for all tasks. r_s values of .122 - .159 correspond to $p < .05$, two-tailed. r_s values between .16 - .173 correspond to $p < .01$, two-tailed. N = 260. Complete Composite Face Effect Task (Complete Composite Task), Partial Composite Face Effect Task (Partial Composite Task), Configural/Featural Difference Detection Task (Configural Featural Task), Part Whole Effect Task (Part/Whole Task).

Correlational data revealed some interesting patterns. We began by investigating correlations between all task conditions from the four face processing tasks, and found that most were significantly correlated with each other (mean correlation coefficient $r_s = .31$, $p < .001$, 95% CI [.20, .42], range $r_s = .05$ to $r_s = .83$). However, task conditions were more strongly correlated within each task (mean correlation coefficient $r_s = .55$, $p < .001$, 95% CI [.46, .63], range $r_s = .29$ to $r_s = .83$) than they were between tasks (mean correlation coefficient $r_s = .26$, $p < .001$, 95% CI [.14, .37], range $r_s = .05$ to $r_s = .56$). The average within-task correlation coefficient is much stronger than the average between-task correlation coefficient, $r_s = .55$ and $r_s = .26$ respectively. This suggests that while there is some commonality between what these tasks (and their respective four conditions) are measuring, these tasks are also measuring separate constructs or components thereof. This evidence supports the notion that these tasks should not be used as interchangeable measures of face processing, and that results from previous studies do not generalize beyond the specific task(s) that were used.

Next, we investigated how performance on these four face-processing tasks is related to facial recognition ability as measured by performance on the CFMT. The r_s values, the p values, and the 95% confidence intervals around the correlation coefficient are presented in Table 1 below. Performance on all task conditions within the *Complete* and *Partial Composite Face Effect Task* was significantly correlated with performance on the CFMT, as was performance on three of the four *Configural/Featural Difference Detection Task* conditions (Configural upright, Featural upright, and Featural inverted), and two task conditions within the *Part Whole Effect Task* (Whole upright and Part upright).

Table 1

Results from Spearman's Rank Order Correlation Analysis (Accuracy) Between Performance on Processing Task Conditions and the CFMT (N = 260)

Task Condition	r_s	*p*-value	95% CI
Aligned Congruent	.33	$p < .001$	[.22, .43]
Aligned Incongruent	.20	$p < .01$	[.08, .31]
Misaligned Congruent	.28	$p < .001$	[.16, .39]
Misaligned Incongruent	.23	$p < .001$	[.11, .34]
Aligned Upright	.23	$p < .001$	[.11, .34]
Aligned Inverted	.25	$p < .001$	[.13, .36]
Misaligned Upright	.22	$p < .001$	[.10, .33]
Misaligned Inverted	.18	$p < .01$	[.06, .30]
Configural Upright	.17	$p < .01$	[.05, .29]
Configural Inverted	.08	$p = .20$	[-.04, .20]
Featural Upright	.15	$p < .05$	[.03, .27]
Featural Inverted	.16	$p < .01$	[.04, .28]
Whole Upright	.25	$p < .001$	[.13, .36]
Whole Inverted	.04	$P = .52$	[-.08, .16]
Part Upright	.27	$p < .001$	[.15, .38]
Part Inverted	.09	$p = .15$	[-.03, .21]

The average correlation coefficient between accuracy performance on the CFMT and performance on all task conditions for which there was a significant correlation was $r_s = .22$, $p < .001$, 95% CI [.10, .33] range $r_s = .15$ to $r_s = .35$. Taken together, these results indicate that there is a small positive association between accuracy performance on the CFMT and 13 (of 16) face processing task conditions in this experiment.

Moreover, the correlational analysis provides evidence suggesting that the *Complete Composite Face Effect Task* may be the best task to use in the interim until it is empirically established what each task is actually measuring. This is because its across-task correlations are stronger in magnitude than are those of the other three tasks (average $r_s = .34$, $p < .001$, 95% CI [.23, .44] range $r_s = .10$ to $r_s = .57$; *Partial Composite Face Effect Task* average $r_s = .27$, $p < .001$, 95% CI [.15, .38] range $r_s = .05$ to $r_s = .56$; *Configural/Featural Difference Detection Task*

average $r_s = .18$, $p < .001$, 95% CI [.06, .30] range $r_s = .05$ to $r_s = .31$; *Part Whole Effect Task* average $r_s = .23$, $p < .001$, 95% CI [.11, .34] range $r_s = .07$ to $r_s = .41$). This suggests that the *Complete Composite Face Effect Task* is doing the best job of capturing the largest piece of the common mechanism (or component thereof) across these four tasks. While it remains ideal to include as many measures as is feasible in future research until researchers establish what each task is measuring, if one task must be selected this is evidence suggesting that the *Complete Composite Face Effect Task* be chosen.

The task conditions on which accuracy performance was not significantly correlated with performance on the CFMT were: 1) Inverted Configural within the *Configural/Featural Difference Detection Task*, 2) Inverted Whole within the *Part Whole Effect Task,* and 3) Inverted Part within the *Part Whole Effect Task*. Interestingly both inverted conditions within the *Partial Composite Face Effect Task*, and the inverted featural condition within the *Configural/Featural Difference Detection Task* did significantly correlate with performance on the CFMT. Because performance on some of the inverted task conditions were significantly correlated with performance on the CFMT and some other inverted task conditions were not, we can conclude that there is more to the small positive association between performance on these face-processing tasks and general facial recognition ability than stimulus orientation alone.

Recall that previous research suggested the significant correlations between the *Complete Composite Face Effect Task* and CFMT depended on the extent of stimulus repetition (Richler et al., 2015). Interestingly our *Partial Composite Face Effect Task* had very little stimulus repetition (60 face halves repeated 4 times each throughout the experiment) and yet our data still indicated that performance on its four conditions was significantly correlated with performance on the CFMT (average $r_s = .22$, $p < .001$, 95% CI [.10, .33], ranging from $r_s = .18$ to $r_s = .25$).

We can compare this association with that between the task conditions of the *Complete Composite Face Effect Task* and the CFMT (r_s = .26, p < .001, 95% CI [.14, .37] ranging from r_s = .20 to r_s = .33). From the present findings, we can conclude that the extent of stimulus repetition does not wholly account for the significant correlations between holistic face processing as measured by accuracy performance on the *Complete* and *Partial Composite Face Effect Tasks* and face recognition ability as measured by the CFMT. We investigated this same question via individual difference scores below in their respective sections.

Exploratory Factor Analysis

To further investigate how performance is related across these four face processing tasks and the CFMT, we also conducted an exploratory factor analysis. This analysis identifies how many latent variables the various tasks are measuring. Our data showed signs of multicolinearity, as determined by the obtained variance inflation factor values that were above 5.00. Variance inflation factor values were calculated by conducting iterative linear regressions. Multicolinearity was also evident from the results of our correlational analysis. Because our data are moderately to highly correlated within each task and weakly correlated between each task we elected to use a direct oblimin rotation (Osborne & Costello, 2009). Rotation cleans the factor structure and provides a simpler factor solution than the unrotated factor structure does because rotation eliminates cross-loadings. Both the unrotated and rotated factor structures are presented below in order to provide readers with the most complete picture regarding the associations between accuracy scores across these five commonly used tasks.

In terms of deciding how many factors to retain we followed the protocol outlined in Achim (2017)'s paper, which detailed the Next Eigenvalue Sufficiency Test (NEST). Many of our task conditions' data were near-normally distributed which allows us to be confident in the

accuracy of the NEST suggested factor solution. We opted to use the NESTip variant. This factor extraction technique is the one that retains the correct number of factors 76.6% of the time when all retained eigenvalues are above 1.0. NESTip is also very conservative and is the least likely of all NEST variants to overestimate the number of factors that should be retained (Achim, 2017). The NESTip model suggested retaining a five-factor solution with the critical eigenvalue above 1.0. Using SPSS, and forcing a five-factor solution, we elected to use principle axis factoring as our factor extraction method because our accuracy score data did not show multivariate normality (Costello & Osborne, 2005). Following Tabachnick and Fidell (2001)'s guidelines we have suppressed factor loadings below .32. This cutoff value corresponds to variances that overlap by 10% approximately.

Item communalities ranged between low and high, and in the case of low and moderate communalities they increased following extraction (refer to Table 2). The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was high, at .86, and Bartlett's test of sphericity was significant (χ^2 (136) = 2697.06, $p < .001$). The five-factor solution explained 60.75% of the total variance. Factor 1 explained 36.24% of the variance, Factor 2, 11.03%, Factor 3, 5.93%, Factor 4, 4.87%, and Factor 5, 2.68%. Tables 2 and 3 display the unrotated and rotated factor structures, respectively.

Table 2

Unrotated Factor Loadings and Communalities for Accuracy Scores Five-Factor Solution (N = 260)

Task		Factor					Communality	
		1	2	3	4	5	Initial	Ext
CCFE	Aligned Cong	.86		-.33			.85	.87
	Aligned Incong	.84					.80	.78
	Misaligned Cong	.87		-.34			.87	.91
	Misaligned Incong	.86		-.30			.85	.87
PCFE	Aligned Up	.71				-.38	.64	.74
	Aligned Inv	.61	-.46	.58			.74	.97
	Misaligned Up	.73				-.42	.65	.76
	Misaligned Inv	.62	-.40	.43			.74	.74
C/F	Configural Up	.33	.40				.29	.31
	Configural Inv		.44				.31	.33
	Featural Up	.46	.55		-.31		.49	.63
	Featural Inv	.44	.57				.49	.61
PW	Whole Up	.49			.41		.39	.49
	Whole Inv	.43	.35		.34		.39	.45
	Part Up	.46	.31		.32		.36	.42
	Part Inv	.35			.38		.29	.32
CFMT	CFMT	.32					.18	.10

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW), Cambridge Face Memory Test (CFMT). Congruent (Cong), Incongruent (Incong). Extraction (Ext)

Table 3

Pattern Matrix, Rotated Five-Factor Structure of Accuracy Scores (N = 260)

Task		Factor				
		1	2	3	4	5
CCFE	Aligned Cong	.85				
	Aligned Incong	.79				
	Misaligned Cong	.90				
	Misaligned Incong	.92				
PCFE	Aligned Up					-.79
	Aligned Inv			.99		
	Misaligned Up					-.81
	Misaligned Inv			.76		
C/F	Configural Up		.53			
	Configural Inv		.59			
	Featural Up		.75			
	Featural Inv		.73			
PW	Whole Up				.67	
	Whole Inv				.63	
	Part Up				.59	
	Part Inv				.60	
CFMT	CFMT					

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW), Cambridge Face Memory Test (CFMT).

Beginning with the unrotated five-factor solution, accuracy performance loaded onto a common first factor for 16 of the 17 task conditions, accounting for 36.24% of the variance explained by the model. This suggests that these tasks are measuring a common construct to a modest degree but also measure different constructs, albeit to lesser extents. This finding is in contrast with the results from the accuracy correlational analysis, which indicated that the magnitudes of correlations were stronger within tasks than they were between tasks suggesting that these tasks are more unique than they are similar. This discrepancy is important to note because it illustrates the different conclusions that can be drawn depending on the type of analysis that was run.

Interestingly, the latent variable represented by the first factor is not sensitive to differences in stimulus orientation because both upright and inverted task conditions load onto it positively with comparable magnitudes. Notably, orientation does seem to play a clear role in how performance is related to the latent variables represented by the second, third, and fourth factors in the accuracy unrotated factor structure. For example, we will draw the reader's attention to the third factor first because three of the four task conditions in the *Complete Composite Face Effect Task* cross loaded onto it as a collective and were negatively related to the latent variable that it measures. Recall that stimuli in this task are always presented upright. Interestingly, performance on the Inverted Aligned and Inverted Misaligned conditions of the *Partial Composite Face Effect Task* loaded onto this third factor with similar magnitude but they were positively related to this same latent variable. This finding suggests that stimulus orientation plays an important role with regard to this aspect of performance as captured by the third factor, whatever that may be. Labeling the factors would be necessarily speculative, and we are uncertain as to if processing mechanisms conceptually have opposite poles to them as differences in positive and negative factor loadings would suggest.

We noted that orientation also appeared to differentially impact performance as it is related to the latent variable that underlies the second factor. Factor loadings on the second factor showed that performance on both the Inverted Aligned and Inverted Misaligned conditions within the *Partial Composite Face Effect Task* was negatively related to this latent variable. Also loading onto this second factor was performance on all trials within the *Configural/Featural Difference Detection Task* with all tasks conditions being positively related to this latent variable. So too was performance on Whole Inverted and Part Upright. This suggests that inverted task conditions within these three tasks are measuring something beyond

the impact that inversion has on integrative processing providing additional evidence to suggest that these tasks are measuring different constructs (or likely different components of a construct (e.g., integrative processing)).

In the rotated five-factor solution data demonstrated that three of the four tasks loaded together onto their respective factors. The exception is the *Partial Composite Face Effect Task* in which inverted conditions loaded together and upright conditions loaded together onto their respective factors. This rotated solution simplifies interpretation by minimizing cross loadings, but it does not explain more variance than the unrotated solution, and we cannot conclude that the rotated factors explain the same percentage of variance explained as the unrotated factors do. With that in mind the pattern is clear that these tasks are differentially related to five latent variables suggesting that they are measuring different constructs, or different components of a common construct. Regardless, they are not interchangeable measures of face processing and should not be used or interpreted as such.

In the unrotated factor solution the CFMT loads positively onto the first common factor at the .32 cutoff. This is a notable difference between the unrotated and rotated factor solutions in which CFMT did not load with respect to accuracy data. CFMT did load however on the rotated factor solution for reaction time. This is the notable difference between the two performance measures with respect to exploratory factor analyses. In reaction time data CFMT loads positively onto the fifth factor with Whole Upright and Part Upright conditions. This finding is another example of why it is important to report both behavioural measures until research has determined how integrative processing is best captured on each of the tasks, and why it is important to report both rotated and unrotated factor solutions. While accuracy and reaction time data tell a similar story with respect to the associations between face processing

and face recognition in the unrotated factor solution, they tell a slightly different story with respect to this same association in the rotated factor solutions.

Difference Scores Obtained via Subtraction and Regression

While group-level analyses of performance on the conditions within these four face processing tasks is informative, holistic and configural processing have typically been measured in terms of differences between pairs of conditions within each task. We used two methods to calculate individual difference scores for each task: subtraction and regression. As explained above, we subtracted and regressed across orientation within task manipulation such that the control condition was subtracted from, or regressed against, the condition of interest to create the individual difference scores for accuracy data. Table 4 identifies the control condition and condition of interest for each difference score.

Table 4

Task Conditions Used to Calculate Difference Subtraction Scores and D-Residuals

Task	Calculation of Difference Scores Condition of Interest - Control Condition	Difference Score Label
CCFE	Aligned Congruent – Aligned Incongruent	Aligned
CCFE	Misaligned Congruent – Misaligned Incongruent	Misaligned
PCFE	Aligned Upright – Aligned Inverted	Aligned
PCFE	Misaligned Upright – Misaligned Inverted	Misaligned
C/F	Configural Upright – Configural Inverted	Configural
C/F	Featural Upright – Featural Inverted	Featural
PW	Whole Upright – Whole Inverted	Whole
PW	Part Upright – Part Inverted	Part

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Task Reliability of Individual Difference Scores

We calculated the reliability of the accuracy subtraction difference scores and regression d-residuals using Guttman's λ_2 (Table 5). According to DeGutis et al. (2013) it is a common

mistake to enter the obtained difference scores into a statistical software package to calculate the reliability scores. They note that when calculating the reliability of difference scores one must take into account the correlation between the two variables. In their supplemental materials they provide the formula needed to do so, and this is the protocol that we have followed.

DeGutis et al. (2013) noted that the reliability scores of both sets of individual difference scores are expected to be lower than the reliabilities for performance data. This too was the case in the present study. DeGutis et al. (2013) also found that d-residual reliability scores were more reliable than the individual difference subtraction scores were. We found this to be the case in our data also for three of the four tasks (see Table 5). The exception was the Configural and Featural difference scores. In both cases the subtraction-based difference scores were more reliable than the regression-based scores. Importantly we would like to note that the reliability score is negative for Configural and Featural subtraction-based difference scores (as well as regression-based Featural individual difference scores). If this were to occur in group-level performance data it would be due to the mean of the inter-item correlations being negative, and the researcher would infer that the items (in our case individual trials) should not be combined because they are not part of a common scale. In the case of this occurring for the reliability of individual difference scores the reason and interpretation is less clear. To the best of our knowledge based on the extensive literature review in the field no published papers have reported negative reliability scores for integrative face processing tasks, or interpreted their meaning. In both cases within our data the magnitude of the negative reliability scores is quite low. Nevertheless, what this could indicate is that for the Configural and Featural difference scores collapsing across orientation is not advisable. This is because, due to the negative reliability score, stimulus orientation appears to result in items that reflect different constructs.

We proceeded with analyses for all difference scores as the nature of this manuscript is largely exploratory.

Table 5

Task Reliabilities of Subtraction Difference Scores and Regression Difference D-residuals for Accuracy Data

Task	Individual Difference Scores	Guttman's λ_2 Subtraction	Guttman's λ_2 Regression
CCFE	Aligned	.78	.79
CCFE	Misaligned	.71	.74
PCFE	Aligned	.56	.63
PCFE	Misaligned	.68	.73
C/F	Configural	-.06	.02
C/F	Featural	-.21	-.03
PW	Whole	.39	.46
PW	Part	.27	.39

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Part Whole Effect Task (PW), Configural/Featural Difference Detection Task (C/F).

Correlation Analysis for Individual Difference Scores

As with the group-level accuracy data, we conducted non-parametric correlational analyses on the individual difference scores obtained via subtraction and via regression. This was done because the scores were not normally distributed and the individual difference scores demonstrated a monotonic relationship. Again we felt that transforming these scores in an attempt to achieve univariate normality would complicate the interpretation of the results and instead elected to use the non-parametric Spearman's Rank-Order correlation. Tables 6 and 7 depict these results for individual difference scores obtained via subtraction and regression respectively.

Table 6

Non-parametric Spearman Rho Correlations for Accuracy Subtraction Difference Scores (N = 260)

Task	Variables	1	2	3	4	5	6	7	8	9
CCFE	1. Aligned	-								
	2. Misaligned	-.71**	-							
PCFE	3. Aligned	.02	-.04	-						
	4. Misaligned	-.16**	.13*	.03	-					
C/F	5. Configural	-.05	-.10	-.05	.18**	-				
	6. Featural	-.02	.04	-.07	.10	.08	-			
PW	7. Whole	-.02	.09	-.07	-.02	-.20	-.05	-		
	8. Part	-.09	-.08	-.01	.01	-.05	.08	-.02	-	
CFMT	9.CFMT	.16**	-.10	.00	.10	.00	-.10	-.01	.19**	-

Note. *$p < .05$, two-tailed. ** $p < .01$, two tailed. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Table 7

Non-parametric Spearman Rho Correlations of Accuracy Regression D-Residual Scores (N = 260)

Task	Variables	1	2	3	4	5	6	7	8	9
CCFE	1. Aligned	-								
	2. Misaligned	.27**	-							
PCFE	3. Aligned	.18**	.19**	-						
	4. Misaligned	.25**	.21**	.56**	-					
C/F	5. Configural	.13*	.05	.10	.09	-				
	6. Featural	.04	.05	.11	.12	.33**	-			
PW	7. Whole	.20**	.16*	.04	.08	.12	.04	-		
	8. Part	.12*	.14*	.10	.14*	.13*	.14*	.32**	-	
CFMT	9.CFMT	.25**	.15*	.14*	.12	.16**	.05	.25**	.25**	-

Note. *$p < .05$, two-tailed. **$p < .01$, two tailed. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Part Whole Effect Task (PW), Configural/Featural Difference Detection Task (C/F).

The correlational analyses revealed a different pattern across both sets of individual difference scores (i.e., subtraction and regression). This is likely partially due to the lower reliability of subtraction difference scores compared to regression difference scores. This may indicate that the "control" conditions are pure control conditions and that therefore their variability should be removed from (regressed out of) the equation.

Analysis of Subtraction-Based Individual Difference Scores.

Regardless of the reasons why, the subtraction difference scores yielded very few significant correlations. This is the case between integrative face processing tasks' difference scores (four significant correlations) and between these difference scores and performance on the CFMT (two significant correlations). There does not seem to be a discernable pattern between the task conditions that are significantly correlated nor with respect to the direction of the correlations. With respect to the significant correlations between performance on the CFMT and the two difference scores (Complete Aligned, and Part) we see a low positive association.

Analysis of Regression-Based Individual Difference Scores.

Importantly, in the d-residual accuracy difference scores correlational data revealed the anticipated patterns. D-residual difference scores within each task were more strongly correlated with each other than were d-residual difference scores across tasks. This replicated the pattern found in the accuracy performance data correlation analysis, albeit with smaller effect sizes. The smaller effect sizes can be explained by the fact that, consistent with DeGutis et al., (2013), reliability scores of difference scores were lower than the reliability scores for group-level performance data. Regardless of the smaller effect sizes, these patterns in the correlational analyses provide supporting evidence that each of these tasks is primarily measuring something unique and that therefore researchers should not use these tasks as interchangeable measures of

face processing. From the d-residuals we also see a greater number of significant correlations between difference scores in the *Complete Composite Face Effect Task* and other task conditions than we do between *Partial Composite Face Effect Task* and the other task conditions. This suggests that the *Complete Composite Face Effect Task* is the task that is most similar to all others meaning it is capturing aspects that the other tasks capture. Therefore, in the absence of empirical evidence pertaining to what each of these integrative tasks is measuring in terms of face processing and face recognition ability that we would therefore suggest that researchers use this task if they do not want to present a number of tasks to participants.

The pattern of performance across all d-residual difference scores was such that there was a small positive association between performance on these four integrative face processing tasks (r_s = .16, p < .005, 95% CI [.04, .28], ranging from r_s = .04 to r_s = .56). Importantly, individual difference scores were more strongly correlated within tasks (r_s = .37, p < .001, CI [.26, .47],) ranging from r_s = .27 to r_s = .56) than they were across tasks (r_s = .12, p < .05, 95% CI [.00, .24], ranging from r_s = .04 to r_s = .25). Again this pattern suggests these tasks are more unique than they are similar and therefore researchers should not use these tasks interchangeably. Furthermore, results in the literature should not be generalized beyond the task that was used in a given experiment.

All regression d-residual scores were significantly related to performance on the CFMT except for two: Partial Misaligned (although it was approaching significance, r_s = .12, p = .06), and Featural (r_s = .05, p = .46). The average correlation for all individual difference scores that significantly correlated with the CFMT was (r_s = .20, p < .001, 95% CI [.08, .31], ranging from r_s = .14 to r_s = .25) which indicates a small positive association between performance as measured by d-residual individual difference scores and accuracy on the CFMT. This is

comparable to the results from the accuracy group-level performance data and CFMT, which also demonstrated a small positive association between face processing as measured by these tasks and face recognition ability as measured by the CFMT $r_s = .22$, $p < .001$, 95% CI [.10, .33] range $r_s = .15$ to $r_s = .35$.

Taken together, the patterns that emerged from correlation analyses show that d-residual results are more in line with the accuracy group-level performance data than the subtraction-based individual difference scores are. Subtraction based scores did not demonstrate significant within task correlations, and very few significant correlations across integrative tasks. Moreover, only a two subtraction-based scores were significantly correlated with CFMT. This is in contrast to the majority of accuracy performance scores and d-residual scores showing a significant correlations with CFMT scores. This may not be surprising because the reliability scores were higher for d-residuals than subtraction difference scores for two of the tasks (*Part Whole Effect Task, and Partial Composite Face Effect Task*) and comparable for the *Complete Composite Face Effect Task*. Therefore, with respect to capturing patterns of performance, our data suggest that d-residuals seem to be the better measure of individual difference scores.

Exploratory Factor Analysis of Individual Difference Scores

We conducted exploratory factor analyses to further investigate if and how performance (as measured by two types of individual difference scores) is related between these four integrative face-processing tasks and performance on the CFMT. In order for NESTip to most accurately extract the correct number of factors the data must be normally, or near-normally, distributed. We investigated normality via histograms with a superimposed normal distribution curve as well as via Q-Q plots. Our subtraction-based individual difference scores and regression-based individual difference scores showed minor violations of normality. Therefore

we weighed several considerations in our decisions regarding factor extraction as will be discussed below. Both the unrotated and rotated factor structures for each type of difference score are presented below in their respective section.

Analysis of Subtraction-Based Individual Difference Scores. In order to determine which factor rotation method to select we checked our data for signs of multicollinearity. Variation inflation factors were calculated by conducting iterative linear regressions. Our data did not show signs of multicollinearity. This finding coincides with the lack of significant between-task correlations as reported above (Table 6). Despite the fact that our data did not show signs of multicollinearity we wanted to select rotation that would allow for the possibility of correlation between factors, therefore we elected to use an oblique rotation. Research suggests that oblique rotation options produce comparable results (Fabrigar, Wegener, MacCallum, & Strahan, 1999), so in order to be consistent with the accuracy data exploratory factor analysis we chose to use a direct oblimin rotation. Because our data showed violations of multivariate normality we elected to use principle axis factoring as the factor extraction method.

NESTip suggested a two-factor solution; however, NESTip is conservative, therefore we also elected to visually inspect the scree plot, and then force two, three, four, and five-factor solutions comparing the percentage of variance explained, the critical eigenvalues, as well as factor stability in terms of the item loadings. Weighing the considerations put forward by Osborne and Costello (2009) we chose to report the four-factor solution for subtraction-based individual difference scores.

The Kaiser-Meyer-Olkin measure of sampling adequacy was above .50 (.53). Because sampling adequacy was close to the cutoff of .50 we investigated the anti-image correlation matrix. The correlations along the diagonal were all significant with the exception of Partial

Aligned. Bartlett's test of sphericity was significant (χ^2 (36) = 301.73, $p < .001$). Communality values were initially low but increased following extraction. The four-factor solution explained 34.82% of the total variance. Factor 1 explained 17.94% of the variance, Factor 2, 9.05%, Factor 3, 4.60%, and Factor 4, 3.24%. Unrotated and rotated factor matrices are presented in Table 8 and 9, respectively.

Table 8

Unrotated Factor Loadings and Communalities for Accuracy Subtraction Difference Scores Five-Factor Solution (N = 260)

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
CCFE	Aligned	-.89				.58	.83
	Misaligned	.82				.57	.69
PCFE	Aligned					.03	.02
	Misaligned		.49			.17	.37
C/F	Configural		.62			.18	.41
	Featural					.04	.09
PW	Whole		-.33		.45	.05	.34
	Part			-.43		.05	.21
CFMT	CFMT			.37		.06	.18

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Table 9

Rotated Pattern Matrix for Accuracy Subtraction Difference Scores (N = 260)

Task		Factor			
		1	2	3	4
CCFE	Aligned	-.91			
	Misaligned	.82			
PCFE	Aligned				
	Misaligned		.59		
C/F	Configural		.61		
	Featural				
PW	Whole				.57
	Part			-.45	
CFMT	CFMT			.41	

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

The unrotated four-factor solution did not show evidence of a common factor. This is in contrast to the common first factor in the accuracy group-level data (as shown above in Table 2) or as it was in the d-residual unrotated factor analysis (as shown in Table 10 below). This weak evidence of a common factor is not surprising given the small number of significant between-task correlations in the subtraction-based correlation analysis for individual difference scores reported above (Table 6). From this we can conclude that, according to the unrotated factor analysis of subtraction-based difference scores, performance across these tasks is more unrelated than it is related. This is in contrast to results from the unrotated factor analyses of accuracy scores and d-residuals which both clearly demonstrate a common first factor that accounted for a substantial percentage of variance explained (36.24% and 35.56% respectively). This discrepancy highlights the importance of noting which type of score is being reported and subsequently analyzed because in this case different conclusions can be drawn.

Beyond the fact that there was no evidence of a common factor, the unrotated factor solution showed performance on the two difference scores from the *Complete Composite Face*

Effect Task in the subtraction-based unrotated factor solution loading together onto the first factor which accounts for 34.82% of the variance explained by the model. The rest of the tasks did not have their difference scores loading together. CFMT loaded onto the third factor along with the Part difference scores of the *Part Whole Effect Task*. The Part difference scores loaded negatively suggesting that they are measuring different aspect of the construct that Factor Three represents than CFMT is.

With respect to the four-factor rotated solution, the pattern matrix mirrors the pattern of the unrotated factor solution. It removed the cross-loading Whole difference scores which not loads alone onto the fourth factor instead of also loading onto the third factor with Configural and Misaligned of the *Partial Composite Face Effect Task*. In this case there were not many cross-loadings so the unrotated and rotated factor solutions tell a similar story which is that according to subtraction-based individual difference scores accuracy data suggests that they tasks are measuring different constructs, or components thereof, and therefore should not be used interchangeably as measures of integrative processing. Moreover, results should not be generalized beyond the task(s) that were used in various studies.

Analysis of Regression-Based Individual Difference Scores. We began by conducting iterative linear regressions to investigate multicolineraity and determined that are data did not contain variance inflation factor values above 2.0. Nevertheless, we did have significant correlations across tasks as well as within tasks, and wanted to allow for correlation between factors, so we opted for the direct oblimin rotation. Our data showed violations of multivariate normality therefore we elected to use principle axis factoring as the factor extraction method. The conservative NESTip approach suggested a three-factor solution. Again, we also considered

elements such as visually inspecting the scree plot, critical eigenvalues, percent of variance explained, and stability of factors. We opted for a three-factor solution.

The Kaiser-Meyer-Olkin measure of sampling adequacy was .65, and Bartlett's test of sphericity was significant (χ^2 (36) = 334.66, $p < .001$). Communalities were low initially, and increased following extraction. The three-factor solution explained 35.56% of the total variance. Factor 1 explained 20.58% of the variance, Factor 2, 8.61%, and Factor 3, 6.37%. The unrotated and rotated three-factor solution is presented below in Table 10 and 11 respectively.

Table 10

Unrotated Factor Loadings and Communalities for Accuracy D-residual Difference Scores Four-Factor Solution (N = 260)

Task		Factor			Communality	
		1	2	3	Initial	Extraction
CCFE	Aligned	.45			.20	.29
	Misaligned	.41			.17	.23
PCFE	Aligned	.63	-.44		.42	.60
	Misaligned	.73	-.41		.46	.70
C/F	Configural			.38	.14	.28
	Featural			.55	.12	.37
PW	Whole	.36	.36		.17	.28
	Part	.35			.16	.21
CFMT	CFMT	.38			.18	.23

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Part Whole Effect Task (PW), Configural/Featural Difference Detection Task (C/F).

Table 11

Rotated Pattern Matrix for Accuracy D-residual Difference Scores (N = 260)

Task		Factor		
		1	2	3
CCFE	Aligned	.51		
	Misaligned	.44		
PCFE	Aligned		-.77	
	Misaligned		-.81	
C/F	Configural			.48
	Featural			.62
PW	Whole	.54		
	Part	.42		
CFMT	CFMT	.46		

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

The d-residual unrotated factor solution demonstrated that 6 of the 8 difference scores and the CFMT loaded onto the first factor, which explains 35.56% of the variance. Configural difference scores just missed the suppression cutoff value of .32 but would load onto the common first factor with a value of .31. Featural difference scores would load with a value of .24. The notion of a common first factor echoes the results from the d-residual correlational analysis (Table 7) which demonstrated a considerable number of significant between-task correlations as well as the correlational analysis for the 16 task conditions (Figure 5) along with the accuracy unrotated factor solution (Table 2). Taken together these results suggest that these four face-processing tasks and the CFMT are measuring a common mechanism, or aspect thereof, perhaps a higher order visual processing ability.

Importantly though, the rotated factor solution clearly shows that the regression-based individual difference scores load together onto their respective factors for two of the tasks (*Partial Composite Face Effect Task*, and *Configural/ Featural Difference Detection Task*). *Complete Composite Face Effect Task* and *Part Whole Effect Task* d-residual scores loaded

together onto the first factor (along with CFMT). This result complements the patterns observed in the correlational data for d-residual difference scores such that the within-task correlations were stronger in magnitude than were the between-task correlations. As such, results pertaining to d-residual scores suggest that some of these tasks are also measuring unique aspects of a common face processing mechanism, and/or separate processing mechanisms. Regardless of which scenario it is, these tasks should not be used, or be interpreted, as interchangeable measures of face processing.

With respect to the association between performance as measured by regression-based individual difference scores and the CFMT within the rotated factor solution, the CFMT task loaded along with difference scores from the *Part Whole Effect Task* and the *Complete Composite Face Effect Task*. This pattern is different from that of the rotated accuracy performance data (Table 3) where CFMT failed to load in the rotated factor structure suggesting that d-residual difference scores are capturing performance differently than accuracy alone (or subtraction-based scores for that matter) which highlights the importance of reporting all performance measures during this exploratory phase of analyzing integrative face processing tasks and their association with face recognition ability (as measured by performance on the CFMT or any other task).

Discussion

In the research to date pertaining to the associations between integrative face processing tasks, as well as pertaining to the extent to which integrative face processing is important to face recognition ability, researchers have used a limited number of tasks and a heterogeneous mixture of performance measures (e.g., DeGutis et al., 2013; Richler et al., 2012; Richler et al., 2014; Wang et al., 2012). The purpose of this experiment was to investigate these associations using a

wider range of commonly used face processing tasks as well as a commonly used measure of face recognition ability (CFMT). The face processing tasks that were included were the *Complete Composite Face Effect Task*, the *Partial Composite Face Effect Task*, the *Configural/Featural Difference Detection Task*, and the *Part Whole Effect Task*. Our central questions were:

1) Is performance related across these four face-processing tasks? That is, are they interchangeable measures of face processing and do they measure the same construct?

2) Is there evidence to suggest that one version of the *Composite Face Effect Task* is superior to the other? That is, does one of them have better reliability than the other, and/or does one of them relate to face recognition ability more than the other?

3) Does integrative face processing play an important role in face recognition ability? That is, are there significant associations within the correlational analyses? And does CFMT load with any task condition or difference score onto one or more factors in the factor analyses?

4) Is there evidence to suggest that subtraction-based or regression-based individual difference scores are superior to the other? That is, how do reliability scores compare between scores across tasks? And how do patterns of performance compare?

We investigated these questions using correlational analyses and factor analyses after first conducting 2 × 2 repeated measures ANOVAs in order to determine that performance levels on our tasks were demonstrating the expected interaction effects. Patterns of performance across accuracy scores on each task condition were analyzed in this way, as were individual difference accuracy scores calculated via subtraction as well as via regression. This allowed us to compare patterns across these three outcome measures for all four of these commonly used integrative face-processing tasks and a face recognition ability task (as measured by the CFMT).

Results from the present study demonstrated that accuracy group-level performance data and d-residual individual difference scores exhibited similar patterns of performance. For both dependent variables the magnitude of correlations were stronger for within task associations compared to across task associations. This finding suggests that these tasks are measuring their own construct(s), and they are not in fact interchangeable measures of face processing. There is still evidence of across-task associations, albeit to a lesser extent.

This pattern of results can be summarized by examining the differences in the amount of shared variance in ranking explained within-task compared to between-task for each of the three dependent variables. For example, for accuracy performance data average shared variance in ranking explained across task conditions was approximately 6.60%, whereas it was 30.52% for the within task conditions. The same pattern holds true for regression-based individual difference scores. Shared variance in ranking explained across d-residual individual difference scores was on average 1.44%, whereas the same figure for within task shared variance was 13.70%. Therefore, our answer to the first question is that these face processing tasks do not capture the same aspects of performance and therefore researchers should not use them as equivalent measures of integrative face processing.

Regarding the second question, evidence from our group-level accuracy data as well as d-residual individual difference scores suggests that the *Complete Composite Face Effect Task* is superior to the *Partial* version. First, there are more, and stronger, significant correlations between this task and the other integrative face-processing tasks, suggesting it captures aspects of all three of the other tasks used in this experiment. Second, the *Complete* version had a higher reliability score than did the *Partial* Version. Because neither task was specifically designed for

individual difference analyses we based our comparison solely on group-level performance data to address this second question.

In response to the third question, it seems that integrative face processing does play a role in general face recognition ability. This conclusion stems from the fact that we see small but significant correlations between performance on the CFMT and performance on our four integrative face processing tasks. This is true for both group-level accuracy data and d-residual difference data. However, the fact that the magnitude of the correlations is low suggests that these tasks are capturing only one aspect of the mechanism(s) required to accurately complete the CFMT.

In accordance with the correlation data, the factor analysis showed that CFMT loaded with the group-level accuracy data onto the common first factor in the unrotated matrix (see Table 2). This suggests that all of these task conditions are measuring something in common with the CFMT. This was also the case for d-residual difference scores in the unrotated factor solution (see Table 10). In the rotated factor solution for accuracy d-residuals, CFMT loaded with d-residuals from the *Part Whole Effect Task* and the *Complete Composite Face Effect Task,* (see Table 11) again suggesting that these tasks are measuring something similar.

Regarding our fourth question, about the relative utility of subtraction vs. residual-based difference scores, data from the present experiment supports DeGutis et al. (2013) in their assertion that d-residuals are superior to subtraction-based difference scores. This conclusion stems from the reliability analysis, where d-residuals were generally more reliable, as well as patterns of performance on the correlation and factor analyses. D-residual results were similar to those of accuracy group-level data whereas results based on subtraction scores were different.

Task Reliability

One factor that may account for the low correlation magnitudes between task conditions for difference scores is their low reliability compared to the reliability of group-level performance data on each condition. This could arise because these tasks are not designed to measure individual differences. Indeed, only one task thus far has been designed specifically for the study of individual differences in face processing, the *Vanderbilt Holistic Processing Test-Face* (Richler et al., 2014), and it has not been widely used. Reliability was even lower for the subtraction individual difference scores than it was for regression-based difference scores. This suggests that leaving in the variability from the control condition in the calculation of reliability for the difference scores, or regressing this variability out, has a meaningful impact. Therefore, if researchers are only including one type of difference score in their study they should adequately justify their reasons for selecting one over the other. For the time being, we would suggest that authors report both sets of scores.

The fact that d-residual difference scores' results more closely mirrored those of group-level accuracy data, in terms of correlational analysis and factor analysis, suggests that using this method may be superior to that of the subtraction based method. However, recall that reliability does not reflect the task per se but rather reflects measurements (Ross et al., 2015). Regressing out variability from the control condition resulted in more reliable difference scores than did subtracting out variability, but this may not be the case in other data sets. In addition to reliability analyses, researchers should be encouraged to report their hypotheses regarding what the control conditions are actually capturing, and thereby justify their decision regarding whether to regress or subtract out the variability of the control condition from that of the condition of interest.

Our group-level correlation and factor analysis results suggest that, while these integrative face-processing tasks measure something in common, they do not overlap considerably in terms of between-task correlations, and factors beyond the first one contribute significant additional explained variance to the model. Therefore, we caution researchers against generalizing results across studies that have used different integrative face processing paradigms. While these tasks do measure something in common, they seem to primarily measure different aspects of integrative processing. These differences are important and possibly have contributed to some discrepant findings within the literature (e.g., DeGutis et al., 2013, Konar et al., 2010, Richler et al., 2011; Wang et al., 2012).

Limitations

As with any study, several limitations must be considered in interpreting the results presented here. For instance, as with Nelson et al. (in preparation), what we have learned regarding integrative face processing and face recognition pertains more to processing unfamiliar faces than familiar ones. Integrative processing has been shown to play a smaller role when the task involves processing and recognizing familiar, well-known, faces (Burton et al., 2015).

Another limitation of the present study is that the face stimuli all represented Caucasian males, and we did not exclude participants based on their race or degree of exposure to Caucasians. Therefore, our results may have been made more variable by the "Other Race Face Effect" (e.g., Lindsay, Jack, & Christian, 1991). A body of literature has shown that differences between same-race and other-race face identification seem to stem from qualitatively different processing styles during encoding (e.g., Rhodes, Brake, Taylor, & Tan 1989; Goffaux & Rossion, 2006). That is, these studies argue that same-race faces are treated with holistic and/or configural processing mechanisms, whereas other-race faces are treated in a more piecemeal

featural manner. However, findings from other lines of inquiry contest this position and assert that it is a quantitative rather than qualitative difference in processing that accounts for this effect. That is, it is argued that the same processing mechanisms are used in same-race and other-race facial recognition, but with different degrees of effectiveness (e.g. Hayward, Rhodes, & Schwaninger, 2008; Hayward, Crookes, & Rhodes 2013; DeGutis et al., 2013).

Important research questions that remain unanswered relate to developing a better understanding of what the various integrative face processing tasks are actually measuring. For instance, regarding our factor analysis results, it remains to be determined what the constructs or mechanisms are that the factors represent. Referring to Richler et al. (2012)'s framework of proposed hypothetical mechanisms is an excellent first start. Results from Nelson et al. (in preparation), as well as the present study, may help to inform this conceptualization of integrative face processing by offering additional information pertaining to how performance overlaps between this subset of commonly used tasks; however, these two studies cannot on their own provide definitive evidence regarding what each of these tasks is actually measuring.

Conclusion

In conclusion, accuracy group-level performance data and d-residual individual difference scores data provided evidence suggesting that four integrative face processing tasks (*Complete Composite Face Effect Task*, *Partial Composite Face Effect Task*, *Configural/Featural Difference Detection Task*, and the *Part Whole Effect Task*) each measure a unique component of face processing in addition to measuring a common component. This cautions against using or interpreting these tasks in an interchangeable manner.

Until such a time as research can establish precisely what and how each task measures a particular construct, we advise researchers to use a battery of tasks or, if that is not feasible, we suggest that researchers use the *Complete Composite Face Effect Paradigm* because it is the task that exhibited the most and strongest across-task correlations. Therefore it is the best proxy with respect to the subset of tasks studied within this experiment.

Both accuracy performance data and accuracy d-residuals suggest that there is a small but significant positive association between integrative face processing and face recognition ability on the majority of task conditions. This was also the case when analyzing difference scores, although to a lesser extent and less consistently. In the factor analyses, CFMT loaded onto a common first factor in the unrotated factor structures with the majority of task conditions and difference scores. This suggests that these tasks do capture a meaningful aspect of what is required for accurate face recognition. However, the integrative aspect is obviously far from the entire story, because subsequent factors explain additional variance.

Finally, we found that accuracy performance data and d-residuals were adequately reliable, but that subtraction difference scores had lower relative reliability. This finding emphasizes the importance of reporting and considering reliability data in addition to performance measures. Similarly, our data point to the need to use multiple data analysis techniques in order to provide the most complete picture of how performance across tasks is related. It remains to be determined if accuracy, reaction time, or difference scores best capture integrative processing, and it may vary by task or across participants. Therefore, complete reporting is essential in establishing a theoretically sound base from which to delve deeper into the area of face processing and face recognition.

References

Achim, A. E. (2017). Testing the number of required dimensions in exploratory factor analysis. *The Quantitative Methods for Psychology*, *13*(1), 64-74.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., ... & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423-455.

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science*, *10*(4), 482-496.

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*(14), 4057-4092.

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*(2), 89-99.

Carbon, C. C., & Leder, H. (2005). When feature information comes first! Early processing of inverted faces. *Perception*, *34*(9), 1117-1134.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, *195*(4275), 312-314.

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*(6), 1327-1336.

Cho, S. J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., ... & Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, *27*(2), 552-566.

Chua, K. W., Richler, J. J., & Gauthier, I. (2015). Holistic processing from learned attention to parts. *Journal of Experimental Psychology: General*, *144*(4), 723 – 729.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1-9.

DeGutis, J., Cohan, S., Mercado, R. J., Wilmer, J., & Nakayama, K. (2012). Holistic processing of the mouth but not the eyes in developmental prosopagnosia. *Cognitive Neuropsychology*, *29*(5-6), 419-446.

DeGutis, J., Mercado, R. J., Wilmer, J., & Rosenblatt, A. (2013). Individual differences in holistic processing predict the own-race advantage in recognition memory. *PLoS One*, *8*(4), e58253.

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*(1), 87-100.

Dennett, H. W., McKone, E., Edwards, M., & Susilo, T. (2012). Face aftereffects predict individual differences in face recognition ability. *Psychological Science*, *23*(11), 1279-1287.

Donnelly, N., Cornes, K., & Menneer, T. (2012). An examination of the processing capacity of features in the Thatcher illusion. *Attention, Perception, & Psychophysics*, *74*(7), 1475-1487.

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive neuropsychology*, *24*(4), 419-430.

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*(4), 576-585.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272-299.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception? *Psychological review*, *105*(3), 482-498.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, *26*(2), 105-109.

Fitousi, D. (2015). Composite faces are not processed holistically: Evidence from the Garner and redundant target paradigms. *Attention, Perception, & Psychophysics*, *77*(6), 2037-2060.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception*, *29*(2), 159-170.

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition*, *103*(2), 322-330.

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature neuroscience*, *6*(4), 428.

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*(2), 431.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'greeble'experts: a framework for studying expert object recognition processes. *Vision research*, *38*(15-16), 2401-2428.

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*(4), 1023-1039.

Goffaux, V., & Rossion, B. (2007). Face inversion disproportionately impairs the perception of vertical but not horizontal relations between features. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 995 – 1001.

Hayward, W. G., Crookes, K., & Rhodes, G. (2013). The other-race effect: Holistic coding differences and beyond. *Visual Cognition*, *21*(9-10), 1224-1247.

Hayward, W. G., Rhodes, G., & Schwaninger, A. (2008). An own-race advantage for components as well as configurations in face recognition. *Cognition*, *106*(2), 1017-1027.

Hobson, R. P., Ouston, J., & Lee, A. (1988). Emotion recognition in autism: Coordinating faces and voices. *Psychological Medicine*, *18*(4), 911-923.

Hogg, R. V. (1979). An introduction to robust estimation. *Robustness in Statistics*, 1-17.

Huber, P.J. (1981). *Robust Statistics*. New York, NY: John Wiley and Sons.

Joseph, R. M., & Tanaka, J. (2003). Holistic and part- based face recognition in children with autism. *Journal of Child Psychology and Psychiatry*, *44*(4), 529-542.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science*, *21*(1), 38-43.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(2), 513-536.

Lindsay, D. S., Jack, P. C., & Christian, M. A. (1991). Other-race face perception. *Journal of Applied Psychology*, *76*(4), 587-589.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255-260.

Matheson, H. E., & McMullen, P. A. (2011). A computer-generated face database with ratings on realism, masculinity, race, and stereotypy. *Behavior Research Methods*, *43*(1), 224-228.

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences*, *109*(42), 17063-17068.

McKone, E., Davies, A. A., Darke, H., Crookes, K., Wickramariyaratne, T., Zappia, S., ... Fernando, D. (2013). Importance of the inverted control in measuring holistic face processing with the composite effect and part-whole effect. *Frontiers in Psychology*, *4*, 1-33.

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences*, *109*(42), 17063-17068.

Meinhardt-Injac, B., Boutet, I., Persike, M., Meinhardt, G., & Imhof, M. (2017). From development to aging: Holistic face perception in children, younger and older adults. *Cognition*, *158*, 134-146.

Mercure, E., Dick, F., & Johnson, M. H. (2008). Featural and configural face processing differentially modulate ERP components. *Brain Research*, *1239*, 162-170.

Michel, C., Rossion, B., Han, J., Chung, C. S., & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, *17*(7), 608-615.

Mondloch, C. J., Pathman, T., Maurer, D., Le Grand, R., & de Schonen, S. (2007). The composite face effect in six-year-old children: Evidence of adult-like holistic face processing. *Visual Cognition*, *15*(5), 564-577.

Nelson, E.A., Boutet, I., Watier, N., & Collin, C.A. (2018). *Investigating the Orthogonality of Performance on Four Commonly Used Integrative Facial Processing Tasks.* Manuscript submitted for publication.

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology*, *10*(1), 56-67.

Osborne, J. W., & Costello, A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review*, *12*(2), 131-146.

Palermo, R., Willis, M. L., Rivolta, D., McKone, E., Wilson, C. E., & Calder, A. J. (2011). Impaired holistic coding of facial expression and facial identity in congenital prosopagnosia. *Neuropsychologia*, *49*(5), 1226-1235.

Renzi, C., Schiavi, S., Carbon, C. C., Vecchi, T., Silvanto, J., & Cattaneo, Z. (2013). Processing of featural and configural aspects of faces is lateralized in dorsolateral prefrontal cortex: A TMS study. *Neuroimage*, *74*, 45-51.

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1961-1973.

Rhodes, G., Brake, S., Taylor, K., & Tan, S. (1989). Expertise and configural coding in face recognition. *British journal of psychology*, *80*(3), 313-331.

Richler, J. J., Bukach, C. M., & Gauthier, I. (2009). Context influences holistic processing of nonface objects in the composite task. *Attention, Perception, & Psychophysics*, *71*(3), 530-540.

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), 1- 13.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011b). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464-471.

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.

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition ability and holistic processing. *Journal of Vision*, *15*(9), 1-12.

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254-276.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281–1302.

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision research*, *49*(23), 2856-2861.

Richler, J. J., Mack, M. L., Palmeri, T. J., & Gauthier, I. (2011). Inverted faces are (eventually) processed holistically. *Vision Research*, *51*(3), 333-342.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology*, *3*, 1-6.

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*(6), 1356-1368.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, *103*(1), 34-79.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*, *47*(3), 736–743.

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

Russell R, Duchaine B, Nakayama K (2009) Super-recognizers: People with extraordinary face recognition ability. Psychonomic Bulletin Review, 16(2), 252–257.

Sunday, M. A., Richler, J. J., & Gauthier, I. (2017). Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Attention, Perception, & Psychophysics*, *79*(5), 1453-1465.

Susilo, T., Rezlescu, C., & Duchaine, B. (2013). The composite effect for inverted faces is reliable at large sample sizes and requires the basic face configuration. *Journal of Vision*, *13*(13), 14-14.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Fournal of Experimental Psychology*, *46*(2), 225-245.

Van Belle, G., De Graef, P., Verfaillie, K., Rossion, B., & Lefèvre, P. (2010). Face inversion impairs holistic perception: Evidence from gaze-contingent stimulation. *Journal of Vision*, *10*(5), 1-13.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169-177.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of sciences*, *107*(11), 5238-5241.

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*(9), 1108-1117.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141-145.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747-759.

Yovel, G., Wilmer, J. B., & Duchaine, B. (2014). What can individual differences reveal about face processing? *Frontiers in Human Neuroscience*, (*8)*, 1-9.

Appendix B

Supplemental Materials: Experiment 2

This section presents additional analyses in order to provide readers with a thorough review of performance patterns across a number of behavioural measures. All analyses that were reported in the manuscript were repeated for reaction time data, and the results are available below.

The first results presented in this section are those from the ANOVAs for accuracy and reaction time, which replicate expected interaction patterns thought to index integrative processing. Next, we report reliability analyses for group-level reaction time data. Following this, we present correlational analyses for reaction time group-level data and reaction time difference scores across orientation. Lastly, we present the results from exploratory factor analyses. We begin with results for group-level reaction time data, followed by reaction time difference scores across orientation.

Analysis of Variance of Performance Data

We report effect sizes (η_p^2) and their 95% confidence intervals using the same method as Richler and Gauthier (2014). The error bars around the means in Figures B1 to B8 are correlation-adjusted 95% confidence intervals (O'Brien & Cousineau, 2014).

Accuracy

Complete Composite Face Effect Task. In the *Complete Composite Face Effect Task* the factors were Alignment (aligned vs. misaligned) and Congruency (congruent vs. incongruent). A 2 × 2 within-subjects analysis of variance showed that a main effect of Alignment was not significant, $F(1,259) = .649$ $p = .42$, $\eta_p^2 = .00$ (95% CI: -.01, .01). The main effect of Congruency, $F(1,259) = 181.42$, $p < .001$, $\eta_p^2 = .41$ (95% CI: .32, .50), and the interaction term

were significant, $F(1,259) = 169.41$, $p < .001$, $\eta_p^2 = .40$ (95% CI: .30, .84). Figure B1 depicts these results.

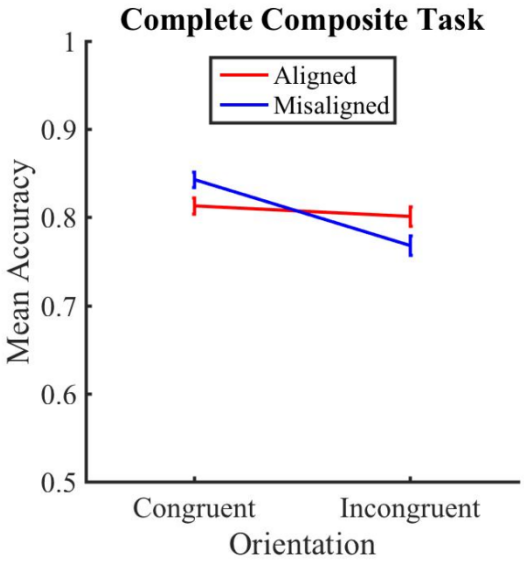

Figure B1. The main effect of Congruency and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Participants performed better on congruent trials than Incongruent trials, and this effect of Congruency was more pronounced on Aligned trials than on Misaligned trials.

Partial Composite Face Effect Task. In this task the factors were Alignment (aligned vs. misaligned) and Orientation (upright vs. inverted). A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,259) = 112.51$, $p < .001$, $\eta_p^2 = .30$ (.95% CI: .20, .38), a main effect of Orientation, $F(1, 259) = 91.71$, $p < .001$, $\eta_p^2 = .26$ (.95% CI: .17, .36), and the interaction term was significant, $F(1, 259) = .31.04$, $p < .001$, $\eta_p^2 = .11$ (95% CI: .04, .19). Figure B2 depicts these results.

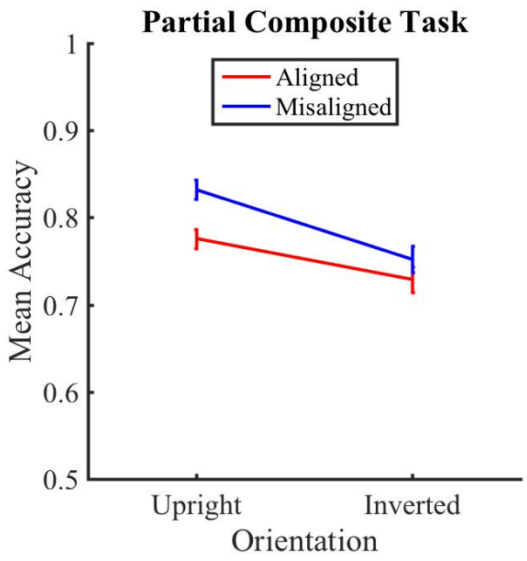

Figure S2. The main effect of Stimulus, of Orientation, and the interaction are significant. The graph presents within-subjects 95% confidence intervals around the mean values. As expected, participants performed better on Misaligned trials compared to Aligned trials, and they performed better on Upright trials compared to Inverted trials. The effect of alignment was more pronounced for upright trials than inverted ones.

Configural Featural Difference Detection Task. In this task the factors were difference type (featural vs. configural) and orientation (upright vs. inverted). A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,259) = 525.44$, $p < .001$, $\eta_p^2 = .67$ (95% CI: .60, .73) a main effect of Orientation, $F(1,259) = 91.37$, $p < .001$, $\eta_p^2 = .26$ (95% CI: .17, .36) and the interaction term was significant, $F(1,259) = 13.12$, $p < .001$, $\eta_p^2 = .05$ (95% CI: .01, .11). Figure B3 depicts these results.

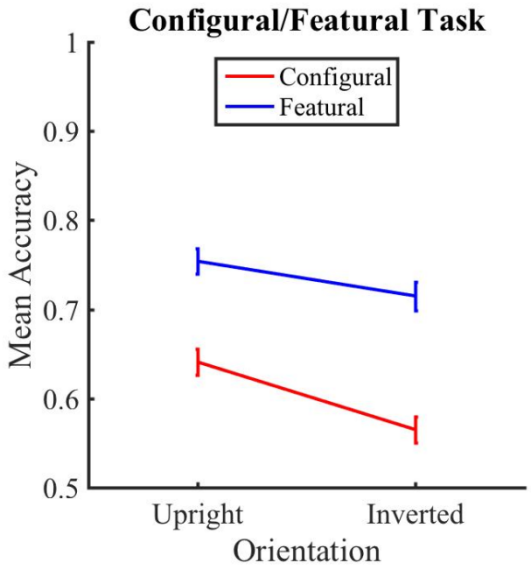

Figure B3. The main effects of Stimulus and Orientation, and the interaction are all significant. The graph presents within-subjects 95% confidence intervals around the mean values. Participants performed better on Featural trials compared to Configural trials, and performed better on Upright trials compared to Inverted trials. The impact of inversion of performance accuracy was more pronounced on Configural trials than on Featural trials.

Part Whole Effect Task. In this task the factors were stimulus type (Whole vs. Part) and orientation (Upright vs. Inverted). A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,259) = 257.70$, $p < .001$, $\eta_p^2 = .50$ (95% CI: .41, .59), a main effect of Orientation, $F(1, 259) = 176.92$, $p < .001$, $\eta_p^2 = .41$ (95% CI: .32, .50), and the interaction term was significant, $F(1, 259) = 33.11$, $p < .001$, $\eta_p^2 = .11$ (95% CI: .05, .18). Figure B4 depicts these results.

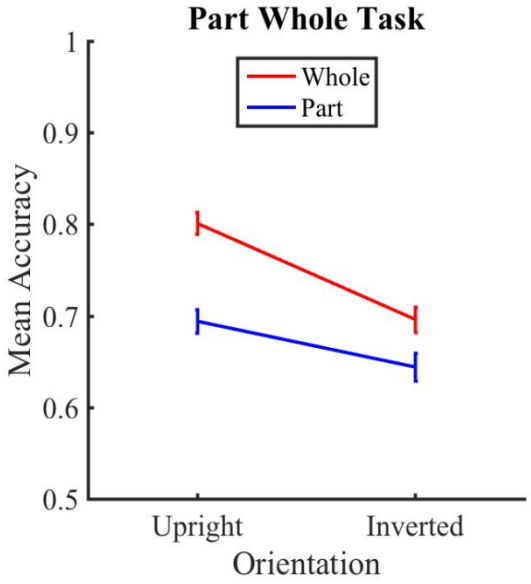

Figure S4. The main effects of Stimulus and Orientation, and the interaction are all significant. The graph presents within-subjects 95% confidence intervals around the mean values. Participants performed better on Whole trials compared to Part trials, and performed better on Upright trials than on Inverted trials. The effect of inversion was more pronounced on Whole trials than on Part trials.

Reaction Time

Complete Composite Face Effect Task. In this task the factors were Alignment (aligned vs. misaligned) and Congruency (congruent vs. incongruent). A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Alignment, $F(1,259) = 8.65$ $p < .01$, $\eta_p^2 = .03$ (95% CI: .00 .08), a main effect of Congruency, $F(1,259) = 84.10$, $p < .001$, $\eta_p^2 = .25$ (95% CI: .16, .34), and the interaction term was significant, $F(1,259) = 29.07$, $p < .001$, $\eta_p^2 = .10$ (95% CI: .04, .18). Figure B5 depicts these results.

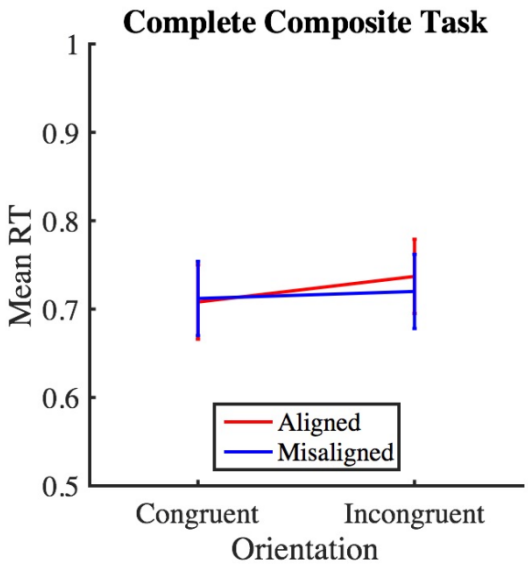

Figure B5. The main effect of Stimulus Orientation, and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. People respond to Aligned faces more slowly than Misaligned faces, and more slowly on Incongruent trials than Congruent trials. The effect of Congruency on reaction time is more pronounced on Aligned trials than Misaligned trials.

Partial Composite Face Effect Task. In this task the factors were alignment (aligned or misaligned) and orientation (upright or inverted). A 2 × 2 within-subjects analysis of variance showed that there was a main effect of task manipulation, $F(1,259) = 27.29$, $p < .001$, $\eta_p^2 = .10$ (95% CI: .05, .18), the main effect of orientation was not significant, and the interaction term was significant, $F(1,259) = 9.74$, $p < .01$, $\eta_p^2 = .04$ (95% CI: .00, .10). Figure B6 depicts these results.

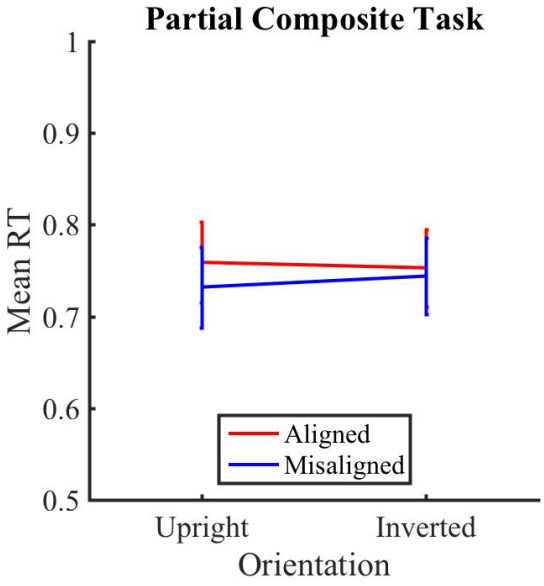

Figure B6. The main effect of Stimulus Manipulation and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Participants responded to Aligned faces more slowly than Misaligned faces. Stimulus orientation impacted response time to a greater degree on Misaligned Trials than on Aligned trials. This interaction is not in the expected direction, although the effect size is quite small.

Configural Featural Difference Detection Task. In this task the factors were difference type (configural vs. featural) and orientation (upright vs. inverted). An analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,259) = 62.44$, $p < .001$, $\eta_p^2 = .19$ (95% CI: .11, .28). The main effect of orientation and the interaction term were not significant. Figure B7 depicts these results.

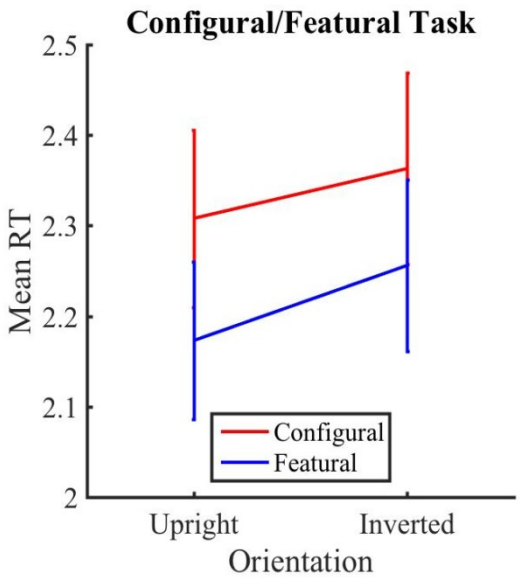

Figure B7. The main effect of Stimulus was significant. The graph presents within-subject 95% confidence intervals around the mean values. Participants responded more quickly on Featural trials compared to Configural trials.

Part Whole Effect Task. In this task the factors were stimulus type (Part or Whole) and orientation (upright or inverted). A 2 × 2 within-subjects analysis of variance showed that there was a main effect of Stimulus Manipulation, $F(1,259) = 242.90$, $p < .001$, $\eta_p^2 = .48$ (95% CI: .39, .56, and the interaction term was significant, $F(1,259) = 12.57$, $p < .001$, $\eta_p^2 = .05$ (95% CI: .01, .11). Figure B8 depicts these results.

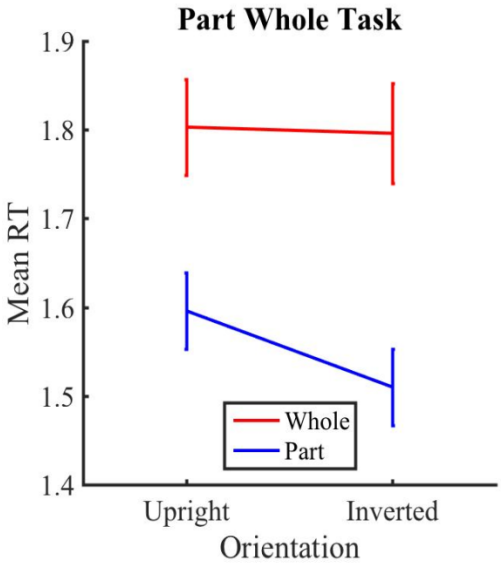

Figure B8. The main effect of Stimulus and the interaction are significant. The graph presents within-subject 95% confidence intervals around the mean values. Participants took longer to respond on Whole trials compared to Part Trials. Impact of inversion on reaction time was more pronounced on Part trials compared to Whole trials. This interaction effect is not in the expected direction; however, the effect size is small.

Reliability Analysis

Reaction Time

Group-level data. We used Guttman's λ_2 to measure task reliability for the four face processing tasks, and the CFMT. The reliability of group-level performance was strong for all tasks: *Complete Composite Face Effect Task* (.98), *Composite Effect Task* (.97), *Configural/Featural Difference Detection Task* (.95), *Part Whole Effect Task* (.87), and CFMT (.90).

Correlational Analyses

Reaction Time – Group-Level Data

		Aligned Con.	Aligned Incon.	Misaligned Con.	Misaligned Incon.	Aligned Up	Aligned Inv.	Misaligned Up	Misaligned Inv.	Configural Up	Configural Inv.	Featural Up	Featural Inv.	Whole Up	Whole Inv.	Part Up	Part Inv.	CFMT
Complete Composite Task	Aligned Con.		0.94	0.94	0.94	0.48	0.42	0.48	0.41	0.21	0.15	0.21	0.17	0.34	0.32	0.30	0.32	0.43
	Aligned Incon.	0.94		0.94	0.94	0.45	0.40	0.45	0.39	0.20	0.17	0.21	0.18	0.33	0.32	0.31	0.34	0.44
	Misaligned Con.	0.94	0.94		0.96	0.47	0.43	0.47	0.43	0.22	0.20	0.23	0.22	0.35	0.34	0.34	0.34	0.47
	Misaligned Incon.	0.94	0.94	0.96		0.47	0.43	0.49	0.43	0.22	0.18	0.23	0.19	0.36	0.33	0.34	0.35	0.48
Partial Composite Task	Aligned Up	0.48	0.45	0.47	0.47		0.87	0.92	0.87	0.33	0.32	0.36	0.31	0.31	0.24	0.32	0.25	0.22
	Aligned Inv.	0.42	0.40	0.43	0.43	0.87		0.87	0.93	0.29	0.31	0.31	0.29	0.28	0.24	0.32	0.29	0.21
	Misaligned Up	0.48	0.45	0.47	0.49	0.92	0.87		0.89	0.27	0.23	0.29	0.23	0.30	0.19	0.29	0.21	0.19
	Misaligned Inv.	0.41	0.39	0.43	0.43	0.87	0.93	0.89		0.30	0.31	0.33	0.30	0.30	0.24	0.34	0.28	0.23
Configural/ Featural Task	Configural Up	0.21	0.20	0.22	0.22	0.33	0.29	0.27	0.30		0.76	0.93	0.78	0.43	0.50	0.47	0.50	0.33
	Configural Inv.	0.15	0.17	0.20	0.18	0.32	0.31	0.23	0.31	0.76		0.75	0.96	0.42	0.56	0.49	0.56	0.33
	Featural Up	0.21	0.21	0.23	0.23	0.36	0.31	0.29	0.33	0.93	0.75		0.77	0.44	0.50	0.47	0.50	0.37
	Featural Inv.	0.17	0.18	0.22	0.19	0.31	0.29	0.23	0.30	0.78	0.96	0.77		0.44	0.55	0.49	0.56	0.34
Part-Whole Task	Whole Up	0.34	0.33	0.35	0.36	0.31	0.28	0.30	0.30	0.43	0.42	0.44	0.44		0.55	0.77	0.49	0.44
	Whole Inv.	0.32	0.32	0.34	0.33	0.24	0.24	0.19	0.24	0.50	0.56	0.50	0.55	0.55		0.53	0.84	0.40
	Part Up	0.30	0.31	0.34	0.34	0.32	0.32	0.29	0.34	0.47	0.49	0.47	0.49	0.77	0.53		0.61	0.50
	Part Inv.	0.32	0.34	0.34	0.35	0.25	0.29	0.21	0.28	0.50	0.56	0.50	0.56	0.49	0.84	0.61		0.43
CFMT		0.43	0.44	0.47	0.48	0.22	0.21	0.19	0.23	0.33	0.33	0.37	0.34	0.44	0.40	0.50	0.43	

Complete Composite Task | Partial Composite Task | Configural/ Featural Task | Part-Whole Task

Figure B9. Non-parametric correlations of accuracy scores for all tasks. r_s values of .122 - .159 correspond to $p < .05$, two-tailed. r_s values between .16 - .173 correspond to $p < .01$, two-tailed. N = 260. Complete Composite Face Effect Task (Complete Composite Task), Partial Composite Face Effect Task (Partial Composite Task), Configural/Featural Difference Detection Task (Configural Featural Task), Part Whole Effect Task (Part/Whole Task).

Reaction Time – Individual Difference Level Data

Table B1

Non-parametric Spearman Rho Correlations for Reaction Time Subtraction Difference Scores (N = 260)

Task	Variables	1	2	3	4	5	6	7	8	9
CCFE	1. Aligned	-								
	2. Misaligned	.09	-							
PCFE	3. Aligned	.00	-.04	-						
	4. Misaligned	.07	-.18**	.39**	-					
C/F	5. Configural	.04	-.18**	.04	.14*	-				
	6. Featural	.06	-.08	.02	.14*	.78**	-			
PW	7. Whole	.05	-.05	.02	.10	.12	.09	-		
	8. Part	.02	.09	.08	.03	.03	.38	.70**	-	
CFMT	9.CFMT	-.14*	-.16**	-.05	-.06	-.03	-.03	.00	.05	-

Note. *$p < .05$, two-tailed. ** $p < .01$, two tailed. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Table B2

Non-parametric Spearman Rho Correlations of Reaction Time Regression D-Residual Scores (N = 260)

Task	Variables	1	2	3	4	5	6	7	8	9
CCFE	1. Aligned	-								
	2. Misaligned	.06	-							
PCFE	3. Aligned	.07	-.01	-						
	4. Misaligned	.17*	-.14*	.40**	-					
C/F	5. Configural	.10	-.10	.07	.16*	-				
	6. Featural	.11	-.02	.11	.18**	.77**	-			
PW	7. Whole	.04	-.00	.05	.13*	.17**	.14*	-		
	8. Part	-.28	-.09	.07	.07	.11	.11	.70**	-	
CFMT	9.CFMT	.08	.04	-.00	.02	.15*	.17**	.28**	.29**	-

Note. *$p < .05$, two-tailed. **$p < .01$, two tailed. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Part Whole Effect Task (PW), Configural/Featural Difference Detection Task (C/F).

Factor Analysis of Performance Data

Reaction Time

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was high, .85. Bartlett's test of sphericity was significant (χ^2 (136) = 5788.95, $p < .001$). We made the same decisions regarding rotation and item suppression that we made and explained in the manuscript regarding accuracy scores. The scree test approach and the conservative NESTip approach both suggested a four-factor solution. The four-factor solution explained 79.91% of the total variance. Factor 1 explained 43.45% of the variance, Factor 2, 28.40%, Factor 3, 12.94% and Factor 4, 5.11%. Table B3 and B4 display the unrotated and unrotated factor structures respectively.

Table B3

Unrotated Factor Matrix for Reaction Time Data

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
CCFE	Aligned Cong	.72	-.45	-.44		.91	.93
	Aligned Incong	.72	-.44	-.45		.91	.92
	Misaligned Cong	.73	-.43	-.46		.94	.95
	Misaligned Incong	.73	-.45	-.46		.94	.95
PCFE	Aligned Up	.65	-.44	.53		.90	.89
	Aligned Inv	.65	-.38	.55		.88	.88
	Misaligned Up	.65	-.47	.51		.90	.90
	Misaligned Inv	.67	-.36	.56		.88	.89
C/F	Configural Up	.67	.52		-.32	.89	.84
	Configural Inv	.65	.56			.92	.82
	Featural Up	.68	.49			.89	.82
	Featural Inv	.67	.56			.92	.84
PW	Whole Up	.60			.33	.66	.57
	Whole Inv	.62	.39			.78	.65
	Part Up	.62	.33		.40	.70	.67
	Part Inv	.63	.40		.34	.80	.69
CFMT	CFMT	.54				.41	.37

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW), Congruent (Cong), Incongruent (Incong), Upright (Up), Inverted (Inv).

Table B4

Pattern Matrix, Rotated Four-Factor Structure of Reaction Time Data

Task		Factor			
		1	2	3	4
CCFE	Aligned Cong				-.96
	Aligned Incong				-.94
	Misaligned Cong				-.96
	Misaligned Incong				-.97
PCFE	Aligned Up			.92	
	Aligned Inv			.94	
	Misaligned Up			.93	
	Misaligned Inv			.93	
C/F	Configural Up		.94		
	Configural Inv		.86		
	Featural Up		.92		
	Featural Inv		.88		
PW	Whole Up	.74			
	Whole Inv	.74			
	Part Up	.85			
	Part Inv	.81			
CFMT	CFMT	.39			

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW), Congruent (Cong), Incongruent (Incong), Upright (Up), Inverted (Inv).

Factor Analysis: Individual Difference Scores Within Task Across Orientation

Reaction Time

Subtraction Difference Scores. Below is the factor analysis for reaction time subtraction difference scores across orientation and within stimulus manipulation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was adequately high, .53. Bartlett's test of sphericity was significant (χ^2 (36) = 569.75, $p < .001$). We determined the number of factors to retain using the conservative NESTip approach, which suggested retaining four factors. The

scree test and parallel analysis corroborated this result. The four-factor solution explained 51.66% of the total variance. Factor 1 explained 21.41% of the variance, Factor 2, 104.57%, Factor 3, 10.44%, and Factor 4, 5.23%. Table B5 and B6 display the unrotated and rotated factor structures respectively.

Table B5

Unrotated Factor Matrix for Reaction Time Subtraction Difference Scores Within Task Manipulation Across Orientation

Task		Factor				Communality	
		1	2	3	4	Initial	Extraction
CCFE	Aligned					.03	.05
	Misaligned				.53	.10	.37
PCFE	Aligned			.53		.25	.38
	Misaligned	.43		.65		.30	.65
C/F	Configural	.81	-.38			.65	.89
	Featural	.72	-.34			.63	.72
PW	Whole	.48	.63			.52	.63
	Part	.48	.77			.52	.82
CFMT	CFMT				-.38	.06	.16

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Table B6

Rotated Factor Matrix for Reaction Time Subtraction Difference Scores Within Task Manipulation Across Orientation

Task		Factor			
		1	2	3	4
CCFE	Aligned				
	Misaligned				.56
PCFE	Aligned			.62	
	Misaligned			.79	
C/F	Configural	.95			
	Featural	.85			
PW	Whole		.79		
	Part		.92		
CFMT	CFMT				-.36

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Regression D-Residual Scores. Below is the factor analysis for regression-based reaction time difference scores (i.e., d-residuals) within task manipulation across orientation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was adequately high, .56. Bartlett's test of sphericity was significant (χ^2 (36) = 597.22, $p < .001$). We determined the number of factors to retain using the conservative NESTip approach, which suggested retaining three factors, as did parallel analysis. The scree test suggested a four-factor solution; however, CFMT did not load into the four factor solution. We elected to present the three-factor solution. The three-factor solution explained 48.23% of the total variance. Factor 1 explained 22.46% of the variance, Factor 2, 24.25%, and Factor 3, 11.53%. Table B7 and B8 display the unrotated and rotated factor structures respectively.

Table B7

Unrotated Factor Matrix for Reaction Time Regression Difference D- Residual Scores Within Task Manipulation Across Orientation

Task		Factor			Communality	
		1	2	3	Initial	Extraction
CCFE	Aligned				.06	.04
	Misaligned				.03	.03
PCFE	Aligned			.52	.29	.38
	Misaligned		-.40	.69	.32	.74
C/F	Configural	.77	-.35	-.35	.66	.84
	Featural	.74	-.35	-.33	.66	.77
PW	Whole	.57	.56		.50	.67
	Part	.53	.65		.50	.72
CFMT	CFMT				.14	.16

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Table B8

Rotated Factor Matrix for Reaction Time Regression Difference D-Residual Scores Across Orientation Within Task Manipulation

Task		Factor		
		1	2	3
CFE	Aligned			
	Misaligned			
CFE	Aligned			.62
	Misaligned			.86
C/F	Configural	.93		
	Featural	.89		
PW	Whole		.82	
	Part		.87	
CFMT	CFMT		.33	

Note. Complete Composite Face Effect Task (CCFE), Partial Composite Face Effect Task (PCFE), Configural/Featural Difference Detection Task (C/F), Part Whole Effect Task (PW).

Chapter 4: General Discussion

To briefly re-iterate the rationale for the work in this thesis, we felt that it was imperative to empirically examine the validity of a fundamental assumption within face processing literature. Specifically, a variety of research paradigms are assumed to capture integrative face processing in comparable ways. Researchers have interpreted results from a number of different tasks as though the tasks were measures that reflected the same underlying construct. If this assumption is not correct, then it is likely contributing to the confusion that persists with respect to the nature of integrative face processing mechanisms and their association with face recognition ability. This confusion arises, we stipulate, because these tasks do not in fact measure featural, configural, and/or holistic processing in the same way and therefore these paradigms should not be used or interpreted as interchangeable measures of the same constructs.

Findings from several previous studies give credence to this postulate (e.g., DeGutis, Wilmer, Mercado, & Cohan, 2013; Richler, Palmeri, & Gauthier, 2012; Richler & Gauthier, 2013; Wang, Li, Fang, Tian, & Liu, 2012); However, only a small subset of integrative face processing tasks have been investigated in these prior studies. Therefore, we endeavored to compare performance across four of the most commonly-used face processing tasks in order to investigate if and how performance on them is related. We chose to include the most commonly-used face processing tasks in order for our results to have the broadest possible applicability to the existing literature.

The fact that research on the different mechanisms underlying face processing lacks a solid conceptual foundation has had negative consequences for the face recognition literature. Attempts to investigate the role that featural, configural and holistic face processing mechanisms might play in face recognition are hampered by the same issue because again a variety of tasks

have been used to examine this question, under the implicit assumption that they measure the same construct(s). If these paradigms are not capturing the same aspect(s) of face processing then researchers will not be able to accurately ascertain the role that the various purported processing mechanisms might play in face recognition. For this reason, in Experiment 2, we included a measure of general face recognition ability, the *Cambridge Face Memory Test* (CFMT). This allowed us to compare patterns of performance between face processing tasks and general face recognition performance. This in turn helped us to examine if and how what these tasks measure is related to face recognition ability.

Another factor limiting the research on face processing mechanisms and face recognition are disagreements regarding how to analyze the data from the various tasks. For instance, different authors report accuracy, reaction time, difference scores, d', or any of a number of other performance measures, under the implicit assumption that the measure reported is the best reflection of the construct under study. However, this assumption is rarely empirically tested. The different ways that researchers address these methodological questions, or fail to address them, all contribute to the state of confusion that exists. For this reason, we designed our experiments to contribute data to as many of these ongoing debates as possible. We included many behavioural performance measures, including both accuracy and reaction time group-level data. This allowed us to compare them to see if the measures agree with one another. Our findings indicate that they do, somewhat assuaging concerns regarding differences in the literature on this point. We nevertheless believe it is essential for future research to report both of these aspects of the data.

We also analyzed our data using an individual differences approach. While recent research suggests that the face processing paradigms we investigated are better designed to

capture group-level effects than individual differences (Sunday, Richler, & Gauthier, 2017), the majority of research to date investigating our main research questions have used individual differences data to do so. Therefore, we included individual difference analyses in the present thesis, and paid particular attention to their measures of reliability. Because there are two types of individual difference scores, with different underlying assumptions, and different authors advocating for one type over the other (e.g., DeGutis et al., 2013; Ross, Richler, & Gauthier, 2015), we included both measures. Again, this allowed us to compare across them, and again we found that regression-based individual difference scores are either just as reliable, or more reliable, than were subtraction-based ones. Results from the regression-based data mirrored those of group-level accuracy performance data.

Another ongoing debate concerns which of two versions of the *Composite Face Effect Task* is the better measure of holistic processing. To address this question we included both versions in our second experiment and compared patterns of performance across them as well as their respective associations with the CFMT. Briefly, we found the *Complete Composite Face Effect Task* to be more reliable than the *Partial Composite Face Effect Task*. Within-task correlations were stronger in magnitude in the *Complete* version compared to the *Partial* Version. Furthermore, the across-task correlational analyses revealed that there were more, and stronger, significant correlations with the other task conditions and those of the *Complete* version of the task. Performance levels on both versions of the task were significantly positively correlated with performance on the CFMT, though the magnitudes were low.

To reiterate, our central questions were:

1) Do commonly used face-processing tasks capture performance in comparable ways? Put another way, are they indexing the same underlying construct?

2) Is there evidence of separable configural and holistic aspects of face processing? Addressing this question will inform whether or not there are several orthogonal integrative face-processing mechanisms.
3) Do patterns of performance and reliability measures indicate that one of the two versions of the *Composite Face Effect Task*, the Complete or the Partial version, is a superior measure of integrative processing? And if so, which version?
4) To what extent is integrative face processing important for general face recognition ability as indexed the CFMT?

In order to address these research questions we conducted two experiments. Both were within-subjects designs with large sample sizes ($N = 223$ and $N = 260$). Correlational analyses and factor analyses were the methods that we used to investigate patterns of performance.

With respect to the first research question, regarding the extent to which these integrative face-processing tasks could be seen as interchangeable paradigms, there were a number of common findings across the two experiments. Within-task correlation coefficients were stronger in magnitude compared to across-task correlations, which had a very low average magnitude. This pattern suggests that the conditions within each task are measuring a common aspect of face processing to a larger degree than they are measuring this same aspect across tasks. In practical terms, we interpret this result to indicate that researchers are making a mistake if they use and interpret these tasks as equivalent measures of underlying face processing mechanisms. Our results suggest that, while these tasks may be measuring a common aspect of face processing, they do so only to a small degree. Shared variance in ranking explained in the accuracy group-level performance data between tasks was 7.80% in Experiment 1 and 6.60% in Experiment 2. This was low compared to the same values within-tasks, which were 23.00% in Experiment 1

and 30.25% in Experiment 2. The same pattern was evident in the individual differences data, though with smaller magnitudes overall and subsequently lower shared variance in ranking explained.

The differences in overall correlation magnitudes between group-level data and individual differences data highlight the importance of reporting both these kinds of data, as well as assessing reliability scores. It was expected that the reliability of individual difference scores would be lower than the reliability of group-level performance on the task conditions because of the nature of individual difference scores. Recall that difference scores are composed of data from two task conditions, each with their respective reliability score. The calculated reliability of a difference score (DeGutis et al., 2013) takes into account the reliability of the two task conditions as well as the correlation between the two task conditions. Because individual differences data has lower reliability scores compared to group-level performance data, presenting individual difference data without group-level data may be misleading. For the present experiments, presenting individual difference data alone would have provided very strong evidence for the position that the various face-processing tasks measure separate constructs. But doing so would have been misleading, because their low correlations are in part due to their low reliabilities. Thus, presenting only the individual differences data would have obscured the (modest) extent to which the various integrative face-processing tasks do measure something in common.

One approach to the problem of differences in reliability is to report disattenuated correlation coefficients, which have been scaled up to "correct" for low reliability. Another is to also report group-level data. Given that most of the commonly-used face processing tasks were not developed specifically for individual differences analyses we advise researchers to report group-

level data. Disattenuation of correlation coefficients can lead to difficulties in interpretation, including coefficients greater than 1.

Results from the factor analyses in Experiment 1 and Experiment 2 concurred with correlational analyses. In both experiments, there was evidence of a common first factor in the unrotated solutions, suggesting that the various conditions across all tasks are measuring something in common. In contrast, the rotated factor solutions demonstrated evidence that each of the tasks (i.e. each set of four task conditions) tended to load onto its own respective factor. This pattern mirrors that of the correlational analyses, and provides additional evidence that these tasks are not interchangeable measures of a unified integrative face processing construct.

Regarding the second research question, across both Experiments we found no clear evidence of separable configural and holistic processing mechanisms. From our factor analyses, as previously mentioned, there was evidence of a common first factor, but patterns were not evident across the other factors in the unrotated matrix, and the rotated matrix provided strong evidence that each task loaded separately rather than together. The data from the present thesis is not sufficient to answer this question definitively though, because it remains to be determined if, and to what extent, these tasks are measuring featural, configural, and holistic processing mechanisms. Without knowing what each task is measuring, all we can currently conclude is that they are not measuring the same thing.

In terms of comparing the two versions of the *Composite Face Effect Task*, correlational data demonstrated that within-task correlation magnitudes were stronger in the *Complete* version compared to the *Partial* version. There were also a greater number of significant correlations of the other integrative processing tasks with the *Complete Composite Face Effect Task* conditions, and their magnitudes were higher. Also, the reliability of the *Complete* version was higher. The

fact that the *Complete* version of the *Composite Face Effect Task* correlates consistently well across tasks and has higher reliability argues that it is the best candidate for a measure of holistic face processing amongst those tested here. However, this assertion is based on a narrow set of criteria which do not include a consideration of the theoretical underpinnings of the task (see, e.g., Rossion, 2013 on this issue). .

The last of our main research questions pertained to the extent that integrative face processing is related to face recognition ability (as indexed by performance on the CFMT). Data converged across behavioural measures and analysis techniques to indicate that holistic and/or configural face processing plays a small but significant role in face recognition ability. It remains to be determined what else, other than what is captured by these tasks, contributes to the human ability to individuate faces.

An ancillary research question we examined alongside our central ones concerned the best method for calculating difference scores. To address this question, we collected and analyzed individual difference data in order to be able to compare differences in patterns of performance between data obtained via the subtraction-based method and the regression-based method. Reliability scores were comparable between subtraction and regression methods on some tasks, and it was higher for regression-based scores on others. Reliability is only one aspect that should be considered when choosing between the two measures, however. Recall that the fundamental difference between the two methods is the assumption regarding what the control condition is measuring (i.e., the extent to which it is a pure control condition, or to which it contains elements of the condition of interest). This decision is necessary for researchers to make and justify if they elect to report one set of individual difference scores over the other. Until such a time as researchers are able to determine what each task and its subsequent task conditions are

measuring, we would recommend that researchers report both sets of individual difference scores. For example, in Study 2 there were many more significant correlations between face processing and CFMT performance on d-residual difference scores compared to subtraction-based difference scores. This discrepancy indicates that reporting either type of difference score in isolation could be problematic and contribute additional confusion to this field of research.

Future research in the field of face perception and face recognition may wish to replicate Experiment 2 with a different measure of face recognition ability, or face processing ability. While the CFMT task is commonly used, it is not without its critics and it is possible that another measure will better reflect general face recognition ability. Indeed, the correlation coefficients we measured between CFMT and the various integrative face-processing tasks were small in magnitude. It may be that this reflects a weakness in the integrative tasks, but it could also be that it reflects problems with the CFMT. It is possible that a task like the Vanderbilt Holistic Face Processing Test (Richler, Floyd, & Gauthier, 2014) can capture additional useful information pertaining to the association between face processing and face recognition. If we are investigating and possibly modifying face processing tasks in order to better capture integrative processing (Sunday et al., 2017) then we should do the same for the face recognition paradigms.

Another potential area of future research pertains to the level of familiarity of the face stimuli. In the present thesis, face stimuli were all unfamiliar to the participants. Integrative face processing is thought to play a more prominent role in the recognition of unfamiliar faces compared to familiar ones (Burton, Schweinberger, Jenkins, & Kaufmann, 2015). Because it remains to be determined to what extent face processing tasks are tapping into integrative face processing mechanisms, it may be of interest to use these paradigms with familiar faces. If patterns of performance demonstrate even weaker between-task correlations when using familiar

faces, then this may provide support for the idea that whatever construct each of these tasks are measuring is in fact integrative in nature. If we see stronger between-task correlations then this may indicate that these processing tasks are tapping into a more featural or texture-based strategy (Burton, et al., 2015).

Another future direction to consider involves the use of electrophysiological measures. It may be that featural, configural, and holistic processing mechanisms are not best captured by behavioural measures, but rather such measures Event Related Potentials (ERP). For example, it has been found that two ERP components appear to index neural processing that is required for humans to detect faces and to recognize them (Schweinberger, Pickering, Jentzsch, Burton, & Kaufmann, 2002; Tanaka, Curran, Porterfield, & Collins 2006; Bentin, Allison, Puce, Perez, & McCarthy, 1996). These are, respectively, the N170—possibly localized to the Occipital Face Area (OFA)—and the N250r –possibly localized to the Fusiform Face Area (FFA) (Schweinberger et al., 2002). It is thought that activity in the OFA is associated with perceiving facial features whereas the FFA is thought to be important for a later stage of visual processing, during which configural encoding takes place (Haxby, Hoffman, & Gobbini, 2000). Bentin et al. (1996) demonstrated that the N170 indexes feature detection for faces while Tanaka et al. (2006) demonstrated that the N250r ERP is associated with analyzing the structure of faces. The latter process would likely include analysis of configural information. Recording both ERPs and behavioural measures when participants complete these tasks would provide additional useful information pertaining to when they are using which type of processing.

While the results from the present thesis provide a detailed picture of patterns of performance across several configural/holistic processing tasks and with the CFMT, and have contributed data to important ongoing debates regarding how best to address existing methodological limitations,

many important questions remain unanswered. Determining what each of these tasks is measuring and to what extent it captures featural, configural, and holistic processing is perhaps the most important next step.

Conclusion

The results of the experiments presented in this dissertation indicate that commonly used face-processing tasks should not be used as interchangeable measures of integrative processing, and that researchers and consumers of research should not generalize results across studies that have used different tasks, or that have reported different behavioural measures. Moreover, the type(s) of visual processing that these tasks capture is related to face recognition ability as measured by the Cambridge Face Memory Test, albeit with low to moderate positive correlation coefficients. Therefore, what other visual information is important to face recognition ability as indexed by the CFMT remains unknown and is not captured by any of these four face processing tasks (*Complete Composite Face Effect Task*, *Partial Composite Face Effect Task*, *Part Whole Face Effect Task*, and the *Configural/Featural Difference Detection Task*). Finally, because it remains to be determined what each task is measuring regarding featural, configural, and holistic processing mechanisms, our data cannot draw conclusions regarding the orthogonality of these processing mechanisms; however, we can conclude that performance on these tasks is orthogonal.

References

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, *8*(6), 551-565.

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science*, *10*(4), 482-496.

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*(1), 87-100.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223-233.

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.

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254-276.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology*, *3*, 1-6.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*, *47*(3), 736–743.

Schweinberger, S. R., Pickering, E. C., Jentzsch, I., Burton, A. M., & Kaufmann, J. M. (2002). Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. *Cognitive Brain Research*, *14*(3), 398-409.

Sunday, M. A., Richler, J. J., & Gauthier, I. (2017). Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Attention, Perception, & Psychophysics*, *79*(5), 1453-1465.

Tanaka, J. W., Curran, T., Porterfield, A. L., & Collins, D. (2006). Activation of preexisting and acquired face representations: the N250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience*, *18*(9), 1488-1497.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169-177.