Journal of Applied Research in Memory and Cognition

Fixing the Stimulus-as-a-Fixed-Effect Fallacy in Forensically Valid Face-Composite Research

Michael B. Lewis

Online First Publication, June 29, 2023. https://dx.doi.org/10.1037/mac0000128

CITATION

Lewis, M. B. (2023, June 29). Fixing the Stimulus-as-a-Fixed-Effect Fallacy in Forensically Valid Face-Composite Research. *Journal of Applied Research in Memory and Cognition*. Advance online publication. https://dx.doi.org/10.1037/mac0000128

AMERICAN PSYCHOLOGICAL ASSOCIATION SARMAC Journal of Applied Research in Memory and Cognition

© 2023 The Author(s)
ISSN: 2211-3681

https://doi.org/10.1037/mac0000128

REVIEW

Fixing the Stimulus-as-a-Fixed-Effect Fallacy in Forensically Valid Face-Composite Research

Michael B. Lewis
School of Psychology, Cardiff University, United Kingdom

Face composites from eyewitnesses' memories are a valuable resource in tackling crime. Many studies have focused on identifying the best system to produce a nameable composite. In this article, it is described that how many of these studies do not provide reliable conclusions because they fail to treat the faces constructed as being a random factor and so make the stimulus-as-a-fixed-effect fallacy. Simulations are reported in which the statistical methodologies typically employed in these studies are performed on random data generated by a null effect. The first simulation shows that the typical analysis of variance (ANOVA) analysis in this field produces a significant effect (i.e., Type 1 error) 20% of the time. A further simulation shows that using generalized estimating equations (GEE) analysis (recently employed in this type of research) does not resolve the problem. Recommendations are made for the analysis of face-composite experiments to best evaluate and hence improve the quality of the face composites made by eyewitnesses.

General Audience Summary
Eyewitness memories are never perfect, so reconstructing a person's face from memory is never entirely accurate. The goal of face-composite research is to identify the best techniques and technologies to help eyewitnesses produce an image that is most likely to be recognized by someone familiar with the target. This field advances by comparing different methods of face-composite construction to determine which performs best. The present article reviews research that compares these methods and reveals that the variability between the constructed images is often not considered. Data simulations show that typical methods of analysis can lead to incorrect conclusions. Alternative data analysis methods are proposed to accelerate the discovery of better ways to support witnesses in face-composite construction.

Keywords: face composites, witnesses, forensic face reconstructions, mixed-effect models, fixed-effect fallacy

Facial composite construction from memory is an important process in the forensic identification of offenders. Systems designed to improve and support witnesses' composites of faces have developed from sketches, through Identikit/Photofit systems to automated self-evolving systems (see Zahradnikova et al., 2018, for a review). There have been many successes attributed to these systems (see for example C. D. Frowd, Pitchford, et al., 2012, and C. D. Frowd, 2021), and the development and refinement of the processes of face composite construction have involved a great deal of scientific research. This research aims to inform us how to best use face-composite systems to reproduce most accurately a nameable image from memory. It has been found, for example, that the use of a cognitive interview (C. D. Frowd, Nelson, et al., 2012) and the blurring of the external features during construction (C. D. Frowd et al., 2013) both help to make the resulting face composite more likely to be correctly identified by another person who knows the target. Such findings are clearly of forensic importance.

While the experiments that evaluate face composites are useful and important, I argue here that there is a recurring error in their design that limits their conclusions and hence their utility. This does not concern the way the research is carried out, which is often excellent and forensically valid (see C. D. Frowd, Carson, Ness, Richardson, et al., 2005); rather, it concerns the statistical analysis from which conclusions are drawn about the relative importance of manipulations that are made in the construction of composites. It is argued here that much of the work that has compared different methods for face-composite construction has committed the

Michael B. Lewis https://orcid.org/0000-0002-5735-5318

The additional data and study materials are available at https://osf.io/me4w7/.

Michael B. Lewis played a lead role in conceptualization, formal analysis, investigation, methodology, project administration, resources, software, visualization, writing–original draft, and writing–review and editing.

Open Access funding provided by Cardiff University: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; http://creativecommons.org/licenses/by/4.0). This license permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Correspondence concerning this article should be addressed to Michael B. Lewis, School of Psychology, Cardiff University, Park Place, Cardiff CF10 3AT, United Kingdom. Email: lewismb@cf.ac.uk

stimulus-as-a-fixed-effect fallacy (Clark, 1973). Examples of studies that have committed this fallacy are highlighted below. Further, in two data simulations, it is shown that this fallacy can lead to a large inflation in Type 1 errors leading to many more significant differences being reported than are justified by the data. The consequence of this is that the scientific development of effective composite construction could have been delayed through misleading conclusions. The scale of this error within the field is explored. Recommendations are made to improve the processing of results in the field of face-composite construction.

Stimulus-as-a-Fixed-Effect Fallacy

Fifty years ago, Clark (1973) alerted the psycholinguistic community to the language-as-a-fixed-effect fallacy: A statistical error that was common within that field of research. The fallacy can be illustrated as follows. A researcher might be interested in whether reading time is faster for one type of word than another type (nouns and verbs for example). They may sample words from each type and measure reading times with a large group of participants. The analysis would proceed by looking at the participants' average reading speed for one sample and their average reading speed for the other sample of words. A researcher might believe that the statistical comparisons of these participants' averages would reveal whether the difference between the two types of words reach significance; however, the researcher would have committed the language-as-a-fixed-effect fallacy.

The analysis described above ignores that the two sets of words tested are each sampled from two populations of words. As such, the size of the effect between the two groups is not fixed but is based on the samples selected. Selecting a different sample of words would have revealed different size of effect. The by-participants analysis ($F1$) can be used to generalize to new participants and so a significant result would suggest that a new set of participants would show the same difference as the sample selected. What this analysis fails to do is to allow the generalization of any observed effect to newly sampled words. The conclusions drawn only apply for the current sample of words and, as psychologists are typically interested in the general features of words, this is of little value. By ignoring the nature of the variability of the items, the analysis commits the language-as-a-fixed-effect fallacy, and it means that any significant effect observed cannot be reliably generalized to all items of that type.

Several solutions to the issue have been suggested. One solution is to average performance over the participants for each item and conduct the statistics on those averages. This is a by-item analysis and provides an $F2$ measure. Clark (1973) suggested that this measure tells us whether we can generalize any observed effect to a new set of sampled words for the same set of participants. From $F1$ and $F2$, it is possible to generate min$F$', which will indicate whether the effect can be generalized to both items and participants. It has become common practice, however, to merely report a by-item and by-participant analysis ($F1$ and $F2$) and to claim significance based on both these being significant. This is an improvement over just reporting $F1$ (and is something that the current author has been guilty of in the past—e.g., Lewis, 1999; Lewis et al., 2002); however, this is less conservative than min$F$' and was only ever intended as an intermediate step in the calculation of min$F$' (J. G. Raaijmakers, 2003).

The discussion around the correct F to use has been overtaken by an uptake in the use of linear mixed models analysis (Brysbaert, 2007; Hutchinson et al., 2014) and generalized linear mixed models (GLMM) when the data are not normally distributed (Jaeger, 2008). These analyses consider each participant-by-item data point and evaluate both the by-item and by-participant variability in order to judge the overall effects. Arguably, this is the best method of analysis but any of these methods, ($F1$ and $F2$; min$F$'; GLMM) are superior to just testing $F1$ in terms of addressing the language-as-a-fixed-effect fallacy.

The language-as-a-fixed-effect fallacy is not restricted to experiments using language and so should be better called the stimulus-as-a-fixed-effect fallacy. Indeed, within studies assessing the effectiveness of therapies, it is known as the therapist-as-a-fixed-effect fallacy (Martindale, 1978). Any experiment that explores the differences between items sampled from populations need to consider the difference between these populations as a random effect rather than a fixed effect. For example, the information content of landscapes has been addressed using this method (Antes, 1977). In many experiments looking at the recognition of faces, the items (i.e., faces) need to be treated as a random effect. So, sampling from two different races of faces needs a by-items analysis (e.g., Byatt & Rhodes, 1998) in order to generalize a race effect to other faces of the same race. Even within the brain imaging analysis of face processing, it is possible to treat the stimulus as a random effect rather than fixed effect (e.g., Westfall et al., 2016). Failure to consider stimuli as a random effect has led at least one case of an article being retracted (e.g., Fisher et al., 2015). Despite the well-established consideration of stimuli as a random variable, there remains an area of psychology where the stimulus-as-a-fixed-effect fallacy is still routinely committed.

Face-Composite Research

Identifying the best way for witnesses to construct a face from memory has generated a considerable amount of research and a range of technological advances. Many experiments that have explored improving facial-composite construction have preceded in the following way. Individual "witnesses" generate a face composite in one of several conditions, 8–12 witnesses for each condition and each witness seeing a different face within that condition. These composites are produced of a target who is unfamiliar to them but were viewed some time (e.g., 24 hrs) before for a few minutes or less. These composites are then shown to a set of participants familiar with the target and the participants make a recognition attempt to each of the 8–12 faces from one condition. This method mimics many of the elements of the forensic-witness setting and so, in terms of evaluating face-composite construction, it is considered forensically valid (see C. D. Frowd, Carson, Ness, Richardson, et al., 2005).

While the procedure may have a high degree of validity, the data analysis can be problematic. Often, the analysis reported takes the mean performance for each participant, averaging over the composites from one condition, and compares them across conditions (e.g., Brace et al., 2006; Fodarella et al., 2021; C. D. Frowd, Bruce, et al., 2007; C. D. Frowd, Carson, Ness, McQuiston-Surrett, et al., 2005). These by-participant analyses, if used alone, commit the stimulus-as-a-fixed-effect fallacy, as explained above. They ignore the variability of recognizability of the items themselves. Indeed, there is good reason to believe that there will be variability in the recognizability of the constructed images: First, it has been demonstrated that witnesses differ in their ability to reproduce a seen face (Ellis et al., 1977), and second, it has been shown that some faces are easier to reproduce than others (Richardson et al., 2020)—although this second issue is less important when the same faces are being used for composites in different conditions. As will be shown in the simulation below, this type of by-participant-only analysis greatly increases the chances of getting a significant result when there is no effect. Simulation 1, below, quantifies the problem by showing the proportion of times one is likely to get a significant result when there is a null effect if one were to use by-participant-only analyses in a typical face-composite experiment.

Simulation 1: By-Participant-Only Analysis of Variance (ANOVA)

Fodarella et al. (2021) investigated the effects of context reinstatement and composite construction method on the naming of face composites. This article was chosen for illustrative purposes; however, it is the case that this article originally had additional analyses that were removed at the request of a reviewer (Hancock, personal communication) and so the fallacy may only be a feature of the published version and not the original article. In their Experiment 1, a set of 10 faces was used for the construction of 60 composites by 60 different witnesses in six different conditions. In an assessment of the naming of these composites, 48 participants were tested on one of the sets of 10 faces constructed in one of the six conditions so there were eight participants in each. The differences between the conditions were evaluated between-participants so that a difference between any two sets of eight participants revealed a difference between those conditions. Comparisons between the six difference conditions were evaluated using a by-participant-only analysis.

Fodarella et al.'s (2021) analysis committed the stimulus-as-the-fixed-effect fallacy and so we are unable to generalize these results beyond the composites employed. It is possible to demonstrate the size of the problem by carrying out a simulation on data of the same format but with no difference between the conditions (i.e., a null effect). To do this, two theoretical populations of "composites" were generated with equal mean recognizability of .179 (i.e., the approximate overall naming performance in the study being simulated) and a standard deviation of .19.[1] That is, each of the 20 items were generated to have a probability of naming defined by a normal distribution with mean of .179 and standard deviation of .190 (values were bounded by 0 and 1). The simulation looked at comparing just two conditions because this is the simplest case that demonstrates the issue, and from this it is possible to scale up to six conditions. From these two identical populations, samples of 10 items were randomly selected for each of the two conditions. Participant-level data were generated as eight sets of binary responses for the 10 items for each of the two samples. These binary responses were categorized as recognized if a uniform random number between 0 and 1 fell below the item's probability of naming measure. This models the situation where all participants have the same level of ability and so the probability of recognizing each composite is determined by the recognizability of the composite. From these randomly generated 16 participants' data, a between-participant independent samples *t* test was conducted, and it was recorded whether the *p* value was less than .05. This by-participants analysis models the type of analysis reported in Fodarella et al. (2021). A second by-items analysis was carried out where the means for the performance of each of the 20 items (over the 16 participants) were analyzed using an independent samples *t* test assessing whether the *p* values was less than .05. These two analyses were carried out 10,000 times for randomly selected samples of composite faces. The details of the data simulations are available at https://osf.io/me4w7/.

From the 10,000 iterations of the random samples, it was assessed what proportion gave a significant result ($p < .05$). The spread of the exact *p* values for the two types of analyses are shown in Figure 1. From the by-participants analyses, the results would be classified as being significant 2018 times out of 10,000 iterations. This represents a Type 1 error rate of .20 or one false significant result in every 4.96 experiments (this is assuming a two-tailed test was being used—as is the case in Fodarella et al., 2021). In contrast, the by-item analyses led to significant results just 462 times (just slightly less than 5% as would be expected using a $p < .05$ criteria). This simulation shows that the by-participant-only analysis employed by Fodarella et al.'s greatly inflates the possibility of showing a significant effect when there is no difference between two populations of composite faces.

The purpose of this simulation was to quantify the potential dangers of committing the stimulus-as-a-fixed-effect fallacy. It shows that failing to consider the random effect of the stimulus leads to an increased probability of getting false significant result. For Fodarella et al.'s (2021) study, several of the critical *p* values for comparing conditions were only just less than .05 and so this fourfold increase in false significant results must draw these conclusions into doubt. This is problematic for this article, but the issues go well beyond a single article.

Generalized Estimating Equations (GEE) Analyses

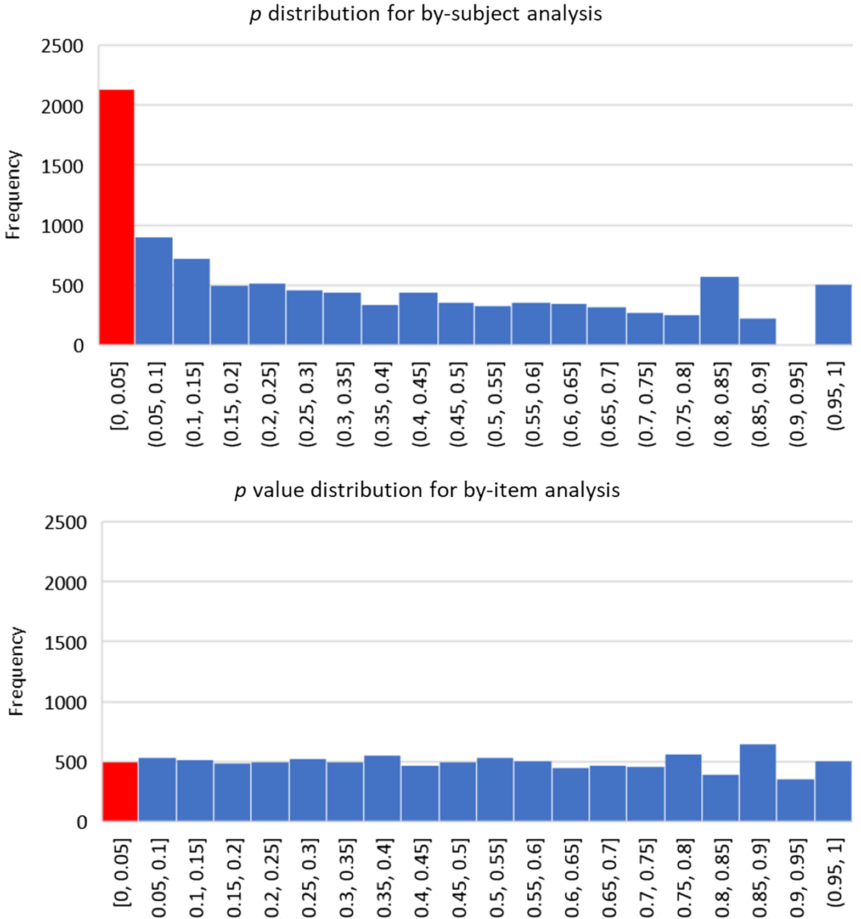
Since 2013, the face-composite-research community has moved toward the use of generalized estimating equations (GEE) as a method of analysis (e.g., C. D. Frowd et al., 2013). This is a method of analysis that can be used to model dichotomous outcomes based on multilevel factors and so is useful for longitudinal studies (e.g., Ballinger, 2004). It has been argued that GEE provides a "regression-type approach that is statistically more powerful than ANOVA and provides a combined by-subjects and by-items model appropriate for the repeated observations" (Richardson et al., 2020, p. 384).

GEE can, in fact, be used to carry out by-participant and by-item analysis, but it does so in much the same way as ANOVA can be used to find *F*1 and *F*2. That is, it needs a two-stage process where the analysis specifically focuses on the by-participant variation and then by-item variation. For example, Kootstra and Şahin (2018) used parallel by-participant and by-item GEE analyses to test their analysis of linguistic priming in nonnative speakers. A simple GEE analysis considers the items as a fixed effect and so the analysis can be just as likely to fall victim of the stimulus-as-a-fixed-effect fallacy

[1] The standard deviation of a set of composites' performance in naming task can be estimated from C. D. Frowd, Nelson, et al.(2012) data who provided by-item data. They report that for sets of 10 items, when analyzed over a range of participants, composites showed that in a naming task where performance was between .03 and .13, the standard error was between .11 and .51. From these values, the measure of by-item standard deviation for naming performance at the .179 level was .19 (based on a standard error of .059).

Figure 1

The Distributions of the Observed p Values, for the Two Types of Analyses, Comparing the Two Groups of Items When There Is No Difference Between the Two Populations



Note. Each figure is based on 10,000 replications. The red bar shows the *p* values that were less than 0.05. The top figure shows the by-subject analysis inflates the Type 1 error by having a greater number of values below 0.05 than would be expected by chance. See the online article for the color version of this figure.

as an ANOVA unless a separate by-items GEE is also reported. Conversely, a single GLMM analysis can handle both fixed- and random-effects and so can provide an analysis that does not fall foul of the stimulus-as-a-fixed-effect fallacy (see Quené & Van den Bergh, 2008). This GLMM method of analysis, however, has not previously been employed to evaluate face-composite-naming experiments.

The potential dangers of using a GEE analysis in repeated measures situations, where the items exert a random effect, can be demonstrated by the following simulation. As in Simulation 1, a null situation was investigated and tested using the standard GEE procedures typically employed in more recent face-composite research. The increase in Type 1 error was assessed by measuring what proportion of these simulations produced significant results when the samples were taken from populations that were identical. A GLMM analysis was also performed on the same data. The data simulated two conditions of the experiment reported in Portch et al. (2017), which employed a GEE analysis to explore the effects of interview type on naming of the face composite produced.

Simulation 2: GEE

In Portch et al. (2017), eight different faces were constructed in four different conditions by 32 participants. The naming test then employed 10 people for each condition familiar with the targets to evaluate whether they would be able to name the target of the eight composites. For the simulation, only two conditions were included, and it was set that there was no difference between the two populations of face composites such that they each had average probability-of-naming scores of .45, and there was a standard deviation of .316 (based on a by-items standard error estimate of .10 from C. D. Frowd, Nelson, et al., 2012). From each of these two populations, eight items were sampled and were tested whether participants would recognize them based on the sampled items' probability-of-naming scores. Each participant-by-item recognition value was scored as 1 if a uniform random number was below the items' probability-of-naming score, else it was 0. Ten participants were modeled in each condition giving 160 dichotomous scores over the two conditions for the 16 items and 10 participants. These scores were subjected to the GEE analysis that was used by Portch et al. (2017) to assess whether the data suggest that there is a significant difference between the two groups. In addition, a separate GLMM analysis (as recommended by Quené & Van den Bergh, 2008, to solve the stimulus-as-a-fixed-effect fallacy) was carried out on the same data to assess the significance of the difference between the two conditions with composites and participants as random effects. This whole process was repeated 1,000 times, using randomly selected samples of face composites, with the expectation that, as there was no difference in the two populations, only 5% of the simulations would produce significant difference between the two conditions with $p < .05$. The details of the data creation and analyses are available at https://osf.io/me4w7/.

The 1,000 iterations of the simulation produced 433 significant results using the GEE analysis. That means that when there is no difference between two conditions, the analysis employed by Portch et al. (2017) would show a significant difference ($p < .05$) between any two conditions 43.3% of the time. This inflation of the Type 1 error is due to GEE not being able to handle the different items as a random effect. The GLMM analysis, however, produced just 48 significant results (similar to the 50 that would be expected with .05 α level) demonstrating that this analysis is a suitable inferential method to assess the data.

A method of analysis that produces a significant result 43% of the time when there is no difference between groups is not a useful tool. The use of GEE is a continuation of the stimulus-as-a-fixed-effect fallacy. GEE does not allow for random-effects in the data model and so it is not suited to the types of analysis it has been employed in within the face-composite literature. GEE has been a popular methodology recently for assessing composite naming even after a by-items ANOVA has been found to be marginally nonsignificant (e.g., Giannou et al., 2021).

Simulation 2 shows that GLMM analysis could be usefully employed to allow the generalization of observed effects to both new participants and new face composites. This method solves the stimulus-as-a-fixed-effect fallacy and so offers a way forward for face -composite research.

The Scale of the Problem in Face-Composite Research

To assess the scale of the stimulus-as-a-fixed-effect fallacy in composite construction research, a literature search was conducted. A search was carried out searching for studies that assessed composites using naming as an evaluation method in a forensically valid manner published since 2000. Only peer reviewed research was included. Table 1 lists the 33 studies identified in this search, indicating the types of analysis employed and whether the research considered stimulus as a random factor.

Across the field, as Table 1 shows, there are times when by-participant only analyses are employed (e.g., Fodarella et al., 2021) and times when by-items only analyses are employed (e.g., C. D. Frowd, Nelson, et al., 2012). There were five papers out of the 33 that employed both by-item and by-participant analyses and interpreted the combination of these (e.g., C. D. Frowd, Pitchford, et al., 2012).

There are times when it is appropriate, or at least acceptable, to use by-participant-only analysis. This would be in situations where there is a fully counterbalanced design (see J. G. W. Raaijmakers et al., 1999). Such a situation is reported in McIntyre et al. (2016) when the same created composites are assessed under a range of different transformations. In this case, there is clear matching between the items and so the $F1$ by-participant analysis is an appropriate one to report as the researchers do in this case. However, the majority of the papers reported in Table 1 looked at variables at the construction stage and so do not fall into this category.

While Table 1 provides a useful overview, it is worth focusing in one of the studies. C. D. Frowd, Carson, Ness, Richardson, et al. (2005) forensically valid study of composite naming has become the "gold-standard" (see C. D. Frowd, Pitchford, et al., 2011) for research in this area. This article provided the framework for many of the studies that followed that assessed the effectiveness of face-composite construction methods. This study did employ both by-participant and by-item analysis; however, this is not the whole story. It followed a now familiar pattern in which a number of witnesses (50) produced face composite in different conditions (five) providing one composite each and later people familiar with the target faces attempt to name the composites. This study pitted sketches against Photofit, PROfit, E-Fit and EvoFIT. Ten faces were constructed with each method with 26 judges for each set attempting to name the composites. The by-participant analysis found a significant effect of construction style and post hoc analysis revealed that E-FIT was significantly better than the sketches, Photofit and EvoFIT and PROfit was significantly better than EvoFIT. Highlighting the potential for variability among the items, the researchers also carried out a by-item analysis. The overall effect of construction methods was significant (albeit at a considerably reduced level). Unlike for the by-participant analysis, pairwise comparisons were not presented for the by-item analysis, so it was not possible to conclude whether specific methods significantly outperformed others. Within the discussion of the article, the reader is told to trust the details of by-participant analysis only. Indeed, there is subsequent analysis of pooled data in the discussion that relies entirely on the by-participant analysis. There is one further stimulus-as-a-fixed-effect fallacy in the discussion, where a further study is reported using the same images (presumed) with 18 new participants tested within-participant. The by-participant analysis showed that E-FIT was superior to Photofit. However, if one tests the same items repeatedly using new sets of participants, this does not mean that it is possible to generalize any effects to new sets of items. It is probably the case that E-FIT is better than Photofit, but without testing by-items then we cannot say that the effect is robust across new situations and simply testing the same composites with new participants does not add to the strength of the argument. So, in this report, by-items analysis is included but only the by-participant results are used to inform the conclusions.

Moving Forward

The two simulations presented here demonstrate that the stimulus-as-a-fixed-effect fallacy is not just a technical issue against which an analysis is robust most of the time. The fallacy can make a large difference to the confidence that one can have in the results by effectively increasing the real α level from purported .05 to either

Table 1

Papers Between 2000 and 2022 Exploring Naming Accuracy of Face Composites Constructed in Forensically Valid Ways (Not Necessarily Exhaustive) Including Whether They Included a By-Participant and/or By-Item Analysis

Research paper	By-participant analysis	By-item analysis	Type of analysis	Notes
Davies et al. (2000)	No	Yes	ANOVA or *T* test	
C. D. Frowd et al. (2004)	Yes	No	ANOVA or *T* test	
Frowd, Carson, Ness, McQuiston et al. (2005)	Yes	No	ANOVA or *T* test	
C. D. Frowd, Carson, Ness, Richardson, et al. (2005)	Yes	Yes	ANOVA or *T* test	The by-items analysis did not include post hoc analyses and differences between specific conditions could not be evaluated.
Brace et al. (2006)	Yes	No	ANOVA or *T* test	
Tredoux et al. (2006)	Yes	No	ANOVA or *T* test	Each item-by-participant data point was treated as independent.
Frowd and Hepton (2009)[a]	Yes	Yes	ANOVA or *T* test	
C. D. Frowd, Bruce, et al. (2007)	Yes	No	ANOVA or *T* test	
C. D. Frowd, McQuiston-Surrett, et al. (2007)	Yes	No	ANOVA or *T* test	
C. D. Frowd et al. (2008)	Yes	Yes	ANOVA or *T* test	A secondary analysis looking at gender used only by-item analysis.
Paine et al. (2008)	Yes	No	ANOVA or *T* test	
C. D. Frowd, Pitchford, et al. (2012)[a]	Yes	Yes	ANOVA or *T* test	
Schmidt (2010)	Yes	No	ANOVA or *T* test	
C. D. Frowd, Pitchford, et al. (2011)[a]	Yes	Yes	ANOVA or *T* test	
C. D. Frowd, Skelton, et al. (2011)[a]	Yes	Yes	ANOVA or *T* test	
Hancock et al. (2011)	Yes	No	ANOVA or *T* test	A by-item analysis was carried out using resampling that did not resolve the fixed effect fallacy.
C. D. Frowd, Nelson, et al. (2012)	No	Yes	ANOVA or *T* test	
C. D. Frowd, Skelton, et al. (2012)	No	Yes	ANOVA or *T* test	
Taylor (2012)[a]	Yes	Yes	ANOVA or *T* test	
C. D. Frowd et al. (2013).	Yes	No	GEE	
Fodarella et al. (2015)	Yes	No	Logistic regression	
Skelton et al. (2015)	No	Yes	ANOVA or *T* test	
McIntyre et al. (2016)	Yes	No	ANOVA or *T* test	Composites were matched between conditions and so by-item level analysis was not necessary.
Brown et al. (2017)	Yes	No	GEE	
Fodarella et al. (2017)	Yes	No	Logistic regression	
Martin et al. (2017)	Yes	No	ANOVA or *T* test	
Pitchford et al. (2017)	Yes	No	GEE	
Portch et al. (2017)	Yes	No	GEE	Simulated in Simulation 2
Martin et al. (2018)	Yes	No	ANOVA or *T* test	
Brown et al. (2019)	Yes	No	GEE	
Brown et al. (2020)	Yes	No	GEE	GEE used for naming but by-participants and by-item ANOVA used for likeness measures.
Skelton et al. (2020)	Yes	No	GEE	
Fodarella et al. (2021)	Yes	No	ANOVA or *T* test	Simulated in Simulation 1
Giannou et al. (2021)	Yes	Yes	ANOVA or *T* test and GEE	GEE was employed after by-item analysis gave marginal nonsignificant results

Note. Also noted is whether they included a by-participant and/or by-item analysis and the type of analysis employed. ANOVA = analysis of variance; GEE = generalized estimating equations.

[a] Studes that fully considered participant and items as random effects.

.2 (from Simulation 1) or .4 (from Simulation 2). The simulations also demonstrate that there are robust solutions to this fallacy. Simulation 1 demonstrates that a by-items and by-participant analysis together can resolve the fixed-effect fallacy, as has been employed for many years in other fields, or alternatively GLMM can allow an analysis where both participants and items are random effects. Moving forward, either of these analyses would provide robust assessment of measures designed to improve the construction of composites. GLMM is not a simple procedure, but it is available on SPSS and R using the *glmer* command in the *lme4* package (Bates et al., 2015). Alternatively, JASP offers a user interface to *lme4* that allows a simple way to select variables as either fixed or random in a GLMM analysis (JASP Team, 2020). The data set can be constructed such that there is a column “item” (coding each of the face composites generated), “participant” (coding each of the participants naming the faces), “condition” (coding the manner of composite construction) and finally “naming accuracy” (coding whether a specific trial was positive or not—including only cases where the participant was familiar with the individual). From this data set, JASP can execute a GLMM using the “naming accuracy” as dependent variable and the “condition” as a fixed variable with “item” and “participant” both selected as random variables.

Power is an important consideration for establishing effects in psychology. Underpowered experiments can be problematic for the research literature (Button et al., 2013). When evaluating the power of a study, it is important to identify its power to generalize to new items; a by-item power analysis should be conducted for experiments where it is intended to generalize results to other items (Brysbaert & Stevens, 2018). Where the plan is to use GLMM to analyze the data, Westfall et al. (2014) describe how a power analysis should be carried out with stimuli as a random effect using a calculator available via a web page (https://jakewestfall.org/power/).

Westfall et al.'s (2014) calculator by can be used to provide a rough estimate of the power of previous designs. In their nomenclature, a study like Portch et al. (2017) is a "stimuli and participant within condition" design, with eight items per condition and 10 participants looking at each condition. With a power of .8, this means that the minimum effect size uncoverable is $d = 1.107$; albeit, this is using the default estimates of participant and stimuli variance. Using this type of design, to find a moderate effect size of $d = 0.5$ with power .8 and 100 participants (50 rating each of two conditions) would require 64 composites (32 in each of two conditions). These numbers can be reduced by having participants look at all the stimuli rather than just one set. Using this kind of "stimulus-within-condition" design, the number goes down to 46 composites (23 in each condition). Refinements to these estimates can be made using more specific measures of participant and stimulus variability but in general, more stimuli and more participants are required for properly powered experiments to assess difference in methods for generating face composites.

Most of the face-composite studies I have mentioned here have been underpowered, but they still measured useful information. The stimuli-as-a-fixed-effect may have artificially increased the significance of the differences observed, but the differences were still there and can still provide useful information if analyzed over several studies. The fact that these studies were published with incomplete analyses is probably a result of the pressure that journals exert on researchers to publish significant findings (see Fanelli, 2012). Publishing only half of the analysis does not serve the journals well, and it does not serve the authors well, but, most of all, it does not serve the people who would make use of these results well. All the studies that I have critiqued warranted publication regardless of the level of significance of their findings: After all, $p < .05$ is an artificial, arbitrary and potentially harmful cutoff (Hubbard et al., 1997). If these results are worth publishing if they were significant, then they are still worth publishing if they are not significant; retraction of partially analyzed data (as happened with Fisher et al., 2015) does not service science well and only adds to the idea that only significant findings are of value. A full description of the data, regardless of whether differences between conditions reach some arbitrary threshold, allow for a more rapid development of methods to improve face composites.

Going forward, I make three recommendations and a suggestion for carrying out research into the effectiveness of face-composite construction. First, ensure that the face constructions are dealt with as a random effect in the analysis. This could be either with a by-items and by-participants design or by using GLMM. Second, evaluate the number of face composites required for the effect size expected. Under powered experiments can be problematic when drawing conclusions. Third, deposit raw data, analysis code, and even the composites constructed in an open-data repository to allow follow-up analysis and comparison between data sets. By following these steps, the speed of discovery of methods that improve the quality of face composites will be accelerated. I also suggest that researchers look into using the registered reports publication model (Chambers & Tzavella, 2022) to ensure that the research results reach the correct audiences regardless of whether the results are significant or not.

Conclusions

The scientific field of face-composite construction has developed through hard work and innovative experiments testing new ideas to advance knowledge. Simulations presented here demonstrate that the analyses in many of these experiments have inflated significance levels. The fact that some of the analyses are problematic does not detract from the valuable work that was conducted. The data collected are valuable but need to be interpreted in a robust fashion. Advances in the ability to reconstruct a person's face from the memory of an eyewitness offer real forensic payoffs. Indeed, the real test of face-composite systems is in the real world and modern systems have been shown to have successes (e.g., C. D. Frowd, Pitchford, et al., 2012; and C. D. Frowd, 2021). The scientific study of the process of face-composite construction from memory should increase these successes. Many of the research articles on face-composite construction present only half of the analyses necessary to fully evaluate the processes involved even though the research was well carried out. Currently, there are a number of face-composite systems that appear to work well, and much is known about memory in order to apply science in a meaningful way to improve face composites. To best drive this research forward, it is important that all findings and data in the domain are published openly regardless of whether they are significant or not and also the analyses of these data need to consider the stimulus-as-a-fixed-effect fallacy.

References

Antes, J. R. (1977). Recognizing and localizing features in brief picture presentations. *Memory & Cognition*, *5*(1), 155–161. https://doi.org/10.3758/BF03209208

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*(2), 127–150. https://doi.org/10.1177/1094428104263672

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brace, N. A., Pike, G. E., Allen, P., & Kemp, R. I. (2006). Identifying composites of famous faces: Investigating memory, language and system issues. *Psychology, Crime & Law*, *12*(4), 351–366. https://doi.org/10.1080/10683160500151159

Brown, C., Frowd, C. D., & Portch, E. (2017). *Tell me again about the face: Using repeated interviewing techniques to improve feature-based facial composite technologies* [Conference session]. 2017 Seventh International Conference on Emerging Security Technologies (EST), Canterbury, United Kingdom. https://doi.org/10.1109/EST.2017.8090396

Brown, C., Portch, E., Nelson, L., & Frowd, C. D. (2020). Reevaluating the role of verbalization of faces for composite production: Descriptions of offenders matter! *Journal of Experimental Psychology: Applied*, *26*(2), 248–265. https://doi.org/10.1037/xap0000251

Brown, C., Portch, E., Skelton, F. C., Fodarella, C., Kuivaniemi-Smith, H., Herold, K., Hancock, P. J. B., & Frowd, C. D. (2019). The impact of external facial features on the construction of facial composites. *Ergonomics*, *62*(4), 575–592. https://doi.org/10.1080/00140139.2018.1556816

Brysbaert, M. (2007). *The language-as-fixed-effect-fallacy: Some simple SPSS solutions to a complex problem.* Royal Holloway, University of London.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), Article 9. https://doi.org/10.5334/joc.10

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Byatt, G., & Rhodes, G. (1998). Recognition of own-race and other-race caricatures: Implications for models of face recognition. *Vision Research*, *38*(15–16), 2455–2468. https://doi.org/10.1016/S0042-6989(97)00469-0

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Davies, G., van der Willik, P., & Morrison, L. J. (2000). Facial composite production: A comparison of mechanical and computer-driven systems. *Journal of Applied Psychology*, *85*(1), 119–124. https://doi.org/10.1037/0021-9010.85.1.119

Ellis, H. D., Davis, G. M., & Shephard, J. W. (1977). *An investigation of the photofit system for recalling faces. Report of work carried out under grant HR 3123/1*. Social Science Research Council.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fisher, C. I., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2015). Women's preference for attractive makeup tracks changes in their salivary testosterone. *Psychological Science*, *26*(12), 1958–1964. https://doi.org/10.1177/0956797615609900

Fodarella, C., Brown, C., Lewis, A., & Frowd, C. D. (2015). Cross-age effects on forensic face construction. *Frontiers in Psychology*, *6*, Article 1237. https://doi.org/10.3389/fpsyg.2015.01237

Fodarella, C., Frowd, C., Warwick, K., Hepton, G., Stone, K., & Heard, P. (2017). Adjusting the focus of attention: Helping witnesses to evolve a more identifiable composite. *Forensic Research & Criminology International Journal*, *5*(1), 221–230. https://doi.org/10.15406/frcij.2017.05.00143

Fodarella, C., Marsh, J. E., Chu, S., Athwal-Kooner, P., Jones, H. S., Skelton, F. C., Wood, E., Jackson, E., & Frowd, C. D. (2021). The importance of detailed context reinstatement for the production of identifiable composite faces from memory. *Visual Cognition*, *29*(3), 180–200. https://doi.org/10.1080/13506285.2021.1890292

Frowd, C. D. (2021). Forensic facial composites. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks* (pp. 34–64). Taylor and Francis. https://doi.org/10.4324/9781003138105-5

Frowd, C. D., Bruce, V., Ness, H., Bowie, L., Paterson, J., Thomson-Bogner, C., McIntyre, A., & Hancock, P. J. (2007). Parallel approaches to composite production: Interfaces that behave contrary to expectation. *Ergonomics*, *50*(4), 562–585. https://doi.org/10.1080/00140130601154855

Frowd, C. D., Bruce, V., Smith, A. J., & Hancock, P. J. (2008). Improving the quality of facial composites using a holistic cognitive interview. *Journal of Experimental Psychology: Applied*, *14*(3), 276–287. https://doi.org/10.1037/1076-898X.14.3.276

Frowd, C. D., Carson, D., Ness, H., McQuiston-Surrett, D., Richardson, J., Baldwin, H., & Hancock, P. (2005). Contemporary composite techniques: The impact of a forensically-relevant target delay. *Legal and Criminological Psychology*, *10*(1), 63–81. https://doi.org/10.1348/135532504X15358

Frowd, C. D., Carson, D., Ness, H., Richardson, J., Morrison, L., Mclanaghan, S., & Hancock, P. (2005). A forensically valid comparison of facial composite systems. *Psychology, Crime & Law*, *11*(1), 33–52. https://doi.org/10.1080/10683160310001634313

Frowd, C. D., Hancock, P. J., & Carson, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Perception*, *1*(1), 19–39. https://doi.org/10.1145/1008722.1008725

Frowd, C., & Hepton, G. (2009). The benefit of hair for the construction of facial composite images. *British Journal of Forensic Practice*, *11*(4), 15–26. https://doi.org/10.1108/14636646200900025

Frowd, C. D., McQuiston-Surrett, D., Anandaciva, S., Ireland, C. G., & Hancock, P. J. (2007). An evaluation of U.S. systems for facial composite production. *Ergonomics*, *50*(12), 1987–1998. https://doi.org/10.1080/00140130701523611

Frowd, C. D., Nelson, L., Skelton, F., Noyce, R., Atkins, R., Heard, P., Morgan, D., Fields, D., Henry, J., McIntyre, A., & Hancock, P. J. (2012). Interviewing techniques for Darwinian facial-composite systems. *Applied Cognitive Psychology*, *26*(4), 576–584. https://doi.org/10.1002/acp.2829

Frowd, C. D., Pitchford, M., Bruce, V., Jackson, S., Hepton, G., Greenall, M., McIntyre, A., & Hancock, P. J. (2011). The psychology of face construction: Giving evolution a helping hand. *Applied Cognitive Psychology*, *25*(2), 195–203. https://doi.org/10.1002/acp.1662

Frowd, C. D., Pitchford, M., Skelton, F., Petkovic, A., Prosser, C., & Coates, B. (2012). *Catching even more offenders with EvoFIT facial composites* [Conference session]. 2012 Third International Conference on Emerging Security Technologies, Lisbon, Portugal. https://doi.org/10.1109/EST.2012.26

Frowd, C. D., Skelton, F., Atherton, C., Pitchford, M., Hepton, G., Holden, L., McIntyre, A. H., & Hancock, P. J. (2012). Recovering faces from memory: The distracting influence of external facial features. *Journal of Experimental Psychology: Applied*, *18*(2), 224–238. https://doi.org/10.1037/a0027393

Frowd, C. D., Skelton, F. C., Butt, N., Hassan, A., Fields, S., & Hancock, P. J. (2011). Familiarity effects in the construction of facial-composite images using modern software systems. *Ergonomics*, *54*(12), 1147–1158. https://doi.org/10.1080/00140139.2011.623328

Frowd, C. D., Skelton, F., Hepton, G., Holden, L., Minahil, S., Pitchford, M., McIntyre, A., Brown, C., & Hancock, P. J. (2013). Whole-face procedures for recovering facial images from memory. *Science & Justice*, *53*(2), 89–97. https://doi.org/10.1016/j.scijus.2012.12.004

Giannou, K., Frowd, C. D., Taylor, J. R., & Lander, K. (2021). Mindfulness in Face Recognition: Embedding mindfulness instructions in the face-composite construction process. *Applied Cognitive Psychology*, *35*(4), 999–1010. https://doi.org/10.1002/acp.3829

Hancock, P. J., Burke, K., & Frowd, C. D. (2011). Testing facial composite construction under witness stress. *International Journal of Bio-Science and Bio-Technology*, *3*(3), 65–71.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the *Journal of Applied Psychology*, 1917–1994. *Theory & Psychology*, *7*(4), 545–554. https://doi.org/10.1177/0959354397074006

Hutchinson, S., Wei, L., & Louwerse, M. (2014). Avoiding the language-as-a-fixed-effect fallacy: How to estimate outcomes! of linear mixed models. *Proceedings of the annual meeting of the cognitive science society* (Vol. 36, pp. 2293–2398). Cognitive Science Society. https://escholarship.org/content/qt65z86895/qt65z86895.pdf

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

JASP Team. (2020). *JASP* (Version 0.14.1) [Computer software].

Kootstra, G. J., & Şahin, H. (2018). Crosslinguistic structural priming as a mechanism of contact-induced language change: Evidence from Papiamento-Dutch bilinguals in Aruba and the Netherlands. *Language*, *94*(4), 902–930. https://doi.org/10.1353/lan.2018.0050

Lewis, M. B. (1999). Age of acquisition in face categorisation: Is there an instance-based account? *Cognition*, *71*(1), B23–B39. https://doi.org/10.1016/S0010-0277(99)00020-7

Lewis, M. B., Chadwick, A. J., & Ellis, H. D. (2002). Exploring a neural-network account of age-of-acquisition effects using repetition priming of faces. *Memory & Cognition*, *30*(8), 1228–1237. https://doi.org/10.3758/BF03213405

Martin, A. J., Hancock, P. J. B., & Frowd, C. D. (2017). *Breathe, relax and remember: An investigation into how focused breathing can improve identification of EvoFIT facial composites* [Conference session]. 2017 Seventh International Conference on Emerging Security Technologies (EST), Canterbury, United Kingdom.

Martin, A. J., Peter, J. H., Frowd, C. D., Heard, P., Gaskin, E., Ford, C., & Hewett, T. (2018). *EvoFIT composite face construction via practitioner interviewing and a witness-administered protocol* [Conference session]. 2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS), Edinburgh, United Kingdom. https://doi.org/10.1109/AHS.2018.8541464

Martindale, C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *46*(6), 1526–1530. https://doi.org/10.1037/0022-006X.46.6.1526

McIntyre, A. H., Hancock, P. J., Frowd, C. D., & Langton, S. R. (2016). Holistic face processing can inhibit recognition of forensic facial composites. *Law and Human Behavior*, *40*(2), 128–135. https://doi.org/10.1037/lhb0000160

Paine, C. B., Pike, G. E., Brace, N. A., & Westcott, H. L. (2008). Children making faces: The effect of age and prompts on children's facial composites of unfamiliar faces. *Applied Cognitive Psychology*, *22*(4), 455–474. https://doi.org/10.1002/acp.1374

Pitchford, M., Green, D., & Frowd, C. D. (2017). *The impact of misleading information on the identifiability of feature-based facial composites* [Conference session]. 2017 Seventh International Conference on Emerging Security Technologies (EST). https://doi.org/10.1109/EST.2017.8090421

Portch, E., Logan, K., & Frowd, C. D. (2017). *Interviewing and visualisation techniques: Attempting to further improve EvoFIT facial composites* [Conference session]. 2017 Seventh International Conference on Emerging Security Technologies (EST), Canterbury, United Kingdom. https://doi.org/10.1109/EST.2017.8090406

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002

Raaijmakers, J. G. W. (2003). A further look at the" language-as-fixed-effect fallacy." *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *57*(3), 141–151. https://doi.org/10.1037/h0087421

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416–426. https://doi.org/10.1006/jmla.1999.2650

Richardson, B. H., Brown, C., Heard, P., Pitchford, M., Portch, E., Lander, K., Marsh, J. E., Bell, R., Fodarella, C., Taylor, S. A., Wortington, M., Ellison, L., Charters, P., Green, D., Minahil, S., & Frowd, C. D. (2020). The advantage of low and medium attractiveness for facial composite production from modern forensic systems. *Journal of Applied Research in Memory and Cognition*, *9*(3), 381–395. https://doi.org/10.1016/j.jarmac.2020.06.005

Schmidt, H. C. (2010). *Determinants of estimated face composite quality* [Doctoral dissertation]. Department of Psychology, University of Cape Town.

Skelton, F. C., Frowd, C. D., Hancock, P. J. B., Jones, H. S., Jones, B. C., Fodarella, C., Battersby, K., & Logan, K. (2020). Constructing identifiable composite faces: The importance of cognitive alignment of interview and construction procedure. *Journal of Experimental Psychology: Applied*, *26*(3), 507–521. https://doi.org/10.1037/xap0000257

Skelton, F. C., Frowd, C., & Speers, K. E. (2015). The benefit of context for facial-composite construction. *Journal of Forensic Practice*, *17*(4), 281–290. https://doi.org/10.1108/JFP-08-2014-0022

Taylor, D. A. (2012). *Featural and holistic processing in facial composite construction: The role of cognitive style and processing sets* [Doctoral dissertation]. University of Westminster.

Tredoux, C., Nunez, D., Oxtoby, O., & Prag, B. (2006). An evaluation of ID: An eigenface based construction system: Reviewed article. *South African Computer Journal = Suid-Afrikaanse Rekenaartydskrif*, *37*, 90–97. https://journals.co.za/doi/10.10520/EJC28017

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

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, *1*, Article 23. https://doi.org/10.12688/wellcomeopenres.10298.1

Zahradnikova, B., Duchovicova, S., & Schreiber, P. (2018). Facial composite systems. *Artificial Intelligence Review*, *49*(1), 131–152. https://doi.org/10.1007/s10462-016-9519-1

Received April 12, 2022
Revision received April 28, 2023
Accepted May 25, 2023 ■