

This is a repository copy of *Viewers extract mean and individual identity from sets of famous faces*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/83704/>

Version: Submitted Version

---

**Article:**

Neumann, Markus F., Schweinberger, Stefan R. and Burton, A. Mike [orcid.org/0000-0002-2035-2084](https://orcid.org/0000-0002-2035-2084) (2013) *Viewers extract mean and individual identity from sets of famous faces*. *Cognition*. pp. 56-63. ISSN 0010-0277

<https://doi.org/10.1016/j.cognition.2013.03.006>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

1 Running Title: average set identity

2

3 **Viewers extract mean and individual identity from sets of famous faces**

4

5 Markus F. Neumann<sup>1</sup>, Stefan R. Schweinberger<sup>1,2</sup>, & A. Mike Burton<sup>3</sup>

6

7 <sup>1</sup>Department of General Psychology, University of Jena, Germany

8 <sup>2</sup>DFG Research Unit Person Perception, University of Jena, Germany

9 <sup>3</sup>University of Aberdeen, UK

10

11

12

13 Correspondence to: Markus F. Neumann, School of Psychology, The University of  
14 Western Australia; 35 Stirling Highway, Crawley, WA 6009, Australia; Phone +61 8 6488  
15 113; Fax +61 8 6488 1006

16 E-mail: [markus.neumann@uwa.edu.au](mailto:markus.neumann@uwa.edu.au)

17

18 **Abstract**

19 When viewers are shown sets of similar objects (for example circles), they may extract  
20 summary information (e.g., average size) while retaining almost no information about the  
21 individual items. A similar observation can be made when using sets of unfamiliar faces:  
22 Viewers tend to merge identity or expression information from the set exemplars into a  
23 single abstract representation, the set average. Here, across four experiments, sets of  
24 well-known, famous faces were presented. In response to a subsequent probe, viewers  
25 recognized the individual faces very accurately. However, they also reported having  
26 seen a merged 'average' of these faces. These findings suggest abstraction of set  
27 characteristics even in circumstances which favour individuation of the items. Moreover,  
28 the present data suggest that, although seemingly incompatible, exemplar and average  
29 representations co-exist for sets consisting of famous faces. This result suggests that  
30 representations are simultaneously formed at multiple levels of abstraction.

31

32 **Keywords:** set representation, ensemble coding, face, identity, averaging

33

34

35

## Introduction

36 “Set representations” have recently attracted increasing research interest. When  
37 seeing groups of perceptually similar objects, information such as size, or motion, may  
38 be coded via summary statistics in terms of a mean value across exemplars (Albrecht &  
39 Scholl, 2010; Chong & Treisman, 2003). Whenever observers can capitalize on  
40 redundancy of information – a common observation in structured sets – they can  
41 compress this information into a single representation such as the set average (Alvarez,  
42 2011). In a seminal investigation, Ariely (2001) investigated size representations from  
43 sets containing differently sized circles. Critically, participants tended to identify a test  
44 circle as having been presented when it had a similar size to the mean of the whole set,  
45 even when such an item had not been present. Moreover, participants were near  
46 chance when they had to choose which of two circles had been presented. Taken  
47 together, these findings suggest that i) mean size information was computed and  
48 retained for the set and ii) size information of individual set members was unavailable.  
49 There are different potential explanations for weak exemplar representations. First,  
50 encoding of precise exemplar representations may not routinely occur, or may simply  
51 contain too much noise, perhaps due to the lack of focal attention to set exemplars.  
52 Alternatively, an individual representation may initially be computed but may then be  
53 discarded extremely fast.

54

55

56

57

Recently, statistical representations have been demonstrated for sets of  
perceptually complex stimuli, such as faces. When asked to compare the emotional  
intensity of a single image with that of a set (up to 16 face photographs varying in  
emotional intensity), participants performed highly accurately (Haberman & Whitney,

58 2007, 2009). Performance was actually comparable to a control “exemplar” condition, in  
59 which participants compared an image with a homogeneous set with constant emotional  
60 intensity. Beyond extraction of mean emotion (and gender, see Haberman & Whitney,  
61 2007) information from sets of faces, a similar mechanism may compute the mean  
62 *identity* from sets of *unfamiliar* faces. In one recent study (de Fockert & Wolfenstein,  
63 2009), participants initially saw sets containing photographs of 4 unfamiliar faces from  
64 different individuals. In a “match” condition, a subsequent single image could either be  
65 an exemplar image from the previous set, or an average morph created from the four set  
66 images. Strikingly, the set averages (which had never been seen) received more  
67 ‘present’ responses than the (seen) exemplars. The authors concluded that averaging  
68 identity information might serve as the “default mode” for generating mental  
69 representations from groups of faces.

70         Given that facial representations should serve person recognition, this is a  
71 surprising finding, since mean identity representations should actually *prevent*  
72 identification of a specific person in a group. It is relatively straightforward to understand  
73 how superficial averaging of abstract shapes might take place in the visual system, but  
74 much harder to account for averaging over such high-level characteristics as someone’s  
75 identity. For this reason, it is important to note that the authors used unfamiliar faces.  
76 Crucially, unfamiliar face recognition is strongly image-dependent and sensitive to  
77 superficial picture similarity (Bruce et al., 1999), and is thus based on very different  
78 mechanisms than familiar face recognition. For example, viewers are very good at  
79 matching different images of a familiar person, but very poor at matching unfamiliar  
80 faces (Bruce, Henderson, Newman, & Burton, 2001; Burton, Bruce, & Hancock, 1999;

81 Kemp, Towell, & Pike, 1997; Clutterbuck & Johnston, 2004). This discrepancy suggests  
82 a qualitative difference in perception of familiar and unfamiliar face identities (Hancock,  
83 Bruce, & Burton, 2000), which may also have consequences for the interpretation of the  
84 identity set averaging data. Accordingly, increased percentages of “present” responses  
85 to matching averages in the study of de Fockert and Wolfenstein (2009) could reflect  
86 *image* averaging across similar pictures, rather than *identity* averaging. If viewers are  
87 failing to differentiate between the unfamiliar people shown to them, they might plausibly  
88 construct a set average combining these images. So, while this study certainly  
89 demonstrates set averaging for a class of high-level stimuli (faces), we argue that  
90 evidence for *identity* set averaging would be much more compelling if it could also be  
91 shown to exist for familiar faces sets.

92 Another important characteristic of previous studies examining set averaging for  
93 faces was relatively small image variability within sets. For instance, set averaging for  
94 facial expressions was generally investigated by assembling sets from a single identity,  
95 using slightly different emotional intensities from a morph continuum between two  
96 veridical expressions (Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2007,  
97 2010). One study on set *identity* averaging actually involved 4 true set photographs, but  
98 had sets deliberately arranged to comprise similar identities (de Fockert & Wolfenstein,  
99 2009). Therefore, low recognition rates for set exemplars may have originated from  
100 participants being unable to differentiate between exemplars at encoding. It is important  
101 to see if the use of more naturally diverse sets could increase exemplar memory, and  
102 whether this would in turn affect the quality and strength of set representations.

103           In sum, previous studies have investigated set averaging using face sets that  
104 varied little on either identity or image properties. In the present study, we tested facial  
105 identity averaging by using diverse pictures from highly familiar identities, for which  
106 participants have rich pre-existing mental representations. We further encouraged  
107 identity processing for half of the participants by instructing them to indicate whether a  
108 specific *person* had been seen in a set of faces, while the other half indicated whether a  
109 specific *image* had occurred. We expected that set averaging would be strongly reduced  
110 or absent for highly familiar faces, and that performance would reflect accurate  
111 representation of exemplars instead; Since viewers know these identities, and faces in  
112 the set were quite diverse, there appears to be no advantage in averaging across them.

113

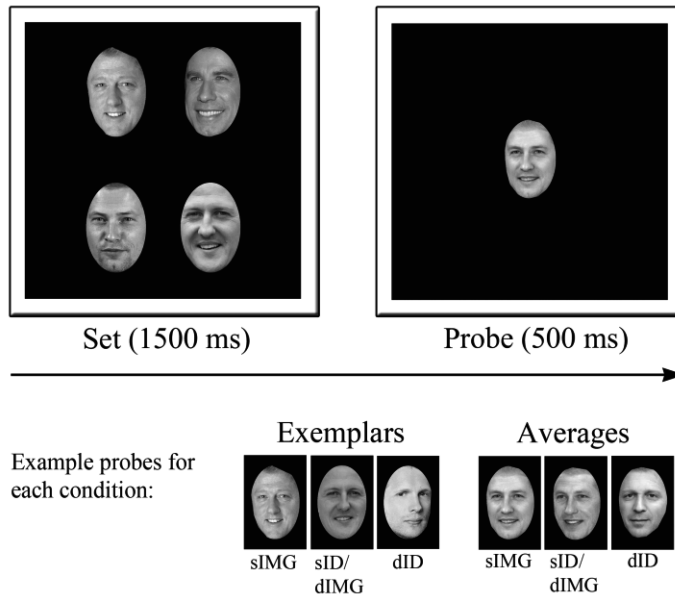
114

**Material and methods**

115           The present article includes 4 experiments that share the following aspects.  
116       Stimuli were 240 original faces collected from various internet sources, 10 each from 24  
117       well-known celebrities (12 German and 12 International). Sixty gender-homogeneous  
118       sets were created from these photographs, each containing 4 images of different  
119       identities. Images contributing to a set were chosen to be roughly similar with respect to  
120       head angle and gaze direction. Five sets from 12 different identity combinations were  
121       assembled. Note that as a result of obtaining the images from the internet, image  
122       variation within the sets was large. All images were taken under entirely non-  
123       standardized conditions, causing considerable variation on image parameters such as  
124       lighting. Additional *set averages* were created for each of the 60 sets by morphing  
125       across the respective 4 set images. Image size was 247 x 387 pixels, all images were  
126       presented grey-scaled and fitted in an oval mask, excluding most of the hair.

127           Set displays contained 4 images randomly assigned to 4 specified positions on  
128       the screen (cf. Fig. 1), and were presented for 1500 ms. Immediately following the set  
129       display (ISI = 0), probe images were displayed for 500 ms, in smaller size than the set  
130       images (200 x 300 pixels). Participants used both index fingers to indicate via button  
131       press (“f” and “j” on a standard German keyboard) whether or not the probe image had  
132       been present in the previous S1 set. Probe images were: i) a set exemplar (i.e., an  
133       image from the previous set); ii) a new exemplar of one of the 4 identities of the previous  
134       set; iii) a new exemplar of a different familiar identity; iv) the average of the 4 set  
135       images; v) the average of 4 different images of the set identities; or vi) the average of 4  
136       images of different familiar identities.





137

138 **Fig. 1: Example of a set, followed by a probe (sIMG average). Sets were presented simultaneously in**  
 139 **Experiments 1-3, and sequentially in Experiment 4. Celebrities in the example set depict (top left to bottom**  
 140 **right): Bill Clinton, John Travolta, Till Schweiger (German actor), and Michael Schumacher (German race car**  
 141 **driver). Examples for all probe conditions of this set are given below.**

142 In each of these six conditions, 60 trials were presented, with 10 trials per  
 143 condition in each of 6 experimental blocks. Response button assignment for “present”  
 144 and “absent” was counterbalanced across participants. A blank screen for 2200 ms  
 145 allowed for a total response window of 2700 ms.

146 Experiments were run in two versions, varying in task requirements. Version a)  
 147 required participants to indicate whether a particular *image* had been a set member,  
 148 whereas version b) required participants to match *identity* (i.e., whether a person had  
 149 been a set member). Participants in version a) were explicitly informed that a different  
 150 image for one of the set identities could occur as a probe stimulus and were instructed  
 151 to respond “absent” in this case. Overall, 84 young adult participants (mean age = 22.01,  
 152  $SD = 3.38$ ; 19 male) were tested and received monetary compensation or course credit.  
 153 Participants gave written informed consent and reported normal or corrected-to-normal

154 visual acuity. Experiments 1a and 1b each comprised 18 participants, and all remaining  
155 experiments (2a, 2b, 3a, 3b, 4a, 4b) comprised 8 participants each.

## 156 **Experiment 1 - main study**

### 157 **Method**

158 Experiments 1a and 1b followed the procedure laid out above, differing only in  
159 the response required by participants (image-present, or person-present). These and  
160 subsequent experiments followed a 2 (Probe Type) x 3 (Match Type) design. Probe  
161 types were either exemplars (i.e., original images), or set averages. Match Type referred  
162 to the relation of the probe face to the set images in that it involved either one, or an  
163 average of all i) image(s) from the set (sIMG), ii) different image(s) from the same set  
164 identities (sID/dIMG), or iii) image(s) of different identities (dID).

165 Prior to the experiment proper, participants were given 24 practice trials, and  
166 provided with trial-by-trial feedback on accuracy. Note that the correct answer to  
167 average probes is always “absent”. In order to prevent participants from learning this  
168 association, averages were not presented in the practice phase. In order to assess  
169 familiarity of the identities used, new pictures of the 24 celebrities were shown following  
170 the main procedure in Experiment 1b. Participants were presented images consecutively  
171 in the middle of the screen for an unlimited duration, and for each face they indicated by  
172 button press whether or not they were familiar with the person. For a “familiar” response,  
173 participants were additionally asked to indicate the name, or if they were unable to do  
174 so, some identifying semantic information for that person (i.e., occupation, nationality).

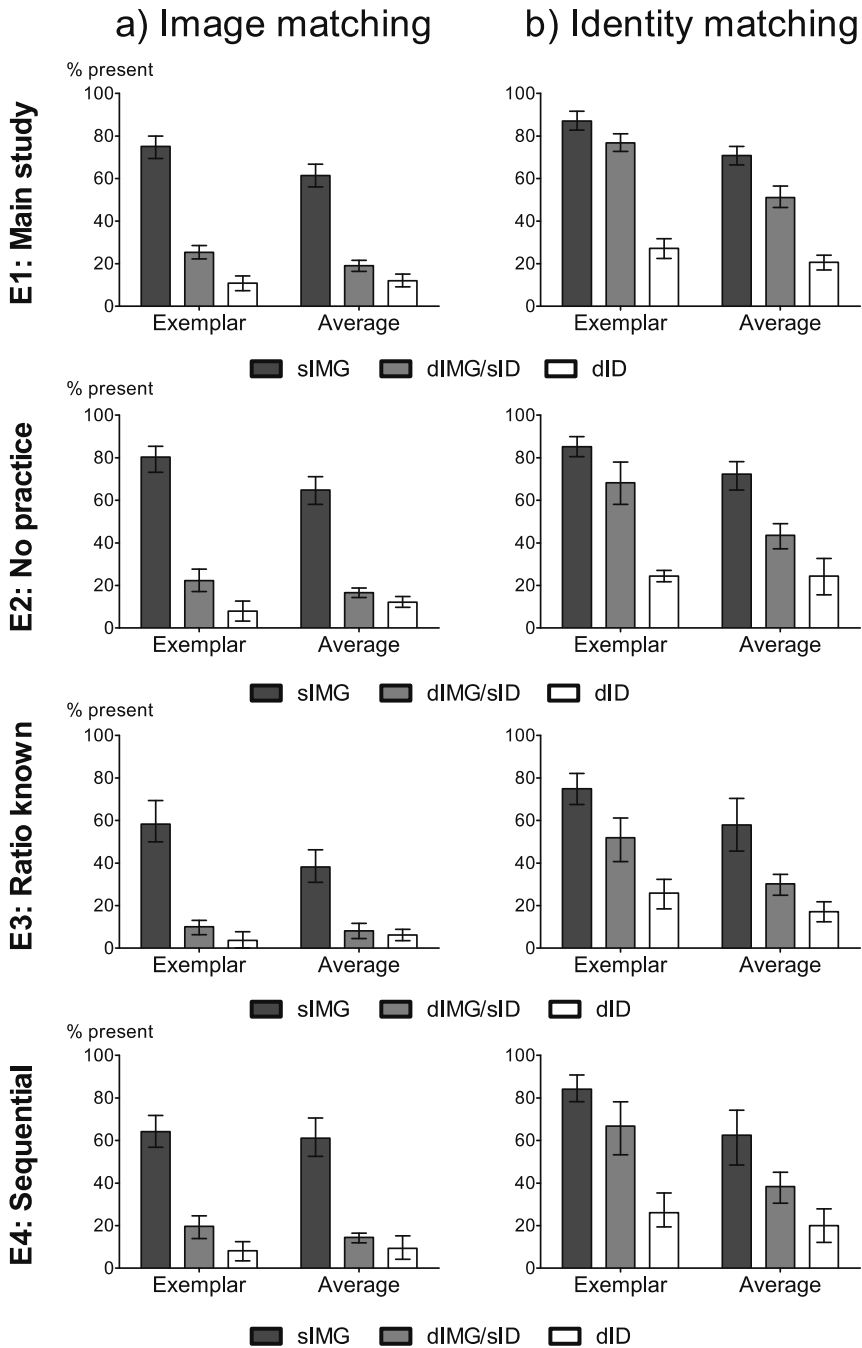
175

**176 Results****177 Face familiarity task**

178           Written responses from one participant were unavailable. Overall, recognition  
179 performance was high, and nearly all the celebrities used in the Experiment could be  
180 spontaneously named. On average, celebrities were successfully identified by unique  
181 semantic information or name in 92.4% of the cases.

**182 “Present” responses to probe faces**

183           Figure 2 (row 1) shows the proportion of present responses for each of the probe  
184 types in Experiment 1 for the image matching (left) and the identity matching group  
185 (right). First, and as expected, participants performed very accurately on probe  
186 exemplars. Proportions of “present” responses during both tasks were clearly largest for  
187 sIMG conditions, indicating good matching performances when a probe image was  
188 identical to one of the set images. New identities in dID conditions received few  
189 “present” responses overall, i.e., false positives were rare. During image matching, new  
190 images from one of the set identities (sID/dIMG) were rejected quite accurately, but less  
191 well than dID images. During identity matching, “present” hits to sID/dIMG images were  
192 frequent, though reduced compared to present responses to identical images (sIMG).



193

194 **Fig. 2: Percentage “present” responses to probe images in all 4 Experiments. Left column: image**  
 195 **matching; Right column: identity matching. Error bars represent 95% CI based on normalized data**  
 196 **(see Cousineau, 2005).  $N = 18$  in Experiments 1a and 1b,  $N = 8$  in all control Experiments 2a,b;**  
 197 **3a,b; and 4a,b, respectively.**

198 Unexpectedly, a strikingly similar pattern was elicited by set average images.  
199 sIMG averages elicited remarkably large proportions present responses, indicating that  
200 participants erroneously identified the set average as an actual set member. This was  
201 not because averages per se tended to elicit responses (e.g., due to inflated typicality),  
202 since averages of different identities (dID) were reliably rejected. During image  
203 matching, averages across 4 new images from the set identities (sID/dIMG) were  
204 rejected - correctly - almost as accurately as averages from new identities (dID). By  
205 contrast, a much larger and intermediate level of incorrect present responses occurred  
206 during identity matching (incorrect, because an average *never* represented an identity  
207 from a set).

208 Statistical analyses were performed by entering data from both experiment  
209 versions to separate 2 by 3 ANOVAs. These revealed reliable main effects of Match  
210 Type and Probe Type (all  $F > 8$ , all  $p < .011$ , all  $\eta^2_P > .320$ ), and significant interactions  
211 of both factors,  $F(2, 34) = 9.55$ ,  $p < 0.01$ ,  $\eta^2_P = .360$ ;  $F(2, 34) = 17.50$ ,  $p < 0.01$ ,  $\eta^2_P =$   
212  $.507$ , for Experiments 1a and 1b, respectively. Accordingly, present responses were  
213 most frequent for sIMG, intermediate for sID/dIMG, and infrequent for dID conditions. In  
214 addition, while present responses occurred overall more often for exemplar than  
215 average probes, the amount of the difference varied with Match Type, and was rather  
216 small (Exp. 1b) or absent (Exp. 1a) in dID conditions. Critically, even when averages  
217 were analysed separately, Match Type was still highly significant (both  $F > 170$ ,  $p <$   
218  $.001$ ,  $\eta^2_P > .810$ ), as were all pairwise contrasts between sIMG and sID/dIMG, and  
219 between sID/dIMG and dID conditions in both tasks (all  $t(17) > 3.81$ , all  $p < .002$ ).  
220 Importantly, this confirms that averages were more often selected not only when created

221 from the identical set images (sIMG), but also when created from different images of the  
222 set identities (sID/dIMG), compared to averages from new identities (dID). More detailed  
223 descriptions of all 2 by 3 ANOVAs and follow-up paired comparisons are detailed in  
224 Table 1 in the appendix.

225 In a second step, we examined differences between task conditions by including  
226 Task as between-subjects factor in an ANOVA on combined data from Experiments 1a  
227 and 1b. The 3-way interaction of Task by Probe Type by Match Type,  $F(2, 68) = 6.88$ ,  $p$   
228  $< .01$ ,  $\eta^2_p = .360$  was significant, indicating differences in patterns elicited during image  
229 and identity matching, respectively. Fig.2 suggests that a main source for this interaction  
230 were large differences in sID/dIMG exemplar conditions in both tasks. This was  
231 unsurprising, because a “present” response had been the correct answer during identity  
232 matching, but the incorrect response during image matching.

233 Of greater theoretical interest were differences in present responses elicited by  
234 average probes across the two tasks. ANOVA on data from average probes with Match  
235 Type and Task revealed a significant interaction,  $F(2,68) = 20.34$ ,  $p < .001$ ,  $\eta^2_p = .374$ ).  
236 Independent sample t-tests carried out on corresponding Match Type conditions  
237 between the two tasks indicated comparable proportions present responses for sIMG  
238 averages in image and identity matching,  $t(34) = 1.670$ ,  $p = .104$ , and slightly more  
239 present responses during identity than image matching for dID conditions  $t(34) = 2.619$ ,  
240  $p = .014$ . Most importantly, sID/dIMG present responses differed substantially between  
241 task conditions,  $t(34) = 6.460$ ,  $p < .001$ , with more present responses given in the  
242 identity than in the image matching task. Thus, sID/dIMG averages were not easily  
243 mistaken as a *set image*, but were frequently mistaken as a *person* occurring in a set.

244

**Control Experiments 2-4**

245           Considering that set averaging was typically observed in combination with  
246 impaired exemplar memory, the finding from Experiment 1 is particularly challenging,  
247 because it suggests that viewers are extracting identity-average information from a set,  
248 while simultaneously representing individual exemplar information. Moreover, while it  
249 seems reasonable to suppose that viewers might code a set of circles using summary  
250 statistics, or even a set of unknown faces, there seems no reason why one should  
251 extract an average of, for instance, Bill Clinton and John Travolta. In the following  
252 control experiments, we tested for a number of possible alternative explanations for this  
253 effect.

**254 Method**

255           Experiments 2-4 were identical to Experiment 1 except as follows. Experiment 2  
256 did not include practice trials. During practice in Experiment 1, the ratio of correct  
257 “present” responses was larger than in the actual experiment, such that one might be  
258 concerned that participants developed exaggerated expectations about the required  
259 ratio of present responses. To exclude this possibility, practice trials were omitted in  
260 Experiment 2 and all further experiments. In Experiment 3, participants were additionally  
261 informed, correctly, that present responses were required in 16.6 % (Exp. 3a), or 33.3 %  
262 (Exp. 3b) of the trials. In Experiment 4, set images were presented sequentially rather  
263 than at the same time (order: top left, top right, bottom left, bottom right). Each image  
264 was shown for 375 ms, such that total presentation duration was equivalent to  
265 Experiments 1-3 (i.e., 1500 ms).

266 **Results**

267 Control Experiments 2-4 yielded results completely consistent with Experiment 1  
268 (cf. Fig. 2, rows 2-4). Most importantly, performance in sIMG conditions was in each  
269 case quite accurate for exemplars, and very inaccurate for averages, with large  
270 proportions present responses to both sIMG exemplars and, only slightly reduced, to  
271 sIMG averages.

272 Separate 2 by 3 ANOVAs for each experiment corroborated the pattern of  
273 Experiment 1. Again, more present responses were given to exemplars than to  
274 averages (except for Experiment 3a, where the main effect of Probe Type only  
275 approached significance,  $p = .076$ , and in Experiment 4a,  $p = .334$ ). Main effects of  
276 Match Type indicated more present responses to sIMG vs. sID/dIMG conditions, and to  
277 sID/dIMG vs. dID conditions throughout. Probe Type interacted with Match Type in all  
278 experiments except for Experiment 3b and 4a. Further descriptions of 2 by 3 ANOVAs  
279 for all control experiments are detailed in Table 1 in the appendix.

280 Experiments 2 and 3 controlled for possible expectation effects in Experiment 1a  
281 regarding the correct proportion present responses. Such expectations could either  
282 originate from practice trials, or from a more general expertise with psychological  
283 experimentation methods. However, Experiment 2 replicated all key results of  
284 Experiment 1 in virtually identical form, despite excluding practice trials (cf. Fig. 2).  
285 Similarly, informing participants about the correct ratio of present trials in Experiment 3  
286 did not differentially affect responses to set averages, although it led to an overall  
287 decrease in present responses, indicating that this information successfully induced a



288 more conservative response criterion. We conducted additional ANOVA on combined  
289 data from Experiments 2 and 3, and included “Ratio Information” (Experiment 2: not  
290 informed, Experiment 3: informed) as an additional between-subjects factor. No  
291 significant 4-way interaction was found,  $F < 1$ , and no other interaction including Ratio  
292 Information, all  $p > .05$ , except for an interaction of Match Type by Ratio Information,  
293  $F(2, 56) = 8.09$ ,  $p = .002$ ,  $\eta^2_P = .224$ . The latter interaction simply reflects the fact that  
294 informing participants about correct ratio led to a greater reduction of present responses  
295 in SIMG matching (18.3%) conditions (critically, both for exemplars and averages), and  
296 less reduction in the other two conditions (sID/dIMG = 12.6%; dID = 4.0%), in which  
297 present responses were already less frequent. Importantly, Experiment 3 provides no  
298 evidence that the ratio of present responses might explain the remarkably large  
299 proportions of present responses to “matching” set averages.

300 Experiment 4 addressed a different possibility. Specifically, when presented  
301 simultaneously, set images could have been processed to a different extent (e.g., with a  
302 focus on the top two faces, and only brief inspection of the bottom faces). By presenting  
303 the set faces sequentially for the same amount of time, participants are encouraged to  
304 process all faces equivalently. Note that simultaneous presentation is not essential for  
305 statistical processing (Chong & Treisman, 2005b; Haberman & Whitney, 2009). In the  
306 ANOVA on combined data from Experiments 4a and 4b, the 3-way interaction only  
307 approached significance,  $F(2, 28) = 2.75$ ,  $p = .086$ ,  $\eta^2_P = .164$ , possibly due to relatively  
308 low power. However, interactions of Task by Probe Type,  $F(2, 28) = 4.66$ ,  $p = .049$ ,  $\eta^2_P$   
309 = .250, and Task by Match Type,  $F(2, 28) = 8.75$ ,  $p = .002$ ,  $\eta^2_P = .385$ , were revealed.  
310 Overall, the pattern of results strikingly resembles the previous findings. Most

311 importantly, sequential presentation caused no selective reduction in present responses  
312 to sIMG set averages compared to Experiment 2. If anything, sIMG *exemplar* detection  
313 was slightly compromised during image matching in Experiment 4a: Exemplars received  
314 comparable proportions present responses as averages, and neither the main effect of  
315 Probe Type,  $F(1, 7) = 1.08$ ,  $p = .334$ ,  $\eta^2_P = .134$ , nor the interaction of Probe Type and  
316 Match Type  $F(1, 7) = 2.62$ ,  $p = .111$ ,  $\eta^2_P = .273$  were significant.

317

318

## General Discussion

319 We examined set averaging for identity information in face sets. In contrast to  
320 previous work, sets in the present study involved both familiar faces, and large image  
321 variability. Compared to earlier work, we used an extended experimental procedure by  
322 including both an image-change condition (SID/dIMG) and an additional task (identity  
323 matching) to promote identity processing of sets exemplars. Across four experiments,  
324 we consistently received two key results that extend the current knowledge regarding  
325 set representations for complex stimuli, and that can be summarized as follows.

326 First, and as predicted, the use of familiar faces in briefly presented sets  
327 produces good memory for set exemplars. Second, and surprisingly, viewers  
328 nevertheless show clear and consistent evidence for averaging identity information in  
329 faces, even across highly familiar set exemplars. Three control studies ruled out  
330 alternative explanations based on participants' expectations, or a potential selective  
331 processing of a subgroup of set items. We will first discuss these novel findings in the  
332 context of our specific approach to create variable sets from familiar faces, and then  
333 relate these results to the concepts of set averaging and individual face recognition  
334 more generally.

335 Previous studies had used low image variability within sets. Set images were  
336 either taken from standardized databases and set identities were chosen to resemble  
337 each other (de Fockert & Wolfenstein, 2009), or – more commonly – sets comprised  
338 perceptually similar levels from a morph continuum (e.g. happy to neutral expression,  
339 see Haberman & Whitney, 2007, 2009). One reason why participants in previous studies

340 were almost unable to recall individual set exemplars may have been simply because  
341 when presented in the set, they all looked alike. By contrast, sets in the present study  
342 employed images from different internet sources, and therefore varied more naturally on  
343 various dimensions including lighting, viewing angle, head posture, and expression. We  
344 expected that set exemplars would consequently be easier to discriminate and that this  
345 would lead to improved exemplar memory, which was the pattern we observed in the  
346 present study. However, we also assumed that increased exemplar memory would  
347 coincide with little if any evidence for set average representations. This assumption was  
348 based on our understanding of set averages as an efficient process to capture the  
349 essential information from a set in situations where accurate encoding of the set  
350 constituents is impossible, for instance by short presentations of crowded displays. Such  
351 an idea seemed intuitively plausible and was supported by many previous studies using  
352 both simple and complex stimulus material (for a recent review, cf. Alvarez, 2011).

353         Here we observed a strikingly different pattern: Despite the expected good  
354 performance in exemplar memory, set averaging was remarkably robust. In actual fact,  
355 present response rates for sIMG averages of about 60% in the present study were even  
356 higher when compared to a analogous condition of a different study, where unfamiliar  
357 faces had been used (approximately 40%, de Fockert & Wolfenstein, 2009).  
358 Accordingly, set averaging of facial identity appears robust to substantial image  
359 variability within sets, at least for familiar faces.

360         Importantly, the use of familiar faces enabled us to address alternative low level  
361 explanations for this identity set averaging effect, which previous work could not  
362 completely rule out. Specifically, it was unclear whether participants generated average

363 *identity*, or rather *average image* representations from sets. Here, we tested separate  
364 groups of participants either with an image matching task as in previous work (e.g., de  
365 Fockert & Wolfenstein, 2009), or with an identity matching task. Such a task should have  
366 promoted identity processing for the set exemplars, and participants could not simply  
367 rely on matching certain low-level aspects of an image due to the potential image  
368 change in sID/dIMG conditions.

369         Critically, we found clear evidence for set averaging in the identity matching  
370 group. This suggests that the abstraction of identity information into a summary statistic  
371 is not simply a low-level stimulus-driven process, but includes averaging of actual  
372 identity information from several faces. This argument receives further support when  
373 taking into account the results from sID/dIMG conditions, where participants of the  
374 identity matching group often misinterpreted an average across 4 different identities as  
375 an actual person from the previous set, even though the probe average involved  
376 different images of these identities! Note that this was not a result of inaccurate person  
377 memory due to the rather short presentation duration: Identity recognition for exemplars  
378 was generally accurate even across the image change in the present experiments:  
379 Participants in the identity matching group very accurately accepted sID/dIMG  
380 exemplars, while the very same sID/dIMG exemplars were rejected – again very  
381 accurately – by participants from the image matching group.

382         We had expected that both using familiar faces and more variable images would  
383 increase exemplar recognition, but reduce or abolish set averaging. By contrast, while  
384 accurate exemplar recognition was indeed observed, set averaging for facial identity  
385 was also robust. This is remarkable since **compelling evidence for set averaging was**

386 previously associated with absent or noisy memory for instances, irrespective of  
387 stimulus type (Ariely, 2001; Chong & Treisman, 2005b; Alvarez & Oliva, 2008;  
388 Haberman & Whitney, 2007, 2009). Accordingly, average set processing has been  
389 thought of as an effective and efficient method to extract only the most important  
390 information from a complex visual scene (Alvarez, 2011). Supporting this idea, it has  
391 been shown that abstractive representations are more precise under distributed than  
392 under focused attention (Chong & Treisman, 2005a), and summary coding of high-level  
393 information can proceed even in the near absence of attention (Alvarez & Oliva, 2009).  
394 In fact, set averaging seems to be so efficient that it can be performed almost as  
395 accurately as coding of a single exemplar (Chong & Treisman, 2003). This research  
396 suggests that precise exemplar and set average representations are incompatible to the  
397 extent that only one representation is extracted at a time, according to task needs. Most  
398 research on set averaging employs settings in which it is difficult for viewers to extract  
399 precise exemplar representations for their experience. Sets were usually quite crowded  
400 or perceptually very similar. Here, task conditions (distinct, familiar faces) allowed  
401 forming of precise exemplar representations, accompanied with strong set average  
402 representations. To our knowledge, this is the first demonstration of robust simultaneous  
403 exemplar and average representations.

404 In our experiments, “present” responses for exemplars exceeded those for sIMG  
405 averages, a pattern that contradicts the commonly described preponderance of average  
406 over exemplar representations. This is clearly not reflecting weak average  
407 representations in the present study, but rather a consequence of increased recognition  
408 of familiar face exemplars (approximately 80%, compared to 30-35% for unfamiliar faces

409 in de Fockert & Wolfenstein, 2009). Our data demonstrate that robust set average  
410 representations can co-exist with precise exemplar representations.

411         Given that ensemble coding is supposed to foster efficient extraction of  
412 information, as suggested by previous studies, a simultaneous extraction of exemplar  
413 and set average representations does not appear to be particularly efficient. The extent  
414 to which exemplar and average representations may draw upon identical or distinct  
415 resources is a matter of current debate. Of particular interest, a recent study suggested  
416 that hierarchical representations in working memory may simultaneously be formed on  
417 multiple levels of abstraction (Brady & Alvarez, 2011). In this study, participants  
418 remembered the size of an individual circle at clearly above-chance precision, but size  
419 judgements were consistently biased towards the average size in the set. Accordingly,  
420 items in working memory could be represented via a combination of set ensemble  
421 statistics and individual exemplar information, with statistical representations increasing  
422 accuracy in situations of inaccurate exemplar memory. Data from the present  
423 experiments are in line with the general idea of a hierarchical representation system.

424         In the experiments reported here, there seems no obvious advantage to be  
425 gained from constructing a representation that merges the individuals. For example,  
426 when interacting with groups, there is no communicative advantage to forming a single  
427 visual representation of all faces. A tentative suggestion is that set averaging could  
428 serve compensatory purposes. For instance, while impaired at recognizing individuals,  
429 participants with developmental prosopagnosia nevertheless showed preserved identity  
430 and expression set averaging for unfamiliar faces (Leib et al., 2012). Additionally, face  
431 recognition performance did not correlate with set averaging performance in that study,

432 suggesting that both tap into distinct processes. While this is an important finding, it  
433 remains unclear how set averaging could compensate for poor *individual face*  
434 recognition. Further research is needed to clarify the relation between the different  
435 coding mechanisms (individual exemplars versus set averages) and their respective  
436 relevance for typical and impaired identity processing of both unfamiliar and familiar  
437 faces.

438           It remains to be seen whether the accurate simultaneous computation of  
439 exemplar and average representations – which were expected to be incompatible – is a  
440 feature of categories beyond faces. These have made a good starting-point, because it  
441 is simple to manipulate familiarity without affecting stimulus structure, and because there  
442 are well-understood technical mechanisms for combining different images. However,  
443 even within the class of faces, a thorough understanding of this phenomenon will require  
444 further investigation into the role of encoding time to test efficiency of set  
445 representations, set characteristics (e.g., male vs. female, own-race vs. other-race) and  
446 other operational variables.

447



448           **Acknowledgements**

449    The authors would like to thank Kathrin Rauscher, Stefanie Luttmann, and Michaela  
450    Kessler for their assistance with data collection, and Stefanie Broncel, Kristin  
451    Gottschlich, Marlena Itz, and Jan Rehbein for helping with stimulus preparation and data  
452    collection in Experiment 1a. This work is supported by a young researchers grant  
453    awarded by the University of Jena to MFN.

454

455

## References

- 456  
457  
458 Albrecht, A. R., & Scholl, B. J. (2010). Perceptually Averaging in a Continuous Visual World:  
459       Extracting Statistical Summary Representations Over Time. *Psychological Science*,  
460       21(4), 560-567.
- 461 Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition.  
462       *Trends in Cognitive Sciences*, 15(3), 122-131.
- 463 Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside  
464       the focus of attention. *Psychological Science*, 19(4), 392-398.
- 465 Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be  
466       represented with reduced attention. *Proceedings of the National Academy of Sciences of*  
467       *the United States of America*, 106(18), 7345-7350.
- 468 Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*,  
469       12(2), 157-162.
- 470 Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory:  
471       Ensemble Statistics Bias Memory for Individual Items. *Psychological Science*, 22(3), 384-  
472       392.
- 473 Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999).  
474       Verification of face identities from images captured on video. *Journal of Experimental*  
475       *Psychology-Applied*, 5(4), 339-360.
- 476 Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar  
477       and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology-*  
478       *Applied*, 7(3), 207-218.
- 479 Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar  
480       face recognition. *Cognitive Science*, 23(1), 1-31.
- 481 Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*,  
482       43(4), 393-404.
- 483 Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual  
484       displays. *Perception & Psychophysics*, 67(1), 1-13.
- 485 Chong, S. C., & Treisman, A. (2005b). Statistical processing: computing the average size in  
486       perceptual groups. *Vision Research*, 45(7), 891-900.
- 487 Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual*  
488       *Cognition*, 11(7), 857-869.

- 489 Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to  
490 Loftus and Masson's method. *Tutorial in Quantitative Methods for Psychology*, 1(1), 42-  
491 45.
- 492 de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces.  
493 *Quarterly Journal of Experimental Psychology*, 62(9), 1716-1722.
- 494 Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of*  
495 *Vision*, 9(11), -.
- 496 Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of  
497 faces. *Current Biology*, 17(17), R751-R753.
- 498 Haberman, J., & Whitney, D. (2009). Seeing the Mean: Ensemble Coding for Sets of Faces.  
499 *Journal of Experimental Psychology-Human Perception and Performance*, 35(3), 718-  
500 734.
- 501 Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when  
502 extracting average expression. *Attention Perception & Psychophysics*, 72(7), 1825-1838.
- 503 Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in*  
504 *Cognitive Sciences*, 4(9), 330-337.
- 505 Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs,  
506 credit cards and fraud. *Applied Cognitive Psychology*, 11(3), 211-222.
- 507 Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd  
508 perception in prosopagnosia. *Neuropsychologia*, 50(7), 1698-1707.
- 509
- 510
- 511

512

## Appendix

| <i>Exp.</i>                             | <i>Effect</i>         | <i>F-statistics</i>          | <i>Effect size (partial <math>\eta^2</math>)</i> | <i>Description</i>   |                           |
|---|-----------------------|------------------------------|--|--|---------------------------|
| E1a                                     | ProbeType             | $F(1,17) = 8.28, p = .010$   | $\eta^2_p = .328$                                | Exemplars ( $M = 37.0\%$ ) > Averages ( $M = 30.8\%$ )                   |                           |
|   | MatchType             | $F(2,34) = 303.84, p < .001$ | $\eta^2_p = .947$                                | sIMG ( $M = 68.2\%$ ) > sID/dIMG ( $M = 22.2\%$ ) > dID ( $M = 11.4\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,34) = 9.67, p = .003$   | $\eta^2_p = .363$                                | Exp <sub>sIMG</sub> > Avg <sub>sIMG</sub>                                | $t(17) = 3.24$ $p = .005$ |
|   |                       |                              |  | Exp <sub>sID/dIMG</sub> > Avg <sub>sID/dIMG</sub>                        | $t(17) = 2.86$ $p = .011$ |
| Exp <sub>dID</sub> = Avg <sub>dID</sub> |                       |                              |  | $t(17) = -0.70$ $p = .493$   |                           |
| E1b                                     | ProbeType             | $F(1,17) = 37.16, p < .001$  | $\eta^2_p = .686$                                | Exemplars ( $M = 63.6\%$ ) > Averages ( $M = 47.5\%$ )                   |                           |
|   | MatchType             | $F(2,34) = 244.53, p < .001$ | $\eta^2_p = .935$                                | sIMG ( $M = 78.9\%$ ) > sID/dIMG ( $M = 63.9\%$ ) > dID ( $M = 23.9\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,34) = 17.51, p < .001$  | $\eta^2_p = .507$                                | Exp <sub>sIMG</sub> > Avg <sub>sIMG</sub>                                | $t(17) = 4.88$ $p < .001$ |
|   |                       |                              |  | Exp <sub>sID/dIMG</sub> > Avg <sub>sID/dIMG</sub>                        | $t(17) = 6.31$ $p < .001$ |
| Exp <sub>dID</sub> > Avg <sub>dID</sub> |                       |                              |  | $t(17) = 3.26$ $p = .005$  |                           |
| E2a                                     | ProbeType             | $F(1,7) = 6.24, p = .041$    | $\eta^2_p = .471$                                | Exemplars ( $M = 36.8\%$ ) > Averages ( $M = 31.1\%$ )                   |                           |
|   | MatchType             | $F(2,14) = 224.48, p < .001$ | $\eta^2_p = .970$                                | sIMG ( $M = 72.5\%$ ) > sID/dIMG ( $M = 19.4\%$ ) > dID ( $M = 10.0\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,14) = 10.42, p = .006$  | $\eta^2_p = .598$                                | Exp <sub>sIMG</sub> > Avg <sub>sIMG</sub>                                | $t(7) = 3.41$ $p = .011$  |
|   |                       |                              |  | Exp <sub>sID/dIMG</sub> = Avg <sub>sID/dIMG</sub>                        | $t(7) = 2.09$ $p = .075$  |
| Exp <sub>dID</sub> = Avg <sub>dID</sub> |                       |                              |  | $t(7) = -1.67$ $p = .134$  |                           |
| E2b                                     | ProbeType             | $F(1,7) = 7.11, p = .032$    | $\eta^2_p = .504$                                | Exemplars ( $M = 59.2\%$ ) > Averages ( $M = 46.7\%$ )                   |                           |
|   | MatchType             | $F(2,14) = 111.43, p < .001$ | $\eta^2_p = .941$                                | sIMG ( $M = 78.7\%$ ) > sID/dIMG ( $M = 55.9\%$ ) > dID ( $M = 24.3\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,14) = 18.26, p < .001$  | $\eta^2_p = .723$                                | Exp <sub>sIMG</sub> > Avg <sub>sIMG</sub>                                | $t(7) = 2.77$ $p = .028$  |
|   |                       |                              |  | Exp <sub>sID/dIMG</sub> > Avg <sub>sID/dIMG</sub>                        | $t(7) = 3.70$ $p = .008$  |
| Exp <sub>dID</sub> = Avg <sub>dID</sub> |                       |                              |  | $t(7) < 0.01$ $p > .999$   |                           |
| E3a                                     | ProbeType             | $F(1,7) = 4.33, p = .076$    | $\eta^2_p = .382$                                | Exemplars ( $M = 24.0\%$ ) = Averages ( $M = 17.5\%$ )                   |                           |
|   | MatchType             | $F(2,14) = 109.16, p < .001$ | $\eta^2_p = .940$                                | sIMG ( $M = 48.2\%$ ) > sID/dIMG ( $M = 9.1\%$ ) > dID ( $M = 4.9\%$ )   |                           |
|   | ProbeType x MatchType | $F(2,14) = 10.50, p = .010$  | $\eta^2_p = .600$                                | Exp <sub>sIMG</sub> > Avg <sub>sIMG</sub>                                | $t(7) = 2.89$ $p = .023$  |
|   |                       |                              |  | Exp <sub>sID/dIMG</sub> = Avg <sub>sID/dIMG</sub>                        | $t(7) = 0.78$ $p = .460$  |
| Exp <sub>dID</sub> = Avg <sub>dID</sub> |                       |                              |  | $t(7) = -1.84$ $p = .108$  |                           |
| E3b                                     | ProbeType             | $F(1,7) = 6.94, p = .034$    | $\eta^2_p = .498$                                | Exemplars ( $M = 50.9\%$ ) > Averages ( $M = 35.0\%$ )                   |                           |
|   | MatchType             | $F(2,14) = 57.55, p < .001$  | $\eta^2_p = .892$                                | sIMG ( $M = 66.4\%$ ) > sID/dIMG ( $M = 41.0\%$ ) > dID ( $M = 21.5\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,14) = 3.14, p = .077$   | $\eta^2_p = .310$                                |  |                           |
| E4a                                     | ProbeType             | $F(1,7) = 1.08, p = .334$    | $\eta^2_p = .134$                                | Exemplars ( $M = 30.7\%$ ) = Averages ( $M = 28.3\%$ )                   |                           |
|   | MatchType             | $F(1,7) = 78.58, p < .001$   | $\eta^2_p = .918$                                | sIMG ( $M = 68.2\%$ ) > sID/dIMG ( $M = 22.2\%$ ) > dID ( $M = 11.3\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,14) = 2.62, p = .111$   | $\eta^2_p = .273$                                |  |                           |
| E4b                                     | ProbeType             | $F(1,7) = 6.77, p = .035$    | $\eta^2_p = .492$                                | Exemplars ( $M = 59.0\%$ ) = Averages ( $M = 40.2\%$ )                   |                           |
|   | MatchType             | $F(2,14) = 63.82, p < .001$  | $\eta^2_p = .901$                                | sIMG ( $M = 73.2\%$ ) > sID/dIMG ( $M = 52.6\%$ ) > dID ( $M = 23.0\%$ ) |                           |
|   | ProbeType x MatchType | $F(2,14) = 6.42, p = .013$   | $\eta^2_p = .478$                                | Exp <sub>sIMG</sub> = Avg <sub>sIMG</sub>                                | $t(7) = 2.33$ $p = .052$  |
|   |                       |                              |  | Exp <sub>sID/dIMG</sub> > Avg <sub>sID/dIMG</sub>                        | $t(7) = 3.17$ $p = .016$  |
| Exp <sub>dID</sub> = Avg <sub>dID</sub> |                       |                              |  | $t(7) = 1.12$ $p = .301$   |                           |

513 **Table S1: Results from all four Experiments' 2x3 ANOVAs and, where applicable, post-hoc comparisons.**

514