1   **Statistical model specification and power: recommendations on the use of test-qualified pooling in**

2   **analysis of experimental data**

3

4   Nick Colegrave[1] and Graeme D Ruxton[2]

5   1. School of Biological Science, University of Edinburgh, Edinburgh EH14 4AJ, UK

6   2. School of Biology, University of St Andrews, St Andrews KY16 9TH, UK

7   **Abstract**

8   A common approach to the analysis of experimental data across much of the biological sciences is

9   test-qualified pooling. Here non-significant terms are dropped from a statistical model, effectively

10   pooling the variation associated with each removed term with the error term used to test

11   hypotheses (or estimate effect sizes). This pooling is only carried out if statistical testing on the basis

12   of applying that data to a previous more complicated model provides motivation for this model-

13   simplification; hence the pooling is test-qualified. In pooling, the researcher increases the degrees of

14   freedom of the error term with the aim of increasing statistical power to test their hypotheses of

15   interest. Despite this approach being widely adopted and explicitly recommended by some of the

16   most widely-cited statistical textbooks aimed at biologists, here we argue that (except in highly

17   specialised circumstances that we can identify) the hoped-for improvement in statistical power will

18   be small or non-existent, and there is likely to be much reduced reliability of the statistical

19   procedures through deviation of type I error rates from nominal levels. We thus call for greatly

20   reduced use of test-qualified pooling across experimental biology, more careful justification of any

21   use that continues, and a different philosophy for initial selection of statistical models in the light of

22   this change in procedure.

23   **Key words**: experimental design, pseudoreplication, model simplification

Introduction

A common approach to the analysis of experimental data across disparate parts of the biological

sciences is test-qualified pooling. A common manifestation of this approach can be summarised as

follows: the researcher fits their data to a model that they select on the basis of the design of their

study and the hypotheses they are interested in testing. After examining the significance of terms in

the model that are not specifically related to the hypothesis currently under investigation, the

researcher then removes non-significant terms from the model, and re-fits their data to this

simplified model. That is, some terms were included in the original model not because they allow an

interesting hypothesis to be tested but because (on the basis of the specifics of the experimental

design allied to previous knowledge of the system) they were expected to explain substantial

portions of the variation. If the data generated in this particular experiment do not suggest that one

or more of these terms are strongly influential then they are dropped from the model, and further

analysis is performed based on a simplified model. Such a simplification process is often seen as

attractive in making presentation of results more compact, in highlighting more influential variables,

and/or in increasing statistical power for exploring the significance of remaining terms. By

simplifying the model in this way, the researcher is effectively *pooling* the variation associated with

each removed term with the error term that will ultimately be used to test their hypotheses. This

pooling is only carried out if statistical testing on the basis of applying that data to a previous more

complicated model provides motivation for this approach, hence the pooling is *test-qualified*. In

pooling, the researcher increases the degrees of freedom of the error term with the aim of

increasing statistical power to test their hypotheses of interest. Despite this approach being widely

adopted and explicitly recommended by some works on data analysis (e.g. [1]), other influential

authors explicitly warned against this practice (e.g. [2]). Here we want to offer some resolution of

this apparent conflict in the literature, in order to help authors, reviewers, editors and readers

evaluate the consequences of pooling in different circumstances. Note that although we couch this

discussion in terms of null-hypothesis statistical testing, the arguments transfer naturally to

50 approaches based on estimation of effect size; our discussion is however focussed on the analysis of

51 data from planned experiments rather than from purely observational studies. The costs and

52 benefits of test-qualified pooling are more clear-cut for planned experiments where potential

53 confounding factors can often be eliminated or controlled for by careful experimental design,

54 removing the need to deal with these factors statistically. Also, planned experiments generally are of

55 what is termed a "confirmatory" nature, where the study specifically aims to test one or more

56 hypotheses known from the outset. Observational studies more often have an "exploratory"

57 motivation involving measuring a broad range of variables and then seeking to rank them in terms of

58 potential importance and influence. We return to these issues in the *Discussion*.

59 **Being clear what pooling is and why you might want to do it**

60 To clarify the issues we consider a specific example. You are interested in the effect of an

61 experimental treatment (a new humidification system) on the growth of individually-potted tomato

62 plants. Your experiment will be conducted in ten small greenhouses at your research station, and the

63 nature of the treatment means that it has to be applied to whole greenhouses. You install the

64 humidification system in five (randomly selected) greenhouses, leaving the other five as controls,

65 and you assay the growth of 40 tomato plants in each greenhouse. In this design the greenhouse is

66 the experimental unit, and any hypothesis test of the treatment should use an error based on the

67 variation amongst greenhouses rather than variation amongst the individual plants. In this case the

68 simplest means of analysis would be to calculate a mean growth rate across the 40 plants in each

69 greenhouse and carry out a one-way ANOVA using these 10 independent data points.

70 However, as a thought experiment, suppose that we somehow knew for a fact that growth

71 conditions (in the absence of our treatment manipulation) were absolutely identical amongst our

72 greenhouses. In this imaginary situation we might argue that, since greenhouse-to-greenhouse

73 variation is not confounded with any treatment effect we can use the growth measures from the

74 individual plants as independent data points in our analysis. This will result in a substantial increase

75   in our degree of freedom, and consequently our statistical power to detect treatment effects. Of

76   course in reality, we cannot usually know with certainty whether our greenhouses vary, and this has

77   led to the development of methods for test-qualified pooling. In this case, we would start by fitting

78   the nested model defined by the design of our study (with individual plants being nested within

79   greenhouse). This would include the treatment term, a nested term for the variation amongst

80   greenhouses in the same treatment group, and a second error term corresponding to the variation

81   amongst plants in the same greenhouse. The key to test-qualified pooling is that the set of data itself

82   influences the nature of the analyses performed on it. If initial analysis of the full model indicates

83   substantial variation amongst greenhouses, then the significance of the treatment term is tested

84   using the variation amongst greenhouses as its error term with 8 df. However, if there is no evidence

85   of substantial greenhouse-to-greenhouse variation in this initial analysis then the among-

86   greenhouse and the true error variations are pooled, and this combined error term with 398 df is

87   used then to provide a test of the treatment effect that is expected to benefit from higher statistical

88   power (see [3-5] for commonly-cited texts that recommend this approach). The justification that

89   advocates of test-qualified pooling give for this approach is that in the absence of any greenhouse

90   effect, the among-greenhouse and the within-greenhouse error terms are both estimating the same

91   thing, and so by combining them we get a better estimate than we would estimating the two

92   separately.

93   However pooling is not limited to nested designs. Continuing with tomatoes and greenhouses, you

94   now want to compare the effects of four different growing media in individually-potted tomato

95   plants rather than the effect of humidity. To gain a sufficient sample size for the experiment you

96   have to use three different greenhouses to keep all the plants, but because your treatments can now

97   be applied randomly to individual plants, you randomly allocate equal numbers of plants to each

98   treatment in each greenhouse leading to a randomised block design (with specific greenhouse

99   identity as the blocking factor, with three levels). The statistical model implied by this design would

100   include terms for both treatment applied to a plant and the specific greenhouse a plant was kept in,

4

101   as well as a treatment-by-greenhouse interaction and an amongst-plant error term based on the

102   variation amongst individual plants within the same treatment-greenhouse combination. Depending

103   on the exact hypothesis we wish to test, the appropriate error term for our treatment effect will be

104   either the interaction term, or the amongst-plant error term [6], but in either case, if the interaction

105   term is not significant, we might chose to pool its variation with the amongst-plant error term prior

106   to testing the treatment effect. Similarly, we might then decide that if the greenhouse term is also

107   non-significant, we would add that source of variation and its associated degrees of freedom to our

108   error pool. In either case, we would be carrying out test-qualified pooling.

109   Another form of pooling can involve the initial test that triggers whether pooling is used or not being

110   entirely separate to the model testing the hypotheses of interest. To illustrate this, we return to the

111   experiment above comparing the effects of four different growing media on individually-potted

112   tomato plants. Imagine that, because of a change of supplier at your institute, you ended up using

113   two different but broadly similar types of pots to grow the tomatoes in. Plants are randomised to

114   pot type as well as to growth medium and greenhouse. You really do not expect type of pot to

115   influence growth rates, but just to be careful you first of all perform a t-test comparing growth rates

116   across the two types of pot. Your plan is that if (as you expect) this t-test reveals no evidence of a

117   difference, you report this and use this test as justification for pooling data across the two pot types

118   in your subsequent analyses. However if it does reveal evidence of a difference then you will either

119   add pot-type as a factor in subsequent analyses or carry out separate analyses for the two types of

120   pot. Again, there is the potential for pooling driven by the results of a pre-test, so this scenario is

121   another manifestation of test-qualified pooling.

122   **Why is test-qualified pooling controversial?**

123   The case against pooling was made most forcefully and explicitly in the biological literature by Stuart

124   Hurlbert primarily in relation to its use in nested designs [2]. Hurlbert coined the expression

125   *pseudoreplication* for the situation where authors treat data-points that are not independent as if

126 they were independent in their data analysis. His original paper on this [7] has been cited over 6000

127 times and has been hugely influential in the design of data collection and the analysis of data

128 spanning all of biology. Hurlbert considers the pooling of errors in a nested analysis to be a form of

129 pseudoreplication, a form that he calls *test-qualified sacrificial pseudoreplication*. He argues that

130 pooling biases p-values downwards and biases confidence intervals towards being too narrow. He

131 further argues that demanding a higher p-value than 0.05 in the initial test before pooling (a process

132 often called "sometimes pooling") reduces but does not eliminate these problems. An analogous

133 argument can be made against pooling interaction terms with error terms when analysing

134 randomised block designs [6]. However, even in situations where pooling might not be regarded as

135 analogous to pseudoreplication (e.g. pooling an interaction between two fixed factors prior to

136 testing the main effects), type 1 error rates can be increased (as we will see below). Despite this,

137 pooling is still regularly practiced, and is recommended in influential statistics textbooks aimed at

138 biologists (e.g. [3-5]) and research papers on statistical methodology (e.g. [2,8]). In the next section

139 we argue that both philosophically and pragmatically there are strong arguments for siding with

140 Hurlbert.

141 **The philosophy and pragmatics of pooling**

142 The two main philosophical arguments against pooling are well articulated by Newman et al. [7], and

143 can be explained in the context of our greenhouses and growth media example. Firstly, if we use

144 pooling, then the way that we test for an effect of growth medium becomes conditional on the data,

145 but that conditionality is not acknowledged in the associated p-values. That is, whether we test the

146 effect of medium in a model with or without a *greenhouse* term will be determined by the data.

147 Philosophically, p-values are probabilities based on a very large number of notional replicates of

148 exactly the experiment under investigation. So imagine that we repeat the full experiment and

149 analysis of the resulting data again and again. In replicates of this experiment, if we adopt a test-

150 qualified pooling approach then sometimes the analysis will test the main hypothesis one way and

151 sometimes the other. For each form of the analysis, that particular analysis will be implemented only

152 for a specific subset of replicate experiments determined by the patterns of data in that replicate

153 experiment. Importantly, this is a biased sample of all the possible replicate experiments in terms of

154 properties of the sample. Yet the test is predicated on the assumption that it is applied to data from

155 an experiment drawn without bias from the population of all possible replicates of this experiment.

156 It is this mismatch that leads to lack of control of type I error and of confidence intervals. Secondly,

157 by pooling (no matter what critical value we compare the calculated p-value against) we are

158 accepting that the null hypothesis that there is no effect of *greenhouse* is true, and the whole

159 philosophy of null-hypothesis statistical testing is that the null hypothesis is never accepted as true,

160 rather we might either reject it or find that we do not have sufficient grounds to reject it. Thus, from

161 a purist philosophical perspective pooling should not be recommended.

162 We next ask if there is a pragmatic argument that says that pooling may have some less-than-ideal

163 properties, but pooling leads to relatively mild misbehaviours that are sometimes outweighed by the

164 (enhanced power) benefits of pooling. There is no underlying theory to give general and definitive

165 answers to the issue of pragmatics raised above; all we have to go on are a number of numerical

166 explorations of specific cases. However, the consensus in this literature is that (i) pooling can cause

167 actual type one error rates to be very different from the nominal value, and (ii) there is no consistent

168 and substantial increase in power to compensate. Walde-Tsadik & Afifi [9] explore the effect of

169 always pooling when one factor is associated with a p-value above 0.05, and also of "sometimes

170 pooling" when the required critical value was higher than 0.05 in two-way ANOVA random effects

171 models. They found that both procedures very rarely offered adequate control of type-1 error rate

172 and even less commonly lead to significant improvement in power to test for an effect of the other

173 factor. Hines [10] performed extensive simulations and concluded that for multifactorial ANOVA

174 "the conditions for pooling to be even potentially rewarding are more restrictive than might be

175 expected, and power improvements are generally lower". Janky [11] performed a similar analysis of

176 split-plot designs and concluded that "pooling generally inflates Type I error and offers at best

177 insubstantial gain in power (and often power loss) relative to the nominal test." Even when using a

178 conservative "sometimes pooling" value of $\alpha = 0.35$ to trigger pooling, Janky found the type I error

179 rate in subsequent tests on pooled data rose from the nominal 5% to generally somewhere between

180 7% and 11%. This study was interesting for highlighting that pooling actually led to a reduction of

181 power more often than it lead to a substantial gain in power; this occurs because the increase in

182 inherent variation caused by pooling dominates any effect of increased degrees of freedom devoted

183 to exploring remaining factors. Figure 1 shows examples of deviations in both directions from the

184 nominal 5% level for type I error rates generated by simulations of our whole-greenhouse-treatment

185 thought experiment. In exploring our model we found that small changes in parameter values could

186 lead to substantial change in the magnitude and direction of deviations from the nominal level. It is

187 difficult to make generalisations about the circumstances under which deviations will be strongest.

188 In common with the other studies discussed directly above, we found that the direction and

189 magnitude of deviations are driven by a complex interaction between structure of the experimental

190 design, aspects of the shape of the underlying "population" from which sample values are obtained,

191 and sample sizes. Also, as the highest line in Figure 1 illustrates, relationships with parameter values

192 can be non-monotonic.

193 **Discussion and Conclusion**

194 Use of test-qualified pooling is widely adopted, but its prevalence across biological sciences is

195 patchy. For example, it is much less commonplace in clinical trials; where often statistical analyses

196 have to be specified in pre-registration of trials, and thus scope for flexibility in data analysis is

197 reduced. Test-qualified pooling is also relatively uncommon in the agricultural sciences, where

198 particular designs and modes of analysis that avoid issues of pooling are traditional; and the

199 statistical software package *Genstat* is commonly used, which is particularly suited to forms of

200 analyses that avoid test-qualified pooling.

201 We do still consider that test-qualified pooling is over-used in biology. Simply, in "confirmatory

202 studies" based on designed experiments where we aim to test specific hypotheses (or estimate

203 specific effect sizes) we do not recommend pooling under any circumstances. The often-modest

204 expected increases in power from pooling do not make it an attractive option when its drawbacks

205 are taken into account. Apart from statistical power, the other attraction to pooling is simplification

206 of the presentation of results, but we feel that this will never be sufficient grounds for justifying the

207 process. We would only recommend pooling in such a study if the decision to consider test-qualified

208 pooling was made on the basis of a prior simulation study that aimed at evaluating the

209 consequences of pooling for Type I and Type II error rates. We have yet to see an example of a study

210 that provided such a justification for pooling.

211 As we mentioned in the *Introduction*, it is not as easy to offer clear and simple guidance on pooling

212 in purely observational studies, and studies where the researchers' aims are more focussed on

213 exploration or prediction than on testing specific hypotheses. However, in such situations pooling

214 can be seen as a facet of *model selection* – which is an area of considerable activity in applied

215 statistics. A particularly useful introduction to the concepts involved is that of Chatfield [12]. He

216 makes the point that if the same data-set is used to both select the most appropriate model from a

217 suite of alternatives and also to fit that model, then the interpretation of the fitted model should be

218 quite different from circumstances where the form of the model is decided upon first and only then

219 is the data applied to fit that model. Where there is uncertainty as to the most appropriate model,

220 then there are methodological developments in *model averaging* that can acknowledge this ([13]

221 and [14] offer good introductions for the biologist). A failure to properly acknowledge model

222 uncertainty when the same data is used to select and fit the model can read to very unreliable

223 inferences ([12],[15],[16]).

224 Despite the complexity of the literature on model selection and model uncertainty, we feel that we

225 can offer a general opinion on the utility of test-qualified pooling outside designed experiments. For

226   more exploratory studies where the intention is to identify factors that might be of interest, rather

227   than to test specific hypotheses, then test-qualified pooling might be more attractive; since

228   researchers may be willing to live with loss of control of type I error rates if this helps boost their

229   statistical power to flag up factors of interest. That is, they may be prepared to suffer higher rates of

230   false positives to boost their likelihood of detecting real effects. We expect that these power gains

231   may sometimes be considerable for nested-designs. However for other types of design the literature

232   discussed in the last section should serve as a caution that power gains from pooling may be small or

233   non-existent. Our view is that even in exploratory studies, test-qualified pooling cannot really be

234   recommended except perhaps where the design is nested and where the size of the experiment was

235   reduced from its ideal size by practical constraints or unforeseen adverse circumstances.

236   Where does this leave the experimenter in our tomato plant example who just wanted to be diligent

237   and reassure themselves and their readers that there was no effect due to two different types of

238   pots being used? They have to make a decision about how important this check is to them. If they

239   feel that it is worth investing a few degrees of freedom in, then they should include *type-of-pot* as a

240   factor in their analysis and pay a modest cost in reduced power to test the hypothesis (comparing

241   different growth media) that they are really interested in. Alternatively, they may decide that careful

242   experimental design and explanation of that experimental design should allay concerns about

243   differential effects of pot types sufficiently that there is no need for formal statistical testing. More

244   generally, we all have to accept that there are no free statistical analyses, and think hard about

245   which factors to include in any model. This is analogous to the decision to block on a given variable

246   in experimental design. It is only advantageous to block on variables that explain a substantial

247   fraction of variation between experimental units, otherwise the degrees of freedom lost in including

248   that blocking term are not compensated for by effective partitioning of variation into error and other

249   terms.

Sometimes we can make a strong enough case based on careful experimental design (especially use of randomisation), biological intuition, and logical reasoning for why we can safely assume that some potentially influential factors are in fact very unlikely to be important in our study, and so we omit them from our statistical procedures. In fact, we do this all the time. In our example the researcher felt no need to test whether which shelf on a greenhouse a pot was placed on had an effect, or what side of the greenhouse, or how near to the door of the greenhouse it was. Sometimes we will feel that we cannot make a sufficiently strong case this way, and we should then include that factor in our model and explore its effects statistically. As so often in the design and analysis of scientific studies, there are no black-and-white rules for which factors to include in your statistical model; we need to think hard about it and justify our choices in terms of experimental design, understanding of underlying biology and logical reasoning. This should be good news: model selection should be much more about biology than about mathematics and probability theory – and biology is what we are interested in.

Author contributions: This article was conceived, developed and written equally by both authors.

References

[1] Schank, J. C., & Koehnle, T. J. (2009). Pseudoreplication is a pseudoproblem.*Journal of Comparative Psychology*, *123*(4), 421.

[2] Hurlbert, S. H. (2009). The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology*, *123*(4), 434.

[3] Sokal, R. R., & Rohlf, F. J. (1995). Biometry: the principals and practice of statistics in biological research. *WH Freeman and Company, New York*.

[4] Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

275   [5] Zar, J. H. (1999). *Biostatistical analysis*. Pearson Education India.

276   [6] Newman, J. A., Bergelson, J., & Grafen, A. (1997). Blocking factors and hypothesis tests in

277   ecology: is your statistics text wrong?. *Ecology*, *78*(5), 1312-1320.

278   [7] Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological*

279   *monographs*, *54*(2), 187-211.

280   [8] Crits-Christoph, P., Tu, X., & Gallop, R. (2003). Therapists as fixed versus random effects-some

281   statistical and conceptual issues: a comment on Siemer and Joormann (2003).Psychological Methods

282   8, 518-523.

283   [9] Wolde-Tsadik, G., & Afifi, A. A. (1980). A comparison of the "sometimes pool","sometimes switch"

284   and "never pool" procedures in the two-way ANOVA random effects model. *Technometrics*, *22*(3),

285   367-373.

286   [10] Hines, W. G. S. (1996). Pragmatics of pooling in ANOVA tables. *The American Statistician*, *50*(2),

287   127-139.

288   [11] Janky, D. G. (2000). Sometimes pooling for analysis of variance hypothesis tests: A review and

289   study of a split-plot model. *The American Statistician*,*54*(4), 269-279.

290   [12] Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal*

291   *Statistical Society, Series A, 158, 419–466.*

292   [13] Whittingham, M. J., Stephens, P. A., Bradbury, R. B. & Freckleton, R. P., (2006). Why do we still

293   use stepwise modelling in ecology and behaviour?. *Journal of Animal Ecology*, *75*,.1182-1189.

294   [14] Richards, S. A., Whittingham, M. J. & Stephens, P. A., (2011). Model selection and model

295   averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and*

296   *Sociobiology*, *65*, 77-89.

297   [15] Blanchet, F. G., Legendre, P. & Borcard, D., (2008). Forward selection of explanatory

298   variables. *Ecology*, 89, 2623-2632.

299   [16] Mundry, R. & Nunn, C. L., (2009). Stepwise model fitting and statistical inference: turning noise

300   into signal pollution. *The American Naturalist*, *173*, 119-123.

301

302

303   **Figure legend**

304   Figure 1: To illustrate how the type 1 error rate can be affected by test qualified pooling we

305   examined simulated data sets for both 4 (broken line) and 10 (solid line) greenhouses. In both cases,

306   equal numbers of greenhouses were allocated to control or treatment conditions (but condition had

307   no effect on plant growth), and 40 plants were measured in each greenhouse. We also examined the

308   effect of two different alpha levels for the pooling decision (recommended in [3]: open circles = 0.25

309   and closed circles = 0.75), and several different levels of among-greenhouse variation (σ^2). Under

310   many different parameter combinations the actual type 1 error rate differs from the desired value of

311   0.05, sometimes substantially.

312   Plant growth rates were calculated as a baseline value (10) plus an individual deviation drawn from

313   N(0,1) plus a greenhouse-deviation drawn from N(0,σ) and the same for all plants in a given

314   greenhouse. We analysed each data set in two ways. First we carried out a nested analysis of

315   variance in which the treatment mean square was tested over the among-greenhouses within-

316   treatment mean square. The same analysis tested for variance among greenhouses by comparing

317   the among-greenhouses mean square to the amongst-plants error mean square. Second we carried

318   out an analysis in which data from all greenhouses was pooled. The decision as to which P value to

319   use for our actual hypothesis test for the effect of the treatment was based on the significance of

320   the among-greenhouse test at one of two alpha levels. If this test was significant at the appropriate

321   alpha level we used the P value from the nested model, otherwise we used the P value from the

322   second model. This process was repeated 100000. The proportion of these runs that gave a P value
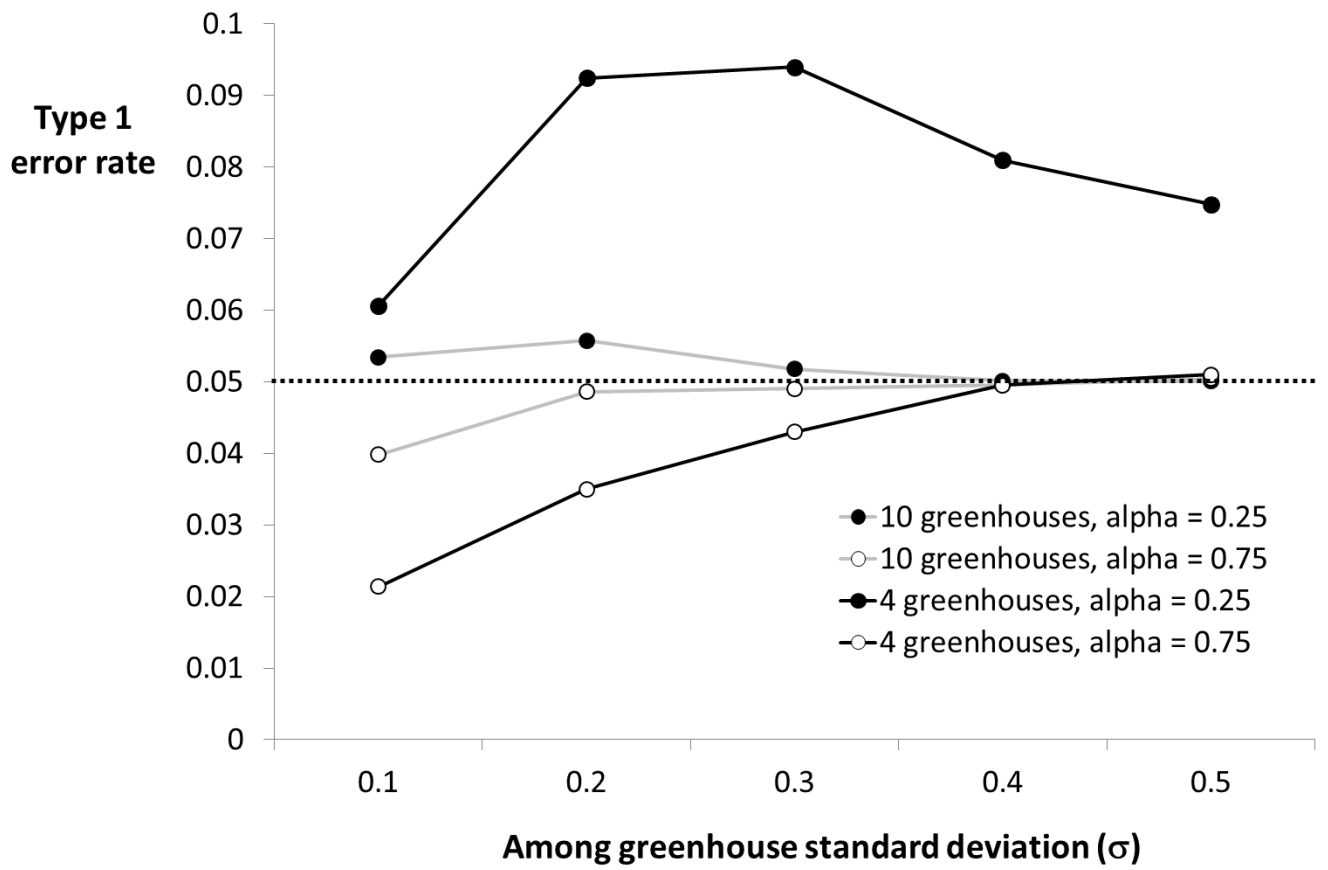
323   of less than 0.05 (i.e. a false positive at alpha = 0.05) is an estimate of the type 1 error rate. The

324   simulations were carried out in R, with the AOV function being used for the analyses.

325

326

327

328   **Figure 1:**

329