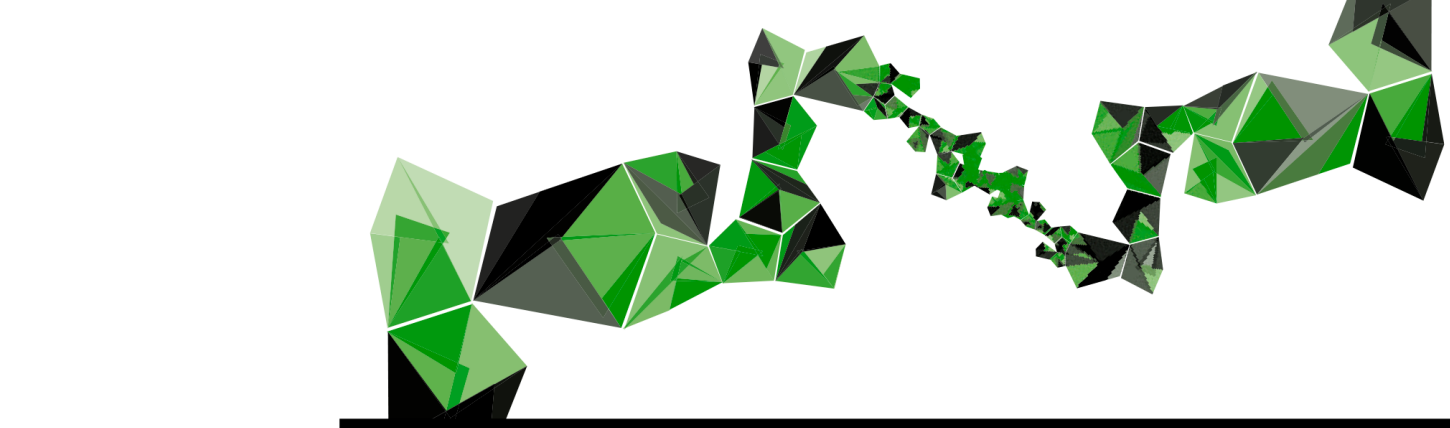

BACHELOR THESIS

Intra-Class Correlation Testing to Examine Intra-Group Differences

Iris Gabriëlle Maria Jongmans

DEPARTMENT OF RESEARCH METHODOLOGY, MEASUREMENT AND DATA ANALYSIS (OMD)

EXAMINATION COMMITTEE
dr. ir. G. J. A. Fox
K. Klotzke, MSc

Enschede, January 2021

UNIVERSITY OF TWENTE.

Abstract

Hierarchical data structures are the norm in educational and social sciences. The intraclass correlation coefficient (ICC) quantifies the relative variation explained by clustering observations in groups (Snijders & Bosker, 2011). The ICC is used to measure the strength of the hierarchical dependence, but the ICC might have more to offer than its conventional use. The ICC is studied as a tool for the measurement for intra-group differences.

In a simulation study, the performance of the ICC Bayes factor (BF) test of Mulder and Fox (2019) is examined for different sample sizes. For small sample sizes, the BF test showed accurate results. Furthermore, a real data trial run for the ICC test was performed on the FOCUS data (van Geel, Keuning, Visscher & Fox, 2016).

The conclusion, the ICC does convey intra-group differences and offers additional information to the mean scores in nested data. Furthermore, the ICC test can process the ICC, even when it becomes negative, to a sufficient degree. Bearing in mind that, the evidence categories used and thus the BF values of the test require further investigation when testing the ICC.

Introduction

Within the social sciences, data often has a hierarchical structure (Paterson & Goldstein, 1991). Especially in academic areas such as student performance and/or growth, this hierarchical structure is the norm (Goldstein in Socha, 2013). It is important to account for this nested data structure in statistical tests (Dyer, Hanges & Hall, 2005; Socha, 2013). When a grouping effect is ignored the standard errors are generally too small, thus conclusions about the statistical significance of a treatment effect might be overestimated (Mulder & Fox, 2019). A grouping effect is present when the intraclass correlation coefficient (ICC) is greater or lower than zero, indicating the independence assumption is broken (Snijders & Bosker, 2011).

The ICC was introduced by Fisher as a measure of reliability and has received attention as such in the research community (Bartko, 1976). However, it can also be used as a value for the degree of resemblance between groups nested in clusters (Snijders & Bosker, 2011), in that context it is also known as the variance partitioning coefficients. The ICC, often written as ρ, quantifies the relative variation explained by clustering observations over the groups. In a standard two-level model, the ICC is the proportional variance explained by the second level (the group) in comparison to the total variance, as per the formula.

$$\rho = \frac{variance\ between - group}{variance\ between - group + variance\ within - group} = \frac{\tau_c}{\tau_c + \sigma^2_c}$$

When there are multiple sets of grouped observations, for instance children are nested in schools in each country, then the school will be referred to as a group and the country as a cluster. The ICC in cluster *c* indicates the proportion of variance explained by the grouping of observations in cluster *c*. The larger the ICC, the more variance is explained by the grouping of observations. An ICC of 1 indicates that all variance is explained by the groups.

This structure can also be viewed as a three-level hierarchy, observations define level-1 units, the groups level-2 units and the clusters the level-3 units (Hox, Moerbeek & van de Schoot, 2017). However, in this study, the level-3 units are not considered to be randomly sampled units and do not introduce another variance component. For example, a measurement (level-1 observation) of children who are nested within a family (level-2 groups). When different instruments are used, to measure the same construct, the families whose children are measured with the same instrument are assigned to cluster *c*. Then, the ICC of cluster *c*

represents the proportional variance explained by differences between families assigned to cluster *c*. This view on the hierarchical design is used in the current study.

The ICC can offer additional information about intra-group differences on top of the mean component because, it deals with the ratio of between-groups variance and within-group variance. Which can show a change in e.g., the growth of a latent ability and on which level the growth was present. For instance, when groups receive a different treatment, then the ICC for groups receiving the same treatment indicates the presence of random variation in the treatment across groups. The treatment works better for some groups than for others. The difference in ICCs across groups receiving different treatments can be used to measure differences in the variability of the treatment effects.

Furthermore, following the approach of Nielsen, Smink and Fox (2021) the ICC can become negative e.g., on occasions involving competition between groups for scares resources. Regrettably, few researchers are equipped to manage a statistical test despite a negative value and discard them by changing them to zero (Nielsen et al., 2021). This unknowingly leads to incorrect standard errors, Type-I errors and confidence intervals which can lead to wrong conclusions. Towards that end, the ICC should be carefully considered when used for hypothesis testing. Especially since ICC testing has only recently emerged in the field of statistics and therefore it must be further explored (Mulder & Fox, 2019: Zhang, 2019).

To explore the functionality of the ICC in hierarchical data analysis, the performance of the ICC test was examined through a simulation study. The ICC test is also tried on a real data study from van Geel, Keuning, Visscher and Fox (2016). The objective of this study is to examine whether the ICC test can be used to examine hypotheses based on the ICC.

Method

Background ICC test

The object of the simulation study was to analyse the performance of the ICC test for different hierarchical data sets. In other words, how sensitive the ICC test is and if results lead to correct conclusions. For the confirmatory tests, data was generated for various ICCs. Special attention was given to scenarios where the ICC was zero or negative. Furthermore, the ICC test was tried on a real data study wherein a multilevel data structure was present. A linear mixed effects (LME) model was fitted to estimate between-cluster variance. The ICC test used the LME results to evaluate hypotheses about the ICC. In the real data trial run, a school intervention was performed to improve student achievement. The ICC of schools prior to and during the intervention were evaluated to examine the presence of an intervention effect across schools during the intervention period (FOCUS study of van Geel et al., 2016).

The ICC test is a Bayes factor (BF) test and evaluates the data evidence of the null hypothesis against an alternative hypothesis. The BF is the ratio of marginal likelihoods of the data given the model in comparison to a competing model (Kosheleva, Kreinovich, Trung & Autchariyapantikul, 2019; Page & Satake, 2017). For example, if hypothesis A is compared to hypothesis B and the BF is 5 then hypothesis A is 5 times more likely under the data than hypothesis B. However, using the BF contrasts with the current standard in social sciences and related educational programs, the Null Hypothesis Statistical (or Significance) Testing (NHST). Its popularity has, among other reasons, come from its inclusion in standard statistical packages like SPSS (Andraszewicz et al., 2015; Quintana & Williams, 2018), unlike the BF.

However, NHST has received criticism over the years. A fundamental problem is that NHST is based on two incompatible theories, the theories of Fisher and Neyman-Pearson (Page & Satake, 2017; Wasserstein & Lazar, 2016). Page and Satake (2017) describe that NHST is taught in academic settings as derived from one coherent theory, despite that the founders of the initial theories disputed the others approach. Furthermore, NHST is designed to find support for an alternative hypothesis. The NHST does not allow for a calculation of evidence in favour of the null hypothesis. However, the BF enables such findings (Hoijtink, Mulder, van Lissa & Gu, 2019). Especially when concerning the ICC, the accuracy of the BF is recommended over NHST (Mulder & Fox, 2019).

Furthermore, Andraszewicz et al. (2015) and Page and Satake (2017) state that the use of *p*-values is sensitive to the intention of the sampling plan whereas the BF is not. In Bayesian analysis the interpretation of separate entities such as *p*-values is not necessary, which makes it less complicated to evaluate and thus less prone to interpretation mistakes (Kass & Raftery, 1995), if the prior probabilities were chosen after considering the empirical context of the hypotheses (Morey, Romeijn & Rouder, 2016). Often those prior probabilities are the default uniform priors i.e., all hypotheses have an equal prior probability (Hoijtink, Mulder, van Lissa & Gu, 2019).

Another feature of the BF is that it has the ability to change per added datum until it finally reaches a point where all factors point towards the most likely hypothesis. However, because not all sample sizes are big enough to reach this point some criteria rules of thumb have been agreed upon. In Bayesian analysis the consensus is that there is a minimum level of the BF necessary to decide if the evidence in support of a hypothesis is statistically strong enough. These thresholds for the evidence categories were devised by Jeffreys (1961). The language used in the original table of Jeffreys was adjusted by Andraszewicz et al. (2015) to be a bit shorter, which is why those terms were chosen (see Table 1). While anything higher or lower than 1 indicates a more probable hypothesis based on the data, the BF of between 3 and 10 represents moderate evidence. Therefore, a BF of around 3 was chosen as an aiming point for confident results of the ICC test.

Table 1

Evidence Categories for the Bayes Factor BF_{12} (Adjusted From Jeffreys, 1961)

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for M_1
30	-	100	Very strong evidence for M_1
10	-	30	Strong evidence for M_1
3	-	10	Moderate evidence for M_1
1	-	3	Anecdotal evidence for M_1
	1		No evidence
1/3	-	1	Anecdotal evidence for M_2
1/10	-	1/3	Moderate evidence for M_2

1/30	-	1/10	Strong evidence for M_2
1/100	-	1/30	Very strong evidence for M_2
	<	1/100	Extreme evidence for M_2

Note. Reprinted from "An Introduction to Bayesian Hypothesis Testing for Management Research" by S. Andraszewicz, B. Scheibehenne, J. Rieskamp, R. Grasman, J. Verhagen, and E. J. Wagenmakers, 2015, *Journal of Management*, *41*(2), p.521-543.

The only caveat for interpreting the resulting values is that the true model needs to be included in the candidate models for the BF to be effective (Vrieze, 2012). While a true model can be present in a simulation study, the discussion on this caveat is beyond the scope of this paper please refer to Morey et al. (2016) for more insight. In this simulation study, all possible models were included to ensure effectiveness. The candidate models were hypothesis 1 (H1) the ICC of cluster 1 is smaller than cluster 2 and hypothesis 2 (H2) the ICC of cluster 1 and cluster 2 are equal. Lastly, the complement of H1, hypothesis 3 (H3) the ICC of cluster 1 is greater than cluster 2.

The BF is calculated by comparing all hypothesis with H1. This initially led to every BF to be 1 for H1. Therefore, the BF for H1 is computed with respect to H3. With that in mind, the BF and thresholds can be regarded as effective for this study.

The ICC test

The analyses were performed in the statistical software R (R Core Team, 2013). R-code was made to simulate data, which were analysed using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). The package *Bfpack* (Mulder et al., 2019) was used to compute the ICC test, which can both be downloaded from CRAN. For the generation of data, the ICC was defined under the conceptual multilevel model. The Y_{cij} is the score for item j (j=0, …, n) of student i (i=1, …,N) within cluster c (c=1, …, C), which is represented as

$$Y_{cij} = \mu_c + \beta_{cij} + \mu_{ci} + \varepsilon_{cij}$$

where μ_c is the average of all students in cluster c and μ_{ci} the average of student i in cluster c. The student average in cluster c is assumed to be normally distributed with mean zero and variance τ_c. The β_{cij} represents the difficulty of item j. Due to the positive sign, for higher β

values the item becomes less difficult. The ε_{cij} is the error component, which is assumed to be independently and normally distributed with variance σ^2. The item responses are nested within the student, and the student represents a group which are settled in cluster *c*. The ICC examined in this study represents the proportion of variance explained by the grouping of responses by students. Although this study is focused on an educational setting, this model and the ICC test could be applied to other nested data settings.

Method of analysis

The BF for the ICC was computed under various sample sizes, see Table 2. The group sizes were chosen to examine the performance of the ICC test for small sample sizes. As an exception, for *N*= 2, 50 observations were simulated for each group. This was done to imitate reality whereas the sample size becomes smaller usually more data is gathered per group. For the other sample sizes, 5 observations per group were simulated in the balanced design condition. Although these sample sizes are generally regarded as too small for ICC testing (Maas & Hox, 2005), the performance of the ICC test was examined for the small sample size condition.

Table 2

Sample Sizes per Group (N) with Observations (n) and τ_1 for Cluster 1 (C1) and τ_1 Cluster 2 (C2) over 500 Replications.

Balanced data C1, C2		
N	*n*	(τ_1, τ_2)
2,2	50,50	(.5,.5) (.3,.5) (.1,.5)
5,5	5,5	(.5,.5)[a] (.3,.5) (.1,.5)[b]
10,10	5,5	(.5,.5) (.3,.5) (.1,.5)
50,50	5,5	(-.1,0) (0,.5) (-.1,.5)
Unbalanced data C1, C2		
N	*n* for both clusters	(τ_1, τ_2)
5,5	Centred around 20	(.3,.3)
10,10	Centred around 20	(.3,.3)

20,20	Centred around 20	(.3,.3)
15,20	Centred around 20	(.5,.3) (.5,.1)

[a]For this condition, a similar condition was examined with a similar sample size of N= 5 with n= 10 observations and the τ of (.3,.3) for C1 and C2 respectively over 10,000 replications. [b]For this condition, a similar condition was examined with a similar sample size of N= 5 with n= 10 observations, the same τ_1 and τ_2 were used over the 10,000 replications.

The ICCs across clusters were centred around .3 to avoid lower-bound issues that normally arise when ICCs are close to zero. However, in educational research ICCs are often around .2 or lower (Hox & Maas, 2001). Therefore, special attention was paid to the small ICC values of around .1. For the balanced design condition, τ= .5 was examined to evaluate the accuracy of the test to detect differences between ICCs when they are high.

Furthermore, following Nielsen et al. (2021), data was simulated with a negative ICC and one of zero in comparison to a high ICC, which are considered *special* ICCs. Another point of interest was imbalanced data. Imbalanced data sets were created by generating random missing number of observations, centred around 20 observations per group. Imbalance was studied further by varying the ICC values over different number of groups in the clusters (Table 2).

For each condition, the recommended default uniform priors by Mulder and Fox (2019) were used. Data sets were generated and then the Bayes factors and posterior probabilities were computed for the considered hypotheses. The posterior probabilities were calculated for the hypotheses with the same reference category. The number of replications was at first limited to 500, due to the required computation time. However, to examine the sensitivity of the results to the 500 replications, several runs were made with 10,000 replications. If the 10,000 replications were comparable to the cut-off points of the limited replications the study would be continued with 500 replications per condition.

The percentages of the computed Bayes factors that pointed towards the correct hypothesis were also compared over the different sample sizes. This was done to evaluate if the chances of being pointed towards the correct hypothesis was high enough for the ICC test to be applied to the FOCUS data.

Results

By evaluating the BF over the various sample sizes and number of replications in simulating data the performance of the ICC test was examined. Towards that end, this study was done to see what the advantages are for testing ICCs. The average results of the different sample sizes (see Table 3) show that as the sample size increases the BF increases accordingly. Furthermore, the larger the difference in ICCs became between the two clusters the stronger the evidence pointed towards the correct hypothesis.

However, in Table 3 it shows that discriminating between H1 and H2 proved to be rather difficult when the difference between τ_1 and τ_2 were small (.3,.5). As can also be seen in the changes in posterior probabilities. An explanation for this may be that the estimated values over the generated data contained more similar samples that made distinguishing difficult. For example, as data for hypothesis 1 and 2 occurred equally often after 500 replications. When the averages were calculated, the differences were levelled out.

Table 3

Estimated Posterior Probability and Bayes Factor for Varying Sample Sizes Across 500 Replications. Sorted per τ_1 for Cluster 1 (C1) and τ_2 Cluster 2 (C2).

τ C1,C2	.5,.5			.3,.5			.1,.5		
Hypothesis	N2n50	N5n5	N20n5	N2n50	N5n5	N20n5	N2n50	N5n5	N20n5
Posterior probability									
H1	0.334	0.320	0.241	0.337	0.342	0.311	0.356	0.400	0.578
H2	0.352	0.415	0.548	0.361	0.416	0.553	0.376	0.412	0.369
H3	0.314	0.265	0.211	0.302	0.242	0.136	0.268	0.188	0.053
Bayes Factor									
H1	1.003	1.041	1.001	1.056	1.174	1.571	1.231	1.558	5.307
H2	1.082	1.440	2.617	1.106	1.364	2.378	1.114	1.292	1.123
H3	0.997	0.960	0.999	0.947	0.851	0.637	0.812	0.642	0.188

Note. The BF is calculated by comparing the likelihood of a hypothesis (H2,H3) in comparison to the H1, and the BF for H1 represents H1 in comparison to H3. N represents number of groups, n the number of observations per group.

To confirm that the number of replications was sufficient to grasp the pattern of the ICC test similar parameters were used for the generation of 10,000 replications. A comparison was drawn between the larger data set with $N = 5$ along 10 observations, with $\tau_1 = .1$ and $\tau_2 = .5$ (Figure 1). For the smaller data set $N = 5$ with 5 observations was used (Figure 2). The results did not change with an increase in replications. As illustrated by the Figures, the pattern became more detailed as the number of replications increased, while the cut-off points between hypotheses remained constant. The same was true for the comparison with equal ICCs per cluster, $\tau = .3$ in the larger data set and $\tau = .5$ in the smaller one. Therefore, it was concluded that the 500 replications were sufficient for this study to draw valid conclusions.

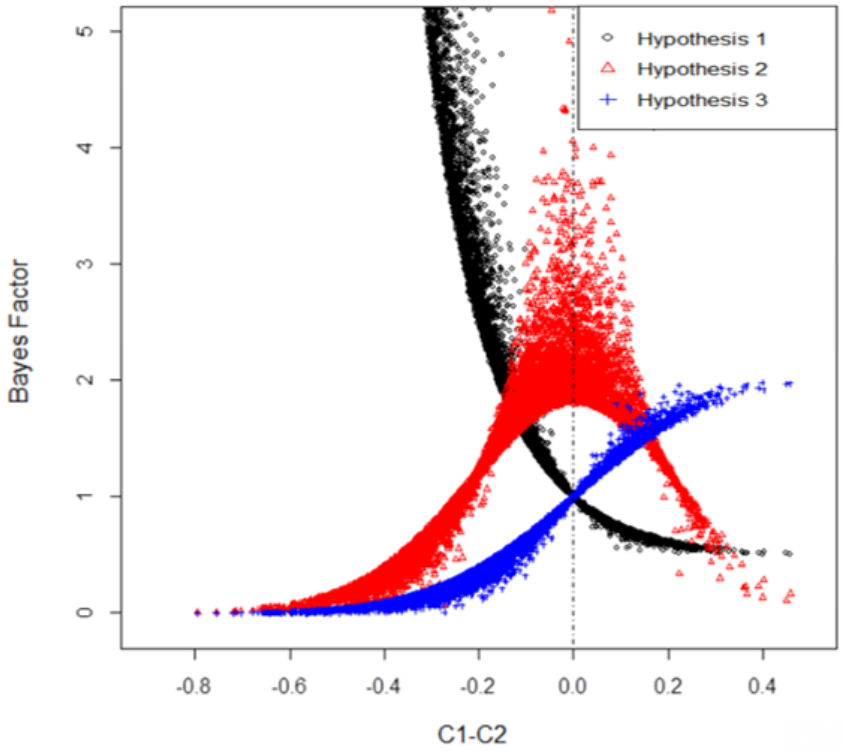

Figure 1. The BF results for the three hypotheses plotted against the difference in estimated ICCs of cluster 2 and 1 for 10,000 replications.

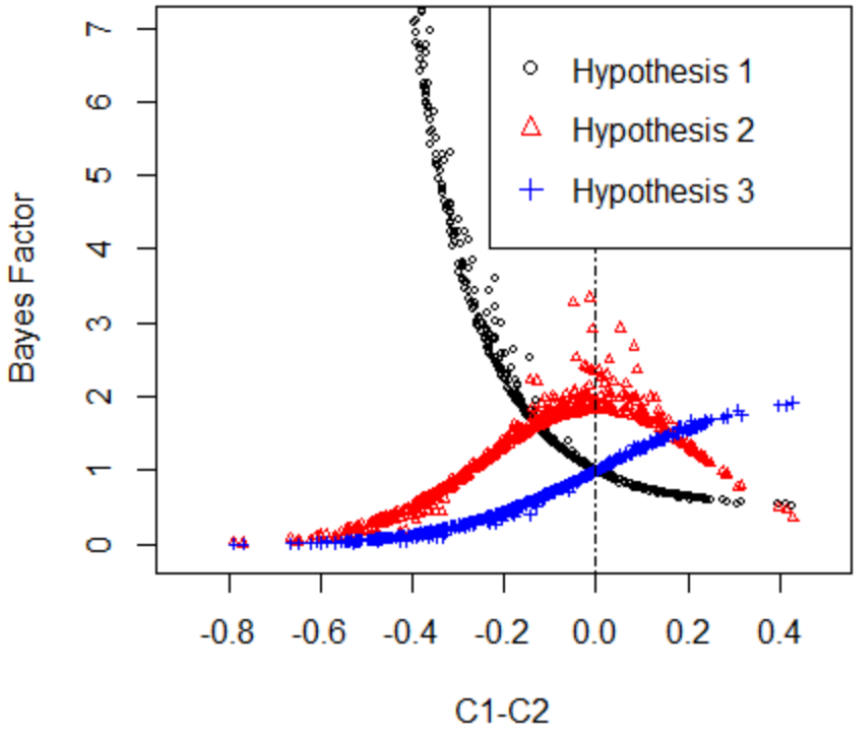

Figure 2. The BF results for the three hypotheses plotted against the difference in estimated ICCs of cluster 2 and 1 for 500 replications.

When looking at the *special* ICCs the averages became prominent (see Table 4). Because the ICC test was seen to not only process negative ICCs but also favour the hypothesis as predicted under the parameters. In the terms of the evidence categories the ICC test found at least strong evidence and extreme evidence for the H1. This feature distinguished the ICC test from other tests which restrict the ICC to be positive and can only process it in that manner.

Table 4

Estimated Posterior Probability and Bayes Factor for a Sample Size of 50 Groups (N) with 5 Observations (n). Sorted per τ_1 for Cluster 1 (C1) and τ_2 Cluster 2 (C2).

		*N*50*n*5	
τ C1,C2	-.1,0	0,.5	-.1,.5
Hypothesis			
Posterior probability			

H1	0.573	0.969	0.999
H2	0.411	0.029	.000
H3	0.017	0.002	.000
Bayes Factor			
H1	10.29	251.5	1.11×10^5
H2	1.977	0.065	.000
H3	0.097	0.004	.000

To further examine the performance of the ICC test unbalanced designs were investigated. Starting with inequal number of observations per group and τ= .3 for both clusters (see Table 5). The results were comparable to those obtained under the balanced design with τ= .5 for both clusters (see Table 3). The more data that was available the more evidence was found for the correct hypothesis, as could be explained by the reduction in sampling error.

In Table 6 the results of unequal group sizes across clusters are depicted. Here, the difference between the τ_1= .5 and τ_2= .3 was not well detected. The BF supported, to a higher degree, H2 than the correct H3. This was most likely due to the amount of data sets which convey equality of the ICCs. Still, as the difference between the ICCs increased in the second column the BF pointed towards the correct hypothesis.

Table 5

Estimated Posterior Probability and Bayes Factor for 500 Replications for an Unbalanced Design by Inequal Number of Observations. Sorted per Sample Size for Cluster 1 (C1) and Cluster 2 (C2) Respectively.

		τ .3,.3	
N C1,C2	5,5	10,10	20,20
Hypothesis			
Posterior probability			
H1	0.314	0.276	0.239

H2	0.408	0.474	0.558
H3	0.278	0.250	0.203
Bayes Factor			
H1	1.005	0.972	1.016
H2	1.399	1.943	2.724
H3	0.996	1.029	0.984

Table 6

Estimated Posterior Probability and Bayes Factor for 500 Replications for an Unbalanced Design by Inequal Number of Observations. Sorted per τ_1 for Cluster 1 N= 15 (C1) and τ_2 for Cluster 2 N= 20 (C2).

	*N*15,20	
τ C1,C2	.5,.3	.5,.1
Hypothesis		
Posterior Probability		
H1	0.236	0.276
H2	0.434	0.201
H3	0.330	0.538
Bayes Factor		
H1	0.707	0.538
H2	2.061	0.940
H3	1.414	1.860

To make an overview of the simulation results, the percentage of times the ICC test asserted the correct hypothesis was computed. From the evidence categories, the middle ground of no evidence could be extended to include anecdotal evidence. Because virtually no sample size created a perfect 1 fitting in the “no evidence” category from Table 1. Thus, the grey area of anecdotal evidence had to be further examined in this study.

When examining the data, the ICC test had most difficulty with judging the data given equal ICCs per cluster, in four of the six sets it leaned towards the incorrect hypothesis, but only for less than 3,6% of the 500 replications. Out of these four, three were in the unbalanced data sets regarding random number of observations with equal ICCs. The only other instance was, again, over the unbalanced clusters and number of observations with $\tau_1 = .5$ and $\tau_2 = .3$, with 4%. However, these incorrect results were only situated in the moderate evidence category. In summary, the ICC test either points towards the correct hypothesis or has trouble choosing between hypotheses with moderate or stronger evidence.

The FOCUS data was used as a trial run for the ICC test to illustrate the performance on real data. When looking at the FOCUS data the ICC test found the same results as in the original paper, there was a positive intervention effect and the extent differed over schools (van Geel et al., 2016). This could be concluded only by looking at the changes in ICCs between prior to (cluster 1) and during (cluster 2) the intervention. All random intercept variances decreased meaning the schools differentiated to a higher degree (see Table 7), that in combination with the mean increase in ability scores per grade (please refer to van Geel et al., 2016) substantiates the hypotheses without the BF.

Table 7

Random Intercept Variances and Standard Deviations (SD) FOCUS data. With Prior to Intervention (C1) and During Intervention (C2) over N= 97 Schools and n= 59.208 Observations in Total Sorted per Grade.

Variance (SD) School grade	Prior to Intervention	During Intervention
Grade 3	30.05 (5.48)	26.18 (5.12)
Grade 4	24.10 (4.91)	18.08 (4.25)
Grade 5	19.77 (4.45)	14.45 (3.80)
Grade 6	31.17 (5.58)	16.39 (4.05)
Grade 7	20.45 (4.52)	10.20 (3.19)
Grade 8	19.92 (4.46)	11.62 (3.41)
Total residual	186.83 (13.67)	

Discussion

The question whether intra-class differences can be properly found and judged with testing the intra-class correlation coefficient lies in the middle. On the one hand, the ICC test from Mulder and Fox (2019) processed imbalanced data to a similar degree as that of the balanced data, removing the need for pair-wise exclusion in cases of drop-out, therefore having the ability to preserve more data than other statistical tests. Furthermore, the ICC test can, unlike most tests, process the ICC whether it was negative or zero. This feature contributes to the accessibility of processing those *special* ICCs and awareness about them.

Furthermore, the ICC test either chose the correct hypothesis or has difficulty deciding between the compared hypotheses. When looking at the percentages the ICC test rejected the right hypothesis, within the frequentist $\alpha = 0.05$ criterion, in only 3,6% of the cases. The threshold used was that of pointing towards the correct hypothesis, deeming anything above a BF of 1 as enough. This revised aiming point became the threshold for the Type-I Error because the lack of other ICC tests to compare to. However, although this frequentist threshold for the Type-I error is more easily applied than the corresponding Type- II error in Bayesian statistics, further research should be done on the power of the ICC test. Moreover, given the small sample sizes, the context in which these results occurred matters more than set criterion in other tests. The increase in posterior probability as the sample size increases (see Table 3 & 5) shows that the evidence for the correct hypothesis increased due to the lessening of sampling errors. Because the sampling error influences the power, this threshold is set until other tests or studies on these sample sizes have been conducted that promote another threshold.

The meaning of the BF was initially derived from the evidence categories (Andraszewicz et al., 2015). However, those categories might not do justice to the intricate ICC test. Especially since the ICC test produced a high rate of accuracy at this point in time, in this relatively new field of testing ICCs. Over such small sample sizes anything including anecdotal evidence and stronger evidence could be considered as either correct or incorrect. Therefore, it might be interesting to further study the evidence categories to judge the BF per sample size.

On the other hand, while true for virtually all statistical tests, small sample sizes distort the realized parameters of generated data in comparison to the true parameters. The percentages of incorrectly supported hypotheses could suggest that the ICC test might have

had to judge unfortunate data sets. Over the 500 replications the minority of data could support, the initially perceived as incorrect results, the unlikely hypothesis given the parameters. Furthermore, the general downside of small sample sizes i.e., the less data the greater the chance of no distinction between hypotheses has influenced this study. Both these downsides of small sample sizes could explain the difficulties the ICC test had choosing the correct hypothesis.

In Figure 1 and 2 it can be seen that the BF is not symmetric around the point zero on the x-axis when τ_1 is equal to τ_2. There is a lower bound for the ICCs and therefore the τ_1- τ_2 cannot become more negative at a certain point, however the evidence in favour of H1 rapidly increases due to the lower bound. At the other end of the scale, when τ_1 - τ_2 is high, the lower bound does not effect the term as much, since τ_2 does not reach the lower bound quickly and also τ_1 is greater than zero. Then, even higher values for τ_1- τ_2 do not rapidly increase the evidence in favour of H3. Therefore, the evidence in favour of H3 is more stretched out on the x-scale.

In future work the longitudinal setting can be examined via ICC testing. While this simulated data was generated with a cross sectional study design in mind, much like the FOCUS data, it would be interesting to see if the ICC test can also perform well in a generated longitudinal setting. Preferably a smaller data set will also be generated for that setting to compare it with current results. Afterwards bigger sample sizes can be tried as well, to prevent the issues mentioned above. It is most likely that also in that setting the ICC test will be able to distinguish intra-group differences by examining the ICC. The ICC can be computed over the time points, because those can be viewed as the cluster where the participants are the groups in which the measurement scores are nested.

Another aspect that the FOCUS data offered was that of real data, the ICC test should be further explored over other real datasets. The ICC test while not being able to have the BF assert the correct hypothesis proved that, by combining the significant mean score difference (van Geel et al., 2016) with the changes in variance, the ICC does offer additional information. This underlines the reason to continue research in the field of ICC testing.

In conclusion, hierarchical data can be tested with the ICC. Changes in the ICC do point towards intra-group differences, which can suggest an intervention effect and on which level in the model the intervention has effect. The ICC test demonstrated that the BF can be employed on ICC testing and that the ICC conveys additional information to the mean scores.

However, the sample size and evidence categories of the BF require the researcher to judge the outcomes per study to a higher degree than what is commonplace in the frequentist approach. Therefore, more studies should be conducted in ICC hypothesis testing.

Literature

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E. J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*(2), 521-543. doi: 10.1177 /0149206314560412

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83*(5), 762-765. doi: 10.1037/0033-2909.83.5.762

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects ¨ Models Using {lme4}. Journal of Statistical Software, 67(1), 1–48. doi:10. 18637/jss.v067.i01

Dyer. N. G., Hanges. P. J., & Hall. R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly. 16*(1), 149-167. doi:10.1016/j.leaqua.2004.09.009

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556. doi:10.1037 /met0000201

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural equation modeling*, *8*(2), 157-174. doi:10.1207/S15328007SEM0802_1

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). Multilevel analysis: Techniques and applications. Routledge. Retrieved from https://books.google.nl/books

Jeffreys, H. (1961). Theory of probability-3rd ed. New York: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795. doi:10.1080/01621459.1995.10476572

Kosheleva O., Kreinovich V., Trung N.D., & Autchariyapanitkul K. (2019) How to Make a Decision Based on the Minimum Bayes Factor (MBF): Explanation of the Jeffreys Scale. Data Science for Financial Econometrics. *Studies in Computational Intelligence*, 898. Springer, Cham. doi:10.1007/978-3-030-48853-6_8

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86-92. https://doi.org/10.1027/1614-2241.1.3.86

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. Journal of Mathematical Psychology, 72, 6-18. https://doi.org/10.1016/j.jmp.2015.11.001

Mulder, J., & Fox, J.-P**.** (2019). Bayes Factor Testing of Multiple Intraclass Correlations. *Bayesian Analysis.* Volume 14, Number 2 (2019), 521-552.

Mulder, J., van Lissa, C., Gu, X., Olsson-Collentine, A., Boeing-Messing, F., Williams, D. R., Fox, J.-P., Menke, J., et al. (2019). BFpack: Flexible Bayes Factor Testing of Scientific Expectations. [Version 0.2.1] https://CRAN.R-project.org/package=BFpack

Nielsen, N. M., Smink, W. A. C., & Fox, J-P. (2021). *Small and Negative Cluster Correlations.* Manuscript submitted for publication.

Page, R., & Satake, E. (2017). Beyond P Values and Hypothesis Testing: Using the Minimum Bayes Factor to Teach Statistical Inference in Undergraduate Introductory Statistics Courses. *Journal of Education and Learning*, *6*(4), 254-266.

Paterson. L., & Goldstein. H. (1991). New statistical methods for analysing social structures: An introduction to multilevel models. *British Educational Research Journal*. 17(4): 387–393.

Quintana, D. S., & Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. BMC psychiatry, 18(1), 178. https://doi.org/10.1186/s12888-018-1761-4

R Core Team. (2019). R: A Language and Environment for Statistical Computing [Version 3.6.1]. R Foundation for Statistical Computing. doi:10.1007/978-3- 540-74686-7

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage. Retrieved from: https://books.google.nl/books?id= N1BQvcomDdQC&printsec=copyright&hl=nl#v=onepage&q&f=false

Socha. A. (2013). A hierarchical approach to students' assessments of instruction. *Assessment and Evaluation in Higher Education. 38*(1). 94-113. doi: 10.1080/02602938.2011.604713

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. American Educational Research Journal, 53(2), 360-394. https://doi.org/10.3102/0002831216637346

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the akaike information criterion (AIC) and the bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228-243. doi:10.1037/a0027127

Wasserstein, R. L., & Lazar, N. A. (2016) The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133, doi:10.1080/00031305.2016.1154108

Zhang, D. (2019). *Bayesian analysis for the intraclass model and for the quantile semiparametric mixed-effects double regression models.* (Dissertation, Michigan Technological University). https://doi.org/10.37099/mtu.dc.etdr/857