# Analyzing Randomized Controlled Interventions:

# Three Notes for Applied Linguists

Jan Vanhove

University of Fribourg

Author Note

I thank Peter Lenz, Winfried Kronig, and Raphael Berthele for their useful comments.

Jan Vanhove, University of Fribourg, Department of Multilingualism, Rue de Rome 1, CH-1700 Fribourg, Switzerland, jan.vanhove@unifr.ch.

## Abstract

I discuss three practices common in analyses of randomized controlled interventions in applied linguistics. These are (a) checking whether randomization produced groups that are balanced on a number of possibly relevant covariates, (b) using repeated-measures ANOVA to analyze pretest-posttest designs, and (c) using traditional significance tests to analyze interventions in which whole groups were assigned to the conditions. These practices are labeled superfluous, overcomplicated, and inappropriate, respectively, and I suggest alternatives for each of them that will be useful to applied linguists and educators who need to design, analyze, or assess intervention studies or other randomized controlled trials. The statistical formalism is kept to a minimum throughout.

*Keywords*: randomized experiments, cluster randomization, pretest-posttest designs, covariates, mixed-effects modeling

# Analyzing Randomized Controlled Interventions:
# Three Notes for Applied Linguists

Intervention studies in which participants are randomly assigned to the treatment or control group are the gold standard for establishing the effectiveness of language learning methods. It is therefore essential that the data that they produce be analyzed optimally and the analyses reported as cogently as possible. In this contribution, I discuss three common practices in the statistic analysis of randomized controlled interventions that obfuscate or even invalidate the insights gained from such studies. Specifically, I focus on superfluous 'balance tests' on covariates, overcomplicated analyses of pretest–posttest data, and the inappropriateness of ignoring the effects of assigning whole groups of participants to the experimental conditions.

The issues I discuss are not new nor are they unique to applied linguistics. But applied linguists that take an interest in them need to foray into different fields where they are often explained using dauntingly looking statistical formalism. My goal is to bring these issues to the attention of applied linguists and educators who need to design, analyze, or assess intervention studies or other randomized controlled experiments, and to furnish them with practical recommendations. Throughout the article, the formalism is deliberately kept to a minimum.

## Superfluous Balance Tests

As I do not wish to single out specific studies, let us consider a hypothetical intervention that investigates the effectiveness of a new method for learning to read genealogically closely related foreign languages through self-study. Fifty participants are recruited and randomly given either the new learning method (treatment) or its predecessor (control) for self-study. After six weeks, all participants are administered a reading test in the related languages they have been studying (*posttest-only randomized controlled experiment*). In addition, the researchers extract several background variables at the onset of the study, e.g., the learners' age, sex, socioeconomic status, and other foreign language skills. (Pretest scores are discussed below.)

## What are Balance Tests?

In addition to analyzing the post-test scores, conscientious researchers will often run statistical tests (e.g., *t*-tests or ANOVAs, or $\chi^2$-tests) on the background variables in order to ensure that the treatment and control groups are comparable in all relevant respect save for the self-study method used. The two groups are then typically deemed comparable if these *balance tests* yield a *p*-value higher than specified threshold (often $p > .05$). If the *p*-value lies below this cut-off, additional analyses to investigate the impact of the potential confound variable are often carried out. For instance, say we observe a significant difference in the reading scores between the control and treatment group (e.g., $t(48) = 2.4$, $p = .02$), but also a significant age difference in the same direction (e.g., $t(48) = 2.7$, $p = .01$). We then might be tempted to include the learners' age as a 'control variable' in an analysis of covariance (or might be spurred to do so by a reviewer). Conversely, had we observed a nonsignificant treatment effect, we might also be inclined to include age as a covariate in an ANCOVA if the control and treatment groups differed in age.

## Problems with Balance Tests

As Mutz and Pemantle (2013) argue, such balance tests are (almost always) superfluous in randomized experiments and may lead to suboptimal analyses down the road. Here, I summarize their main points. First, to understand their superfluousness, consider the null hypothesis that is being tested when the researcher conducts a balance test on (for instance) the age variable: The null hypothesis is that the participants in both the control and the treatment groups were randomly drawn from the same population and that any age differences between them are consequently due to sampling error. Since the researcher randomly assigned the participants to the control or treatment groups, she already knows that this is the case. Thus, rejecting this null hypothesis always constitutes a false positive (‘Type-I error’: finding a significant result even though there is no effect). (Note that the topic is *randomized* experiments; the point does not hold if the allocation of participants to control or treatment groups was not done at random.) Consequently, the *p*-value produced by a “silly” (Abelson, 1995, p. 76) balance test cannot tell us anything that we do not already know.

Second, the use of balance tests suggests that researchers think that randomization is a mechanism for creating samples that are balanced with respect to potential confound variables. Balance tests would then be a method for establishing whether our sample is indeed balanced with respect to the background variables measured or whether we were ‘unlucky’ to have drawn an arguably random but nevertheless unbalanced sample. However, randomization is not meant to ensure background variable balance in any specific sample. Rather, randomization balances out the effect of both measured *and* unmeasured confounds *on average*: If we were to run the same intervention study a large number of times and randomized the assignment of the participants to the control or treatment groups each time, randomization guarantees that the average observed treatment effect corresponds to the true treatment effect. In any given sample, we may observe smaller or larger effects, but the distribution of these effects follow the laws of probability. Thus, statistical tests already account for fluke findings due to randomization (see also Oehlert, 2010, Chapter 2). In other words, the hypothetical *p*-value of .02 cited above has its precise meaning: Assuming that the treatment is equally as effective as the control method (= the null hypothesis), the probability of observing a difference as large as or even larger than the one observed is 2%.

Third, since *p*-values already take chance findings due to randomization into account, it follows that acting on balance tests invalidates them. As illustrated above, *p*-values are conditional probability statements: They reflect the probability of observing a particular (or more extreme) result if the null hypothesis were true. By using balance tests, we introduce an additional set of conditions ('if the null hypothesis were true and background variables *A*, *B*, and *C* differed no more than by a specific margin between the two groups') that is not taken into account by the reported *p*-value. What is more, not only do *p*-values lose their precise meaning after a *significant* balance test: The mere *intention* to act on significant balance tests invalidates the reported *p*-value even if the balance test comes out nonsignificant. The reason is that statistical inference is affected not only by what we do when observing a particular data pattern (e.g., a positive balance test) but also by what we would have done had the data come out differently (see Gelman & Loken, 2013; see also Simmons, Nelson & Simonsohn, 2011, for related issues). Unfortunately, for practical purposes, it is impossible to recalibrate *p*-values to take such conditionalities into account.

**Proper Uses of Background Variables**

Does all of this mean that collecting background variables is a waste of time? No. One valid and indeed highly recommendable procedure is to *block* on variables that are deemed important (see Oehlert, 2010, Chapter 13). For instance, if we have reasons to believe that women and men would differ substantially in their reading test scores, we may want to consider blocking on the sex variable—half of the men would be randomly assigned to the treatment group and half to the control group, and similarly for the women. The decision to block on a given (or several) variables must be made before the randomization, and blocking is difficult if the participant sample is not completely defined at the onset of the data collection (e.g., when participants trickle to the lab over the course of severals weeks; but see Moore & Moore, 2013). When feasible, however, blocking on background variables that are actually related to the outcome increases statistical precision by accounting for a source of residual variance: Less residual variance means narrower standard errors, which in turn translates into a greater probability of finding true treatment effects ('power'). Note that this requires that the blocking variable is included in the analysis. Interestingly, even blocking on poorly chosen or measured background variables does not decrease precision in any but the smallest of samples (see Imai, King & Stuart, 2008).[1]

A second way to make fruitful use of background variables is to use them as covariates in the analysis *regardless* of whether the two groups are balanced with respect to them or not (Mutz & Pemantle, 2013). Entering covariates that are really related to the outcome into the analysis again increases statistical precision by accounting for a source of residual variance, even if the means of the variable is similar in the two groups. However, covariates that are not actually related to the outcome decrease statistical precision since they fit noise in the data at the cost of degrees of freedom. Hence, it pays to be discriminate and to select a small number of background variables that are known to be substantially related to the outcome (e.g., from the literature); then, when analyzing the study's results, one can forgo conducting balance tests and use the selected variables in a multivariate analysis.

## Using Pretest Scores

The fictitious intervention study described above can substantially be improved upon by the inclusion of a pretest, giving rise to a *pretest–posttest randomized controlled experiment*. When pretest scores are available, it is always advantageous to use them in the analysis—even if the control and treatment groups are comparable with respect to them. Pretreatment ability is perhaps the single most important predictor of posttreatment ability, and taking interindividual differences in pretreatment ability into account greatly improves the precision of the study with respect to the treatment effect. However, the analyses of pretest–posttest studies reported in the literature are often needlessly complicated and can be simplified for greater readability without loss of technical accuracy. This point is not new (see Huck & McLean, 1975) but bears repeating.

### Repeated-Measures ANOVA: Too Much Information

In our fictitious pretest–posttest example, researchers might be tempted to conduct a repeated-measures ANOVA (RM-ANOVA) with Condition (control, treatment) as a between-subject factor and Time (pretest, posttest) as well as the interaction between Condition and Time as within-subject terms. The report may then read something like this:

> A repeated-measures ANOVA yielded a nonsignificant main effect of Condition ($F(1, 48) < 1$) but a significant main effect of Time ($F(1, 48) = 154.6$, $p < 0.001$): In both groups, the posttest scores were higher than the pretest scores. In addition, the Condition × Time interaction was significant ($F(1, 48) = 6.2$, $p = 0.016$): The increase in reading scores relative to baseline was higher in the treatment than in the control group. (Fictitious example)

In technical terms, this analysis is defensible (but see below), but its main problem is one of communicative efficiency. As Huck and McLean (1975) showed, the main effect of Condition substantially underestimates the treatment effect: Logically, the treatment effect can only be observed in the posttest data, but the main effect of Condition is computed on the basis of both the pre- and the posttest data. This yields a diluted estimate of the actual treatment effect. Additionally, finding a main effect of Time is trivial: We already know that test scores change with time. Of the three reported results, only the interaction result is actually germane to the research question—we were interested in knowing whether the increase relative to baseline differed between the two groups. This interaction term provides the true treatment effect.

### *t*-Tests: As Correct but Much Simpler

Inundating one's readership with irrelevant or trivial results derived from a rather complicated analysis strikes me as counterproductive. What may not be fully appreciated is that a much simpler analysis provides the same information: A two-sample *t*-test on the gain scores, i.e., the differences between the posttest and pretest scores, *always* yields the same result as the Condition × Time interaction in a RM-ANOVA. In our example, the report would reduce to: "A two-sample *t*-test on the gain scores revealed that the treatment group showed a higher increase in reading scores than the control group ($t(48) = 2.49$, $p = 0.016$)." (Note that for a given analysis, $F = t^2$.)

### A More Flexible and (Slightly) More Powerful Alternative

Both RM-ANOVAs and *t*-tests on gain score assume that the pretest and posttest scores are linearly related with a slope of 1 (Hendrix, Carter & Hintze, 1978; Huck & McLean, 1975). This assumption is clearly violated when the pretest and posttest scores are on different scales, but it can also be violated by mere measurement error. Performance on imperfect tests is a combination of ability and extraneous factors such as form on the day, topic of a reading test, etc. Participants who over- or underperformed due to extraneous factors on the pretest are relatively unlikely to over- or underperform by the same margin on the posttest ('regression to the mean'). As a result, the slope of the relationship between the pretest and posttest scores will often be less than 1 even if both tests are scored on the same scale.

Using analysis of covariance (ANCOVA), the slope parameter can be estimated from the data rather than be assumed to be 1. Doing so costs a degree of freedom and hence some loss of statistical power if the slope actually is equal to 1, but it can lead to more power if the slope is not equal to 1 (e.g., Van Breukelen, 2006). However, as Maris (1998) showed, estimates of the treatment effect that are derived from gain score analyses are not biased in the presence of measurement error. Put differently, gain score analyses and analyses using the pretest data as a covariate yield the same, correct estimation of the treatment effect on average, but ANCOVA has more power. Note, however, that this applies only to studies in which the assignment was done at random (see Van Breukelen, 2006).

In addition to having greater statistical power, ANCOVA comes with increased flexibility compared to gain score analyses. First, the pretest and posttest data need not be expressed on the same scale: If the pretest data have been collected on a 10-point scale and the posttest data on a 25-point scale, there is no need to transform the data to fit on a common scale. Second, gain score analyses (or RM-ANOVA) and ANCOVA both assume that the relationship between the pretest and the posttest data is linear. However, ANCOVA models can be furnished with nonlinear terms (e.g., quadratic or cubic terms, or regression splines) to relax this assumption (see, e.g., Baayen, 2008, pp. 108-111).

In sum, using repeated-measures ANOVA to analyze pretest–posttest data is an unnecessary complication as simpler analyses on gain scores always yield the same result. Using the pretest data as a covariate in an ANCOVA is usually (somewhat) more powerful and is more flexible with respect to the linearity assumption.

## Interventions with Intact Groups

In the example above, the learners were randomly assigned to the treatment or control group on an individual basis. Oftentimes, however, randomization occurs not at the individual level but at a higher level. For instance, whole classes or schools are often assigned to the conditions of the intervention so that within each class or school, all pupils belong to the same condition. Designs in which intact groups are randomly assigned to conditions are known as *group-* or *cluster-randomized interventions* and are a popular choice when practical considerations do not allow random assignment at the individual level, when interactions between participants in different group at the same school may contaminate the results, or when doing so increases the study's ecological validity (e.g., when testing methods that will later only be used with intact groups). Crucially, however, ignoring the fact that randomization took place at the group level drastically affects the insights gained from the study.

To take another hypothetical example, consider a researcher who is interested in comparing the efficacy of two class-based methods (control and intervention) for teaching reading skills in related foreign languages. At her disposal stand ten classes, each taught by a different teacher, of 20 students each.[2] Five classes are randomly assigned to the control method and five to the intervention method. At the end of the semester, all 200 students take a reading test that provides the dependent variable for the analysis.

Unfortunately, a *t*-test (or, equivalently, an ANOVA) on the 200 resultant data points is too likely to yield a statistically significant result when there is, in fact, no effect (see, e.g., Barcikowski, 1981; Walsh, 1947). In statistics parlance, such a test has a higher-than-nominal (i.e., higher than 5%) Type-I error rate and is thus anti-conservative. This anti-conservatism is the result of a violation of the test's assumption that the data points be independent of one another: Participants' scores tend to be more alike within a cluster than between clusters since their performance is shaped by the same or similar contextual factors (e.g., the students' background, teachers). As a result, data points within a cluster do not contribute entirely independent information, and analyzing the data as though they did overstates our confidence in the results. The degree to which our confidence would be overstated depends on the intraclass correlation.

## Intraclass Correlation and its Effects

The degree to which participants' scores within a cluster tend to be alike is expressed by the intraclass correlation coefficient (ICC). It is computed by dividing the between-cluster variance by the sum of the between- and within-cluster variances. The ICC takes on values between 0 and 1, with 0 indicating the complete absence of clustering and 1 indicating that all scores within each cluster are identical.

Importantly, ignoring even small degrees of interrelatedness within clusters can invalidate the analysis. Figure 1 shows how the actual Type-I error rate increases as a function of (a) the ICC and (b) the number of participants per cluster if the data of a cluster-randomized experiment are analyzed by means of *t*-tests on the actual observations. As is clear from this plot, even ICCs as small as .01 give rise to appreciably higher Type-I error rates, especially if the number of participants per cluster is high. For reference, ICC estimates provided by Hedges and Hedberg (2007) for reading achievement tests when whole schools are the level of randomization range from .17 to .27, and Schochet (2008) provides an estimate of about .15 for evaluations of educational programs with randomization occurring at the school or classroom level.

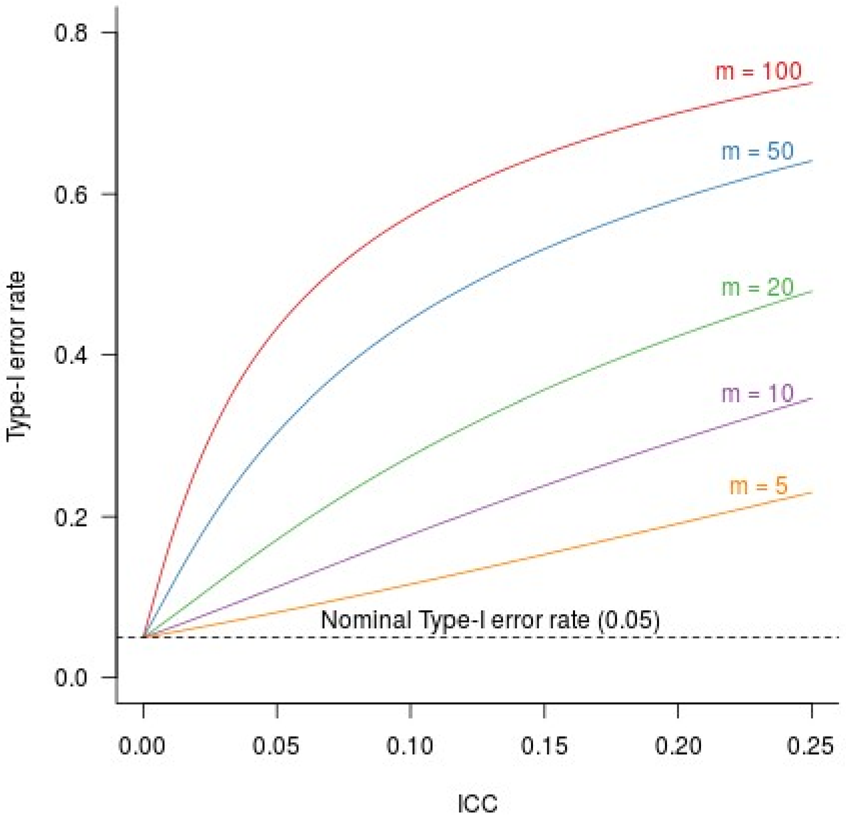

*Figure 1. Type-I error rates for cluster-randomized experiments as a function of the intraclass correlation coefficient (ICC) and the number of participants per cluster (*m*). For this graph, the number of clusters was fixed at 10, but the Type-I error rate differs only slightly for different numbers of clusters.*

Such reference values are useful for planning cluster-randomized experiments (see Hedges & Hedberg, 2007; Killip, Mahfoud & Pearce, 2004; Schochet, 2008; Spybrook et al., 2011). For a given number of participants, clusters, ICC coefficient and expected effect size, the probability of finding a significant intervention effect can be computed. This is referred to as the study's *power*. Figure 2 illustrates the effects of cluster size and ICC on the power of a study with 128 participants to observe a medium-size effect (*d* = 0.5, see Cohen, 1992).[3] As this graph illustrates, a (correctly analyzed; see below) cluster-randomized experiment always has less power than a completely randomized experiment (i.e., *m* = 1) with the same total number of observations.

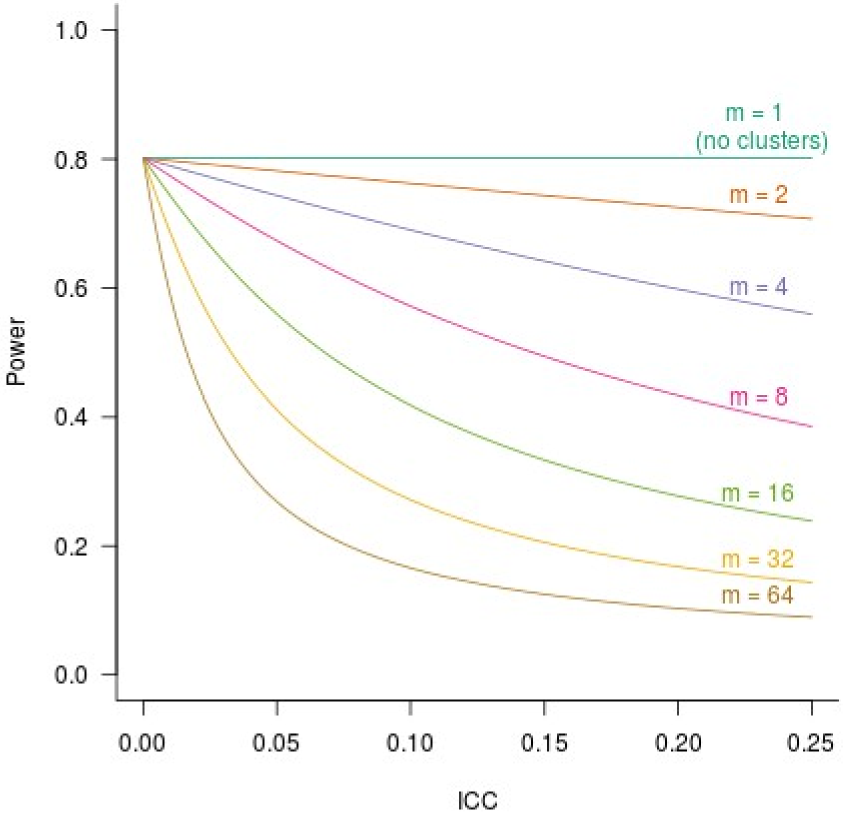

*Figure 2. Power to detect a medium-size effect (*d *= 0.5) for a cluster-randomized experiment with 128 participants in function of the intraclass correlation coefficient (ICC) and the number of participants per cluster (*m*).*

Including strongly predictive covariates (e.g., pretest scores) in the analysis reduces the effect of clustering, i.e., it increases statistical power and reduces Type-I error rates of flawed analyses, but cannot be counted on to make it disappear entirely (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Moerbeek, 2006; Murray & Blitstein, 2003; Schochet, 2008).

## Taking Clustering into Account

When the ICC value is known, it can be used to recalibrate statistical analyses (see Blair & Higgins, 1986, for when the raw data are available, and Hedges, 2007, for when they are not). However, these recalibration methods require that the ICC parameter of the population to which we want to generalize is known, or has at least been estimated with great precision—using sample estimates of the ICC to recalibrate one's results does not work as these estimates are too imprecise (Blair & Higgins, 1986; Hedges, 2007). In most cases, an accurate estimate will not be available, which is why I focus on two analytical tacks that do not require one: analyses on cluster means and mixed-effects modeling.

**(Weighted) *t*-Tests on Cluster Means.** A conceptually straightforward approach is to calculate the mean (or another summary measure) of each cluster and run a *t*-test on them rather than on the original observations. When the number of observations differs from cluster to cluster, a *t*-test in which the cluster means are weighted for cluster size is recommended (see, e.g., Campbell, Donner, & Klar, 2007).[4] This analysis is easy to compute and report, and it perfectly accounts for violations of the independence assumption: The Type-I error rate is at its nominal level (i.e., 5%).

A drawback of this approach is that individual-level covariates (e.g., the participants' age) cannot directly be accounted for. Cluster-level variables (e.g., the teacher's sex) and cluster-level summary measures of individual-level covariates (e.g., the average age of pupils in a class), however, can be entered in a multiple regression model or ANCOVA ran on the cluster means. This comes at the cost of a residual degree of freedom for each covariate, however, which can be prohibitively expensive power-wise in studies with a small number of clusters if the covariate is only weakly related to the dependent variable.

Additionally, researchers may find it psychologically difficult to reduce a dataset to a fraction of its original size—if the analysis is carried out on ten cluster means, why bother recruiting several participants per cluster? However, larger clusters reduce the variance of the cluster means within each treatment group, which in turn makes the intervention effect stand out more clearly (Barcikowski, 1981). Put differently, cluster-level analyses are more powerful when the clusters are large compared to when they are small. That said, when given the choice between running an experiment on ten clusters with 50 observations each or on 50 clusters with ten observations each, the latter is vastly preferred due to its higher power (see Figure 2).

**Mixed-effects modeling.** Mixed-effects modeling has gained popularity among language researchers in recent years and represents an attractive alternative to researchers who have carried out a cluster-randomized experiment. Mixed-effects models are fitted on the original observations and can explicitly account for the effect of clustering by fitting the clusters by means of a random-effect term. This is also known as multilevel or hierarchical modeling. Multilevel modeling is not yet part and parcel of the statistical education of applied linguists. Given its increasing popularity in related disciplines (e.g., psycholinguistics), I think it is nonetheless useful to outline its advantages for analyzing group-randomized experiments. Additionally, I want to highlight that even multilevel models do not necessarily produce entirely accurate *p*-values.

The following discussion is unavoidably more technical in nature and assumes some familiarity with the `lmer` (‘linear mixed-effect regression’) function in the `lme4` package (Bates et al., 2014) for R. In addition to being free, `lme4` combines state-of-the-art modeling with relative user-friendliness and has become the software package of choice in psycholinguistics (for an introduction to `lme4` geared towards language researchers, see Baayen, 2008). Readers who are primarily interested in the bottom line of this more technical discussion can skip to the Recommendations section.

***Advantages.*** One advantage of `lmer`, and multilevel modeling more generally, is that it can cope with several dependency levels simultaneously. Suppose that ten schools with a total of 80 classes and 1360 students participate in an experiment. Assignment to the experimental conditions occurs at the class level, and each student contributes five observations (e.g., test scores). In addition to modeling classes (the level of randomization) as random effects, schools and students can—and indeed should—be modeled as random effects, too. Second, `lmer` fits can include covariates pertaining to different levels (e.g., age of the student, sex of the teacher, and school type). Third, mixed-effects models can be fitted on non-Gaussian data as well (e.g., dichotomous or Poisson data) using the `glmer` function in the `lme4` package. However, `lmer` outputs do not, by default, feature *p*-values. As Bates (2006) explained, the reason is that it is not clear how to compute degrees of freedom for terms in mixed-effects models. However, several strategies exist for 'squeezing' *p*-values out of `lmer` fits.

***Assessing statistical significance in multilevel models.*** The first strategy is to take the *t*-value for the model term and interpret it as a *z*-value (see, e.g., Baayen, 2008, pp. 247-248). The null hypothesis is then rejected when $|t| > 1.96$. As Baayen (2008) points out, however, this approach is anti-conservative (i.e., yields spurious significance) for small samples. What counts as a large enough sample for this approach to be reasonable depends on the design of the study, specifically the number of clusters, as well as on the intraclass correlation.

A different kind of *p*-value can be computed by comparing nested models by means of a likelihood-ratio test (LRT, see, e.g., Faraway, 2006, pp. 156-157). This test, too, is anti-conservative for small samples: It assumes that the distribution of the LRT statistic approximates a $\chi^2$ distribution with one degree of freedom (in the present context) under the null hypothesis, but this approximation is poor for small samples and sometimes even for large samples (Faraway, 2006). Again, what constitutes 'small' and 'large' samples depends on the number of clusters and the intraclass correlation.

A third strategy is to compute approximate *p*-values using a simulation-based approach called *parametric bootstrapping* (see Faraway, 2006, pp. 157-158). In a parametric bootstrap, a large number (say, 1,000) of alternative datasets are produced under the null hypothesis. This is accomplished by fitting a mixed-effects null model to the data that does not contain a term for the intervention effect but is otherwise fully specified (intercept, random effects and covariates). New datasets are then simulated on the basis of this null model: New dependent variables are created by taking predictions on the basis of the intercept and covariates and adding residuals drawn from the distributions of the random effects and the residual standard error. Then, a LRT statistic is computed for each dataset by comparing the model with the intervention effect fitted to the simulated data to the corresponding model without the intervention effect. Since the datasets were simulated in the absence of a treatment effect, any variability in the (1,000) LRT statistics is due to randomness alone. Thus, to an approximation, the distribution of bootstrapped LRT values  serves as a reference distribution for the LRT statistic computed on the basis of the actually observed data (i.e., as per the second strategy): The bootstrapped *p*-value is equal to the proportion of LRT values under the null hypothesis that are larger than the actual one. A user-friendly implementation of this procedure is available in the `pbkrtest` package (Halekoh & Højsgaard, 2014) for R.

Parametric bootstrapping yields more conservative *p*-values than likelihood-ratio tests, and is therefore to be preferred over the other two strategies, which give anti-conservative results. The Achilles' heel of this approach is that the null model needs to be specified correctly. For technical reasons,[5] the null model on the basis of which the simulated data sets are generated tends to underestimate the between-cluster variance (see Faraway, 2006, pp. 154) and hence the ICC. This underestimation may yield anti-conservative *p*-values from parametric bootstrapping, too. Parametric bootstrapping is computationally expensive, however, and it is therefore difficult to verify whether the *p*-values that it yields are at nominal levels. To my knowledge, no Monte Carlo study has yet been run to investigate the properties of *p*-values computed by means of parametric bootstrapping. In any event, the bootstrapped *p*-values will be *more* accurate than those resulting from the other two strategies, but not necessarily *entirely* accurate.

**Recommendations**

When randomization at the level of the participants is not feasible or desirable, cluster-level randomization presents an alternative. The effects of clustering need to be duly account for so that the statistical significance of the intervention be not overstated. This need brings with it an appreciable but unavoidable loss of statistical power. In order to salvage as much power as possible, carefully considered covariates (both at the individual and at higher hierarchical levels) can be included in a mixed-effects analysis. If *p*-values need to be extracted from such an analysis, parametric bootstrapping is the least unduly optimistic of the three methods discussed, but even so, borderline significant results are best enjoyed with a grain of salt.

A second important corollary of what I have discussed is that, unless the ICC is known (and it rarely is), interventions with one cluster per condition do not allow for meaningful statistical inferences (see also Murray, Varnell, & Blitstein, 2004). For instance, a *t*-test on two cluster means would have a *t*-distribution with zero degrees of freedom as a reference—which does not exist. Conceptually, this translates to the realization that we have no way of teasing apart the intervention effect from the cluster effect when the conditions and the clusters are completely conflated, i.e., we have zero power. Such designs are therefore strongly discouraged.

Lastly, clustering can also occur even when randomization occurs at the level of the individual. For instance, if researchers randomly assign individual to conditions, but then proceed to apply the treatments to groups of several participants, they could also induce clustering effects (see Lee & Thompson, 2005). Such forms of clustering need to be taken into account as well.

## Closing Thought

I have discussed three relatively common practices in the analysis of randomized intervention experiments and labeled them as superfluous, overcomplicated, and inappropriate, respectively. I have done so in part by citing the effects of these practices on *p*-values, and specifically with their yielding too many or too few statistically significant results. This may give the impression that the categorization of findings into ‘significant’ and ‘nonsignificant’ results is the goal of experimental studies (“categoritis”, Abelson, 1995, p. 111). However, uncertainty is inherent in statistical analyses, and even randomized experiments—whether they yield *p*-values below or above an arbitrary threshold—do not necessarily produce definite answers. In simple (but not oversimplified) terms, a significant result does not prove the existence of an effect, and a nonsignificant result much less proves that an effect does not exist (see, among many others, Cohen, 1994; Schmidt, 1996). Thus, while this article has offered recommendations for conducting more accurate or more transparent significance tests, it has done so with the understanding that their results are not the be-all and end-all of experimental research.

# References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational and Behavioral Statistics*, *6*(3), 267-285.

Bates, D. (2006). lmer, p-values and all that. Post to the r-help mailing list, May 19, 2006. https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D., Martin, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package, version 1.1-7. Available from http://cran.r-project.org/package=lme4

Blair, R. C., & Higgins, J. J. (1986). Comment on "Statistical power with group mean as the unit of analysis". *Journal of Educational and Behavioral Statistics, 11*(2), 161-169.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Campbell, M. J., Donner, A. & Klar, N. (2007). Developments in cluster randomized trials and *Statistics in Medicine*. *Statistics in Medicine, 26*, 2-19.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Dalton, S., & Overall, J. E. (1977). Nonrandom assignment in ANCOVA: The alternate ranks design. *Journal of Experimental Education, 46*(1), 58-62.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript. Retrieved on 31 August 2014 from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Halekoh, U., & Højsgaard, S. (2014). pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison. R package, version 0.4-0. Available from http://cran.r-project.org/package=pbkrtest

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*(2), 151-179.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hendrix, L. J., Carter, M. W., & Hintze, J. L. (1978). A comparison of five statistical methods for analyzing pretest–posttest designs. *Journal of Experimental Education, 47*(2), 96-102.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*(4), 511.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(2), 481-502.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*(3), 204-208.

Lee, K. J., & Thompson, S. G. (2005). Clustering by health professional in individually randomised trials. *BMJ, 330*, 142-144.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3*(3), 309-327.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin*, *95*(1), 136-147.

McAweeney, M. J., & Klockars, A. J. (1998). Maximizing power in skewed distributions: Analysis and assignment. *Psychological methods*, *3*(1), 117.

Moerbeek, M. (2006). Power and money in cluster randomized trials: when is it worth measuring a covariate? *Statistics in Medicine, 25*(15), 2607-2617.

Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis*, *20*(4), 460-479.

Moore, R. T., & Moore, S. A. (2013). Blocking for sequential political experiments. *Political Analysis, 21*(4), 507-523.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review, 27*(1), 79-103.

Mutz, D., & Pemantle, R. (2013). The perils of randomization checks in the analysis of experiments. Unpublished manuscript. Retrieved on 31 August 2014 from http://www.math.upenn.edu/~pemantle/papers/Preprints/perils.pdf

Oehlert, G. W. (2010). *A first course in the design and analysis of experiments*. Available from http://users.stat.umn.edu/~gary/Book.html

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Spybrook, J., et al. (2011). Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software version 3.0. Available from http://hlmsoft.net/od/od-manual-20111016-v300.pdf

Van Breukelen, G. J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*(9), 920-925.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *The Annals of Mathematical Statistics, 18*(1), 88-96.

# Notes

[1] Blocking on a continuous variable is possible, too, but it is somewhat more involved. One approach, recommended by Dalton and Overall (1977), Maxwell, Delaney, and Hill (1984) and McAweeny and Klockars (1998), is to rank the participants by their covariate score and assign them to the conditions according to an ABBAABB… pattern. Whether 'A' refers to the treatment or the control group is determined at random. The data can then be analyzed in an ANCOVA with the covariate score as a continuous predictor. A discussion on blocking on several covariates simultaneously lies far outside the scope of this article, but see Moore (2012).

[2] The classes can also all be taught by the same teacher. What is important for this example is that the ten classes are not taught by, say, five teachers, each of whom teaches two classes. Such a design would further invalidate the assumption of independence (see below in main text).

[3] These power computations assume that the analyst knows the ICC value and uses it in the analysis; if the ICC value is unknown the power levels will be lower (see Blair & Higgins, 1986).

[4] In statistical packages that do not offer the facility to weight cluster means by cluster size, the same can be accomplished by computing a linear regression on the cluster means with the experimental condition as a (nominal) predictor and cluster sizes as weights.

[5] To compute a LRT statistic, the models being compared need to be fitted using maximum likelihood. Maximum likelihood estimates of variances tend to be too low, however, and more so in small samples.