Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 2

12-1-2017

Experimental Design and Data Analysis in Computer Simulation Studies in the Behavioral Sciences

Michael Harwell
University of Minnesota - Twin Cities, harwe001@umn.edu

Nidhi Kohli
University of Minnesota - Twin Cities

Yadira Peralta
University of Minnesota - Twin Cities

Recommended Citation
Harwell, M., Kohli, N., & Peralta, Y. (2017). Experimental Design and Data Analysis in Computer Simulation Studies in the Behavioral Sciences. Journal of Modern Applied Statistical Methods, 16(2), 3-28. doi: 10.22237/jmasm/1509494520

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 3-28.
doi: 10.22237/jmasm/1509494520

ISSN 1538 − 9472

Experimental Design and Data Analysis in Computer Simulation Studies in the Behavioral Sciences

Michael Harwell
University of Minnesota,
Twin Cities
Minneapolis, MN

Nidhi Kohli
University of Minnesota,
Twin Cities
Minneapolis, MN

Yadira Peralta
University of Minnesota,
Twin Cities
Minneapolis, MN

Treating computer simulation studies as statistical sampling experiments subject to established principles of experimental design and data analysis should further enhance their ability to inform statistical practice and a program of statistical research. Latin hypercube designs to enhance generalizability and meta-analytic methods to analyze simulation results are presented.

Keywords: simulation, experimental design, data analysis

Introduction

Computer simulation studies represent an important tool for investigating statistical procedures difficult or impossible to study using mathematical theory or real data. Descriptors of these studies vary (e.g., statistical experiment, Monte Carlo simulation, computer experiment), but the examples of Hoaglin and Andrews (1975) and Hauck and Anderson (1984) are followed here with use of the term simulation studies. Extensive descriptions of simulation studies can be found in Lewis and Orav (1989) and Santner, Williams, and Notz (2003).

In the behavioral sciences simulation studies have been used to study a wide array of statistical methods (e.g., Cribbie, Fiksenbaum, & Wilcox, 2012; Depaoli, 2012; Enders, Baraldi, & Cham, 2014; Hu & Bentler, 1999; Tomarken & Serlin, 1986). The general goal of these studies is to provide evidence of the behavior of statistical methods under a variety of data conditions that improves statistical practice and informs future statistical research. The goal here is to encourage methodological researchers to treat these studies as statistical sampling

Michael Harwell is a Professor in the Department of Educational Psychology. Email him at harwe001@umn.edu.

experiments subject to established principles of experimental design and data analysis.

An underappreciated facet of simulation studies in statistics is their role in enhancing the reproducibility of scientific findings. The importance of reproducibility has gained momentum in numerous scientific arenas because of growing evidence that many findings cannot be replicated (Stodden, 2015). Concerns over reproducibility and the role of statistics were captured in *Statistics and science: A report of the London workshop on the future of the statistical sciences* (2014) which noted: "The reproducibility problem goes far beyond statistics, of course, because it involves the entire reward structure of the scientific enterprise. Nevertheless, statistics is a very important ingredient in both the problem and the remedy." (p. 27) Simulation studies in statistics can increase the likelihood that scientific findings can be reproduced by providing evidence of the impact of data that are perturbed on estimators, tests, bootstrapping methods, parameter estimation algorithms, model alterations, etc., and subsequent inferences (Stodden, 2015).

Computer simulation studies as statistical sampling experiments

Hoaglin and Andrews (1975) argued that simulation studies should be treated as statistical sampling experiments subject to established principles of research design and data analysis. Special attention is given to experimental design in simulation studies, because of its centrality in a research study and its ability to produce effects of interest, guide analyses of study outcomes, and enhance generalizability of study findings. Hoaglin and Andrews (1975) reviewed a sample of published studies using simulation methods and offered a harsh assessment of the state of the art: "Statisticians (who, of all people, should know better) often pay too little attention to their own principles of design, and they compound the error by rarely analyzing the results of experiments in statistical theory" (p. 124). Gentle (2003) reiterated this point: "A Monte Carlo study uses an experiment, and the principles of scientific experimentation should be observed." (p. vii)

Hauck and Anderson (1984) surveyed studies in five statistics journals and reported that 216 (18%) studies used simulation methods and found little evidence that the recommendations of Hoaglin and Andrews (1975) were being adopted. Harwell, Kohli, and Peralta-Torres (2017) updated the Hauck and Anderson (1984) results by surveying studies in six statistics journals between 1985 and

2012 and found the use of simulation studies had basically doubled since 1984, but less than 5% of 371 simulation studies used an identifiable experimental design. Harwell, Kohli, and Peralta-Torres (2017) also reported that 99.9% of these studies relied exclusively on visual analysis of simulation findings (i.e., "eyeballing" the results).

It is important to emphasize simulation studies have made critical contributions to improving statistical practice; however, the recommendations of Hoaglin and Andrews (1975) imply that treating a simulation study as a statistical sampling experiment can further exploit the ability of these studies to inform statistical practice and a program of statistical research. The latter reflects the case in which a simulation study is part of a research program that includes previous studies whose results inform the conceptualization and execution of a proposed simulation study. The aim of the current study, therefore, is to encourage methodological researchers in the behavioral sciences to routinely treat computer simulation studies as statistical sampling experiments to fully exploit their strengths.

Experimental Design

Experimental design should play a crucial role in simulation studies because of its ability to produce effects of interest, guide analyses of study outcomes, and enhance generalizability of findings. The latter is particularly important because of concerns that generalizability of simulation study findings is frequently limited due to the way that values of simulation factors are selected (Paxton, Curran, Bollen, Kirby, & Chen, 2001; Skrondal, 2000). Modeling realistic conditions such as skewed data and small sample sizes is essential to generalizing simulation results in ways that improve statistical practice; our focus is designs that support generalizing results to simulation factor values beyond those explicitly modeled, which should further enhance generalizability and improve statistical practice.

Santner et al. (2003) defined inputs in a simulation as numerical values of simulation factors that collectively define the experimental region which in turn define the design. Thus experimental design is a specification of values of simulation factors in the experimental region at which we wish to compute an outcome. Input values are sampled from a defined pool of values using one of several sampling methods. The sampling methods are labeled space-filling, because they fill the experimental region in some fashion. More formally, an experimental design is defined by a matrix in which the columns correspond to simulation factors whose elements are researcher-specified numerical values for

the factors, and whose rows represent a combination of input values that define so-called design points. Consider the full factorial case in which all combinations of factor levels are examined. Let m^k represent k factors with m values (levels) which are being investigated using m^k input values; for two factors the experimental region is defined by m^{k_1} by m^{k_2} input values. For example, a binary factor (F_1) with researcher-specified values 10 and 20 crossed with a second binary factor (F_2) with values 18, 29, and 34 produces the values in Table 1. Factor levels are typically recoded for simplicity, for example, −1, 0, and +1 in Table 1, but this is not necessary (Sanchez, 2007).

The above design has six design points defined by the six rows in Table 1 with the coded values in a row representing inputs. In full factorials space-filling is the result of sampling the entire pool of researcher-specified simulation factor values. This practice generates a predictable pattern of space-filling that may answer specified research questions but can limit generalizations.

Table 1. Experimental Design for a 2×3 Full Factorial

	Original Values		Coded Values	
Point	F_1	F_2	F_1	F_2
1	10	18	−1	−1
2	20	18	+1	−1
3	10	29	−1	0
4	20	29	+1	0
5	10	34	−1	+1
6	20	34	+1	+1

An alternative to full factorials are incomplete fractional factorials. Skrondal (2000) described how these designs can be used in simulation studies in ways that enhance generalizability by employing more conditions than would typically be used in a full factorial because higher order interactions (reflected in combinations of factor conditions) are not modeled. These designs are especially appropriate for enhancing generalizability when there are many factors that take only a few values.

Space-filling by random sampling. A related class of designs used to increase the generalizability of simulation findings relies on random sampling methods for space-filling (Santner et al., 2003). In some cases generalizability is increased by spreading points evenly over the experimental region, whereas in other instances points are concentrated on the boundaries of the experimental

region. One sampling method involves defining a pool of design points (with associated input values) assumed to follow a uniform distribution and taking a simple random sample.

Consider an exemplar simulation study investigating the impact of different numbers of clusters, within-cluster sample size, and distribution of cluster residuals when estimating fixed effects and the Type I error rate of tests of these effects for a two-level mixed (linear) model for continuous cross-sectional data. Suppose a pool of number of clusters (J) (J = 10, 11, 12, …, 50) was defined and a simple random sample taken; similarly, we could define design points as pairs of values of J and within-cluster sample size (n_j) that follow a uniform distribution (e.g., J = 10, 11, 12, …, 50; n_j = 5, 6, 7, ..., 100) and take a simple random sample (assuming a normal distribution of cluster residuals for simplicity). This method should enhance generalizability relative to full factorials like that in Table 1 but may not spread design points evenly across the experimental region. Stratified random sampling can potentially enhance generalizability by identifying a stratification variable and selecting a point at random from each stratum. For example, we could define strata using n_j (n_j strata defined as 5-10, 11-15, …, 95-100) with a pool of values of J within each stratum (e.g., J = 10, 11, 12, …, 50) one of which is selected at random from each stratum. The resulting design points ensure space-filling as they include the entire range of values of n_j as captured by the strata.

Perhaps the most widely recommended sampling method for space-filling to increase generalizability of simulation findings is Latin hypercube sampling, which generates a Latin hypercube design (LHD) (Santner et al., 2003). Latin hypercube designs are a variation of traditional Latin squares and spread design points evenly across the range of an input. Santner et al. (2003), Sanchez (2007), and Viana (2013) illustrated the use of LHDs in simulation studies for relatively simple designs and pointed out their benefits generally increase with increases in k; Sanchez (2007) noted the number of points (and potentially the generalizability) increases linearly with increases in k.

Let p denote the total number of design points and assume low and high levels (values) for a factor F_k are coded as 1 and p, and that the set of coded factor levels are 1, 2, …, p. A $p \times k$ design matrix for a LHD can be written as $\boldsymbol{X} = [\boldsymbol{x}_1\, \boldsymbol{x}_2 \ldots \boldsymbol{x}_p]^T$ where each column represents a factor and each row $\boldsymbol{x}_i = (x_i^{(1)}\, x_i^{(2)} \ldots x_i^{(k)})$ for i = 1, …, p represents a design point. In a LHD each factor is divided into p equal levels and one point is sampled at each level using a random procedure. Different optimization algorithms for LHD have appeared such as genetic-type algorithms, simulated annealing, optimum Euclidian distance,

and column-pairwise optimization (Carnell, 2016; Viana, 2013), and specialized software like the `lhs` package in R (R Core team, 2016) is needed to implement even simple LHDs. This software is illustrated below.

Exemplar simulation study. The rationale for our two-level mixed model exemplar comes from a review of statistical theory and previous simulation results (Austin, 2010; Bell, Ferron, & Kromrey 2008; Clarke & Wheaton, 2007; Maas & Hox, 2004, 2005; Maeda, 2007; Moerbeek, van Breukelen, & Berger, 2000). This literature suggests the number of clusters needed to accurately estimate fixed effects and to have tests of these effects control Type I error rates at nominal levels is unresolved for non-normal cluster residuals. This prompted the research question: How many clusters are needed in a two-level model for continuous cross-sectional data with one predictor at each level for conditions of varying within-cluster sample sizes and non-normal cluster residuals to ensure: (a) accurate estimation of fixed effects and (b) statistical tests of these effects control Type I error rates at nominal levels?

For this simulation exemplar the statistical model with one predictor at each level was

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \quad \text{(level 1)} \tag{1}$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + W_{1j}\gamma_{01} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + W_{1j}\gamma_{11} + u_{1j} \end{aligned} \quad \text{(level 2)}$$

which implies the mixed model $Y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j} + (\gamma_{10} + \gamma_{11}W_{1j} + u_{1j})X_{1ij} + r_{ij}$. In equation (1), Y_{ij} represents the (continuous) outcome score of the i^{th} level 1 unit in the j^{th} level 2 unit (cluster), β_{0j} and β_{1j} are the intercept and linear slope for the j^{th} cluster, X_{1ij} is a predictor value sampled from an $N(0,1)$ distribution), r_{ij} is that level 1 unit's residual ($r_{ij} \sim N(0,\sigma^2)$), γ_{00} is the average β_{0j}, γ_{10} is the average X_1,Y slope within clusters, γ_{01} is a slope capturing the effect of the level 2 predictor W_{1j}, γ_{11} is the slope capturing the cross-level interaction effect, and u_{0j} and u_{1j} are cluster residuals for the intercept and slope models (Raudenbush & Bryk, 2002, pp. 100-103). The fixed effects in equation (1) (γ_{00}, γ_{01}, γ_{10}, γ_{11}) were set to zero to reflect the Type I error case meaning the mixed model underlying the data generation was simply $Y_{ij} = u_{0j} + u_{1j} X_{1ij} + r_{ij}$.

To specify simulation conditions we relied on statistical theory (Raudenbush & Bryk, 2002, chpt. 3), previous simulation studies, and documented

characteristics of large multilevel datasets (Hedges & Hedberg, 2007). We assumed $\begin{bmatrix} u_{0t} \\ u_{1t} \end{bmatrix} \sim [\mathbf{0},\mathbf{T}]$ followed a normal or chi-square distribution (see below), where $\mathbf{T} = \begin{bmatrix} 9 & 0 \\ 0 & .75 \end{bmatrix}$ was a 2×2 covariance matrix of random effects with diagonal entries τ_{00} (variance of u_{0j}) and τ_{11} (variance of u_{1j}), and covariance τ_{01}. We specified $\tau_{00} > \tau_{11}$ based on Lee and Bryk (1989) who reported a within-cluster variance for mathematics achievement data of 39.927 for their unconditional model, a between-cluster intercept variance of 9.335, and three between-cluster slope variances whose average was .75. Using values of 40 and 9 for σ^2 and τ_{00} in the unconditional model in our simulation produced an intra-class correlation (ICC) of .19, which is consistent with the results of Hedges and Hedberg (2007). The covariance component τ_{01} was set to 0 based on simulation evidence that this value typically has little impact on the number of clusters (Maas & Hox, 2004, 2005; Zhang, 2005). The resulting pool of inputs in our exemplar study was specified as J = 10, 20, 30, 40, 50 (number of clusters), n_j = 18, 29, 34, 44, 60, 68 (within-cluster sample sizes), and distribution = BVN ~ $\begin{pmatrix} 0 & 9 \\ 40 & 0 \end{pmatrix}$, χ^2_{10} (distribution of cluster residuals). n_j values were selected at random from a range of 5 to 100, because there was no empirical basis for specifying particular values. Data were simulated using the R software.

The estimated fixed effects in the exemplar study served as indicators of bias because the true values equaled zero, and were computed as an average across R = 5,000 replications. Type I error rates of tests of the fixed effects were estimated as the proportion of rejections of the associated statistical null hypothesis across R replications. R = 5,000, a number that generally provides accurate estimates of Type I error rates for general linear model-based statistical tests (Robey & Barcikowski, 1992) and should do the same for bias estimates. Next, the exemplar study is used to illustrate space-filling for a full factorial and LHD, and meta-analysis to analyze simulation results.

Results

The resulting design matrix for the exemplar had three columns and 60 rows (design points) and sampling all design points produced a 5×6×2 full factorial design with 60 cells. We conditioned the design on a particular distribution of cluster residuals (bivariate normal, chi-square); otherwise we must generate a

pool of input values representing distributions. If the focus was exclusively on the two distributions in the exemplar study these define the pool of inputs and the exemplar design matrix would have three columns and 60 rows. If instead the desire is to generalize findings to a family of skewed distributions such as chi-square a pool of input values defined by degrees of freedom could be specified, for example, $df = 1, 2, \ldots, 20$, in which case the design matrix would have three columns and 600 rows. To simplify the graphical display we focus on J and n_j meaning the exemplar study design matrix has two columns and 60 rows. The `lhs` package in R was used to generate the experimental region for the 5×6 full factorial displayed in Figure 1, which is a grid composed of 30 points. Notice the lines of dots for J are equidistant from each other whereas those for n_j vary in distance because the latter vary in value. This figure highlights the non-random nature of space-filling for the 5×6 full factorial which limits generalizability to selected input values.

Employing a LHD signals we are interested in generalizing to design points not explicitly modeled in the simulation. This strategy supports generalizing findings to a pool of design points in ways not possible with a full factorial, and with less uncertainty compared to simple random sampling of points because space-filling throughout the experimental region is not assured.

To construct a LHD for the exemplar simulation study we used the `maximinLHS` function in the lhs package in R, which draws a Latin hypercube sample from a set of uniform distributions that can be rescaled to the range of interest (Carnell, 2016). The `maximinLHS` function optimizes the sample by maximizing the minimum distance between design points (Carnell, 2016). In order to create the LHD we drew a sample of 30 points considering two factors. The resulting design points were then rescaled to the ranges covered by factors one $J = (10, 11, \ldots, 49, 50)$ and two ($n_j = 5, 6, 7, \ldots, 100$) in our exemplar study. That is, F_1 (number of clusters) was rescaled to have values between 10 and 50 and F_2 (within-cluster sample size) to have values between 18 and 68. The number of sampled factor values (inputs) depends on the desired generalizability with more values expected to provide greater space-filling, although this may have to be weighed against available computing resources (Santner et al., 2003).

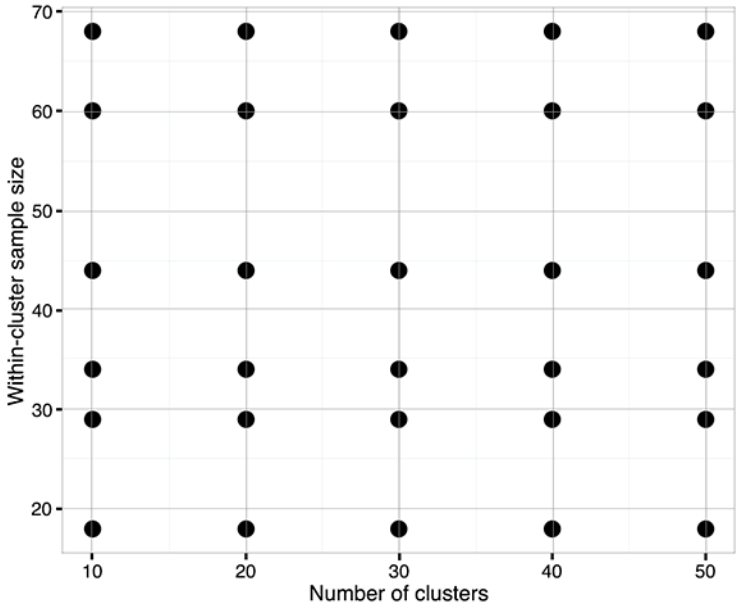

Figure 1. Experimental region for the exemplar simulation study with full factorial design conditioning on distribution of cluster residuals.

Shown in Figure 2 are the design points associated with the LHD for the exemplar simulation study, which are spread evenly across the experimental region. The implication of the LHD in Figure 2 is that findings of our exemplar simulation study are generalizable to the entire pool of researcher-specified values of J and n_j not just those explicitly modeled. Figure 3 contrasts Figures 1 and 2 and illustrates the systematic, non-random space-filling of a full factorial versus the random-sampling-based space-filling of a LHD. R code for generating the experimental regions illustrated in Figures 1-3 appears in Appendix A.

The enhanced generalizability linked to LHDs speaks to their potential to improve statistical practice and inform future statistical research. However, there are areas of statistical research employing simulation methods in which sampling all design points is appropriate because interest is limited to those inputs, perhaps because of theoretical or empirical reasons. For example, interest may be limited to a small number of distributions as was the case for the exemplar, where the space-filling illustrated in Figure 1 is appropriate.

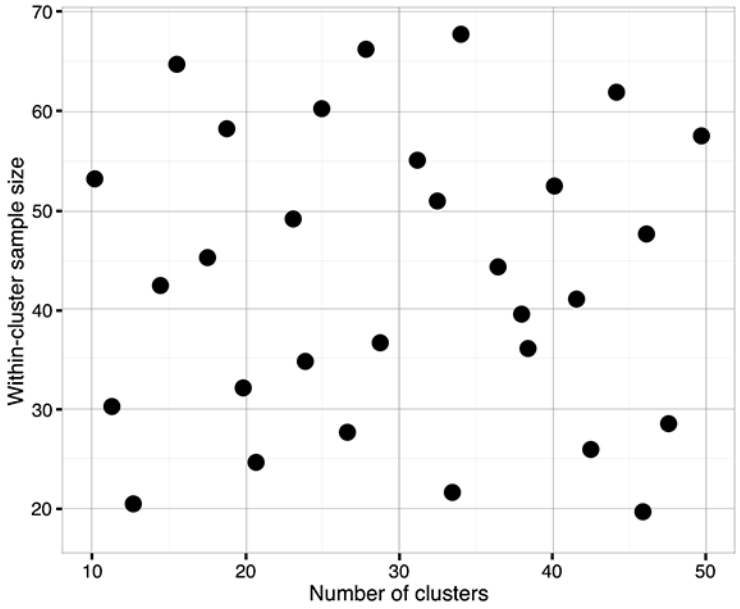

Figure 2. Experimental region with random selection of inputs for Latin Hypercube design conditioning on distribution of cluster residuals.

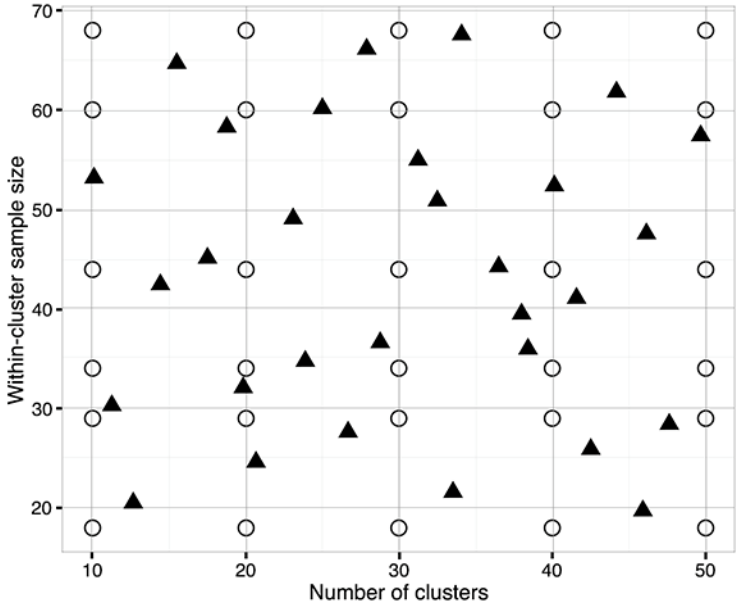

Figure 3. Contrasting the experimental region of the full factorial versus Latin Hypercube design.

Analysis of simulation results

Despite the recommendations of Hoaglin and Andrews (1975), Skrondal (2000), Boomsma (2013), Paxton et al. (2001) and others the analysis and reporting of results continues to rely heavily on visual analyses (Harwell et al., 2017). When there are exceptions they typically involve factorial ANOVA (e.g., Culpepper & Aguinis, 2011; MacCallum, Widaman, Preacher, & Hong, 2001), or less frequently logistic regression (e.g., Skrondal, 2000). Relying on visual analysis of simulation results is reasonable if key patterns and their magnitude are accurately captured such as interaction effects. On the other hand, reliance on tables and plots when summarizing information in dozens, hundreds, or even thousands of simulation results raises questions about how accurately important patterns can be detected and how precisely their magnitude can be estimated. We argue that visual analysis should typically be augmented by inferential analyses of results guided by the experimental design.

Visual analysis of simulation results. Methodological researchers have traditionally relied on visual analyses of simulation results which often appear in tables regardless of the number of simulation outcomes. For example, Wilcox (2009) reported three tables each containing 48 simulation results, Ramsey and Ramsey (2009) reported 1,750 values in five tables, and, as an extreme example, Aaron (2003) reported more than 7,000 values. The accuracy of visual analyses to summarize patterns and estimate the magnitude of effects in studies like Ramsey and Ramsey (2009) has not been tested experimentally, for example, by assembling a group of methodological researchers and assessing their ability to accurately detect patterns in simulation results using artificial sets of findings varying in known ways (e.g., entirely random pattern, only one effect). However, the ability to reliably and validly detect patterns using visual analysis has been studied in other research domains.

Single-case designs in psychology and education (Kratochwill et al., 2010; Kratochwill & Levin, 1992, 2010) involve collecting and plotting repeated measures data to assess the impact of one or more interventions (Smith, 2012). A good deal of research (Bailey, 1984; DeProspero & Cohen, 1979; Jones, Vaught, & Weinrott, 1977; Knapp, 1983; Matyas & Greenwood, 1990) assessing the ability of researchers, clinicians, and others to reliably and validly detect patterns using visual analysis highlighted the difficulties of doing so even for relatively small numbers of data points (e.g., 10-15), and the use of visual and inferential analyses has been recommended (Ferron, 2002; Kratochwill et al., 2010). Consider the estimated Type I error rates in Table 2 generated in the exemplar

study assuming a full factorial design. Values falling outside a 95% confidence interval are treated as sensitive to the conditions modeled. It's clear that a majority of $\hat{\alpha}$ values are inflated and that increases in the number of clusters seem to be associated with $\hat{\alpha}$ values closer to .05; within-cluster sample size and the distribution of cluster residuals do not seem to have much impact. Similarly, a visual analysis of average bias values in Table 3 suggests a chi-square distribution of cluster residuals produces somewhat more bias which generally shrinks as J increases. Careful visual analysis is important but performing inferential statistical analyses and estimating the magnitude of effects can provide additional insight into the impact of simulation factors on outcomes of interest.

Table 2. Estimated Type I error rates for tests of γ_{01} and γ_{11}

	J	***10***	***20***	***30***	***40***	***50***
u_{0j}, u_{1j} distribution	n_j		**Type I Error Rate, γ_{01}**			
$N(0,9)$ and $N(0,0.75)$	***18***	.085 *	.069 *	.055	.058 *	.057 *
	29	.084 *	.069 *	.057 *	.060 *	.057 *
	34	.093 *	.065 *	.059 *	.052	.058 *
	44	.085 *	.062 *	.054	.055	.055
	60	.084 *	.068 *	.057 *	.051	.047
	68	.085 *	.065 *	.059 *	.053	.054
χ^2_{10}	***18***	.090 *	.069 *	.063 *	.054	.059 *
	29	.089 *	.064 *	.062 *	.053	.060 *
	34	.086 *	.057 *	.061 *	.063 *	.052
	44	.086 *	.063 *	.058 *	.058 *	.051
	60	.085 *	.058 *	.063 *	.057 *	.060 *
	68	.093 *	.069 *	.055	.055	.056
u_{0j}, u_{1j} distribution	n_j		**Type I Error Rate, γ_{11}**			
$N(0,9)$ and $N(0,0.75)$	***18***	.051	.053	.049	.050	.052
	29	.060 *	.055 *	.056	.055	.049
	34	.063 *	.060 *	.062 *	.057 *	.050
	44	.067 *	.065 *	.068 *	.063 *	.063 *
	60	.071 *	.064 *	.059 *	.056	.058 *
	68	.074 *	.072 *	.063 *	.063 *	.051
χ^2_{10}	***18***	.086 *	.061 *	.067 *	.055	.055
	29	.083 *	.062 *	.054	.059 *	.064 *
	34	.086 *	.064 *	.055	.059 *	.059 *
	44	.086 *	.068 *	.065 *	.058 *	.055
	60	.091 *	.063 *	.057 *	.055	.054
	68	.093 *	.063 *	.061 *	.054	.055

Note: Tabled values represent estimated Type I error rate across R = 5,000 replications, * = an error rate falling outside the 95% confidence interval limits, u_{0j} and u_{1j} represent cluster residuals, J = number of clusters, n_j = within-cluster sample size.

Table 3. Average bias for γ_{01} and γ_{11}

	J	***10***	***20***	***30***	***40***	***50***
u_{0j}, u_{1j} distribution	n_j			**Average bias for γ_{01}**		
	18	-0.0138	-0.0249	-0.0202	-0.0148	-0.0141
	29	-0.0126	0.0100	0.0045	-0.0054	0.0079
N(0,9) and	***34***	-0.0050	0.0104	0.0101	0.0059	0.0007
N(0,0.75)	***44***	0.0276	-0.0141	0.0028	-0.0004	-0.0069
	60	0.0467	-0.0199	-0.0067	0.0183	0.0042
	68	-0.0180	0.0087	0.0059	0.0012	-0.0074
	18	0.0226	-0.0120	-0.0231	-0.0154	0.0135
	29	-0.0411	0.0154	-8.52E-05	-0.0230	-0.0055
χ^2_{10}	***34***	-0.0432	-0.0450	0.0073	-0.0104	0.0055
	44	0.0411	-0.0132	0.0126	-0.0035	0.0006
	60	0.0571	0.0228	-0.0024	0.0141	0.0069
	68	0.0076	0.0092	0.0247	0.0047	-0.0045
u_{0j}, u_{1j} distribution	n_j			**Average bias for γ_{11}**		
	18	-0.0025	0.0077	-0.0035	-0.0029	-0.0036
	29	0.0041	-0.0002	-0.0015	-0.0054	-0.0008
N(0,9) and	***34***	-0.0019	-0.0011	-0.0076	-0.0025	-0.0047
N(0,0.75)	***44***	0.0111	-0.0058	-0.0055	-0.0049	0.0076
	60	0.0023	-0.0034	-0.0061	0.0051	0.0005
	68	0.0033	0.0008	0.0021	7.04E-05	0.0074
	18	-0.0234	-0.0139	-0.0189	-0.0184	0.0066
	29	-0.0088	-0.0078	0.0203	-0.0149	0.0063
χ^2_{10}	***34***	0.0137	0.0236	-0.0260	0.0085	-0.0052
	44	0.0032	-0.0336	-0.0018	-0.0231	0.0010
	60	0.0146	0.0146	-0.0048	0.0135	0.0145
	68	-0.0269	0.0249	0.0209	0.0216	0.0135

Note: Tabled values represent average bias across R = 5,000 replications, and represent cluster residuals, J = number of clusters, n_j = within-cluster sample size.

Meta-analysis of simulation results. Next, consider the use of meta-analysis to detect patterns in simulation results. Assume the typical case in which simulation outcomes are averaged across R replications in each cell of the design and a fixed effect full factorial design for our exemplar study. However, the method described below can be adapted to LHDs (see below). It is assumed model-checking will be performed to ensure underlying assumptions are plausible.

Meta-analytic methods permit the relationship between simulation factors and outcomes to be assessed and also provide a test of model misspecification. The averaged outcome for each cell serves as an effect size, for example, $\hat{\alpha}$ or

$\overline{bias} = \sum_{s=1}^{R_s} \frac{\left(\hat{\theta}_s - \theta\right)}{R_s}$, $\hat{\theta}_s$ = s^{th} estimated parameter, θ = parameter, and R_s = number of replications $\hat{\theta}_s$ is based on. The mean and variance of outcomes must be available and for $\hat{\alpha}_s$ are well known. The expression $\sum_{s=1}^{R_s} \frac{\left(\hat{\theta}_s - \theta\right)^2}{R_s}$ provides a variance estimate for bias $\left[Var\left(\overline{bias}\right)\right]$ that can serve as an effect size of the impact of simulation factors on the variability of bias estimates. To treat $\left[Var\left(\overline{bias}\right)\right]$ as an effect size $\ln\left[Var\left(\overline{bias}\right)\right]$ is computed under the assumption $\hat{\theta}_s$ values are normally-distributed (Raudenbush & Bryk, 1987). In this case $Var\left\{\ln\left[Var\left(\overline{bias}\right)\right]\right\} = \frac{2}{S - H - 1}$ (S = total number of effect sizes) which allows inferential analyses of $\ln\left[Var\left(\overline{bias}\right)\right]$ values. Similar expressions are available for other outcomes such as statistical power and model convergence rates.

Consider a meta-analytic regression model for Type I error rates:

$$\alpha_s = \beta_0 + \sum_{h=1}^{H} \beta_H X_{SH}, \hat{\alpha}_s = \alpha_s + \xi_s \tag{2}$$

In equation (2), α is the s^{th} effect size (population proportion, s = 1, 2, ..., S) that depends on a set of H predictor variables X_{SH} which could include interactions, β_0 is a population intercept, β_H is a population regression coefficient that captures the linear relationship between a predictor and α_s, ξ_s is a population error term, and $\hat{\alpha}_s$ is an estimated Type I error rate (proportion) (Hedges & Olkin, 1985, p. 169). The fitted model has the form:

$$\hat{\alpha}_s^{'} = \hat{\beta}_0 + \sum_{h}^{H} \hat{\beta}_H X_{SH} \tag{3}$$

In equation (3), $\hat{\beta}_H$ is an estimated slope and $\hat{\alpha}_s^{'}$ is a model-predicted proportion. The relationship between a set of predictors and effect sizes can be tested using the Q_{Reg} statistic presented in Hedges and Olkin (1985, p. 169-171). Assume the distribution of errors is normal with a mean of zero and diagonal

covariance matrix $\Sigma_{\hat{\alpha}}$ with dimensions $S \times S$ and elements $\sigma^2_{\hat{\alpha}}$. The Q_{Reg} test statistic equals the weighted sum of squares due to regression for the model in equation (3) with weights $\left[\sigma^2_{\hat{\alpha}}\right]^{-1} = \frac{R_s}{\hat{\alpha}_s(1-\hat{\alpha}_s)}$, where R_s is the number of replications associated with $\hat{\alpha}_s$. Under the hypothesis $H_0 : \beta = 0$, where β and 0 are $H \times 1$ vectors, Q_{Reg} follows a chi-square distribution with $df = H$. Because $\hat{\alpha}_s$ represents binomial data, a data-analytic alternative is to initially transform each $\hat{\alpha}_s$ using the arcsine transformation (Cox, 1970). The mean and variance of the transformed quantities ($\hat{\alpha}_s^{\text{arcsine}}$) are independent and the assumption of normality is typically plausible even for modest sample sizes. The transformed quantities follow $\hat{\alpha}_s^{\text{arcsine}} \sim N\left[E\left(\hat{\alpha}_s^{\text{arcsine}}\right) = \alpha_s^{\text{arcsine}}, Var\left(\hat{\alpha}_s^{\text{arcsine}}\right) = \frac{1}{S}\right]$ and serve as outcomes in equation (2).

A key feature of the meta-analytic approach is the ability to test model specification i.e., whether all predictors contributing to variation in effect sizes are in the model (Hedges & Olkin, 1985, p. 172). The test for misspecification relies on a weighted error sum of squares associated with the model in equation (2) that is computed using the test statistic $Q_{Error} = \hat{\alpha}' \sum_{\hat{\alpha}}^{-1} \hat{\alpha} - Q_{Reg}$, where $\hat{\alpha}$ is a $S \times 1$ vector of the $\hat{\alpha}_s$. If the model is correctly specified Q_{Error} it is distributed as a chi-square variable with $df = S - H - 1$. Rejection of the hypothesis that the model is correctly specified implies that the weighted error variance is larger than expected, results are subject to misspecification bias (Hedges & Olkin, 1985, p. 172), and adding additional predictors could reduce error and produce less biased estimates.

In all cases the Q tests assume normality and because of the large numbers of replications typically used in simulation the normality approximation for $\hat{\alpha}_s$ should be quite good. Alternatively weighted logistic regression could be used to estimate parameters and test hypotheses for $\hat{\alpha}_s$. The Hedges and Olkin (1985) Q tests were chosen because: (a) these tests can be applied to a variety of effect sizes, (b) this approach provides a widely adopted measure of explained variance (R^2) which is not always the case for weighted logistic regression although it is important to recall that R^2 in weighted least squares represents the variance in the weighted outcomes explained by the weighted prediction model (Willet & Singer, 1988), (c) existing data analysis software can be used to fit the models. Note the meta-analytic regression model in equation (2) assumes predictor values are fixed whereas for LHDs predictor values such as those for J and n_j are sampled at

random. In practice predictors whose values are fixed and those representing random variables produce the same statistical inferences since the former can be considered realizations of the latter (Sampson, 1974). Thus simulation results from LHDs can be analyzed using equation (2) by treating the sampled simulation factor values as realizations from a larger pool of such values.

To illustrate the Q tests consider the results in Table 2. The fixed effects (γ_{00}, γ_{01}, γ_{10}, γ_{11}) could be treated as a within-subjects factor in the analyses but we chose to examine the γ_{01} and γ_{11} results separately (results for γ_{00} and γ_{10} were similar to those for γ_{01} and γ_{11}). The predictors were number of clusters, within-cluster sample size, and distribution of cluster residuals that were centered about their mean, and their two-way interactions. The resulting $Q_{Reg} = 428.2$ ($p < .05$) for the γ_{01} Type I error results signals a statistically significant relationship between Type I error rates and the set of predictors, and the associated R^2 of .66 indicates there is a strong predictive relationship almost all of which ($R^2 = .65$) is attributable to number of clusters. The model-predicted error rates for number of clusters were .077 ($J = 10$), .071 (20), .064 (30), .057 (40), and .051 (50). Post hoc analyses were performed testing each slope against zero (Hedges & Olkin, 1985, p. 174) and controlling for compounding Type I error rates using the method of Sidak (1967) such that the error rate for each test was .05/6 = .0083. Only the slope for the number of clusters predictor was significant (−.001), meaning that Type I error rates for the test of γ_{01} were on average insensitive to within-cluster sample size and cluster residual distribution as well as the three two-way interactions but were sensitive to number of clusters. Testing model misspecification produced a statistically significant test ($Q_{Error} = 223.4$, $p < .05$), implying that the regression findings should be interpreted cautiously and adding predictor variables could potentially reduce error variation and bias in parameter estimates.

The model in equation (3) was then fitted to $\hat{\alpha}_s$ for the test of γ_{11} and obtained $Q_{Reg} = 300.7$ ($p < .05$), meaning there was a statistically significant and, it turns out, strong ($R^2 = .67$) relationship between $\hat{\alpha}_s$ and the set of predictors. Post hoc analyses showed that cluster residual distribution, within-cluster sample size, and the interactions number of clusters × within-cluster sample size and number of clusters × cluster residual distribution were significant predictors. Approximately 18% ($R^2 = .18$) of the variance in was attributable to cluster residual distribution, followed by within-cluster sample size (11%), and the interactions number of clusters × level 2 residual distribution (8%) and within-cluster sample size × cluster residual distribution (6%).

Model-predicted error rates for cluster residual distribution were .059 (normal) and .064 (chi-square) and for number of clusters were .071 (J = 10), .066 (20), .062 (30), .057 (40), and .053 (50); for within-cluster sample size the average model-predicted error rates ranged from .059 to .065. The interaction plot for number of clusters × cluster residual distribution showed a discrepancy for J = 10 with an average error rate of .077 for a chi-square distribution and .065 for a normal distribution and .072, and .063 for J = 20; otherwise average error rates were similar for the remaining conditions. The interaction plot for within-cluster sample size x cluster residual distribution showed a modest difference for J = 10 with an average error rate of .061 for a chi-square distribution and .054 for a normal distribution, and .057 and .062 for J = 20; otherwise average error rates were quite similar. A test of model misspecification produced a significant result (Q_{Error} = 149.8, p < .05) meaning that the findings should be interpreted cautiously and adding predictor variables could reduce error variation and bias in parameter estimates.

Comparing a visual analysis of Table 2 with the inferential results reveals several important differences. For γ_{01} the tabular results showed a majority of $\hat{\alpha}_s$ values were inflated and that increases in the number of clusters seem to be associated with values closer to .05; within-cluster sample size and the distribution of cluster residuals did not seem to have much impact. The inferential analyses supported these inferences but quantified the predictive strength of number of clusters with 65% of the variance attributable to this factor. For γ_{11} a majority of Type I error rates were also inflated but also seemed to move toward .05 as J increased particularly for $J \geq 30$. The inferential analyses demonstrated that error rates were less sensitive to simulation factors than those for γ_{01} and more sensitive to cluster residual distribution than J. The results also showed that combinations of factors impacts Type I error rates although the strength of these effects was modest.

Conclusion

A substantial amount of simulation research is available that has unquestionably made important contributions to improving statistical practice and informing future statistical research, yet the potential of these studies has not yet been fully realized in large part because recommendations to treat them as statistical sampling experiments have not been widely adopted. Adopting the recommendations of Hoaglin and Andrews (1975) should enhance the contributions of simulation studies including their role in increasing the

reproducibility of findings of studies employing statistical analyses. Following Hoaglin and Andrews (1975), the focus was on two key facets of a simulation study: experimental design and analysis of results.

The presence of a literature focused on experimental designs in simulation studies that enhance generalizability, and the availability of software to construct these designs, provides an important resource for methodological researchers. It is argued it is first important to adopt some kind of identifiable experimental design. Of course, simulation studies in some statistical research areas are quite similar, so much so that this may explain why the design is not reported. For example, simulation studies such as Ramsey and Ramsey (2009) typically employed multiple categorical simulation factors and report results in a fashion consistent with a full factorial design but do not identify the design used. Reporting the experimental design used in the study (assuming there is one) and other relevant details is consistent with Hoaglin and Andrews (1975) recommendation "A published report of computation-based results must make it easy for the reader to make reasonable assessments of the numerical quality of the results." (p. 124).

Describing the experimental design also allows readers to assess the generalizability of findings. Simulation studies by their nature offer strong internal validity but require special attention be given to generalizability. Designs in which simulation factor values are randomly sampled from a researcher-specified pool of values, such like Latin hypercube designs, speak to issues of generalizability. Of course, not every simulation study is focused on enhancing generalizability but there appear to be many instances in which adopting designs such as a Latin hypercube can increase their contribution. Construction of a Latin hypercube for our exemplar simulation study highlighted the enhanced generalizability this design offers.

A second facet was analysis of simulation results. Visual analysis of results as illustrated in our exemplar study was useful, but augmenting this approach with inferential methods should improve the accuracy with which patterns are detected and their magnitude estimated. Inferential analysis of simulation results is also consistent with the recommendations of Hoaglin and Andrews (1975). Meta-analytic methods treat simulation outcomes as effect sizes and simulation factors as predictors in a regression model. This approach provides a test of the relationship between the simulation factors and outcomes and an index of explained variance if this relationship is statistically significant. A test of model misspecification provides an important tool for properly modeling variation in outcomes as well as interpreting simulation findings.

What next?

Efforts to encourage methodological researchers to adopt recommendations to increase the impact of simulation studies by treating them as statistical sampling experiments have had limited success in the past four decades. Those who advocated recommendations of Hoaglin and Andrews (1975) be adopted seem to have assumed these recommendations possess a kind of face validity, i.e., their merit is obvious especially to individuals who subscribe to the importance of established principles of experimental design and data analysis. Clearly, this argument has not been sufficiently compelling and changing the conceptualization, execution, and reporting of computer simulation studies in ways consistent with Hoaglin and Andrews (1975) will require continued efforts to convince authors, reviewers, and editors of their merit.

References

Aaron, L. A. (2003). *A comparative simulation of Type I error and power of four tests of homogeneity of effects for random- and fixed-effects models of meta-analysis*. Unpublished doctoral dissertation. University of South Florida.

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics*, *6*(1), Article 16. doi: 10.2202/1557-4679.1195

Bailey, D. B. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis, 17*(3), 359-365. doi: 10.1901/jaba.1984.17-359

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*, 1122-1129.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 518-540. doi: 10.1080/10705511.2013.797839

Carnell, R. (2016). *lhs: Latin Hypercube Samples*. R package version 0.14. https://CRAN.R-project.org/package=lhs

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Research & Methods, 35,* 311-351. doi: 10.1177/0049124106292362

Cox, D. R. (1970). *Analysis of binary data*. London, U.K.: Chapman and Davis.

Cribbie, R. A. A., Fiksenbaum, L., & Keselman, H. J. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology, 65*(1), 56–73. doi: 10.1111/j.2044-8317.2011.02014.x

Culpepper, S. A. & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods, 16*(2), 166-178. doi: 10.1037/a0023355

Depaoli, S. (2012). Measurement and structural model class separation in mixture CFA: ML/EM versus MCMC. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(2), 178-203. doi: 10.1080/10705511.2012.659614

DeProspero A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*(4), 573–579. doi: 10.1901/jaba.1979.12-573

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods, 19*(1), 39–55. doi: 10.1037/a0035314

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers, 34*(3), 324-331. doi: 10.3758/bf03195459

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods* (2nd ed.). New York, NY: Springer. doi: 10.1007/b97336

Harwell, M., Kohli, N., & Peralta-Torres, Y. (2017). A survey of reporting practices of computer simulation studies in statistical research. *The American Statistician*. Advance online publication. doi: 10.1080/00031305.2017.1342692

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician, 38*(3), 214-216. doi: 10.1080/00031305.1984.10483206

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press: Orlando.

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. doi: 10.3102/0162373707299706

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician, 29*(3), 122-126. doi: 10.1080/00031305.1975.10477393

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi: 10.1080/10705519909540118

Jones, R. R., Vaught, R. S., & Weinrott, M. R. (1977). Time series analysis in operant research. *Journal of Applied Behavior Analysis, 10*(1), 151-166. doi: 10.1901/jaba.1977.10-151

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*(2), 155-164.

Kratochwill, T. R., & Levin J. R. (1992). *Single-case research design and data analysis: New directions for psychology and education*. Howe, UK: Lawrence Erlbaum.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). *Single-case designs* [Technical documentation]. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124–144. doi: 10.1037/a0017736

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*(3), 172-192. doi: 10.2307/2112866

Lewis, P. A. W., & Orav, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* (Vol. 1). Pacific Grove CA: Wadsworth and Brooks/Cole.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137. doi: 10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 85-91. doi: 10.1027/1614-1881.1.3.86

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611–637. doi: 10.1207/s15327906mbr3604_06

Maeda, Y. (2007). *Monte Carlo evidence regarding the effects of violating assumed conditions of two-level hierarchical models for cross-sectional data.* Unpublished doctoral dissertation, University of Minnesota Twin Cities.

Matyas T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*(3), 341–351. doi: 10.1901/jaba.1990.23-341

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25*(3), 271-284. doi: 10.3102/10769986025003271

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287–312. doi: 10.1207/S15328007SEM0802_7

R Core Team. (2016). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from: http://www.R-project.org/ (Version 3.3.1).

Ramsey, P. H., & Ramsey, P. P. (2009). Power and Type I errors for pairwise comparisons of means in the unequal variances case. *British Journal of Mathematical and Statistical Psychology, 62*(2), 263-281. doi: 10.1348/000711008X291542

Raudenbush, S. J., & Bryk A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics, 12*(3), 241-269. doi: 10.2307/1164686

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). New York, NY: Sage Publications.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*(2), 283–288. doi: 10.1111/j.2044-8317.1992.tb00993.x

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association, 69*(347), 682-689. doi: 10.2307/2286002

Sanchez, S. M. (2007). Work smarter, not harder: Guidelines for designing simulation experiments. In Henderson, S. G., Biller, B., Hsieh, M. H., et al. (Eds.)

Proceedings of the 2007 Winter Simulation Conference, 9-12 December 2007, Washington DC. pp. 84-94. doi: 10.1109/WSC.2007.4419591

Santner, T., Williams, B., & Notz, W. (2003). *The design and analysis of computer experiments*. New York, NY: Springer Verlag. doi: 10.1007/978-1-4757-3799-8

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of American Statistical Association, 62*(318), 626–633. doi: 10.2307/2283989

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35*(2), 137–167. doi: 10.1207/S15327906MBR3502_1

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510-550. doi: 10.1037/a0029312

Statistics and Science: A Report of the London Workshop on the Future of the Statistical Sciences (2014). [Technical Report]. London, England. Retrieved from http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and its Applications, 2*, 1–19. doi: 10.1146/annurev-statistics-010814-020127

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(1), 90-99. doi: 10.1037/0033-2909.99.1.90

Viana, F. A. C. (2013). *Things you wanted to know about the Latin hypercube design and were afraid to ask*. 10th World Congress on Structural and Multidisciplinary Optimization, Orlando, FL.

Wilcox, R. R. (2009). Robust ANCOVA using a smoother with bootstrap bagging. *British Journal of Mathematical and Statistical Psychology, 62*(2), 427–437. doi: 10.1348/000711008X325300

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. *Communications in Statistics—Simulation and Computation, 15*(4), 933-943. doi: 10.1080/03610918608812553

Willett, J. B., & Singer, J. D. (1988). Another cautionary note about : It's use in weighted least squares regression analysis. *The American Statistician, 42*(3), 236-238. doi: 10.2307/2685031

Zhang, D. (2005). *A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling*. Unpublished doctoral dissertation, Texas A & M University.

Appendix A: R code for Figures 1 - 3

Full Factorial Design (Figure 1)

```
# Libraries needed
library(ggplot2)
library(lhs)
library(scales)
grid.full <- expand.grid(f1 = c(10, 20, 30, 40, 50),
                         f2 = c(18, 29, 34, 44, 60, 68))
# Plot the full factorial design
ggplot(grid.full, aes(x = f1, y = f2)) +
 geom_point(size = 4) +
 xlab("Number of clusters") +
 ylab("Within-cluster sample size") +
 theme_bw()
```

Latin Hypercube Design (Figure 2)

```
# Set seed for reproducibility
set.seed(59832)
# Sample from a [0, 1] LHS design using lhs package
grid.lhd <- maximinLHS(n = 30, k = 2)
# Name columns of grid
colnames(grid.lhd) <- c("f1", "f2")
# Rescale grid to obtain the range of values factor 1 and factor 2 have
in the manuscript
grid.lhd[ , 1] <- rescale(grid.lhd[ , 1],
             to = c(10, 50),
             from = c(0, 1))
grid.lhd[ , 2] <- rescale(grid.lhd[ , 2],
             to = c(18, 68),
             from = c(0, 1))
# Convert the grid to a data frame
grid.lhd.data <- as.data.frame(grid.lhd)
# Plot the LHD
ggplot(grid.lhd.data, aes(x = f1, y = f2)) +
 geom_point(size = 4) +
 xlab("Number of clusters") +
```

```
 ylab("Within-cluster sample size") +
 theme_bw()
```

Full Factorial Design versus Latin Hypercube Design (Figure 3)

```
# Create variable to identify the experimental design
grid.full$factor_data <- c(1)
grid.lhd.data$factor_data <- c(2)
# Combine both data sets
data.all <- rbind(grid.full, grid.lhd.data)
# Create factor variable for experimental design
data.all$factor_data <- factor(data.all$factor_data, levels = c(1,2),
labels = c("Full factorial", "LHD"))
# Plot both experimental designs
ggplot(data.all, aes(x = f1, y = f2, shape = factor_data)) +
 geom_point(size = 4) +
 scale_shape_manual(values=c(1,17)) +
 xlab("Number of clusters") +
 ylab("Within-cluster sample size") +
 theme_bw() +
 theme(legend.position = "bottom", legend.title = element_blank())
```