Between-litter variation in developmental studies of hormones and behavior: inflated false positives and diminished power

Donald R. Williams[1*], Rickard Carlsson[2], and Paul-Christian Bürkner[3]

Abstract— **Developmental studies of hormones and behavior often include litter mates—rodent siblings that share early-life experiences and genes. Due to between-litter variation (i.e., litter effects), the statistical assumption of independent observations is untenable. In two literatures—natural variation in maternal care and prenatal stress—entire litters are categorized based on maternal behavior or experimental condition. Here, we (1) review both literatures; (2) simulate false positive rates for commonly used statistical methods in each literature; and (3) characterize small sample performance of multilevel models (MLM) and generalized estimating equations (GEE). We found that the assumption of independence was routinely violated ($>$ 85 %), false positives exceeded nominal levels (up to 0.70), and power rarely surpassed 0.80 (even for optimistic sample and effect sizes). Additionally, we show that MLMs and GEEs have adequate performance for common research designs. We discuss implications for the extant literature, the field of behavioral neuroendocrinology, and provide recommendations.**

1. INTRODUCTION

Research on rodents sharing litters is at the core of developmental studies of hormones and behavior. Common paradigms take advantage of naturally occurring variation (Champagne et al., 2003), for example differential maternal care (Beery and Francis, 2011; Francis and Meaney, 1999), or experimentally expose entire litters to the same experience such as prenatal stress (Weinstock, 2017). While natural occurring variation and variation due to experimental design seek to answer different questions, each paradigm faces similar statistical challenges due to between-litter variation (Holson and Pearce, 1992; Lazic and Essioux, 2013): both research designs categorize entire litters (i.e., siblings) based on maternal behavior or whether they were exposed to the same experimental condition (Figure 1). Additionally, litters are comprised of siblings that share early-life experiences and genes that can contribute to litter effects (Lazic and Essioux, 2013). Therefore, the statistical assumption that the observations are independent will routinely be violated (Lazic, 2010). The central question is thus the extent to which unaccounted dependencies (e.g., litter effects) can lead to erroneous conclusions in realistic research settings.

In the present paper, we elucidate the importance of this issue for the field of behavioral neuroendocrinology. Specifically, we: (1) review contributions from two influential literatures—natural variation in maternal care and prenatal stress; (2) provide theoretical rationale that the assumption of independence will be violated; (3) examine statistical methods commonly used in both literatures and simulate type I one error (false positive) rates for each approach, in addition to multilevel models (MLM) and generalized estimating equations (GEE); and (5) examine how between-litter variation influences power, and thus experimental design.

1.1. Background

Developmental programming is a process by which early-life experiences influence the phenotype on an organism, including physiological and behavioral trajectories (Gore, 2008). Since the stress axis plays a critical role in survival (Lupien et al., 2009; OConnor et al., 2000) and reproduction (Chatterjee and Chatterjee, 2009; McGrady, 1984), developmental effects on this neuroendocrine system have been thoroughly characterized in laboratory rodents (McEwen, 2008; Sapolsky and Meaney, 1986). The role of maternal care has played a central role in this research. Earlier studies used direct manipulations such as handling (Deitchman et al., 1977) or separation (Hofer, 1973), whereas more recent studies have investigated the role of naturally occurring variation in maternal care (Cameron, 2011; Curley and Champagne, 2016). In addition, the effects of prenatal experiences have been investigated for decades (Bond and di Giusto, 1976; Joffe, 1977). While many aspects of prenatal environment have been investigated, we focus on prenatal stress because of the thoroughness of the literature. For example, several prenatal stress manipulations have been developed and the effects on offspring development described (Weinstock, 2017, 2008).

1.2. Naturally occurring: maternal care

The finding that naturally occurring variation in maternal care can influence development provided a foundation from which an organism can be programmed by their environment (Cameron, 2011). For example, maternal tactile stimulation—licking and grooming (LG)—has been shown to induce changes in the hypothalamic-pituitary-adrenal (HPA) axis of developing offspring (Liu et al., 1997). Behaviorally, this reportedly allows for differential responsiveness to stressful stimuli across the lifespan (Fish et al., 2004). Offspring from so-called high LG mothers demonstrate less fear responsivity (Menard et al., 2004) and more exploratory behavior in novel environments than offspring of low LG mothers (Starr-Phillips and Beery, 2014). These opposing phenotypes are thought to be modulated in part by differential glucocorticoid

[1] Quantitative Psychology, University of California, Davis, One Shields Avenue, Davis, CA 95616

[2] Department of Psychology, Linnaeus University, Sweden

[3] Institute of Psychology, University of Muenster, Fliednerstrae 21, 48151 Muenster, Germany

[*] Corresponding author: drwwilliams@ucdavis.edu

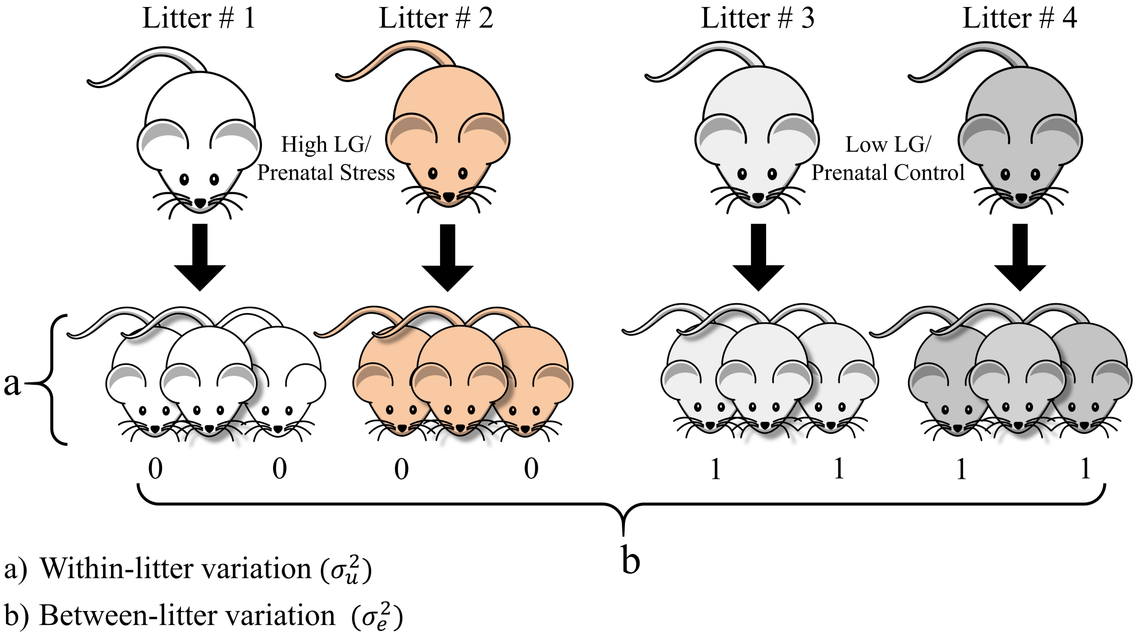

Fig. 1: both research areas—natural variation in maternal care and prenatal stress—categorize entire litters based on maternal behavior or experimental condition. For example, litter mates from High LG/Prenatal Stress and Low LG/Prenatal Control dams are coded the same way. When groups are compared (0 vs. 1), unaccounted for between-litter variation (i.e., dependent measures) violates the statistical assumption of independence. Use of a t-test would be incorrect for this research design.

activity in the hippocampus that promotes feedback inhibition of stress reactivity (Jacobson and Sapolsky, 1991). In support of this notion, high and low LG offspring reportedly differ in HPA responsiveness (Liu et al., 1997), sensitivity to feedback inhibition (Liu et al., 1997), expression profiles of glucocorticoid receptors (GR) (Hellstrom et al., 2012), and epigenetic modifications to NR3c1 (McGowan et al., 2011).

1.3. Experimental: prenatal stress

For several decades, it has been known that prenatal stress can influence offspring development (Archer and Blackman, 1971; Kapoor et al., 2006). More recently, the notion of fetal programming was put forth, where it is hypothesized that the *in-utero* environment can make offspring susceptible to adverse outcomes later in life (Seckl and Holmes, 2007). One aspect of fetal programming is prenatal stress (PNS) which has been investigated by exposing pregnant rodents to stressors including restraint, electrical shock, and social stress across the gestational period (Weinstock, 2017). Increased stress reactivity and anxiety-like behavior have been observed in male and female offspring (Wilson et al., 2013). Later in life, PNS rodents show increased HPA axis reactivity to stressors, such as increases in corticosterone (Koehl et al., 1999) and the adrenocorticotropic hormone (McCormick et al., 1995), as well as up-regulated corticotrophin-releasing factor (CRF) (Cratty et al., 1995). The feedback properties of the hippocampus on the stress response are also affected by PNS (Boersma and Tamashiro, 2015). For example, hippocampal GR are differentially regulated in offspring, but primarily in females (Szuran et al., 2000). In PNS males, increased levels of CRF expression and reductions in GR expression were detected (Mueller and Bale, 2008). Furthermore, the CRF gene had reduced levels of methylation, whereas more methylation was observed on NR3c1 (Gudsnuk and Champagne, 2012).

1.3. Rational for between-litter variation

Although maternal care and prenatal stress are important components of the environment, there are many others factor that can contribute to between-litter differences. For example, litters size influences many aspects of offspring development (Tanaka, 2004), including age at sexual maturity and reproductive behaviors in females (Mendi, 1988). Furthermore, experimentally manipulating pre-weaning litter sizes increased anxiety-like behaviors of adult rodents (Dimitsantos et al., 2007). This has led to routine culling procedures that are often used to control for the effects of variable litter sizes (Agnish and Keller, 1997). In addition, litter mates also share the same prenatal (Marceau et al., 2016) and social environments (von Engelhardt et al., 2015), each of which presents challenges for controlled experiments. Although litter size can be held constant, the hormonal composition of placental fluid or behavioral types within-litter cannot be controlled. There is evidence that *in-utero* hormonal milieus (Fowden and Forhead, 2004) and the early social environment influence development (Turecki and Meaney, 2016). While litter effects are not of primary interest in these studies, they provide indirect evidence for between-litter variation. That is, the fact that the early environment influences development, also suggests that those sharing the same environment (pre- or post-natal) will be more alike than those from different environments (Lazic and Essioux, 2012).

The role of genes on physiological and behavioral phenotypes cannot be understated, and this has been shown in a variety of species (Inoue-Murayama, 2009). Like many questions, laboratory rodents have provided valuable insight into the importance of genetics (Crabbe et al., 1999; Wahlsten et al., 2007). For example, common strains of inbred mice differ in locomotor activity, novelty seeking, fear reactivity, and maternal care (Champagne et al., 2007; Ramos et al., 1997). Neurobiological differences have also been observed such as neurotransmitter levels (Brodkin et al., 1998), gene expression profiles (Kimpel et al., 2007), and structural morphology (Scholz et al., 2016). Whereas inbred mice are genetically identical, outbred rodents from the same litter are effectively dizygotic twins (Lazic and Essioux, 2013). In humans, dizygotic twins show correlations in cognitive ability (Haworth et al., 2010), personality traits (Jang et al., 1996), and brain structure (Scamvougeras et al., 2003). Furthermore, it is sometimes the case that genes contribute more to the adult phenotype than the shared environment in humans (Haworth et al., 2010). Although quantitative genetic approaches are not common in the neuroendocrinology literature, a reasonable assumption is that between-litter variation due to genetics would be found in litter mates that are outbred rodents (Glowa and Hansen, 1994).

natural variation in maternal care and prenatal stress literatures have proven extremely influential. Although an apparently clean picture has emerged, these findings are dependent upon the statistical tests used and the assumptions of those tests (Scariano and Davenport, 1987). An important question is whether group differences were examined without accounting for the fact that individual rodents were litter mates. This would indicate methodological limitations in two prominent research areas in the field of behavioral neuroendocrinology, but would also provide useful information that could improve both fields.

2. Methods and materials

2.1. Literature search

We examined how between-litter variation was accounted for in the natural variation in maternal care and prenatal stress literatures. A search was performed using Web of Science that included all studies published before the search date of May 20, 2017. We sought to understand how litter has been broadly accounted for, which served as a foundation for simulating false positive rates and power, as well as allowing for inferring the extent to which our findings may apply. For naturally occurring variation, the search term was ‘maternal care’ AND ‘licking grooming.’ Only studies that categorized quasi-experimental groups based on the amount of maternal care were considered. The search term for prenatal stress was ‘prenatal stress.’ Because this returned 2,799 hits, we included the 100 most recent studies directly related to the neuroendocrine system. For both literatures, outcomes could be either behavioral or physiological.

The identified original research articles were used to qualitatively describe methods used to account for litter dependencies. Based on previous work (Holson and Pearce, 1992; Lazic and Essioux, 2013; Zorrilla, 1997), we expected aspects of litter to be underreported. Accordingly, we attempted to answer broad questions including: (1) how often multiple animals from the same litter were included in the analyses; (2) whether the paper considered litter effects; and (3) how often litter effects were reported.

To provide realistic simulation conditions, we also obtained the following information: (1) number of litters included in the analyses; (2) number of pups used per litter; (3) methods used to account for litter effects. We documented the search procedure, and provided those documents on the Open Science Framework (https://osf.io/fxy7h/).

2.2. Simulation: false positives

We examined false positive rates for commonly used approaches for dealing with between-litter variation. We were specifically interested in the degree to which between-litter variation inflates false positive rates. From the literature search, we found that litter was not often accounted for, or was considered as a covariate (https://osf.io/fxy7h/). In three papers, a statistical model was used that accounted for non-independence with a random effect (Barha et al., 2007; Neeley et al., 2011) or corrected standard errors (Amugongo and Hlusko, 2014). As such, we compared error rates of four models, including: (1) t-tests (litter not included in the model): (2) analyses of covariance (ANCOVA; litter included as a covariate); (3) multilevel models (MLM; Roux, 2002); and (4) generalized estimating equations (GEE; Hanley et al., 2003). For the MLMs, litter was included as a random effect (varying intercept) that accounts for within-cluster correlations (Gelman and Hill, 2007). A GEE similarly accounts for cluster-related variation, but does so by estimating a population-average model that relaxes many assumptions of MLMs (Hubbard et al., 2010). For example, a MLM assumes that random effects are normally distributed and are uncorrelated with the fixed effects. The latter may or may not be plausible when including litter and maternal care in the same analysis. In contrast, GEEs make no such assumptions. However, in small sample situations, GEEs require standard error corrections to ensure nominal error rates (Gunsolley et al., 1995; Li and Redden, 2015). We determined the appropriate bias correction with simulations (see below for *litterEffects package*).

Reasonable estimates for between-litter variation were obtained from our own data and methodologically oriented papers. We found that litter accounted for upwards of 60 % of the residual variation (Lazic and Essioux, 2013). In our simulations, variability between litters (σ_u^2) was computed as an intra-class correlation coefficient (ICC):

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \quad (1)$$

that is the percentage of residual variation explained by litter (where σ_e^2 is within-litter variation). The ICC can also be

Table 1
Results from literature search

	Used multiple animals from the same litter	Explicitly mentioned litter effects	Reported litter effects	Assumption of independence likely violated
Natural variations in maternal care	24 / 24 (100 %)	7 / 32 (22 %)	0 / 33 (0 %)	24 / 28 (86 %)
Prenatal Stress	84 / 95 (88 %)	30 / 98 (31 %)	0 / 89 (0 %)	81 / 95 (85 %)

Note. These estimates were computed from primary studies (natural variation in maternal care = 35 and prenatal stress = 100) in which sufficient information was provided. For example, 84 / 89 (88 %) indicates that 11 studies did not provide enough information to answer that specific question.

thought of as the correlation between observations within a given litter. The data generating model was a MLM, since it allows for specifying within-cluster correlations. We found that few studies reported the number of litters and the number of animals per litter. As such, we assumed a range of simulation conditions (litters = 4, 8, and 12; pups per litter = 2, 4, 6, and 8; ICC = 0–0.70 by increments of 0.05). For a given condition, observations from half of the litters were dummy coded as 0, whereas observations from the remaining litters were coded as 1 (Figure 1). The average difference between groups (0 vs. 1) was set to zero (a true null hypothesis), thus the expected error rate was 5 %.

2.3. Simulation: conditional false positives

While not an approach we would advocate, common practice is to use non-significance to exclude a variable from a model. As such, we also investigated whether false positive rates are conditional on a significant litter effect. We computed the significance of litter as a MLM random effect, and then analyzed the data with a t-test. In this way, we obtained:

$$P(FE_{p-value} < 0.05 | RE_{p-value} > 0.05) \tag{2}$$

$$P(FE_{p-value} < 0.05 | RE_{p-value} < 0.05) \tag{3}$$

where (2) denotes the probability that the fixed effect $(FE_{p-value})$ is significant, given the random effect is non-significant $(RE_{p-value})$. Alternatively, (3) is conditioned on a significant litter effect.

2.4. Simulation: power

We present two approaches-MLMs and GEEs-to incorporate between-litter variation into experimental design with power calculations. We were specifically interested in how differing ICC values influence power and how this varies with the ratio of litters to observations per litter, holding the total sample size constant. That is, we addressed whether it is more advantageous to increase litters or pups per litter. Based on the literature search, we found that group sizes varied, but were typically small. We chose an extremely optimistic value of 24 observations per group (N = 48) that can be thought of as the best scenario, and varied the composition of the samples (litters = 4, 6, 8, and 12; pups per litter = 12, 8, 6, and 4). Standard effect size measures (Cohen's d) do not exist for MLMs, since variance is partitioned among levels. We thus used an effect size, delta total variance δ_T (Hedges, 2007), defined as:

$$\delta_T = \frac{\beta}{\sqrt{\sigma_u^2 + \sigma_e^2}} \tag{4}$$

where the difference between groups (β) is divided by the square root of the variance components summed. We found that significant effects in the literatures were typically large (d > 1.0), but simulated power for a range of values (δ_T = 0.20, 0.50, 0.80, 1.10). The interpretation of δ_T follows d, so the selected values covered what are considered small (0.20), medium (0.50), and large effects (0.80).

2.5. Simulation: uncertainty due to litter

In addition, we conducted a simulation to demonstrate how between-litter variation influences uncertainty of the fixed effect estimate. This was achieved by computing confidence intervals for an unstandardized group difference (β = 8.0) across a range of ICC values. Since a 95-% interval excluding zero is significant at the α = 0.05 level, this allowed for visualizing how litter impacts false positives and power.

For each combination of litters, observations per litter, and ICC values, 5,000 simulations were performed for each of the models. False positive rates and power were computed as the proportion of simulations with p < 0.05. All computations were done with the R programming language. The MLMs were fitted with the package *lmerTest* (Kuznetsova et al., 2016) that is a front end to *lme4* (Bates et al., 2015), whereas *gee* (Ripley, 2015) was used for the GEEs and *saws* (Fay, 2015) for the bias corrected standard errors (R code: https://osf.io/fxy7h/). To aid applied researchers, we developed a R package (*litterEffects*) that allows for simulating false positive rates, power, determining the optimal GEE bias correction, and includes a tutorial (https://github.com/donaldRwilliams/litterEffects).

3. Results

3.1. Literature search

We identified 35 articles from the natural variations in maternal care (MC) literature and 100 articles from the prenatal stress (PNS) literature. We found that descriptions were too varied to get precise estimates of the number of litters and observations per litter (https://osf.io/fxy7h/), but

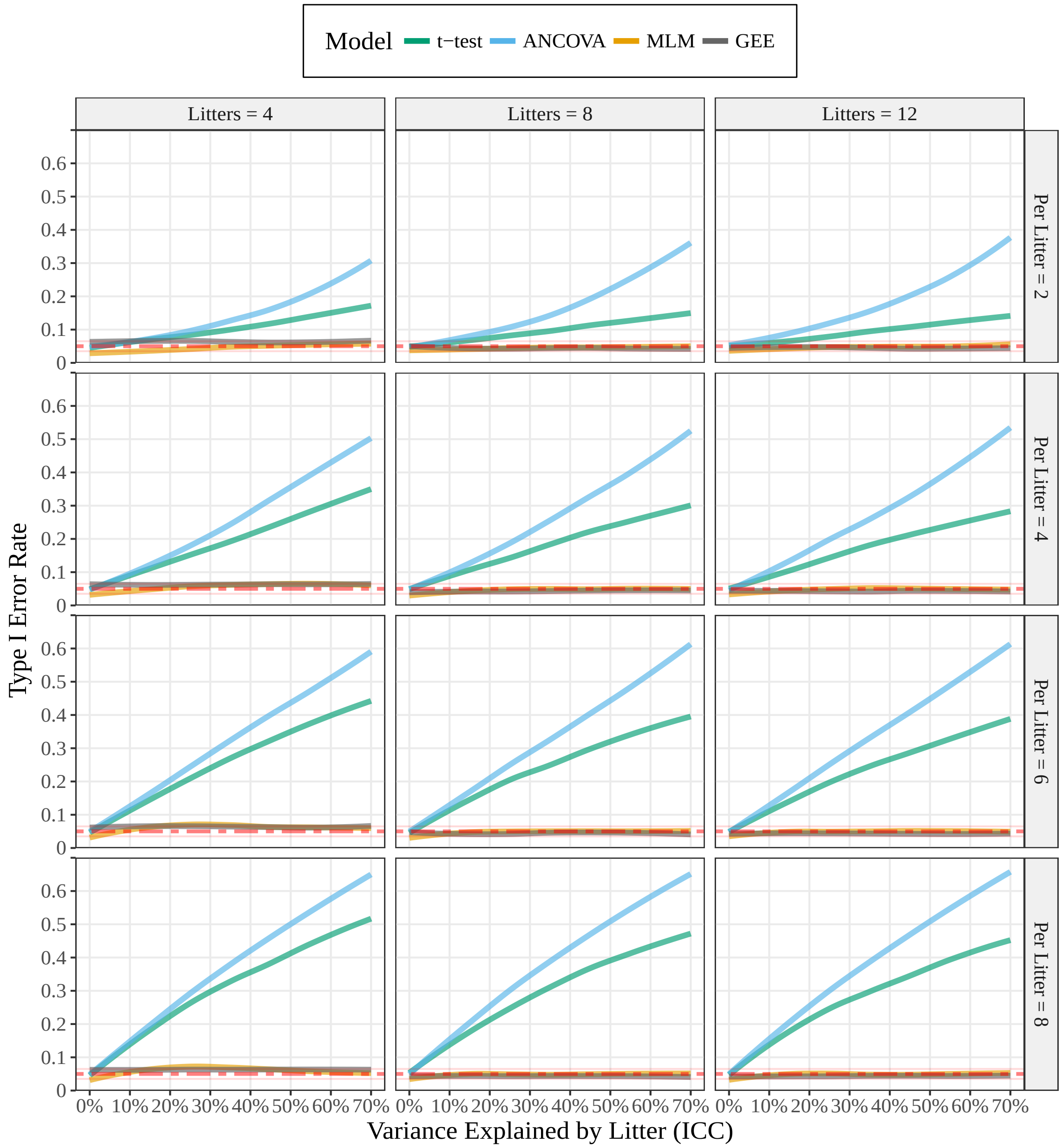

Fig. 2: t-tests and ANCOVAs have inflated type I error rates when between-litter variation in non-zero (ICC $>$ 0 %), and this is directly related to the degree of between-litter variation. That is, error rates are dependent upon similarities among litter mates. MLMs and GEEs—statistical methods that account for dependent measures—have adequate performance across most conditions.

multiple animals from the same litter were used in most studies (MC = 100 % and PNS = 88 %). Although litter effects were explicitly considered (MC = 22 % and PNS = 31 %), this often resulted in reducing the number of litter mates used. That is, dependent measures were still included in the study. In three studies (Amugongo and Hlusko, 2014; Barha et al., 2007; Neeley et al., 2011), a statistical method explicitly for correlated observations was used. In both literatures, we found that the most common statistical approach assumed independence of observations (MC = 86 % and PNS = 85 %). In other words, a large percentage ($>$ 80 %; Table 1) of the reviewed studies likely violated the assumption of independent observations.

We highlight two papers that, while not using a statistical method for dependent measures, considered litter effects by either averaging within litter or using one animal per litter. The former was used in Starr-Phillips and Beery (2014):

> *To avoid the possibility that major findings arise from litter effects rather than maternal care effects, effects of maternal care on social behavior were also analyzed by litter, using litter means in place of individual subject data points.*

Both approaches produced similarly significant effects, but litter means were only used for a subset of outcomes. Interestingly, one prenatal stress paper mentioned that litter mates are siblings and selected one animal per litter:

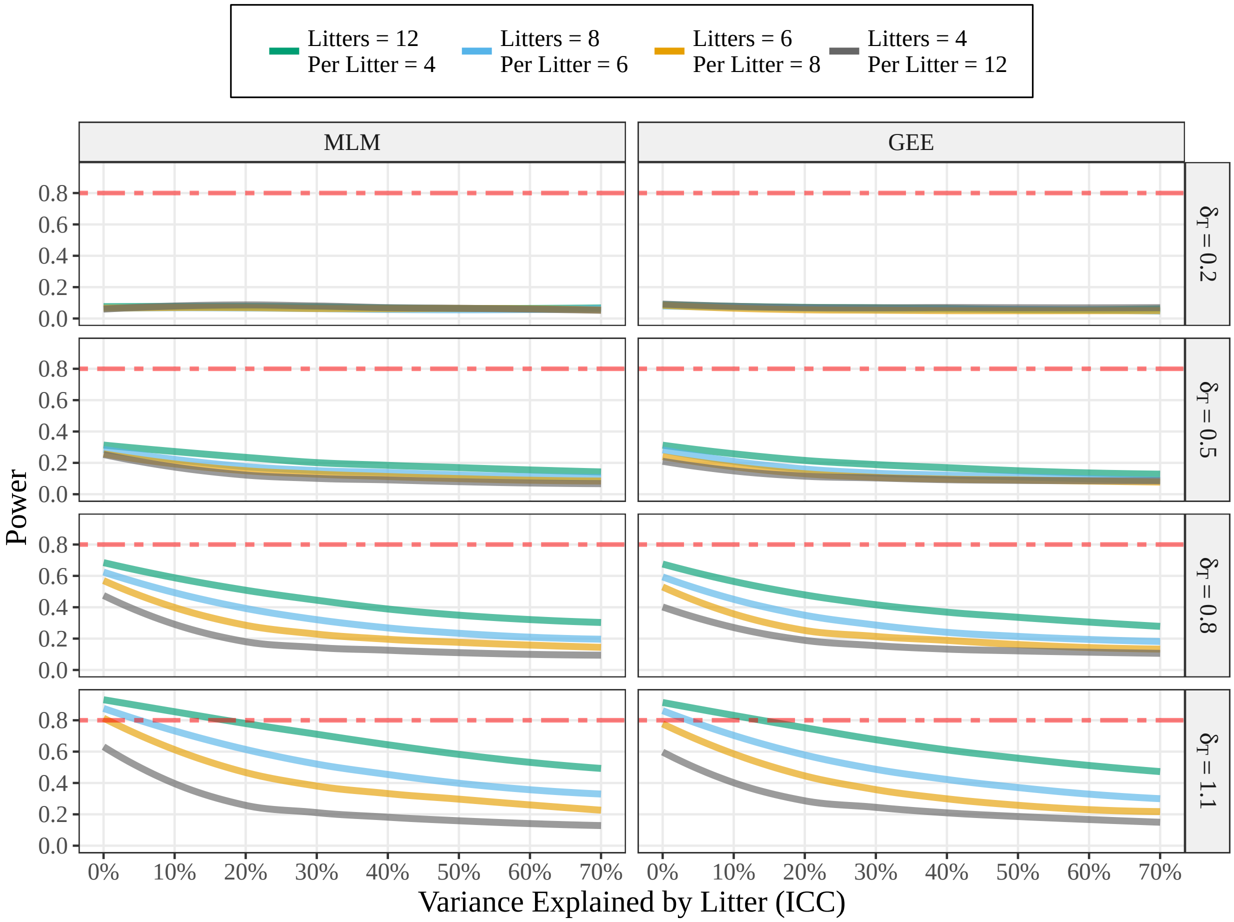

Fig. 3: power is related to the degree of between-litter variation and sample size composition (power is higher with fewer dependent measures [green]). For these simulation conditions, power is rarely at the nominal level of 0.80. Importantly, since we used an optimistic sample size (N = 48; two-groups of 24) these are likely overestimates of typical power in both literatures.

> *To avoid litter effects, only one rat from each of four litters per group was tested in each experiment. Hence, for this study, n implies that four unique (non-siblings) prenatally stressed or control rats were used separately for each method of analysis (Baier et. al).*

3.2. False positive rates

The results are presented in Figure 2, including type I error rates of four models: (1) t-test (green): (2) ANCOVA (blue); (3) MLM (yellow); and (4) GEE (grey). Each model compared mean differences—assuming a true null hypothesis—between two groups, but differed in how litter was accounted for (see section *2.1. Simulation: false positives*). Due to expected sampling variability in simulations, reported quantities were rounded to the hundredth decimal place.

When litter was not included in the model, as in the t-test, type I error rates exceeded nominal levels (α = 0.05; red dashed line). Error rates ranged from approximately 0.05–0.51. The latter indicates an almost 1,000 % increase from 0.05. For all condition in which the litter ICC was 0, nominal levels were achieved. In other words, when litter mates did not resemble one another, the t-test had optimal performance. However, when the ICC was 5 %, error rates approached 0.10 (litters = 12 and per litter = 8). For sample sizes more commonly seen in behavioral neuroendocrinology, error rates approached 0.30 when the litter ICC was 40 % (litters = 4 and per litter = 6). Across all conditions, the magnitude of the ICC was directly related to the increase in error rates and this became more pronounced with larger sample sizes.

With litter modeled as a covariate in an ANCOVA, we observed the same patterns as the t-test. That is, error rates increased with the degree of between-litter variation, and this was influenced by the sample size. Furthermore, when the ICC was 0, nominal levels were achieved. Across all conditions, however, there were substantial differences between the t-test and ANCOVA in that the latter had markedly higher error rates (t-test: 0.05–0.51 vs. ANCOVA: 0.04–0.66). For a total sample size of 24 (litters = 4 and per litter = 6) and an ICC of 40 %, the error rate was 0.37 (640 % increase from the alpha level of 0.05).

Based on the same data generating process as for the t-tests and ANCOVAs, we examined error rates for statistical methods that are specifically for non-independent data. In Figure 2, the MLMs (yellow) and GEEs (grey) showed similar performance. This was expected, and highlights that both methods generally performed well across all conditions. However, we also observed that both methods could be conservative and anti-conservative (MLM: 0.03–0.08 vs. GEE: 0.04–0.07). The conservative estimates (i.e., < 0.05) were observed when samples were small (N = 8; litters = 4 and per litter = 2) and the ICC values were close to zero. However, when the number of litters were more representative of both literatures (litters > 4), both methods had optimal performance in that error rates were close to 0.05 (see here for estimates: https://osf.io/fxy7h/).

3.3. Conditional false positive rates

We examined false positive rates for the fixed effect, conditional on a significant litter effect (random intercept) in a multilevel model (*2.3. Simulation: conditional false positives*). False positive rates were consistently higher when there was a significant litter effect (Figure 4a). However, when the litter effect was non-significant, error rates become problematic when the ICC was greater than 10 %. For an ICC value reported in a related field (60 %; Lazic and Essioux, 2013), false positive rates exceeded 0.20.

3.4. Power

Since only the MLMs and GEEs achieved nominal error rates, power was examined for these methods. We were specifically interested in the degree to which between-litter variation influences the power to detect a difference, and how this varies with the ratio of litters to observations per litter. The results of this simulation are presented in Figure 3, where 0.80 power is indicated with a red dashed line.

For both methods, power was related to the magnitude of between-litter variation. For example, when $\delta_T = 1.1$ power was greater than 0.90 when the ICC was 0 %, but reduced substantially when the ICC was 70 % (MLM = 0.49 vs. GEE = 0.47). Indeed, even with optimistic sample sizes ($N = 48$), power reached 0.80 in few conditions. Specifically, when the effect size was very large ($\delta_T = 1.1$) and the ICC was less than 20 %. For what is considered small (0.2) and medium size effects (0.5), power did not exceed 0.32 for both models. For most conditions, the MLMs had more power than the GEEs and this was the case across the range of ICCs, but this differences was generally small (i.e., < 3 %).

In addition, power was directly related to the composition of the samples for both models. In all conditions, for example, power was higher when less animals per litter were included in the analysis. For a δ_T of 1.1, power exceeded 0.80 for both MLMs and GEEs when there were 12 Litters and 4 observations per litter (grey line), but was substantially lower for 4 litters and 12 observations per litter. In other words, fewer dependent measures resulted in higher power (see here for estimates: https://osf.io/fxy7h/).

3.5. Uncertainty due to litter

To make clear how litter affects power and error rates, we used simulations to compute 95-% confidence intervals for an unstandardized effect across a range of ICC values (0–0.70, Figure 4b). As between-litter variation increases, the confidence intervals become wider. Whereas the effect was significant when the ICC is 0 %, this was no the case (interval includes 0) with an ICC of 20 %. The same logic applies to false positive rates. When between-litter variation is not accounted for, the width of the confidence interval will be too narrow and this increases the rate at which the confidence intervals will exclude zero.

4. DISCUSSION

The present study investigated how between-litter variation has been accounted for in two literatures—natural variation in maternal care and prenatal stress. Specifically, we estimated how often dependent measures (i.e., litter effects) have been considered, as well as the degree to which litter effects can increase false positives and affect power. Although aspects of litter were generally underreported (e.g., total litters included in the study), we found that litter effects were never reported, most studies used several pups from the same litter, and only 15 % used a statistical technique appropriate for data with dependent observations (Table 1). The latter indicates our simulation results apply widely, in that expected error rates ($\alpha = 0.05$) are compromised in a large portion of the published studies in both research areas. Furthermore, since litter effects were never reported, our findings not only apply to analyzing data but also to the design stage of experiments. That is, to accurately compute power for a hypothesized effect, one must consider between-litter variation (Figure 3). This is currently not possible given the current state of both literatures (*3.1. Literature search*).

4.1. False positive rates

Based on the literature search, we computed false positive rates for commonly used statistical approaches for handling litter effects. The most common approach was to assume independent observations, followed by including litter as a covariate. We showed that, across all conditions in which there was between-litter variation, both approaches produced inflated error rates. While this was observed for both t-tests and ANCOVAs, error rates were substantially higher for the latter. The inclusion of covariates in an ANOVA is known to increase power (Borm et al., 2007)–assuming an effect exists. This occurs because residual variance can be reduced (Cox and McCullagh, 1982), thus increasing power to detect an effect for the variable of interest (Borm et al., 2007). However, like ANOVA, and ANCOVA assumes the errors are uncorrelated which is unlikely to be the case when litter mates are included in the analysis (Keselman et al., 1998). In addition, ANCOVA assumes there is no interaction between the independent variable and covariate (Levy, 1980). This may or may not be the case in the published literature, but should be investigated going forward. There is also growing realization that inclusion of covariates can increase type I error rates, and allows for substantial researcher research degrees of freedom. That is, covariates allow for a high degree of flexibility that can be advantageous in certain settings, but not when explored until the $p < 0.05$ threshold is crossed. To address this potential issue, methodologists

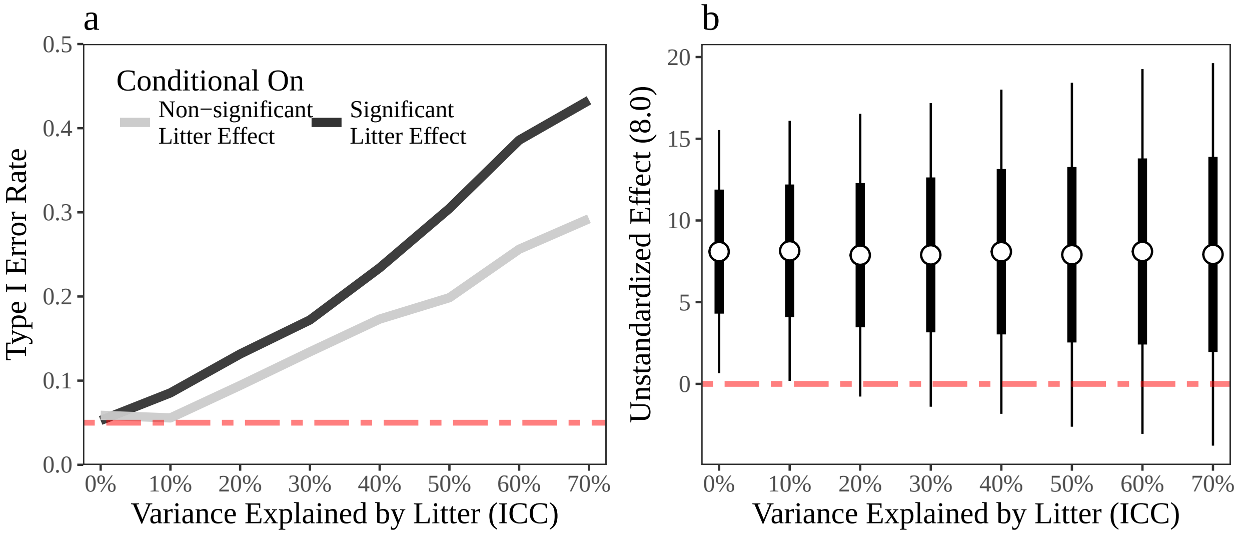

Fig. 4: a) error rates for the fixed effect (e.g., High LG vs. Low LG)–analyzed with a t-test–are not conditional on a significant litter effect. This can be thought of as: 1) testing the significance of litter in a MLM (random intercept) and 2) excluding litter from the model when non-significant. In this case, reliance on statistical significance ($p < 0.05$) can lead to inflated false positive rates. b) between-litter variation increases uncertainty (standard error = thick line; 95-% CI = thin line) of the fixed effect estimate. Results for both figures were obtained with simulations (8 litters and 4 observations per litter).

in human oriented psychology are advising to pre-register covariates (van t Veer and Giner-Sorolla, 2016; Wang et al., 2017).

In addition, we examined error rates of statistical methods–multilevel models and generalized estimating equations–specifically for data with dependent measures. Across most conditions, nominal error rates were achieved for both methods. It should be noted that, when samples were small ($N = 8$), both methods produced error rates problematically above or below the expected level ($\alpha = 0.05$). However, these were not near the levels observed in the t-tests and ANCOVAs. In some literatures, it has been suggested that MLMs and GEEs should not be used with small samples (Callens et al., 2005), such as those that are typical in behavioral neuroendocrinology. When the goal is to account for dependent measures, and not necessarily on estimating the random effect with a certain degree of precision, we show that both methods can be used to ensure nominal error rates when samples are small and clusters (i.e., litters) are few. With a sample of eight ($N = 8$), for example, adequate performance was achieved for mid-range ICC values. Furthermore, GEEs do not provide an estimate of the random effect, but do need small sample corrections to ensure nominal levels for the fixed effect (Gunsolley et al., 1995). We stress this point–bias corrections are a necessity–because default standard error estimators can produce error rates greater than 0.20 for small sample sizes (Gunsolley et al., 1995).

4.2. Conditional false positive rates

We also showed that false positive rates to not depend on a significant litter effect (Figure 4A). This can be thought of as mimicking a two-step procedure:

- the significance of a variable was assessed
- if non-significant, the variable was removed from the model

The important question is not whether the effect of litter is significant, but whether it is exactly–or very close to–zero. Thus, statistical significance is irrelevant in this context and a better solution is to rely on knowledge of experimental design and study subjects, regardless of the observed p-value. It should be noted that the notion of parsimony is often invoked when analyzing data (Bates et al., 2015a). However, all decisions have statistical consequences that need to be considered. In this case, pursuing the most parsimonious model can lead to erroneous conclusions (Barr et al., 2013). Additionally, those who make inferences via the standard error are making statements about expectation over the long run (i.e., hypothetical replications; Greenland et al., (2016)). Accordingly, even if a litter effect is estimated as zero, it is important to consider whether this is a reasonable expectation in future studies.

4.3. Power

Since only MLMs and GEEs achieved nominal false positive rates, we explored power of these models (Figure 3). We were most interested in how power is influenced by between-litter variation, and the composition of the sample (i.e., the ratio of litters to observations per litter). We found that it is more advantageous to reduce the number of dependent measures. For example, power was consistently higher with 12 litters and 4 observations per litters that the reverse (litters = 4; per litter = 12). When the effect size was small (δ_T = 0.2), there was negligible power and this was the case for both models. Given the same sample and effect size on d scale ($N = 48$, $d = 0.20$), power for an independent t-test is 0.10 which parallels the MLM (0.08) and GEE (0.09) estimates when there were 12 litters and 4 observations per litter. Of course, power would be higher if all observations were obtained from separate litters. In addition to optimal power, using one observation per litter would be particularly advantageous in the maternal care literature. For example, in these studies, the mother is the observational unit and within litter differences in maternal care received are not usually considered (van Hasselt et al., 2012).

Our simulations showed that power reached 0.80 in very few conditions. The implications of this are two-fold. First, significance can be thought of a ratio of signal (the effect) to noise (standard error). In small samples noise is generally high, so the effect needs to be very large to reach significance (Walum et al., 2016). This is problematic because the error surrounding a significant effect can make it uninterruptable. To make this point, we selected a significant effect from one papers and computed Cohens *d* (*d* 2.3, 95-% CI = [0.42 4.17]) (Champagne et al., 2003b). Given this effect, we can safely reject values less than 0.42 and 4.17 which shows that the true effect could be small/medium in size to unreasonably large (Kruschke, 2013). Second, non-significant effects are also difficult to interpret. For example, assuming the effect is centered at zero but the interval is the same width (*d* = 0, 95-% CI = [-1.88 1.88]) indicates that even very large effects are still possible and should not be confused with no effect (Lakens, 2017). Together, this lack of power is directly related to small sample sizes and this affects interpretation of significant as well as non-significant effects (Button et al., 2013).

4.4. Comparison to methodological papers

It should be noted that our results parallel methodologically oriented papers on similar topics. For example, Holson and Pearce (1992) showed that between-litter variation inflated false positive rates and Zorrilla (1997) found that the assumption of independence was violated in 85 % of the reviewed papers. Over one decade later, Lazic and Essioux (2013) found that 91 % of studies in the valproic acid literature used an invalid statistical method for analyzing data with litter mates. We built upon these previous papers in several ways, First, our paper addresses these issues specifically in the field of behavioral neuroendocrinology. Second, in addition to varying the sample sizes, we investigated how the magnitude of between-litter variation affects false positives and power. Third, we showed that an ANCOVA inflates false positives more than a t-test. Fourth, as an alternative to MLMs, we characterized the performance GEEs for typical research designs. Fifth, we developed a R package (*litterEffects*) that allows for investigating power, determining the appropriate GEE bias correction, and includes a tutorial. Sixth, we discussed the implications of our findings in the broader context of replication efforts in related fields.

4.5. Implications: natural occurring maternal variation and prenatal stress

The natural variation in maternal care and prenatal stress literatures have proven influential in the field of behavioral neuroendocrinology (Curley and Champagne, 2016; Goldstein et al., 2014). The former has provided a foundation in which developmental programming could occur in nature (Cameron, 2011), whereas the latter has provided insights into the etiology of neurobiological disorders such as autism (Kinney et al., 2008) and schizophrenia (Markham and Koenig, 2011). Although empirical findings are well documented in both literatures, our findings highlight areas for improvement in both research areas. There is substantial evidence that true effects likely exist. However, due to not accounting for between-litter variation caution is warranted when interpreting past research. In addition to our findings, it should be noted that we are unaware of direct replications in either literature. As such, we take the position that the general hypotheses may have support, but do not offer evidence for specific effects. For example, in the general sense, maternal care probably does affect offspring development. However, stating that maternal care can reliably induce epigenetic modifications to a specific gene is not currently supported by the literature. Evidence for this can only be obtained through replication, in addition to using appropriate statistical methods. Revisiting previous data would address the issues we raised here (e.g., potential false positives), but would not address the lack of replications in both literatures.

4.6. Implications: reproducible science

The replication crisis has so far been dominated by human oriented psychology in general, and social psychology in particular (OSC, 2015). Yet, other research areas are also experiencing difficulties replicating findings including biomedical related fields (< 25 %; Prinz et al., 2011). Even if we assume that the reported results in the primary studies are unbiased, using an inappropriate statistical analysis can produce unreliable results. This finding parallels a recent paper that examined clusters in fMRI research in which they concluded that commonly used software could produce false positive rates upwards of 0.70 (Eklund et al., 2016). In contrast to this paper, where it was suggested that interpretation of weakly significant findings was mostly affected, we cannot make this claim. The present simulations show that false positives depends on the number of litters, observations per litter, and between-litter variation, each of which need to be taken on a case-by-case basis (Figure 2). In studies in which the ICC of litter was 0 %, for example, the *p*-value would not change by incorporating a random intercept.

It has been argued that most research findings are false (Ioannidis et al., 2005), the literature is skewed towards positive results due to publication bias (Francis, 2012), individual researchers actively engage in questionable research practices (e.g., HARKing: Kerr, 1998 and *p*-hacking: Simmons et al., 2013), and that seemingly justified choices can make it so results cannot be replicated (e.g., garden of forking paths: Gelman and Loken, 2014) . It should be noted that our analysis does not address these issues. Nevertheless, the take home for replication efforts in neuroendocrinology is that we need not exclusively focus on biases or ill-intent on the part of individual researchers. It is entirely plausible that misspecified statistical models account for many failed replications, in that the original effect may have never been statistically significant to begin with. Furthermore, our simulations showed that power depends on many aspects of litter including the degree of between-litter variation (Figure 3). In other words, detecting a significant finding would prove difficult if the magnitude of between-litter variation was larger in a replication attempt than the original study.

4.7. MLM vs. GEE

In contrast to MLMs, GEEs are less documented in R (Bates et al., 2015; Halekoh et al., 2006; Pinheiro and Bates, 2000), present difficulties for evaluating model fit (Horton et al., 1999), and there are few examples of their use in the hormones and behavior literature (Muth et al., 2016). When sample sizes are small, GEEs require bias corrections to ensure nominal type I error rates (Gunsolley et al., 1995; Li and Redden, 2015). In fact, many bias corrections exist which can introduce substantial researcher degrees of freedom into the analysis (Fay and Graubard, 2001; Pan and Wall, 2002). While GEEs can only consider one source of variation, MLMs offer greater flexibility and provide more information that can be used prospective power analyses (estimates of the random effects that are not provided by GEEs). This flexibility comes with a cost, however, as a misspecified MLM can substantially inflate the error rates. For example, when two treatments are administered to subjects from the same litter, variability in the experimental effect must be considered with a random slope, in addition to the random intercept of litter (Aarts et al., 2015). In small sample situations, this presents challenges in a frequentist framework since convergence issues can arise when the number of estimated parameters exceeds the total number of observations. In these situations, Bayesian methods can be used (Baldwin and Fellingham, 2013)

4.8. Statistical assumptions

Even a simple t-test can be thought of as modeling biological phenomena, in which inference is dependent on many assumptions. To ensure assumptions are met, further statistical tests are often used such as Shapiro-Wilk for normality (Shapiro and Wilk, 1965). In contrast, we do not know of tests explicitly for the assumption of independence. Consideration of the research design, model organism, expert opinion, and reason can all be used. If non-independence is suspected, but not present, inclusion of a random effect will give equivalent estimates to a fixed effect only model (Gelman and Hill, 2006).

Heterogeneous variances can increase false positives in parametric, as well as so-called non-parametric tests (Zimmerman, 1987). However, data that departs from normality can still have the nominal error rates. Non-parametric tests for clustered data are underdeveloped (Noguchi et al., 2012) and expected error rates have not been examined to our knowledge. When distributional assumptions are not met, caution is therefore warranted when choosing which violations require action. Rather than transforming data to normality, for example, a distribution other than Gaussian can be assumed (e.g., Poisson for count outcomes). These questions are challenging and demonstrate the difficulty in modeling hierarchical data structures such as those that include litter mates. By clearly stating the assumptions that the results depend on, however, would allow researchers the opportunity to debate their validity. We consider the assumption of zero between-litter variation untenable and that researchers should always use MLMs or GEEs.

4.9. Statistical expertise vs. improved training

Certain research questions will inevitably require seeking statistical expertise. For example, genomic data is often analyzed by specialists in quantitative oriented fields such as biostatistics. Outside expertise can ensure rigor and correctness for specific problems. However, expertise is also qualified with being highly specialized in a certain area. Therefore, specialists in multilevel modeling, or those that know how to account for clusters more generally, may be a limited resource (Lazic and Essioux, 2013). In addition, fruitful collaborations require a certain amount shared knowledge of one anothers field and specific research question. This would entail communication of the necessary information so that the correct analysis is applied. It is plausible that, if a researcher knew to mention the hierarchical structure of their data (e.g. rodent pups within litter), they would also be familiar with the appropriate statistical methods (Lazic and Essioux, 2013). In turn, if the structure of the data is not shared and a quantitative expert performed the analyses, this can give the false impression of statistical correctness. Since many (possibly most) research designs in behavioral neuroendocrinology have dependencies or additional sources of variation, we see a limited role for statisticians in the most common scenarios.

Alternatively, working towards better quantitative training for researchers may be a viable option to address many of the issues that we highlighted. There are many resources available for individual researchers. While in the past fitting MLMs was a specialized task, free and easy to use statistical packages (Bates et al., 2014; Kuznetsova et al., 2016), tutorials available on blogs (Magnusson, 2016), as well as other social media forums (e.g., Facebook methods groups) has made these methods accessible to all researchers. In addition, since many research designs in behavioral neuroendocrinology are simply factorial with clusters, reading books about multilevel modeling would likely provide a sufficient introduction.

Despite these limitations, statistical expertise and improved training are important and would likely advance quantitative methodology to some degree. However, due to the near ubiquity of litter use, dependencies mostly unaccounted for, deficiencies in reporting between-litter variation (Table 1), and substantially inflated false positive rates (Figure 2), higher-level action by journals and/or funding bodies may be required.

4.10. Limitations

There are several limitations that deserve attention. Simulations entail generating data and numerous assumptions. A valid question is whether our conclusions will generalize to real research situations. While our findings are built upon many assumptions, so are all aspects of inference regardless of whether data was obtained from actual rodents or simulated to have certain characteristics. When using any statistical software, the applied algorithm cannot differentiate between actual and simulated data. This results in all inputted numbers being treated in the same manner. Furthermore, exact false positive rates in the published literatures cannot be determined. Our simulations showed that between-litter variation can inflate error by some degree, and this is important to consider when inferring meaning from the extant literature.

Not providing comparisons between statistical methods with actual data may be viewed as objectionable or incomplete. Although using real data may appear more tangible, this would present several difficulties when examining optimal statistical methods. For example, with actual data, we do not know whether a true effect exists and cannot determine which method is arriving at the correct conclusion. As such, simulations offer clear advantages in that we know the correct conclusion and can therefore determine the appropriate method. In addition, commonly used measures of evidence have meaning in the long run and this makes exploring expected error rates with one data set problematic (Greenland et al., 2016).

It is also possible that estimates of between-litter variation are biased in small sample situations (Maas and Hox, 2005), which may influence false positive rates and power. Indeed, simulation based studies have shown that small samples and few clusters can present challenges for MLMs (Maas and Hox, 2005) and GEEs (Gunsolley et al., 1995). While it is true that small samples may show bias, we do not see this as unique to estimates of litter variation. That is, larger samples are always preferable irrespective of the parameter under investigation, as they provide more precise estimates of the population. We demonstrated that, with only four litters, expected error rates and optimal power can be achieved. Thus, even in small sample situations, we see MLMs or GEEs as methods that should be used for dependent data.

Our results are restricted to specific research designs (observations fully nested within litter), which limits the scope of our findings. We see this as a strength, however, in that we had sufficient focus to answer substantive questions as opposed to general quantitative practices and we provided resources for important research topics in the field of behavioral neuroendocrinology. We also used two different search strategies for each literature: for prenatal stress, we only included the 100 most recent studies. This decision was made because reviewing potentially 1,000's of articles seemed unnecessary to achieve our goal of offering recommendations based on current methodological practices. Said another way, our interest was not on challenging specific findings, but to shed light upon an important topic.

4.11. Conclusion

While there are notable limitations, we conclude that between-litter variation is underappreciated (Table 1), can lead to increased false positive rates (Figure 2), and reduce ones ability to detect an effect (Figure 3). Based on the strength of these findings, we offer several recommendations. First, it should be noted that our recommendations apply to research designs in which entire litters are categorized one way. Litters often receive multiple treatments, and a thorough discussion of the necessary model to account for dependencies in these situations is beyond the scope of the present paper (but see here: Aarts et al., 2015). However, when all litter mates included in a study are categorized based on the same characteristic or treatment, we suggest the following:

- Independence of observations from the same liter cannot be assumed.
- Including litter as a covariate is not appropriate.
- A statistical method specifically for non-independence is necessary.

We prefer a multilevel approach, as generalized estimating equations require small sample corrections for p-values. In addition, MLM's provide estimates of between-litter variation that can be used for prospective power analyses. If a GEE is used, relevant literature and simulations should be used to select the appropriate bias correction. An alternative approach that averages observations within litter can be found in Starr-Phillips and Beery (2014).

- To facilitate prospective power calculations, empirical papers should report the amount of between-litter variation observed.
- Power calculations should assume a range of plausible values for between-litter variation.
- The statistical significance of litter as a random effect should not be used to exclude it from the model.

ACKNOWLEDGMENT

DRW thanks Karen Bales for comments on previous drafts of this manuscript and the Dynamics in Psychological Science lab at UC Davis for providing feedback on this work.

4. References

Aarts, E., Dolan, C. V, Verhage, M., van der Sluis, S., 2015. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. BMC Neurosci 16, 94. doi:10.1186/s12868-015-0228-5

Agnish, N.D., Keller, K. a, 1997. The rationale for culling of rodent litters. Fundam. Appl. Toxicol. 38, 26. doi:10.1006/faat.1997.2318

Amugongo, S.K., Hlusko, L.J., 2014. Impact of Maternal Prenatal Stress on Growth of the Offspring. Aging Dis. 5, 116. doi:10.14336/AD.2014.05001

Archer, J.E., Blackman, D.E., 1971. Prenatal psychological stress and offspring behavior in rats and mice. Dev. Psychobiol. 4, 193248. doi:10.1002/dev.420040302

Baier, C.J., Pallars, M.E., Adrover, E., Monteleone, M.C., Brocco, M.A., Barrantes, F.J., Antonelli, M.C., 2015. Prenatal restraint stress decreases the expression of alpha-7 nicotinic receptor in the brain of adult rat offspring. Stress 18, 435445. doi:10.3109/10253890.2015.1022148

Baldwin, S.A., Fellingham, G.W., 2013. Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Psychol. Methods 18, 151164. doi:10.1037/a0030642

Barha, C.K., Pawluski, J.L., Galea, L.A.M., 2007. Maternal care affects male and female offspring working memory and stress reactivity. Physiol. Behav. 92, 939950. doi:10.1016/j.physbeh.2007.06.022

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Mem. Lang. 68, 255278. doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015a. Parsimonious Mixed Models. doi:arXiv:1506.04967

Bates D, Maechler M, Bolker B, Walker S, 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. doi:10.18637/jss.v067.i01

Baum, M.J., Everitt, B.J., Herbert, J., Keverne, E.B., 1977. Hormonal basis of proceptivity and receptivity in female primates. Arch. Sex. Behav. 6, 173192. doi:10.1007/BF01541126

Beery, A.K., Francis, D.D., 2011. Adaptive significance of natural variations in maternal care in rats: A translational perspective. Neurosci. Biobehav. Rev. 35, 15521561. doi:10.1016/j.neubiorev.2011.03.012

Boersma, G.J., Tamashiro, K.L., 2015. Individual differences in the effects of prenatal stress exposure in rodents. Neurobiol. Stress. doi:10.1016/j.ynstr.2014.10.006

Bond, N.W., di Giusto, E.L., 1976. Effects of prenatal alcohol consumption on open-field behaviour and alcohol preference in rats. Psychopharmacologia 46, 163165. doi:10.1007/BF00421386

Borm, G.F., Fransen, J., Lemmens, W.A.J.G., 2007. A simple sample size formula for analysis of covariance in randomized clinical trials. J. Clin. Epidemiol. 60, 12341238. doi:10.1016/j.jclinepi.2007.02.006

Brodkin, E.S., Carlezon, W.A., Haile, C.N., Kosten, T.A., Heninger, G.R., Nestler, E.J., 1998. Genetic analysis of behavioral, neuroendocrine, and biochemical parameters in inbred rodents: Initial studies in Lewis and Fischer 344 rats and in A/J and C57BL/6J mice. Brain Res. 805, 5568. doi:10.1016/S0006-8993(98)00663-5

Button, K.S., Ioannidis, J.P. a, Mokrysz, C., Nosek, B. a, Flint, J., Robinson, E.S.J., Munaf, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 36576. doi:10.1038/nrn3475

Callens, M., Croux, C., Glazier, R.H., Pebley, A., 2005. Performance of likelihood-based estimation methods for multilevel binary regression models. J. Stat. Comput. Simul. 75, 10031017. doi:10.1080/00949650412331321070

Cameron, N.M., 2011. Maternal programming of reproductive function and behavior in the female rat. Front. Evol. Neurosci. doi:10.3389/fnevo.2011.00010

Champagne, F.A., Curley, J.P., Keverne, E.B., Bateson, P.P.G., 2007. Natural variations in postpartum maternal care in inbred and outbred mice. Physiol. Behav. 91, 325334. doi:10.1016/j.physbeh.2007.03.014

Champagne, F.A., Francis, D.D., Mar, A., Meaney, M.J., 2003. Variations in maternal care in the rat as a mediating influence for the effects of environment on development. Physiol. Behav. 79, 359371. doi:10.1016/S0031-9384(03)00149-5

Chatterjee, A., Chatterjee, R., 2009. How stress affects female reproduction: An overview. Biomed. Res. 20, 7983.
Cox, D.R., McCullagh, P., 1982. A Biometrics Invited Paper with Discussion. Some Aspects of Analysis of Covariance. Biometrics 38, 541561. doi:10.2307/2530040

Crabbe, J.C., Wahlsten, D., Dudek, B.C., 1999. Genetics of Mouse Behavior: Interactions with Laboratory Environment. Science (80-.). 284, 16701672. doi:10.1126/science.284.5420.1670

Cratty, M.S., Ward, H.E., Johnson, E.A., Azzaro, A.J., Birkle, D.L., 1995. Prenatal stress increases corticotropin-releasing factor (CRF) content and release in rat amygdala minces. Brain Res. 675, 297302. doi:10.1016/0006-8993(95)00087-7

Curley, J.P., Champagne, F.A., 2016. Influence of maternal care on the developing brain: Mechanisms, temporal dynamics and sensitive periods. Front. Neuroendocrinol. 40, 5266. doi:10.1016/j.yfrne.2015.11.001

Deitchman, R., Kapusinski, D., Burkholder, J., 1977. Maternal Behavior in handled and nonhandled mice and its relation to later pups behavior. Psychol. Rep. 40, 411420. doi:10.2466/pr0.1977.40.2.411

Dimitsantos, E., Escorihuela, R.M., Fuentes, S., Armario, A., Nadal, R., 2007. Litter size affects emotionality in adult male rats. Physiol. Behav. 92, 708716. doi:10.1016/j.physbeh.2007.05.066

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. U. S. A. 113, 79005. doi:10.1073/pnas.1602413113

Fay, M.P., 2015. Package saws Title Small-Sample Adjustments for Wald tests Using Sandwich Estimators.

Fay, M.P., Graubard, B.I., 2001. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics 57, 11981206. doi:10.1111/j.0006-341X.2001.01198.x

Fish, E.W., Shahrokh, D., Bagot, R., Caldji, C., Bredy, T., Szyf, M., Meaney, M.J., 2004. Epigenetic programming of stress responses through variations in maternal care. Ann. N. Y. Acad. Sci. 1036, 167180. doi:10.1196/annals.1330.011

Fowden, A.L., Forhead, A.J., 2004. Endocrine mechanisms of intrauterine programming. Reproduction 127, 515526. doi:10.1530/rep.1.00033

Francis, D.D., Meaney, M.J., 1999. Maternal care and the development of stress responses. Curr. Opin. Neurobiol. 9, 128134. doi:10.1016/S0959-4388(99)80016-6

Francis, G., 2012. Too good to be true: Publication bias in two prominent studies from experimental psychology. Psychon. Bull. Rev. 19, 151156. doi:10.3758/s13423-012-0227-9

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel. Cambridge University Press. doi:10.1017/CBO9781107415324.004

Gelman, A., Loken, E., 2014. 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. Psychol. Bull. 140, 12721280. doi:dx.doi.org/10.1037/a0037714

Glowa, J.R., Hansen, C.T., 1994. Differences in response to an acoustic startle stimulus among forty-six rat strains. Behav. Genet. 24, 7984. doi:10.1007/BF01067931

Goldstein, J.M., Handa, R.J., Tobet, S.A., 2014. Disruption of fetal hormonal programming (prenatal stress) implicates shared risk for sex differences in depression and cardiovascular disease. Front. Neuroendocrinol. doi:10.1016/j.yfrne.2013.12.001

Gore, A.C., 2008. Developmental programming and endocrine disruptor effects on reproductive neuroendocrine systems. Front. Neuroendocrinol. 29, 358374. doi:10.1016/j.yfrne.2008.02.002

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. 31, 337350. doi:10.1007/s10654-016-0149-3

Gudsnuk, K., Champagne, F.A., 2012. Epigenetic Influence of Stress and the Social Environment. ILAR J. 53, 279288. doi:10.1093/ilar.53.3-4.279

Gunsolley, J.C., Getchell, C., Chinchilli, V.M., 1995. Small sample characteristics of generalized estimating equations. Commun. Stat. - Simul. Comput. 24, 869878. doi:10.1080/03610919508813280

Halekoh, U., Hjsgaard, S., Yan, J., 2006. The R package geepack for generalized estimating equations. J. Stat. Softw.

Hanley, J.A., Negassa, A., Edwardes, M.D. deB., Forrester, J.E., 2003. Statistical analysis of correlated data using generalized estimating equations: An orientation. Am. J. Epidemiol. 157, 364375. doi:10.1093/aje/kwf215

Haworth, C.M.A., Wright, M.J., Luciano, M., Martin, N.G., de Geus, E.J.C., van Beijsterveldt, C.E.M., Bartels, M., Posthuma, D.,

Boomsma, D.I., Davis, O.S.P., Kovas, Y., Corley, R.P., DeFries, J.C., Hewitt, J.K., Olson, R.K., Rhea, S.-A., Wadsworth, S.J., Iacono, W.G., McGue, M., Thompson, L.A., Hart, S.A., Petrill, S.A., Lubinski, D., Plomin, R., 2010. The heritability of general cognitive ability increases linearly from childhood to young adulthood. Mol. Psychiatry 15, 11121120. doi:10.1038/mp.2009.55

Hedges, L. V., 2007. Effect Sizes in Cluster-Randomized Designs. J. Educ. Behav. Stat. 32, 341370. doi:10.3102/1076998606298043

Hellstrom, I.C., Dhir, S.K., Diorio, J.C., Meaney, M.J., 2012. Maternal licking regulates hippocampal glucocorticoid receptor transcription through a thyroid hormone-serotonin-NGFI-A signalling cascade. Philos. Trans. R. Soc. B-BIOLOGICAL Sci. 367, 24952510. doi:10.1098/rstb.2012.0223

Hofer, M.A., 1973. Maternal separation affects infant rats behavior. Behav. Biol. 9, 629633. doi:10.1016/S0091-6773(73)80057-4

Holson, R.R., Pearce, B., 1992. Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species. Neurotoxicol. Teratol. 14, 221228. doi:10.1016/0892-0362(92)90020-B

Horton, N.J., Bebchuk, J.D., Jones, C.L., Lipsitz, S.R., Catalano, P.J., Zahner, G.E.P., Fitzmaurice, G.M., 1999. Goodness-of-fit for GEE: An example with mental health service utilization. Stat. Med. 18, 213222. doi:10.1002/(SICI)1097-0258(19990130)18:2¡213::AID-SIM999¿3.0.CO;2-E

Hubbard, A.E., Ahern, J., Fleischer, N.L., Laan, M. Van der, Lippman, S.A., Jewell, N., Bruckner, T., Satariano, W.A., 2010. To GEE or Not to GEE. Epidemiology 21, 467474. doi:10.1097/EDE.0b013e3181caeb90

Inoue-Murayama, M., 2009. Genetic polymorphism as a background of animal behavior. Anim. Sci. J. doi:10.1111/j.1740-0929.2008.00623.x

Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLoS Med. 2, e124. doi:10.1371/journal.pmed.0020124

Jacobson, L., Sapolsky, R., 1991. The role of the hippocampus in feedback regulation of the hypothalamic-pituitary-adrenocortical axis. Endocr. Rev. 12, 118134. doi:10.1210/edrv-12-2-118

Jang, K.L., Livesley, W.J., Vemon, P.A., 1996. Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study. J. Pers. 64, 577592. doi:10.1111/j.1467-6494.1996.tb00522.x

Joffe, J.M., 1977. Modification of prenatal stress effects in rats by dexamethasone and adrenocorticotrophin. Physiol.

Behav. 19, 601606. doi:10.1016/0031-9384(77)90032-4

Kapoor, A., Dunn, E., Kostaki, A., Andrews, M.H., Matthews, S.G., 2006. Fetal programming of hypothalamo-pituitary-adrenal function: Prenatal stress and glucocorticoids. J. Physiol. 572, 3144. doi:10.1016/j.poly.2005.06.060

Kerr, N., 1998. HARKing: Hypothesizing after the results are known. Personal. Soc. Psychol. Rev.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D.,

Keselman, J.C., Levin, J.R., 1998. Statistical Practices of Educational Researchers: An Analysis of their ANOVA, MANOVA, and ANCOVA Analyses. Rev. Educ. Res. 68, 350386. doi:10.3102/00346543068003350

Kimpel, M.W., Strother, W.N., McClintick, J.N., Carr, L.G., Liang, T., Edenberg, H.J., McBride, W.J., 2007. Functional gene expression differences between inbred alcohol-preferring and -non-preferring rats in five brain regions. Alcohol 41, 95132. doi:10.1016/j.alcohol.2007.03.003

Kinney, D.K., Munir, K.M., Crowley, D.J., Miller, A.M., 2008. Prenatal stress and risk for autism. Neurosci. Biobehav. Rev. doi:10.1016/j.neubiorev.2008.06.004

Koehl, M., Darnaudry, M., Dulluc, J., Van Reeth, O., Moal, M. Le, Maccari, S., 1999. Prenatal stress alters circadian activity of hypothalamo-pituitary- adrenal axis and hippocampal corticosteroid receptors in adult rats of both gender. J. Neurobiol. 40, 302315. doi:10.1002/(SICI)1097-4695(19990905)40:3¡302::AID-NEU3¿3.0.CO;2-7

Kruschke, J.K., 2013. Bayesian estimation supersedes the t test. J. Exp. Psychol. Gen. 142, 573603. doi:10.1037/a0029146

Kuznetsova, A., Brockhoff, P., Christensen, R., 2016. lmerTest: Tests in Linear Mixed Effects Models. R Packag. version 3.0.0, https://cran.r-project.org/package=lmerTest.

Lakens, D., 2017. Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses. Soc. Psychol. Personal. Sci. doi:10.1177/1948550617697177

Lazic, S.E., 2010. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci. 11, 5. doi:10.1186/1471-2202-11-5

Lazic, S.E., Essioux, L., 2013. Improving basic and translational science by accounting for litter-to-litter variation in animal models. BMC Neurosci. 14, 37. doi:10.1186/1471-2202-14-37

Levy, K.J., 1980. A Monte Carlo Study of Analysis of Covariance Under Violations of the Assumptions of Normality and Equal Regression Slopes. Educ. Psychol. Meas. 40, 835840. doi:10.1177/001316448004000404

Li, P., Redden, D.T., 2015. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. Stat. Med. 34, 281296. doi:10.1002/sim.6344

Liu, D., Diorio, J., Tannenbaum, B., Caldji, C., Francis, D., Freedman, A., Sharma, S., Pearson, D., Plotsky, P.M., Meaney, M., 1997. Maternal Care, Hippocampal Glucocorticoid Receptors, and Hypothalamic-Pituitary-Adrenal Responses to Stress. Science (80-.). 277, 16591662. doi:10.1126/science.277.5332.1659

Lupien, S.J., McEwen, B.S., Gunnar, M.R., Heim, C., 2009. Effects of stress throughout the lifespan on the brain, behaviour and cognition. Nat. Rev. Neurosci. 10, 434445. doi:10.1038/nrn2639

Maas, C.J.M., Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. Methodology 1, 8692. doi:10.1027/1614-2241.1.3.86

Magnusson, K., 2016. Using R and lme / lmer to fit different two- and three-level longitudinal models Data format [WWW Document]. github. URL http://rpsychologist.com/r-guide-longitudinal-lme-lmer (accessed 6.7.17).

Marceau, K., McMaster, M.T.B., Smith, T.F., Daams, J.G., van Beijsterveldt, C.E.M., Boomsma, D.I., Knopik, V.S., 2016. The Prenatal Environment in Twin Studies: A Review on Chorionicity. Behav. Genet. 46, 286303. doi:10.1007/s10519-016-9782-6

Markham, J.A., Koenig, J.I., 2011. Prenatal stress: Role in psychotic and depressive diseases. Psychopharmacology (Berl). doi:10.1007/s00213-010-2035-0

McCormick, C.M., Smythe, J.W., Sharma, S., Meaney, M.J., 1995. Sex-specific effects of prenatal stress on hypothalamic-pituitary-adrenal responses to stress and brain glucocorticoid receptor density in adult rats. Dev. Brain Res. 84, 5561. doi:10.1016/0165-3806(94)00153-Q

McEwen, B.S., 2008. Understanding the potency of stressful early life experiences on brain and body function. Metabolism. 57, S11-5. doi:10.1016/j.metabol.2008.07.006

McGowan, P.O., Suderman, M., Sasaki, A., Huang, T.C.T., Hallett, M., Meaney, M.J., Szyf, M., 2011. Broad epigenetic signature of maternal care in the brain of adult rats. PLoS One 6, e14739. doi:10.1371/journal.pone.0014739

McGrady, A. V., 1984. Effects of psychological stress on male reproduction: a review. Arch. Androl. 13, 17. doi:10.3109/01485018408987495

Menard, J.L., Champagne, D.L., Meaney, M.J.P., 2004. Variations of maternal care differentially influence ‘fear reactivity and regional patterns of cFos immunoreactivity in response to the shock-probe burying test. Neuroscience 129, 297308. doi:10.1016/j.neuroscience.2004.08.009

Mendi, M., 1988. The effects of litter size variation on mother-eroffspring relationships and behavioural and physical development in several mammalian species (principally rodents). J. Zool. 215, 1534. doi:10.1111/j.1469-7998.1988.tb04882.x

Mueller, B.R., Bale, T.L., 2008. Sex-Specific Programming of Offspring Emotionality after Stress Early in Pregnancy. J. Neurosci. 28, 90559065. doi:10.1523/JNEUROSCI.1424-08.2008

Muth, C., Bales, K.L., Hinde, K., Maninger, N., Mendoza, S.P., Ferrer, E., 2016. Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data. Educ. Psychol. Meas. 76, 6487. doi:10.1177/0013164415580432

Neeley, E.W., Berger, R., Koenig, J.I., Leonard, S., 2011. Strain dependent effects of prenatal stress on gene expression in the rat hippocampus. Physiol. Behav. 104, 334339. doi:10.1016/j.physbeh.2011.02.032

Noguchi, K., Gel, Y.R., Brunner, E., Konietschke, F., 2012. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. J. Stat. Softw. 50, 123. doi:10.18637/jss.v050.i12

OConnor, T.M., OHalloran, D.J., Shanahan, F., 2000. The stress response and the hypothalamic-pituitary-adrenal axis: from molecule to melancholia. QJM 93, 323333. doi:10.1093/qjmed/93.6.323

OCS, 2015. Estimating the reproducibility of psychological science. Science (80-.). 349, aac4716-aac4716. doi:10.1126/science.aac4716

Pan, W., Wall, M.M., 2002. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat. Med. 21, 14291441. doi:10.1002/sim.1142

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects models in S and S+. Springer Science and Business Media. doi:10.1017/CBO9781107415324.004

Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10, 712712. doi:10.1038/nrd3439-c1

Ramos, A., Berton, O., Mormde, P., Chaouloff, F., 1997. A multiple-test study of anxiety-related behaviours in six inbred rat strains. Behav. Brain Res. 85, 5769. doi:10.1016/S0166-4328(96)00164-7 Ripley, B., 2015. GEE. R Packag. version.

Roux, A.V.D., 2002. A glossary for multilevel analysis. J. Epidemiol. Community Health 56, 588594. doi:10.1136/jech.56.8.588

Sapolsky, R.M., Meaney, M.J., 1986. Maturation of the adrenocortical stress response: Neuroendocrine control mechanisms and the stress hyporesponsive period. Brain Res. Rev. 11, 6576. doi:10.1016/0165-0173(86)90010-X

Scamvougeras, A., Kigar, D.L., Jones, D., Weinberger, D.R., Witelson, S.F., 2003. Size of the human corpus callosum is genetically determined: An MRI study in mono and dizygotic twins. Neurosci. Lett. doi:10.1016/S0304-3940(02)01333-2

Scariano, S.M., Davenport, J.M., 1987. The effects of violations of independence assumptions in the one-way ANOVA. Am. Stat. 41, 123129. doi:10.1080/00031305.1987.10475459

Scholz, J., LaLiberte, C., van Eede, M., Lerch, J.P., Henkelman, M., 2016. Variability of brain anatomy for three common mouse strains. Neuroimage 142, 656662. doi:10.1016/j.neuroimage.2016.03.069

Seckl, J.R., Holmes, M.C., 2007. Mechanisms of Disease: glucocorticoids, their placental metabolism and fetal programming of adult pathophysiology. Nat. Clin. Pract. Endocrinol. Metab. 3, 479488. doi:10.1038/ncpendmet0515

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality(complete samples). Biometrica 52, 591611. doi:10.2307/1267427 Simmons, J.P., Nelson, L.D., Simonsohn, U., 2013. Life after p-hacking. SSRN Electron. J. 41, 775. doi:10.2139/ssrn.2205186

Starr-Phillips, E.J., Beery, A.K., 2014. Natural Variation in Maternal Care Shapes Adult Social Behavior in Rats. Dev. Psychobiol. 56, 10171026. doi:10.1002/dev.21182

Szuran, T.F., Plika, V., Pokorny, J., Welzl, H., 2000. Prenatal stress in rats: effects on plasma corticosterone, hippocampal glucocorticoid receptors, and maze performance. Physiol. Behav. 71, 353362. doi:10.1016/S0031-9384(00)00351-6

Tanaka, T., 2004. The relationships between litter size, offspring weight, and behavioral development in laboratory mice Mus musculus. Mammal Study 29, 147153.

Turecki, G., Meaney, M.J., 2016. Effects of the Social Environment and Stress on Glucocorticoid Receptor Gene Methylation: A Systematic Review. Biol. Psychiatry. doi:10.1016/j.biopsych.2014.11.022

van t Veer, A.E., Giner-Sorolla, R., 2016. Pre-registration in social psychology-A discussion and suggested template. J. Exp. Soc. Psychol. 67, 212. doi:10.1016/j.jesp.2016.03.004

van Hasselt, F.N., Tieskens, J.M., Trezza, V., Krugers, H.J., Vanderschuren, L.J.M.J., Jols, M., 2012. Within-litter variation in maternal care received by individual pups correlates with adolescent social play behavior in male rats. Physiol. Behav. 106, 701706. doi:10.1016/j.physbeh.2011.12.007

von Engelhardt, N., Kowalski, G.J., Guenther, A., 2015. The maternal social environment shapes offspring growth, physiology, and behavioural phenotype in guinea pigs. Front. Zool. 12, S13. doi:10.1186/1742-9994-12-S1-S13

Wahlsten, D., Bachmanov, A., Finn, D.A., Crabbe, J.C., 2007. Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. Pnas 103, 1636416369. doi:10.1073/pnas.0605342103

Walum, H., Waldman, I.D., Young, L.J., 2016. Statistical and Methodological Considerations for the Interpretation of

Intranasal Oxytocin Studies. Biol. Psychiatry 79, 251257. doi:10.1016/j.biopsych.2015.06.016

Wang, Y.A., Sparks, J., Gonzales, J.E., Hess, Y.D., Ledgerwood, A., 2017. Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. J. Exp. Soc. Psychol. 72, 118124. doi:10.1016/j.jesp.2017.04.011

Weinstock, M., 2017. Prenatal stressors in rodents: Effects on behavior. Neurobiol. Stress 6, 313. doi:10.1016/j.ynstr.2016.08.004

Weinstock, M., 2008. The long-term behavioural consequences of prenatal stress. Neurosci. Biobehav. Rev. 32, 10731086. doi:10.1016/j.neubiorev.2008.03.002

Wilson, C.A., Vazdarjanova, A., Terry, A. V, 2013. Exposure to variable prenatal stress in rats: Effects on anxiety-related behaviors, innate and contextual fear, and fear extinction. Behav. Brain Res. 238, 279288. doi:10.1016/j.bbr.2012.10.003

Zimmerman, D.W., 1987. Comparative Power of Student T Test and Mann-Whitney U Test for Unequal Sample Sizes and Variances. J. Exp. Educ. 55, 171174. doi:10.1080/00220973.1987.10806451

Zorrilla, E.P., 1997. Multiparous species present problems (and possibilities) to developmentalists. Dev. Psychobiol. 30, 14150. doi:10.1002/(SICI)1098-2302(199703)30:2¡141::AID-DEV5¿3.0.CO;2-Q [pii]