Frontiers in Neuroendocrinology 47 (2017) 154–166

Contents lists available at ScienceDirect

Frontiers in Neuroendocrinology

journal homepage: www.elsevier.com/locate/yfrne

Review article

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

Donald R. Williams[a,*], Rickard Carlsson[c], Paul-Christian Bürkner[b]

[a] Psychology, University of California, Davis, One Shields Avenue, Davis, CA 95616, United States
[b] Institute of Psychology, University of Muenster, Fliednerstraße 21, 48151 Muenster, Germany
[c] Department of Psychology, Linnaeus University, Sweden

ARTICLE INFO

Keywords:
Litter effects
Hormones and behavior
False positives
Power
Between-litter variation
Maternal care
Prenatal stress

ABSTRACT

Developmental studies of hormones and behavior often include littermates—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 ($\alpha = 0.05$) exceeded nominal levels (up to 0.70), and power ($1 - \beta$) 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., 2003a), 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): litters are comprised of siblings that share early-life experiences and genes that can contribute to litter effects (Lazic and Essioux, 2013). To be clear, between-litter variation is synonymous with litter effects that arise from within-litter similarities. Both research designs categorize entire litters (i.e., siblings) based on maternal behavior or whether they were exposed to the same experimental condition (Fig. 1). 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 for 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 routinely be violated; (3) review statistical methods commonly used in both literatures; (4) simulate type I error (false positive) rates for statistical approaches found in the literature, in addition to multilevel models (MLM), generalized estimating equations (GEE), and analyzing litter means with a *t*-test; 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 of an organism, including physiological and behavioral trajectories (Gore, 2008). Since the stress axis plays a critical role in survival (Lupien et al., 2009; O'Connor 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

* Corresponding author at: One Shields Ave., Department of Psychology, University of California-Davis, Davis, CA 95616, United States.
E-mail address: drwwilliams@ucdavis.edu (D.R. Williams).

http://dx.doi.org/10.1016/j.yfrne.2017.08.003
Received 13 June 2017; Received in revised form 14 August 2017; Accepted 17 August 2017
Available online 22 August 2017
0091-3022/

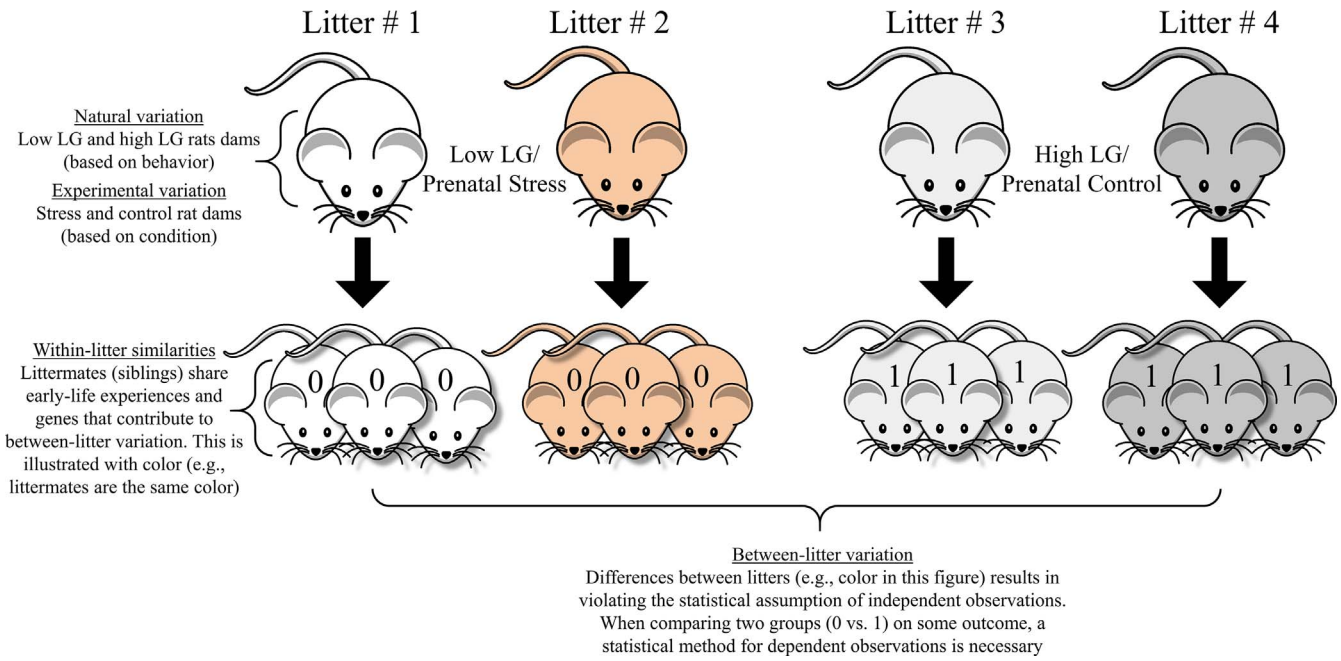

Fig. 1. Both research areas—natural variation in maternal care and prenatal stress—categorize entire litters based on maternal behavior or experimental condition: littermates from High LG/Prenatal Stress and Low-LG/Prenatal Control rat dams are coded the same way. Between-litter variation is illustrated with color. Littermates share the same color, whereas color differs between litters. This shows that within-litter similarities are what produces between-litter variation. Suppose we were interested in some physiological outcome that is hypothesized to differ between pups raised by high and low LG mothers. When groups are compared (dummy coded: 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.

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 the prenatal environment have been examined, we focus on prenatal stress because of the thoroughness of the literature: several prenatal stress manipulations have been developed and the effects on offspring development described (Weinstock, 2008, 2017).

1.2. Natural variation: 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 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 where shown to 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 PNS offspring (Wilson et al., 2013). Later in life, PNS rodents show increased HPA axis reactivity to stressors, such as increased 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).

1.4. Rational for between-litter variation

Maternal care and prenatal stress are important components of the early environment. However, there are many others factor that can contribute to between-litter variation. That is, the fact that the early environment influences development, also suggests that those sharing the same environment (pre or postnatal) will be more alike than those from different environments. Litters size has been shown to influence 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). Littermates 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 substantial evidence that *in utero* hormonal milieus (Fowden and Forhead, 2004) and the early social environment influence development (Turecki and Meaney, 2016). While litter effects were not of primary interest in these studies, they provide indirect evidence for between-litter variation.

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 littermates that are outbred rodents (Glowa and Hansen, 1994).

The the natural occurring maternal variation 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 littermates. 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 May 20, 2017. We sought to understand how litter was 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 rodent studies directly related to the neuroendocrine system. For both literatures, outcomes could be either behavioral or physiological.

The identified research articles were used to describe how litter dependencies were accounted for. 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) the number of litters included in the analyses; (2) the number of observations used per litter; and (3) methods used to account for litter effects. We documented the search procedure and provided these documents on the Open Science Framework (https://osf.io/fxy7h/).

2.2. Simulation: False positives

We examined type I error 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 effects were not often accounted for or were considered as a covariate. In three papers, a statistical method was used that accounted for dependent measures with a random effect (Barha et al., 2007; Neeley et al., 2011) or corrected standard errors (Amugongo and Hlusko, 2014). In two studies litter means were the unit of analysis (Caldji et al., 1998; Starr-Phillips and Beery, 2014). We thus compared five models: (1) *t*-test (litter not included in the model): (2) analysis of covariance (ANCOVA; litter included as a categorical covariate); (3) multilevel model (MLM; Roux, 2002); (4) generalized estimating equation (GEE; Hanley et al., 2003); and (5) *t*-test with litter means (i.e., each litter contributed one observation to the analysis).

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 *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 variance explained by litter (where σ_e^2 is the within-litter variance). The ICC can also be thought of as the correlation between individual 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 or the number of observations used per litter. As such, we assumed a range of simulation conditions (litters = 4, 8, 12, and 24; per litter = 2, 4, 6, and 8; ICC = 0–0.70 by increments of 0.05). For each condition, observations from half of the litters were dummy coded as 0, whereas observations from the remaining litters were coded as 1 (Fig. 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 variables from a model (Barr et al., 2013). We thus investigated whether false positive rates were conditional on a significant litter effect (litters = 8 and per litter = 4). We computed the significance of litter as a MLM random effect (via a likelihood ratio test), then analyzed the data with a *t*-test. In this way,

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	85/96 (89%)	31/99 (31%)	0/90 (0%)	82/96 (85%)

These estimates were obtained from the primary studies (natural variation in maternal care = 35 and prenatal stress = 100) that provided sufficient information for reliable estimates. This means that not all reviewed studies contributed to the percentages: 84/89 (88%) for the prenatal stress literature indicates that 11 studies did not provide enough information to answer that specific question.

we obtained

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

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

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 three approaches to incorporate between-litter variation into experimental design with power calculations: MLM, GEE, and analyzing litter means with a *t*-test. We address two specific objectives: (1) 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; and (2) incremental power increases due to adding additional littermates to the analysis (holding litter number constant). Here, if we assume obtaining more litters is not always feasible, this aim explicitly addresses expected power gains by using more observations from a given litter.

For objective one, we found that group sizes varied but were typically small. We chose an 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; 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

$$\delta_T = \frac{\beta}{\sqrt{\sigma_u^2 + \sigma_e^2}} \quad (4)$$

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 ($\delta_T = 0.20, 0.50, 0.80, 1.10$). The interpretation of δ_T follows Cohen's *d*, so the selected values covered what are considered small (0.20), medium (0.50), and large effects (0.80).

For objective two, we assumed 8 litters in total and varied the number of observations per litter (2–10 by increments of 1). Power for a large effect was investigated ($\delta_T = 0.80$) across a range of ICC values (0–0.80 by increments of 0.20). We also computed power for a *t*-test in which observations were all obtained from different litters. This was done for each sample size and presented with the simulation results.

2.5. Simulation: Uncertainty due to litter

This simulation demonstrated how between-litter variation influences uncertainty of the fixed effect estimate. This was achieved by computing confidence intervals (*CI*) for an unstandardized group difference ($\beta = 8.0$) across a range of ICC values. Since a 95-% *CI* excluding zero is significant at the $\alpha = 0.05$ level, this allowed for visualizing how between-litter variation effects false positives and power (e.g., interval width according to between-litter variation).

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 (R Core Team, 2016). The MLMs were fitted with the package *lmerTest* (Kuznetsova et al., 2016) that is a front end to *lme4* (Bates et al., 2015b), whereas *gee* (Ripley, 2015) was used for the GEEs and *saws* (Fay, 2015) for the bias corrected standard errors. All code and results for the simulations are publicly available (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 (Appendix and 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 for obtaining precise estimates for the number of litters and observations per litter. However, multiple animals from the same litter were used in most studies (MC = 100% and PNS = 89%). Although litter effects were explicitly considered (MC = 22% and PNS = 31%), this often resulted in reducing the number of littermates 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 observations 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 littermates are siblings and thus selected only one animal per litter:

> 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., 2015).

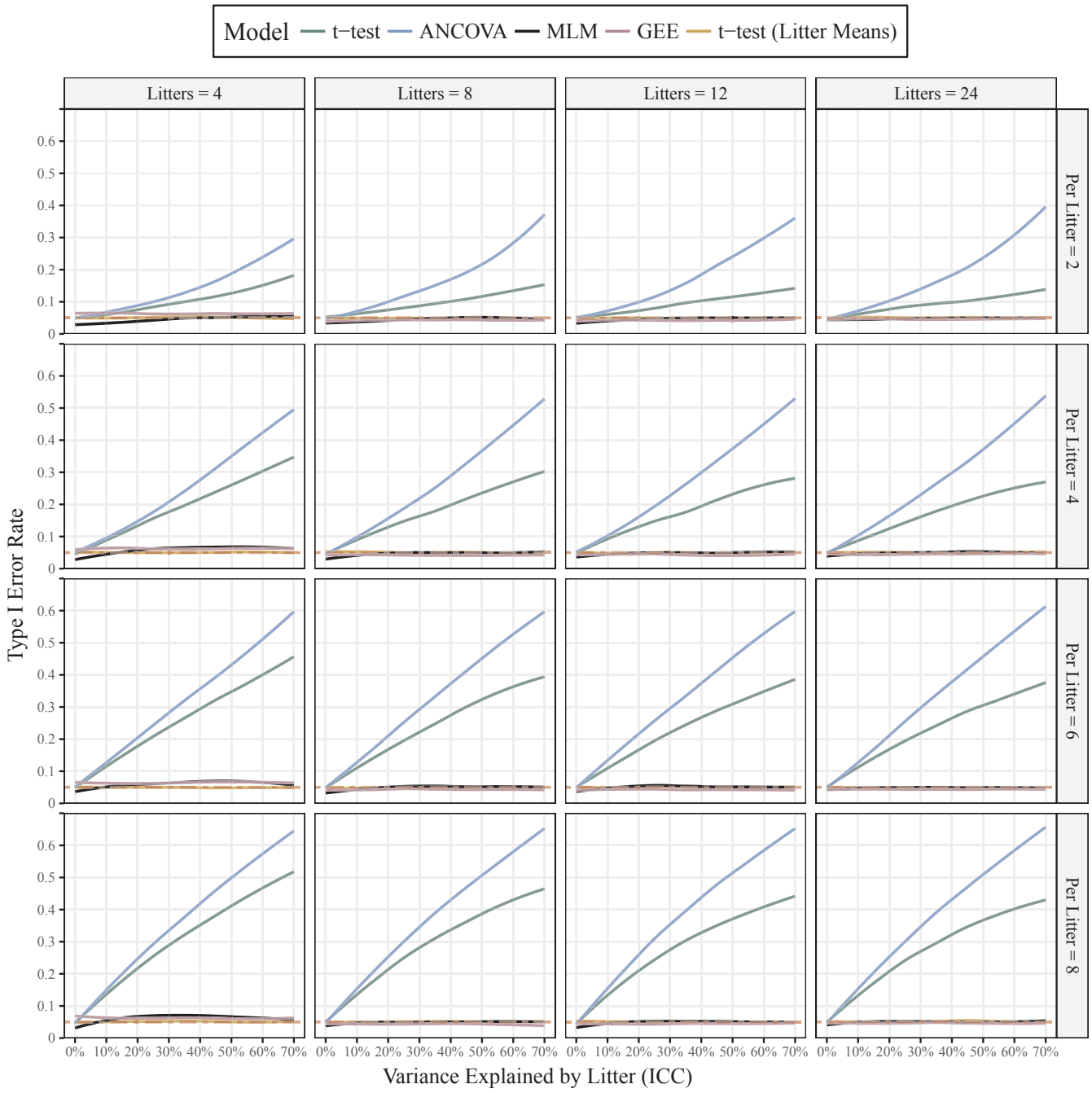

Fig. 2. *t*-tests and ANCOVAs have inflated type I error rates when between-litter variation is non-zero (ICC > 0%), and this is directly related to the degree of between-litter variation. MLM, GEE, and analyzing litter means (*t*-test)—statistical approaches that account for dependent measures—have adequate performance across most conditions.

3.2. False positives rates

The results are presented in Fig. 2 and include type I error rates for five methods: (1) *t*-test (green): (2) ANCOVA (blue); (3) MLM (black); (4) GEE (pink); and (5) *t*-test with litter means (yellow). Each model compared mean differences—assuming a *true* null hypothesis—between two groups but differed in how litter effects were handled (see Section 2.1. *Simulation: false positives*).

The *t*-test (green) did not include litter and type I error rates exceeded nominal levels ($\alpha = 0.05$; red dashed line), ranging from approximately 0.05–0.51. The latter is close to a 1000% increase from 0.05. Nominal error rates were achieved for all conditions in which the ICC was 0%. That is, when littermates did not resemble one another, the *t*-test had optimal performance. However, with an ICC of 5% error rates approached 0.10 (litters = 12 and per litter = 8). For sample sizes more commonly seen in behavioral neuroendocrinology, type I error rates approached 0.30 (ICC = 40%; litters = 4 and per litter = 6). Across all conditions, the ICC was directly related to error rates and this became more pronounced with larger sample sizes. Importantly, for 24 litters and only two observations per litter, type 1error rates were also compromised.

With litter as a categorical covariate in an ANCOVA (blue), we observed the same patterns as the *t*-test: type I 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 ($\alpha = 0.05$) were achieved. Across all conditions, however, there were substantial differences between the *t*-test and ANCOVA in that the latter had 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 ANCOVA had an error rate of 0.37 (640% increase from $\alpha = 0.05$).

We then examined type I error rates for three statistical approaches that are specifically for dependent data: MLM (black); GEE (pink); and *t*-tests with litter means (yellow). They showed similar performance (i.e., substantial overlap in Fig. 2). This was expected and highlights that all three methods generally performed well across conditions. However, we also observed that MLMs and GEEs 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 0%. 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. Additionally, analyzing litter means consistently produced type I error rates around 0.05.

3.3. Conditional false positives rates

We examined type I error rates for the fixed effect (0 vs. 1) using a *t*-test, conditional on a significant litter effect in a multilevel model (*2.3. Simulation: conditional false positives*). The error rates were consistently higher when there was a significant litter effect (Fig. 5a). However, when the litter effect was non-significant error rates were also problematic (ICC > 10%). For ICC values previously reported in the literature (60%; Lazic and Essioux, 2013) error rates exceeded 0.20, even when the litter effect was non-significant.

3.4. Power

Since only MLM, GEE, and *t*-test with litter means achieved nominal type I error rates ($\alpha = 0.05$), power was examined for only these methods. In the first simulation, we held the total sample size constant and varied the ratio of litters to observations per litter (Fig. 3). The second simulation investigated incremental power gains from increasing the number of observations per litter (holding litter size constant; Fig. 4).

The first simulation showed that power was related to the magnitude of between-litter variance and this was the case for all three methods. For the largest effect investigated ($\delta_T = 1.1$) power was greater than 0.90 when the ICC was 0% (litters = 12 and per litter = 4), but reduced substantially when the ICC was 70% (MLM = 0.49; GEE = 0.47; *t*-test = 0.50). Indeed, even with optimistic sample sizes ($N = 48$), power reached 0.80 in few conditions. For small ($\delta_T = 0.2$) and medium size effects ($\delta_T = 0.5$) power did not exceed 0.32. There were some power differences between methods. MLM often had greater power than GEE but this disparity was generally small (i.e., < 3%). In some conditions (ICC = 0%), analyzing litter means had more power than MLM and GEE. However, with larger ICC values, MLM had more power than analyzing litter means. Power was directly related to the sample composition and this was the case for all three methods. For example, power exceeded 0.80 for each method when there was 12 Litters and 4 observations per litter (grey line), but was substantially lower for 4 litters and 12 observations per litter ($\delta_T = 1.1$).

The second simulation also showed that power was influenced by between-litter variation, but in the context of incremental power (holding the effect and litter size constant). When observations were effectively independent (ICC = 0%) including additional littermates increased power for all methods. However, with even some between-litter variance (ICC = 20%) power gains quickly diminished: including 8 additional littermates (64 more observations in total) did not double power. For the largest ICC (80%), 8 additional littermates increased power by less than 2%. In fact, for all ICC values > 20%, each additional littermate increased power by less than 1%. Power was consistently higher when no littermates were included in the analysis (Fig. 4: dotted line).

3.5. Uncertainty due to litter

We used simulations to compute 95-% confidence intervals (from the average standard error across simulations) for an unstandardized effect size across a range of ICC values (0–0.70, Fig. 5b). When between-litter variation increased, the confidence intervals were wider. Whereas the effect was significant with an ICC of 0%, it was no longer statistically significant (interval included 0) with an ICC of 20%. Thus between-litter variance decreased power to detect a *true* effect. The same logic applies to type I error rates. When between-litter variance is ignored, the width of the confidence interval will be too narrow. This can increase the rate at which the confidence intervals erroneously exclude zero.

4. Discussion

The present study investigated how between-litter variation has been accounted for in two literatures—natural occurring maternal variation and prenatal stress. Specifically, we determined how often dependent measures (i.e., litter effects) have been considered and the degree to which between-litter variation effects false positive rates and 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 observations from the same litter, and only 15% used a statistical method 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 likely compromised 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 (Figs. 3 and 4). 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 (ICC > 0%), 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, an ANCOVA assumes the errors are uncorrelated which is not plausible when littermates are included in the same analysis (Keselman et al., 1998). In addition, ANCOVA assumes there is no interaction between the independent variable and the 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 allow for substantial researcher 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 in human oriented psychology are advising to pre-register covariates (van 't Veer and Giner-Sorolla, 2016; Wang et al., 2017).

We also examined type I error rates of statistical methods specifically for dependent measures. Across most conditions, nominal error rates were achieved for all three methods. It should be noted that, with

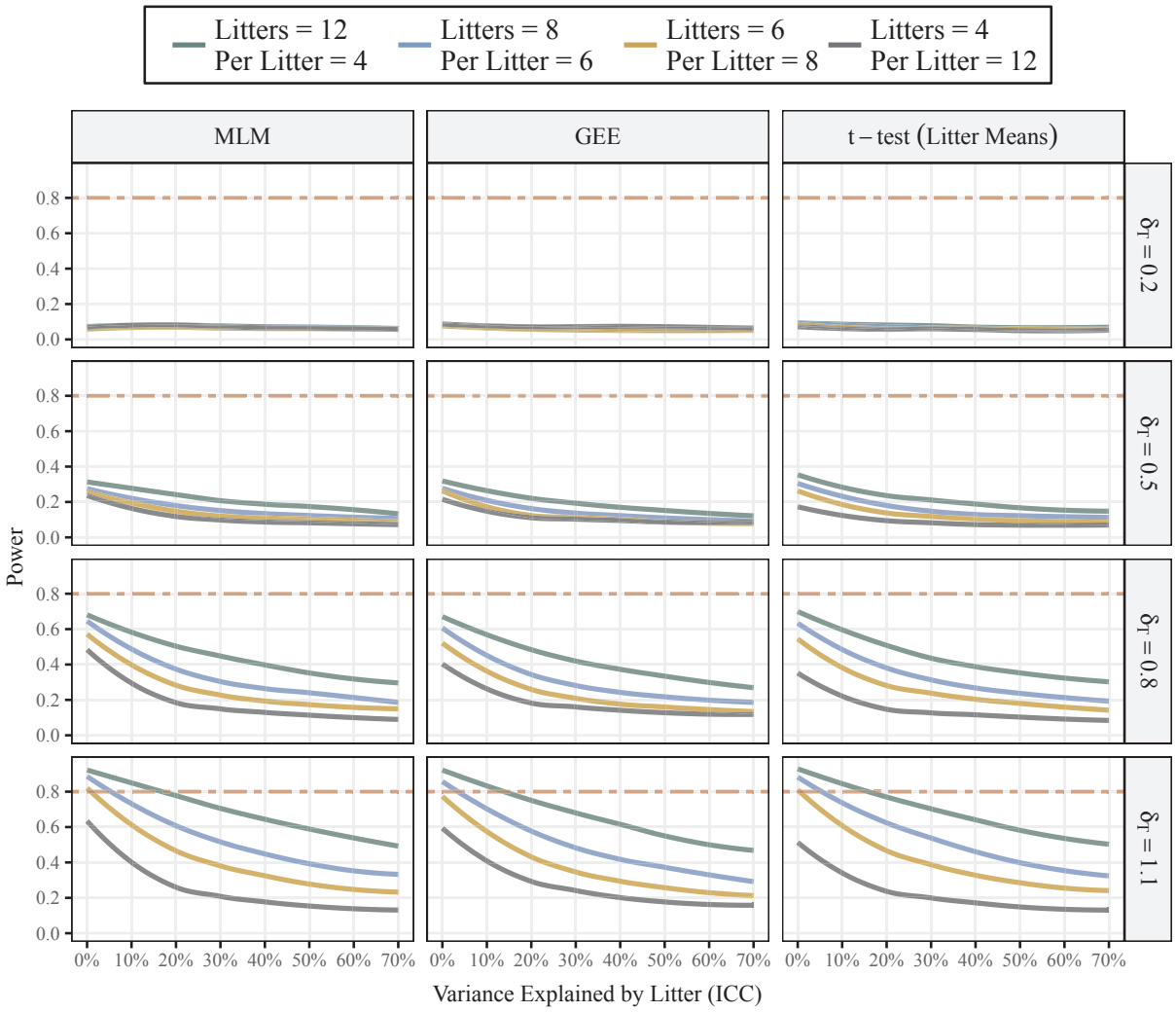

Fig. 3. Power is related to the degree of between-litter variation and sample size composition (power is higher with fewer dependent measures [green]). For these simulation conditions, power rarely reached the optimal level of 0.80. Importantly, since we used an optimistic sample size (N = 48; two-groups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

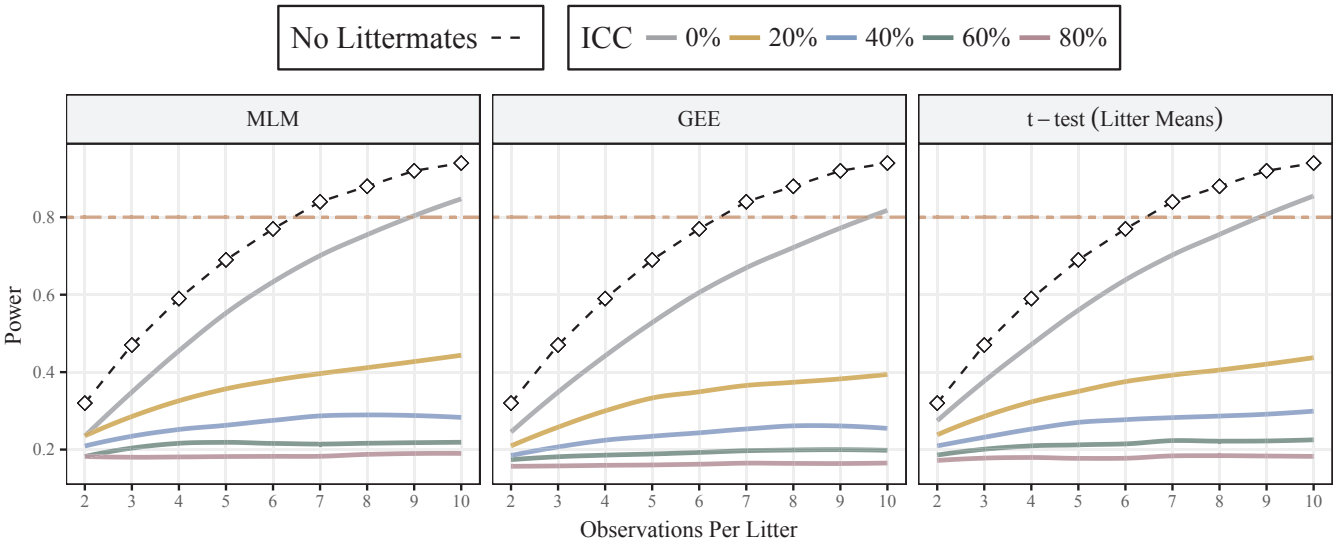

Fig. 4. Incremental power gains from additional littermates (litters = 8 and δ_T = 0.8). When observations within-litter were effectively independent (ICC = 0%) power increased with more littermates. However, power gains quickly diminished with increasing ICC values. For ICC values reported in the literature (ICC = 60%), including an additional 64 observations increased power by less than 5%. The dotted line indicates power for an independent sample *t*-test (e.g., each litter contributed one observation).

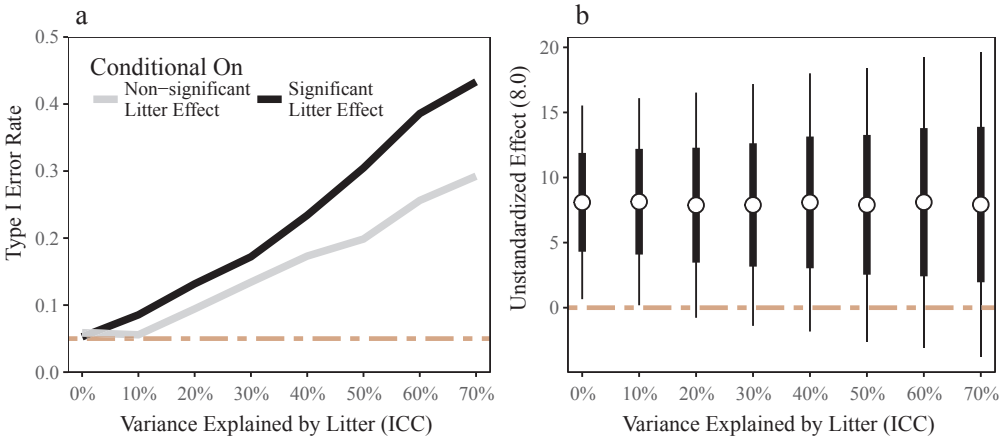

Fig. 5. (a) Error rates for the fixed effect (e.g., High LG vs. Low LG)—analyzed with a *t*-test—are not conditional on a significant litter effect. This can be thought of as: (1) testing the significance of litter in a MLM (random effect); and (2) excluding litter from the model when non-significant. 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 (litters = 8 and per litter = 4).

small sample sizes ($N = 8$), MLM and GEE had error rates above or below the expected level. In contrast, when litter means were the unit of analysis, error rates were consistently close to expectations ($\alpha = 0.05$). 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, we showed that both methods can be used to achieve nominal error rates for the fixed effect. With a total sample of eight ($N = 8$), for example, adequate performance was achieved.

4.2. Conditional false positive rates

We also showed that type I error is not dependent upon a significant litter effect (Fig. 5a). This can be thought of as mimicking a two-step procedure: (1) the significance of litter was assessed; and (2) if non-significant, litter 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 approach is to rely on knowledge of the experimental design and study subjects, regardless of the observed *p*-value. It should be noted that the idea 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 long run expectations (i.e., hypothetical replications; Greenland et al. (2016)). Accordingly, even if between-litter variance is estimated as zero, it is important to consider whether this is a reasonable expectation in future studies.

4.3. Power

Power was investigated for three methods: MLM; GEE; and analyzing litter means (*t*-test). We addressed two specific objectives: (1) how power is influenced by between-litter variation and sample composition (i.e., ratio of litters to observations per litter); and (2) incremental power from increasing the number of littermates.

We found that it is more advantageous to reduce the number of dependent observations. This was the case for all three methods. Importantly, when the ICC was 0%, the *t*-test with litter means showed consistently higher power than the MLMs and GEEs. Power was comparable between analyzing litter means and MLMs when there was some between-litter variation (ICC > 0%) or sample sizes were large (relatively). When the effect size was small ($\delta_T = 0.2$) and medium ($\delta_T = 0.5$) there was negligible power and this suggests that effects often go undetected in both literatures. This reduced power was not attributable to these statistical methods. For example, if we assume a total sample size of 48 and a small effect ($d = 0.20$), power for an independent *t*-test is 0.10. This parallels power for MLM (0.07), GEE (0.08), and litter means (0.10) with12 litters and 4 observations per litter (ICC = 0%). Thus, even when observations are independent, using a method for dependent measures does not substantially reduce power.

The second simulation showed that power gains from increasing littermates is directly related to between-litter variation. For independent observations (ICC = 0%), increasing littermates from 2 to 10 increased power considerably. However, power gains were quickly diminished with between-litter variance. The ethical considerations of this finding cannot be understated. That is, when sacrificing an additional sixty-four animals (assuming 8 litters and increasing littermates from 2 to 10), power was not substantially improved for previously reported ICC values (ICC = 60%; Lazic and Essioux, 2013). These results also show that optimal power is achieved when no littermates are included in the analysis. Importantly, this is not equivalent to an ICC value of 0%. This is a result of the degrees of freedom and corresponding critical values. In an ANOVA framework, the degrees of freedom for 12 litters and 4 observations per litter is 10 ($F_{critical} = 4.97$). In contrast, the degrees of freedom for 48 litters and 1 observation per litter is 46 ($F_{critical} = 4.05$).

Both simulations showed that power rarely reached 0.80. The implications are twofold. First, statistical significance is a ratio between signal (the effect) and noise (standard error). Small samples often produce noisy estimates, so effect needs to be very large to reach statistical significance (Walum et al., 2016). This is problematic because even a significant effect can be uninterpretable. To make this point, we selected a significant effect from a reviewed paper and computed Cohen's *d* ($d = 2.3$, 95-% *CI* = [0.42–4.17]) (Champagne et al., 2003b). Here, we can reject values less than 0.42 and greater than 4.17 which indicates that the *true* effect could be medium to unreasonably large in magnitude (Kruschke, 2013). Second, non-significant effects are also difficult to interpret. For example, assuming the same effect size and interval width centered at zero ($d = 0$, 95-% CI = [−1.88–1.88]) indicates that the *true* effect may be very large (in either direction) 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

Importantly, 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 (psychobiology). Additionally, Lazic and Essioux (2013) found that 91% of studies reviewed in the valproic acid

literature used invalid statistical methods for analyzing data with littermates. We built upon these findings 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 influences false positive rates and power. Third, we showed that ANCOVA inflates error more than a *t*-test. Fourth, as an alternative to MLM, we characterized the performance of GEE for common 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, implications were discussed in the broader context of replication efforts in related fields.

4.5. Implications: Natural occurring maternal variation and prenatal stress

The naturally occurring maternal variation 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, the present findings highlight areas for improvement. 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 strong evidence for specific effects. That is, in the general sense, maternal care probably does influence offspring development. However, stating that maternal care can reliably induce gene-specific epigenetic modifications is not currently supported by the literature. Evidence for this can only be obtained through replication, in addition to using appropriate statistical methods.

To be clear, we are not suggesting all previous studies that failed to report or account for between-litter variation lack scientific value. In fact, we see previous studies as providing a foundation from which to build future research. There is a wealth of findings in both literatures and these provide clear hypotheses to be evaluated going forward. In addition, revisiting past data (where possible) can serve many purposes. These data can be reanalyzed with the methods presented here, and the results made publicly available or published. For the latter, the journal *Meta Psychology* (to be released: https://osf.io/bktc7/) allows for re-evaluating past findings with new methods. Assuming litter information is available, we also view past studies as providing a rich resource for all research areas that use rodent models. For example, estimates of between-litter variance can be systematically quantified and made publicly available. Where data cannot be revisited, these findings should not be automatically discounted. This would underappreciate the limitations of the present paper (*4.10. Limitations*) and ultimately be counterproductive.

4.6. Implications: Reproducible science

The replication crisis has so far been dominated by human oriented psychology in general, and social psychology in particular (OCS, 2015). Yet, other research areas are also experiencing difficulties replicating findings including biomedical related fields (< 25%; Prinz et al., 2011). We showed that using inappropriate statistical methods 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. Nevertheless, the take home for replication efforts in neuroendocrinology is that we need not exclusively focus on biases or ill-intent (e.g., *p*-hacking) on the part of individual researchers. It is entirely plausible that misspecified statistical models will account for many failed replications, in that the original effect was possibly non-significant. Our simulations also showed that power depends on the degree of between-litter variation (Figs. 3 and 4). In other words, detecting a significant effect would prove difficult if the magnitude of between-litter variation was larger in a replication attempt than the original study. Importantly, in studies where the ICC of litter was 0%, the originally reported *p*-value would not change by accounting for between-litter variance.

Moreover, addressing issues surrounding reproducible science requires greater action than improving methodological practices of individual researchers. For example, rodents are often purchased from vendors in which litter information is not always readily available. Standard ordering practice should allow for selecting rodents with consideration for litter of origin. In both literatures, we found that aspects of litter were underreported but this is not necessarily attributable to the study authors. Journals have different guidelines (e.g., some allow for minimal description of statistical methods) and research fields often differ in tradition. Vendors, statistical reporting guidelines, and "cultural" norms may all work in concert to exacerbate the effects of between-litter variance on the literature; addressing these issues will take concerted effort at many levels. Importantly, these issues are not restricted to these two areas (Lazic and Essioux, 2013; Zorrilla, 1997). It is possible that litter effects are adversely impacting most (maybe all) research areas that use rodent models.

4.7. MLM vs. GEE and litter means

In contrast to MLMs, GEEs are less documented in R (Bates et al., 2015c; 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 (Fay and Graubard, 2001; Pan and Wall, 2002). While GEEs can only consider one source of variation, MLMs offer greater flexibility and provide information for prospective power analyses (estimates of between-litter variance). This flexibility comes with a cost, however, as a misspecified MLM can substantially inflate error rates. For example, when sex differences are examined with multiple animals from the same litter, variability in sex differences must be considered with a random slope, in addition to a random intercept of litter (Aarts et al., 2015). In small sample situations, this presents challenges in a maximum likelihood (or restricted maximum likelihood) 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).

MLM was comparable to analyzing litter means with a *t*-test. For some simulation conditions (ICC = 0%), however, power was higher when analyzing litter means. Importantly, when some between-litter variance was present (ICC > 0%) power was almost identical and in some cases MLM had more power. It should be noted that MLM provides richer information for inference than analyzing litter means. For example, estimates of between-litter variance that are essential for prospective power analyses. MLM also allows for answering many research questions: one could examine whether the fixed effect (high vs low LG; prenatal stress vs. control) explains between-litter variance. We see this as especially important for the natural variations in maternal care literature, because there is explicit interest in whether pups from the same dam resemble one another (example provided in the *litterEffects* tutorial). This analysis is not possible with litter means (*t*-test).

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 intercept will give equivalent estimates to a fixed effect only model (Gelman and Hill, 2006). These questions are challenging and demonstrate the difficulty in modeling hierarchical data structures such as those that include littermates. 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 variance untenable and that researchers should always use a statistical method for dependent observations.

4.9. Statistical expertise vs. improved training

Certain research questions will inevitably require seeking statistical expertise. However, expertise is also qualified with being highly specialized in a certain area. 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 another's field and specific research question. This would entail communication of the necessary information so that the correct analysis is applied. In contrast, improved quantitative training for individual research could address many of these issues. There are many resources available for individual researchers. While in the past fitting multilevel models 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.

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, reporting deficiencies, and inflated false positive rates (Fig. 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 the assumed values for between-litter variance are plausible. We think they are, but it should be noted that precise values could not be obtained from either literature. To address this limitation, we assumed a range of values that spanned from 0% to 70%. Additionally, for the ANCOVAs, we coded litter as a categorical covariate. This was an assumption we made, because coding schemes were not reported in the primary studies. We feel this decision was justified, because continuous coding would not make sense for arbitrarily assigned numbers. Although we carried out an extensive search, we were not able to obtain exact estimates for litters or number of littermates included in an analysis. Reporting was too variable, and stating likely values may be misleading. With these limitations in mind, exact false positive rates cannot be determined. Our simulations showed that between-litter variation can increase false positive rates by some degree. This is important to consider when inferring meaning from the extant literature and for planning future studies.

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 statistical quantities have meaning in the long run (hypothetical replications) and this makes exploring expected error rates with one data set impossible (Greenland et al., 2016).

It is also possible that estimates of between-litter variance were biased in our simulations (Maas and Hox, 2005), which may have influenced type I error rates and power. Simulation studies have indicated that small samples and few clusters (e.g., litters) can present challenges for MLM (Maas and Hox, 2005) and GEE (Gunsolley et al., 1995). Here it has been noted that power is low to detect cluster differences (e.g., litter effects) and that estimates of litter variance can be variable. These findings do not suggest that litter should be ignored (excluded from the model). We emphasize that litter must be considered to account for dependent measures and that, with only four litters (Figs. 2 and 3), expected error rates and optimal power can be achieved. Thus, even in small sample situations, MLM and GEE can certainly be used to analyze dependent data.

Our results are restricted to specific research designs (Fig. 1). This limits the generalizability of our findings. We view this as a strength in that specific questions were answered and not general quantitative practices. Additionally, we provided resources for important research topics in the field of behavioral neuroendocrinology. We also used two different search strategies for each literature. The 100 most recent studies were reviewed for the prenatal stress literature, whereas all studies (found in the search) were reviewed in the maternal care literature. This decision was made because including 1000's of studies seemed unnecessary to achieve our goal of offering recommendations based on current methodological practices.

4.11. Recommendations

We conclude that between-litter variation is underappreciated (Table 1), can lead to increased false positive rates (Fig. 2), and reduce one's ability to detect a true effect (Figs. 3 and 4). Based on the strength of these findings, we provide several recommendations. First, it should be noted that these recommendations apply to research designs in which entire littermates are categorized the same way (Fig. 1). Litters often receive multiple treatments or sex differences are of interest. A thorough discussion of the necessary model in these situations is beyond the scope of the present paper (but see here: Aarts et al., 2015). However, when all littermates included in a study are categorized based on the same characteristic or treatment (Fig. 1), we recommend the following:

(1) Independence of observations from the same litter should never be assumed.
(2) Including litter as a covariate in an ANCOVA is not appropriate, since it does not account for dependent measures.
(3) A statistical method specifically for dependent measures is necessary:
    We prefer a multilevel approach. Importantly, estimates of between-litter variance are needed to conduct power analyses. GEEs and analyzing litter means with a *t*-test cannot provide this information. Furthermore, GEEs require small sample for corrections for *p*-values that require referencing relevant literature or using simulations to obtain the appropriate correction. We thus reemphasize our preference for multilevel models.
(4) When using a MLM, the random effect of litter should always be included in the model (irrespective of statistical significance).
(5) To facilitate accurate prospective power calculations future papers must report measures of between-litter variance (only possible with MLM):

We also recommend that researchers revisit previous studies (if possible) and compute between-litter variance (see *litterEffects* package). Establishing a public repository with this information will provide a valuable resource for researchers in these fields, as well as related fields.

(6) Prospective power calculations should assume a range of plausible values for between-litter variance:

Of course, knowing exact values for between-litter variation is not possible. This is like any power calculation in which one assumes a value for the effect size of interest (which is unknown). As such, assuming a range of values will provide richer information that allows for assessing realistic expectations.

(7) Between-litter variation has ethical implications that must be considered going forward.

Acknowledgement

We thank two anonymous reviewers that substantially improved the quality of this paper. 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. DRW is supported by a McNair Scholars graduate fellowship.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.yfrne.2017.08.003.

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. http://dx.doi.org/10.1186/s12868-015-0228-5.

Agnish, N.D., Keller, K.a., 1997. The rationale for culling of rodent litters. Fundam. Appl. Toxicol. 38, 2–6. http://dx.doi.org/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, 1–16. http://dx.doi.org/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, 193–248. http://dx.doi.org/10.1002/dev.420040302.

Baier, C.J., Pallarés, 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, 435–445. http://dx.doi.org/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, 151–164. http://dx.doi.org/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, 939–950. http://dx.doi.org/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, 255–278. http://dx.doi.org/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., Mächler, M., Bolker, B., Walker, S., 2014. Fitting linear mixed-effects models using lme4. eprint arXiv:1406.5823 67, 51. doi: http://dx.doi.org/10.18637/jss.v067.i01.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., Dai, B., Grothendieck, G., 2015b. lme4: Linear Mixed-Effects Models using "Eigen" and S4.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015c. Fitting linear mixed-effects models Using lme4. J. Stat. Softw. http://dx.doi.org/10.18637/jss.v067.i01.

Beery, A.K., Francis, D.D., 2011. Adaptive significance of natural variations in maternal care in rats: A translational perspective. Neurosci. Biobehav. Rev. 35, 1552–1561. http://dx.doi.org/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. http://dx.doi.org/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, 163–165. http://dx.doi.org/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, 1234–1238. http://dx.doi.org/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, 55–68. http://dx.doi.org/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., Munafò, M.R., 2013. Power failure: Why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. http://dx.doi.org/10.1038/nrn3475.

Caldji, C., Tannenbaum, B., Sharma, S., Francis, D., Plotsky, P.M., Meaney, M.J., 1998. Maternal care during infancy regulates the development of neural. Proc. Natl. Acad. Sci. USA 95, 5335–5340.

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, 1003–1017. http://dx.doi.org/10.1080/00949650412331321070.

Cameron, N.M., 2011. Maternal programming of reproductive function and behavior in the female rat. Front. Evolution. Neurosci. http://dx.doi.org/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, 325–334. http://dx.doi.org/10.1016/j.physbeh.2007.03.014.

Champagne, F.A., Francis, D.D., Mar, A., Meaney, M.J., 2003a. Variations in maternal care in the rat as a mediating influence for the effects of environment on development. Physiol. Behav. 79, 359–371. http://dx.doi.org/10.1016/S0031-9384(03)00149-5.

Champagne, F.A., Weaver, I.C.G., Diorio, J., Sharma, S., Meaney, M.J., 2003b. Natural variations in maternal care are associated with estrogen receptor alpha expression and estrogen sensitivity in the medial preoptic area. Endocrinology 144, 4720–4724. http://dx.doi.org/10.1210/en.2003-0564.

Chatterjee, A., Chatterjee, R., 2009. How stress affects female reproduction: An overview. Biomed. Res. 20, 79–83.

Cox, D.R., McCullagh, P., 1982. A biometrics invited paper with discussion. Some aspects of analysis of covariance. Biometrics 38, 541–561. http://dx.doi.org/10.2307/2530040.

Crabbe, J.C., Wahlsten, D., Dudek, B.C., 1999. Genetics of mouse behavior: Interactions with laboratory environment. Science (80-) 284, 1670–1672. http://dx.doi.org/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, 297–302. http://dx.doi.org/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, 52–66. http://dx.doi.org/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 pup's behavior. Psychol. Rep. 40, 411–420. http://dx.doi.org/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, 708–716. http://dx.doi.org/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. USA 113, 7900–7905. http://dx.doi.org/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, 1198–1206. http://dx.doi.org/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, 167–180. http://dx.doi.org/10.1196/annals.1330.011.

Fowden, A.L., Forhead, A.J., 2004. Endocrine mechanisms of intrauterine programming. Reproduction 127, 515–526. http://dx.doi.org/10.1530/rep.1.00033.

Francis, D.D., Meaney, M.J., 1999. Maternal care and the development of stress responses. Curr. Opin. Neurobiol. 9, 128–134. http://dx.doi.org/10.1016/S0959-4388(99)80016-6.

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

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York. http://dx.doi.org/10.1017/CBO9780511790942.

Glowa, J.R., Hansen, C.T., 1994. Differences in response to an acoustic startle stimulus among forty-six rat strains. Behav. Genet. 24, 79–84. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.yfrne.2013.12.001.

Gore, A.C., 2008. Developmental programming and endocrine disruptor effects on reproductive neuroendocrine systems. Front. Neuroendocrinol. 29, 358–374. http://dx.doi.org/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, 337–350. http://dx.doi.org/10.1007/s10654-016-0149-3.

Gudsnuk, K., Champagne, F.A., 2012. Epigenetic influence of stress and the social environment. ILAR J. 53, 279–288. http://dx.doi.org/10.1093/ilar.53.3-4.279.

Gunsolley, J.C., Getchell, C., Chinchilli, V.M., 1995. Small sample characteristics of generalized estimating equations. Commun. Statist.-Simul. Comput. 24, 869–878.

http://dx.doi.org/10.1080/03610919508813280.
Halekoh, U., Højsgaard, 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, 364–375. http://dx.doi.org/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, 1112–1120. http://dx.doi.org/10.1038/mp.2009.55.
Hedges, L.V., 2007. Effect sizes in cluster-randomized designs. J. Educ. Behav. Stat. 32, 341–370. http://dx.doi.org/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, 2495–2510. http://dx.doi.org/10.1098/rstb.2012.0223.
Hofer, M.A., 1973. Maternal separation affects infant rats' behavior. Behav. Biol. 9, 629–633. http://dx.doi.org/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, 221–228. http://dx.doi.org/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, 213–222. http://dx.doi.org/10.1002/(SICI)1097-0258(19990130)18:2<213::AID-SIM999>3.0.CO;2-E.
Hubbard, A.E., Ahern, J., Fleischer, N.L., Van der Laan, M., Lippman, S.A., Jewell, N., Bruckner, T., Satariano, W.A., 2010. To GEE or Not to GEE. Epidemiology 21, 467–474. http://dx.doi.org/10.1097/EDE.0b013e3181caeb90.
Inoue-Murayama, M., 2009. Genetic polymorphism as a background of animal behavior. Anim. Sci. J. http://dx.doi.org/10.1111/j.1740-0929.2008.00623.x.
Jacobson, L., Sapolsky, R., 1991. The role of the hippocampus in feedback regulation of the hypothalamic-pituitary-adrenocortical axis. Endocr. Rev. 12, 118–134. http://dx.doi.org/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, 577–592. http://dx.doi.org/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, 601–606. http://dx.doi.org/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, 31–44. http://dx.doi.org/10.1016/j.poly.2005.06.060.
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, 350–386. http://dx.doi.org/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, 95–132. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.neubiorev.2008.06.004.
Koehl, M., Darnaudéry, M., Dulluc, J., Van Reeth, O., Le Moal, M., 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, 302–315. http://dx.doi.org/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, 573–603. http://dx.doi.org/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. http://dx.doi.org/10.1177/1948550617697177.
Lazic, S.E., 2010. The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis? BMC Neurosci. 11, 5. http://dx.doi.org/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. http://dx.doi.org/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, 835–840. http://dx.doi.org/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, 281–296. http://dx.doi.org/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, 1659–1662. http://dx.doi.org/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, 434–445. http://dx.doi.org/10.1038/nrn2639.
Maas, C.J.M., Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. Methodology 1, 86–92. http://dx.doi.org/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 06.07.17).
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, 286–303. http://dx.doi.org/10.1007/s10519-016-9782-6.
Markham, J.A., Koenig, J.I., 2011. Prenatal stress: role in psychotic and depressive diseases. Psychopharmacology. http://dx.doi.org/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, 55–61. http://dx.doi.org/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–S15. http://dx.doi.org/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. http://dx.doi.org/10.1371/journal.pone.0014739.
McGrady, A.V., 1984. Effects of psychological stress on male reproduction: A review. Arch. Androl. 13, 1–7. http://dx.doi.org/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, 297–308. http://dx.doi.org/10.1016/j.neuroscience.2004.08.009.
Mendi, M., 1988. The effects of litter size variation on mother-offspring relationships and behavioural and physical development in several mammalian species (principally rodents). J. Zool. 215, 15–34. http://dx.doi.org/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, 9055–9065. http://dx.doi.org/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, 64–87. http://dx.doi.org/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, 334–339. http://dx.doi.org/10.1016/j.physbeh.2011.02.032.
O'Connor, T.M., O'Halloran, D.J., Shanahan, F., 2000. The stress response and the hypothalamic-pituitary-adrenal axis: From molecule to melancholia. QJM 93, 323–333. http://dx.doi.org/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, 1429–1441. http://dx.doi.org/10.1002/sim.1142.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects models in S and S+. Springer Science & Business Media. doi: http://dx.doi.org/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. 10http://dx.doi.org/10.1038/nrd3439-c1. 712-712.
R Core Team, 2016. R: A Language and Environment for Statistical Computing.
Ramos, A., Berton, O., Mormède, P., Chaouloff, F., 1997. A multiple-test study of anxiety-related behaviours in six inbred rat strains. Behav. Brain Res. 85, 57–69. http://dx.doi.org/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, 588–594. http://dx.doi.org/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, 65–76. http://dx.doi.org/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. http://dx.doi.org/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, 123–129. http://dx.doi.org/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, 656–662. http://dx.doi.org/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, 479–488. http://dx.doi.org/10.1038/ncpendmet0515.
Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality(complete samples). Biometrica 52, 591–611. http://dx.doi.org/10.2307/1267427.
Starr-Phillips, E.J., Beery, A.K., 2014. Natural variation in maternal care shapes adult social behavior in rats. Dev. Psychobiol. 56, 1017–1026. http://dx.doi.org/10.1002/dev.21182.
Szuran, T.F., Pliška, V., Pokorny, J., Welzl, H., 2000. Prenatal stress in rats: Effects on plasma corticosterone, hippocampal glucocorticoid receptors, and maze

performance. Physiol. Behav. 71, 353–362. http://dx.doi.org/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, 147–153.
Turecki, G., Meaney, M.J., 2016. Effects of the social environment and stress on glucocorticoid receptor gene methylation: a systematic review. Biol. Psychiat. http://dx.doi.org/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, 2–12. http://dx.doi.org/10.1016/j.jesp.2016.03.004.
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. http://dx.doi.org/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, 16364–16369. http://dx.doi.org/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. Psychiat. 79, 251–257. http://dx.doi.org/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, 118–124. http://dx.doi.org/10.1016/j.jesp.2017.04.011.
Weinstock, M., 2017. Prenatal stressors in rodents: Effects on behavior. Neurobiol. Stress 6, 3–13. http://dx.doi.org/10.1016/j.ynstr.2016.08.004.
Weinstock, M., 2008. The long-term behavioural consequences of prenatal stress. Neurosci. Biobehav. Rev. 32, 1073–1086. http://dx.doi.org/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, 279–288. http://dx.doi.org/10.1016/j.bbr.2012.10.003.
Zorrilla, E.P., 1997. Multiparous species present problems (and possibilities) to developmentalists. Dev. Psychobiol. 30, 141–150. http://dx.doi.org/10.1002/(SICI)1098-2302(199703)30:2<141::AID-DEV5>3.0.CO;2-Q [pii].