Hormone Ratios Suffer from Striking Lack of Robustness to Measurement Error

Marco Del Giudice

Steven W. Gangestad

***Psychoneuroendocrinology, 142,* 105802 (2022).**

Department of Psychology, University of New Mexico, Albuquerque, NM USA
Address correspondence to Marco Del Giudice, Department of Psychology, MSC 2022, 1 University of New Mexico, Albuquerque, NM 87111 USA. Email: marcodg@unm.edu

Abstract

Hormone ratios are often used to capture the joint effect (or "balance") of two hormones with opposing or mutually suppressive effects. Despite some statistical and interpretative problems, hormone ratios are being increasingly used to examine associations of testosterone/cortisol, estradiol/progesterone, testosterone/estradiol, and other hormone pairs. Here we discuss a methodological problem that has not been previously recognized, namely, the striking lack of robustness of raw hormone ratios in the face of measurement error. Hormone levels are measured with error, both due to inability of assays to perfectly assess concentrations "in the tube" and due to discrepancies between levels at the time of sample collection and effective levels that produce the physiological and/or behavioral effect of interest. Noise in measured hormone levels can be substantially exaggerated by ratios, especially when the distribution of the hormone at the denominator is positively skewed, as is frequently observed. To evaluate the extent of this problem and explore the conditions that exacerbate it, we present two sets of simulations, one using idealized distributions and one using empirically observed distributions from studies of estrogen and progesterone. Results show that the validity of raw hormone ratios—the correlation between measured levels and underlying effective levels—drops rapidly in the presence of realistic levels of measurement error. Log-ratios are much more robust to measurement error, and their validity is more stable across samples; under some conditions (e.g., moderate amounts of noise with positively correlated hormone levels), they may provide a more valid measurement of the underlying raw ratio than the measured raw ratio itself. These findings have important implications for research that uses hormone ratios as predictors.

Keywords: hormone ratios; log-ratios; measurement error; robustness

1 Introduction

Endocrine systems do not operate completely independent of one another to affect physiology and behavior. In some instances, hormones may have opposing effects on the same outcome, at times due to one hormone impacting the effects of another hormone. For instance, estradiol and progesterone levels may each affect the impacts of the other hormone, with potential interactive effects on behavior (e.g. Eisenbruch et al., 2016). Whereas estradiol may stimulate the proliferation of progesterone receptors (e.g., Wooley & McEwen, 1993), progesterone inhibits the impact of estradiol by reducing receptor densities (e.g., Hseuh et al., 1975; Selcer & Leavitt, 1988). Or, consider testosterone and cortisol, each the end product of their respective axes: testosterone of the hypothalamic-pituitary-gonadal (HPG) axis, and cortisol of the hypothalamic-pituitary-adrenal (HPA) axis. Each may affect events that occur within the other axis (Viau, 2002). For instance, cortisol can decrease pituitary sensitivity to gonadotropins (Tilbrook et al., 2000) as well as act directly upon the gonads to inhibit their endocrine functioning (Johnson et al., 1992), such that, in the presence of cortisol, the HPG axis in men is downregulated. (For a contrasting view of how testosterone and cortisol interact synergistically, see Ketterson & Nolan, 1999.) Some scholars have similarly argued that, in men, the testosterone-estradiol balance predicts outcomes (e.g., sexual desire, cerebrovascular disease) over and above the simple additive impacts of testosterone and estradiol (e.g., Gong et al., 2013). Oxytocin and testosterone too have been conjectured to have opposing effects on reactive aggression and social sensitivity (e.g., Crespi, 2016); each may modulate the impacts of the other, even if the ways they do so are not yet fully understood (e.g., Fragkaki et al., 2018).

A popular way to consider the joint impact of two hormones with opposing effects, some of which are due to one hormone effecting the impact of another hormone, is to calculate a *hormone ratio* and treat it as a predictor of the relevant outcomes. This is especially true when researchers assume that the "balance" between the two hormones importantly reflects their combined actions. Hence, for instance, if the effects of testosterone are suppressed by the presence of cortisol, the testosterone/cortisol ratio may be viewed as a reasonable index of the action of testosterone, taking into account the suppressive effects of cortisol. For example, Terburg et al. (2009) proposed that this ratio may be a hormonal marker of criminal social aggression. Similarly, Roney (2019) argued that the estradiol/progesterone ratio indexes the joint effects of these hormones, as their levels changes across women's ovarian cycles. While the actual processes through which estradiol and progesterone affect neural states and behavior may be quite complex, Roney proposed that the "raw EP ratio is a good index of the outcomes of complex, temporal sequences of hormonal influences" (p. 528) and thereby can usefully be used as a hormonal summary variable in research. Citing Roney (2019), Stern et al. (2021) subsequently used, in primary analyses, the E/P ratio to predict women's sexual desires and preferences for men's muscularity. (For numerous other examples of hormone ratios used in research, see Sollberg & Ehlert, 2016.)

1.1 Previous criticism of hormone ratios

Even though they are popular with researchers, hormone ratios have been criticized on methodological grounds, most notably in a paper by Sollberg and Ehlert (2016). These authors noted a number of statistical and interpretational concerns with using ratios, which are widely recognized in other fields but insufficiently discussed in the neuroendocrine literature. First, distributions of ratios tend to be highly skewed and leptokurtic, with marked outliers. This is true even when the component hormones are normally distributed, and particularly when the coefficient of variation (the standard deviation divided by the mean) of the denominator is relatively large (e.g., Atchley, 1976). This implies the presence of comparatively small denominator values, and as denominator values approach zero, values of the ratio increase exponentially. Second, the ratio A/B is not linearly related to B/A. Hence, results of correlational or regression analysis using ratios will vary depending on whether A/B or B/A is used. In many instances, it is not clear why hormonal balance is better represented as A/B or, alternatively, B/A. Though the choice of one over the other may appear arbitrary, rarely do researchers justify their choice. Third, results using ratios as predictors of outcomes may not be readily understood. Several different possible underlying associations

could give rise to a simple association between a ratio and an outcome: (a) the observed association is driven solely by one of the hormones in the ratio; (b) the observed association is driven solely by additive effects of the two hormones on the ratio; (c) the observed association is driven, at least in part, by statistical interactions between the two hormones. In this last case, interactions may be relatively simple—e.g., a linear × linear interaction—or they may be more complex. In the case of the estradiol/progesterone ratio, it appears that less than half of the total variance can be accounted for by linear main effects and the linear × linear interaction, meaning that most of the variance is likely due to more complex interactions, albeit of unknown (or unspecified) forms (Gangestad et al., 2019). Thus, using ratios may end up obscuring the neurobiological mechanisms that generate the observed empirical associations.

As noted by Sollberger and Ehlert (2016), log-transforming hormone ratios provides a simple alternative that obviates some of the problems with raw ratios. The log of a ratio is simply the difference between the logged components; that is, $\ln(A/B) = \ln(A) - \ln(B)$. Hence, a logged ratio captures equal additive but opposing effects of two hormones, each log-transformed. As distributions of hormone levels often approximate the log-normal rather than normal distribution (e.g., Kletsky et al., 1975), log-transformation of hormone levels often results in near-normal distributions (Sollberger & Ehlert, 2016; see also Gelman, 2019). Furthermore, $\ln(A/B) = -\ln(B/A)$. Hence, results do not hinge on which is used; associations will be identical in magnitude, though different in sign. Of course, log-transformed ratios may still not address the question of what drives an association, one hormone or both. Moreover, logged ratios do not capture interactive effects; they reflect purely additive effects of two logged hormones, constrained to be opposite in sign and equal in magnitude. Hence, another alternative (used as a follow-up to use of logged ratios or as a standalone approach) is to enter the raw or log-transformed levels of each hormone as separate predictors, along with the linear × linear interaction between the two (e.g., Sollberger & Ehlert, 2016).

Despite these concerns, many studies continue to implement the use of hormone ratios. For instance, a Web of Science search (August 19, 2021) yielded 168 published papers with "testosterone-cortisol ratio" in the title, abstract, or keywords. Of these, 60 (36%) appeared just since 2017; more papers appeared in years 2017-2020 than any prior year, with one exception (2013). A total of 131 papers with "testosterone-estradiol ratio" or "estradiol-testosterone ratio," with 49 (37%) appearing since 2017.[1] There are a number of ways in which researchers may defend the use of a hormone ratio. For example, Roney (2019) argued that the raw estradiol/progesterone (E/P) ratio is a good index of the joint hormonal effects of E and P on brain states and behavior, partly based on the association between E/P and conceptive status (probability of conceiving an offspring, given unprotected sex) across the cycle. As noted above, the observation that $A/B \neq B/A$ raises the question of how one decides which ratio better reflects hormonal "balance". Roney's (2019) response is that one should look at which ratio (E/P or P/E) better indexes conceptive status. In a set of archival data, the ratio [mean E/mean P] was associated much more strongly with estimated conceptive status than the ratio [mean P/mean E]. According to his argument, E/P *should* also be expected to predict behavior more strongly than P/E. Likewise, the finding that the raw E/P ratio was more strongly associated with conception status than ln(E/P) may make it a better proxy of the underlying hormonal changes, regardless of which distribution is closer to normal (Roney, 2019).

1.2 The present paper

In this paper, we discuss a limitation of raw hormone ratios that to our knowledge has not been recognized before, namely, their striking lack of robustness to measurement error. In the presence of even moderate amounts of noise, raw hormone ratios rapidly lose validity, in the sense that the measured ratio between the two hormones becomes less and less correlated to their underlying ratio. The effect becomes more dramatic when the distribution of the hormone at the denominator is highly skewed, because a high frequency of small values at the denominator amplifies the impact of error. As we detail below, log-ratios are

[1] These counts are illustrative, and undoubtedly miss many relevant papers that do not use these precise terms to refer to the hormone ratio.

remarkably robust to error compared with raw ratios. Not only do measured log-ratios correlate more strongly with the underlying log-ratios even in presence of large errors; under certain conditions, the measured log-ratio may be a better indicator of the underlying *raw* ratio than the measured raw ratio itself.

In the following sections, we first briefly discuss the various sources of error that add noise to hormone measurements. We then present two sets of simulations—one based on idealized distributions and one based on empirical data from studies of estradiol and progesterone—that examine the validity of raw hormone ratios at different levels of measurement error and compare it with the validity of log-ratios. We conclude with some reflections on the state of the literature and offer some advice for researchers.

2 Sources of Error in Hormone Measurements

The "classic" sources of error that endocrine researchers are used to consider are those due to the sensitivity of hormone assays. Intra-assay CVs reflect variation in measured levels within subsamples taken from the same sample (e.g., tube of saliva); inter-assay CVs reflect variation in measured levels of samples with precisely the same concentration of hormone, when measurement is performed using different assay plates. As implied by the use of CVs to quantify them, the expected magnitude of these errors is roughly proportional to the concentration of the hormone, with larger errors at higher concentrations. However, there are other sources of errors that are likely to disproportionately affect measurements at *low* hormone concentrations. For example, assay cross-reactivity with non-target hormones (e.g., other steroids) becomes more of an issue when the target hormone is present at a low concentration, especially if non-target hormones are much more abundant in the sample (and hence can add significant noise to the measurement even if cross-reactivity is weak in absolute terms). More generally, the signal-to-noise ratio decreases at low concentrations, thus disproportionately increasing the uncertainty associated with small hormone values (e.g., Welker et al., 2016; Prasad et al., 2019).

All the sources of error discussed so far are relative to "what's in the tube"—the concentration of hormone present in the sample. But of course, each individual biological sample (e.g., saliva, serum) is just a noisy proxy for the actual hormone concentration that produces the physiological effects of interest (henceforth *effective level*). For instance, many hormones, including testosterone (e.g., Beaven et al., 2010), estradiol and progesterone (e.g., Rossmanith et al., 1990), cortisol (e.g., Young et al., 2004), oxytocin (e.g., Baskaran et al., 2017), and LH and FSH (e.g., Stamatiades et al., 2018) are secreted in a pulsatile fashion and show substantial fluctuations, not just over the hours of a day but also in the span of a few minutes. Figure 1 illustrates this crucial concept by showing changes of hormone levels over time within individual participants, measured at a high temporal resolution. Even the short-term physiological effects of a hormone are a function of its average levels over a certain span of time; hormone levels in individual samples do not directly reflect this average, but depend on when the sample was collected relative to pulses and other short-lived fluctuations. As can be seen in Figure 1, the magnitude of these fluctuations can be quite large, easily exceeding the classic sources of error "in the tube" that are quantified by intra- and inter-assay CVs.

This problem becomes especially severe when the timing of causal events involving hormones and the collection of samples differ by hours or, potentially, days. Jones et al. (2018), Roney and Simmons (2013), and Righetti et al. (2020) collected saliva samples on women followed longitudinally once a day, and asked women to report events of that day (e.g., levels of sexual desire). In these cases, the effective level of the hormones is the concentration experienced several hours or even days before the collection of the saliva sample (depending on the time required for the physiological and behavioral effects to unfold; see, e.g., Roney & Simmons, 2013); while the measured level of a hormone in the sample is going to covary with its effective level (i.e., the underlying level of interest), it will only do so imperfectly and with considerable error. Other times, researchers may be interested in rapid hormonal effects over much shorter timescales; in such cases, the problem of fluctuations becomes less acute, even though it is unlikely to disappear. Notably, the use of salivary assays as convenient, but indirect (and imperfect) measures of hormone levels in the serum

(see, e.g., Granger et al., 2004; Shirtcliff et al., 2000) introduces one more degree of separation between the effective and measured levels of a hormone, and hence increases the total amount of error.

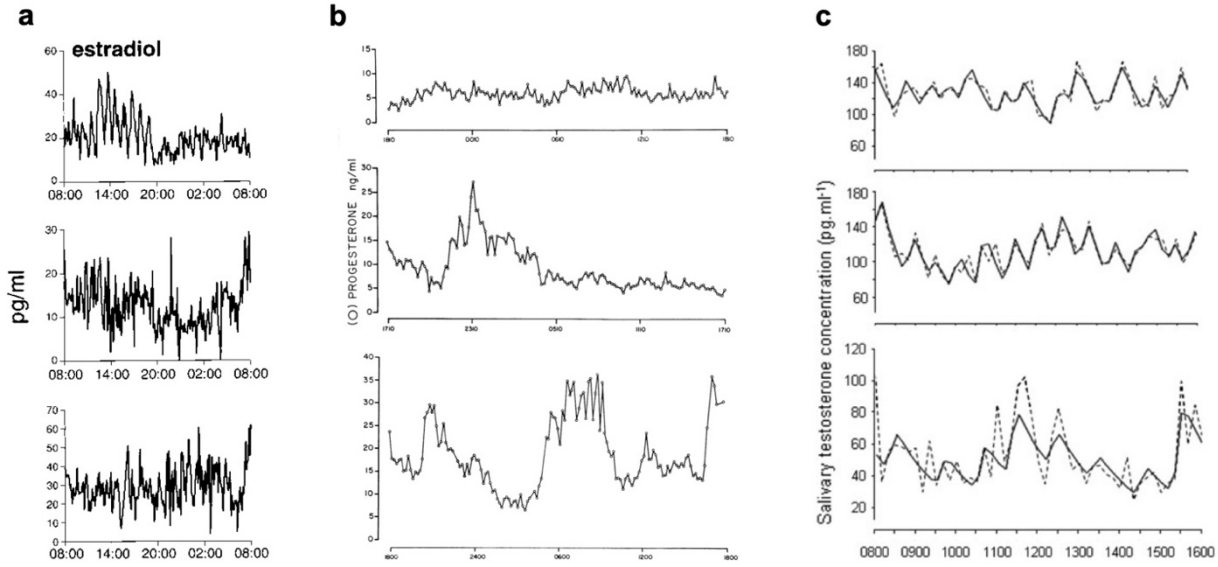

Fig. 1. Time course of frequently sampled hormone levels in individual participants. (a) Estradiol in women over a 24-hour period, sampled at 7-minute intervals (Licinio et al., 1998). (b) Progesterone in women over a 24-hour period, sampled at 10-minute intervals (Filicori et al., 1984). (c) Testosterone in men over an 8-hour period, sampled at 10-minute intervals (Beaven et al., 2010). Each panel shows the first three participants displayed in the respective papers. Reproduced and modified with permission from the original figures.

Recently, scholars have raised concerns about measurement error in immunoassays of a variety of hormones. At times, salivary concentrations measured using enzyme immunoassays are found to covary only weakly with the standard reference method of liquid chromatography tandem mass spectroscopy (especially with regard to testosterone and estradiol, less so with cortisol; e.g., Welker et al., 2016; Prasad et al., 2019; Stern et al., 2019; Schultheiss et al., 2019). Measurement errors, then, may be especially large when enzyme immunoassays are used to measure hormone levels. At the same time, use of mass spectroscopy does not address all the relevant sources of error. Even if these methods very accurately measure concentrations "in the tube," the effective levels of hormones of interest will in many instances covary imperfectly with measured levels.

2.1 Implications for ratios and log-ratios

The validity of a ratio—how strongly the measured ratio in a sample correlates with the underlying ratio—will be affected by the validities of both the individual measures going into the ratio, but not in a straightforward way. Errors of measurement in the denominator hormone can be exaggerated in the ratio, in particular when the true level in the denominator is small. Consider, for instance, the case in which an error leads the measured level of the denominator to be double the true level, not because the error is so large but because the true level is small. The error will then lead the ratio to be halved, where the true ratio and, likely, the measured ratio are themselves very large. In this way, a small error in the denominator can lead to a

massive error in the ratio. Small true values in the denominator are most prevalent when the coefficient of variation of the denominator hormone level is large. (In such cases, the denominator levels are highly positively skewed, with a much larger proportion of values lying below the mean—and relatively close to zero—than lying above the mean.) Thus, the validity of hormone ratios is going to be affected not only by the validities of the hormone measurement themselves, but also by how hormone levels are distributed. By substantially reducing the skew of hormone distributions and effectively turning divisions into subtractions, the log-transformation can greatly attenuate the sensitivity of ratios to noise and fluctuations. As a result, log-ratios can be expected to show higher validity than raw ratios; the difference in performance is going to become more dramatic as measurement error increases and/or the distribution of hormone values at the denominator becomes more skewed.

3 Simulations

As a concrete illustration of our argument, we now present two sets of simulations that explore the robustness of raw versus log-ratios in presence of increasing amounts of error. As we noted in the previous section, hormonal measurements involve multiple sources of noise that behave differently at higher and lower concentrations. To capture this complexity without adding too much detail to the simulations, we modeled measurement error as a mixture of two normally distributed components: a *concentration-dependent error* (CDE) whose standard deviation increases proportionally to the underlying level of the hormone (analogous to a CV, but including all concentration-dependent sources of noise in addition to those "in the tube"); and a *concentration-independent error* (CIE) with a fixed standard deviation across the entire range of values. The CIE accounts for those sources of noise that have a disproportionate effect at low concentrations, and do not become vanishingly small as hormone levels approach zero. We varied the standard deviation of the concentration-dependent error from 10% to 50% of the underlying hormonal value. Note that 50% is not an unreasonably large value, especially when the effective level of the hormone (that is, the underlying level that produces the physiological effect of interest) is not captured by its instantaneous concentration at a given time, but by the average concentration over hours or days (see Figure 1). For the concentration-independent error, we set the standard deviation to 1%, 5%, or 10% of the median of the distribution of hormonal values. (The results are qualitatively similar if the mean is used, but taking the median as a reference point helps reduce the impact of skewness on the magnitude of errors.)

For each combination of parameters, we simulated 100 replicate studies with $N = 500$ hormonal measurements each. We used the simulated data to calculate four validities (correlations ranging from 0 to 1): (a) the validity of the measured raw ratio as an indicator of the underlying raw ratio; (b) the validity of the measured log-ratio as an indicator of the underlying log-ratio; (c) the "crossed" validity of the measured raw ratio as an indicator of the underlying log-ratio; and (d) the "crossed" validity of the measured log-ratio as an indicator of the underlying raw ratio. In addition to the mean validity, we reported a range of variation across studies (5^{th} and 95^{th} percentiles). A wide range of variation indicates that the ratio lacks "second-order" robustness: the achieved validity may fluctuate unpredictably from one study to the next, and hence contribute to a pattern of inconsistent and unreplicable results.

All simulations were performed in R 4.0 (R Development Core Team, 2021). The code and empirical data used to produce the simulations are available at https://doi.org/10.6084/m9.figshare.16840717

3.1 Simulation 1: Idealized distributions

In the first set of simulations, we modeled two idealized hormones A and B by sampling from log-normal distributions with variable amounts of skewness (see Kletsky, 1975). We selected three values of skewness that together cover the range usually observed in hormone studies: 2 (low), 4 (moderate), and 12

(high).[2] For each data point, we calculated the A/B ratio and its logarithm, both before adding measurement error (underlying effective levels of A and B) and after adding measurement error (measured levels of A and B; see above). To avoid zero and negative values, the smallest underlying level of each hormone was used as the lower bound for the measured values of that hormone. The resulting values were used to calculate the validities of individual hormone measurements, their ratio, and their log-ratio. We further considered three scenarios with different correlation patterns between hormones A and B. In the first scenario, A and B are uncorrelated. In the second scenario, A and B are positively correlated, with a correlation of .50 between the logged distributions of the underlying values of A and B. This value corresponds to correlations of approximately .25 to .45 between the measured A and B, depending on the amount of error and the skewness of the distributions. Positive correlations in this range are commonly observed between pairs of functionally related hormones such as estradiol and progesterone, testosterone and cortisol, and testosterone and estradiol.[3] The third scenario is presented in the Supplementary Material, and introduces a negative correlation between A and B. We chose a correlation of −.60 between the logged distributions of the underlying values, corresponding to correlations of approximately −.20 to −.40 between measured values.

Figures 2, 3 and 4 show simulation results for the scenario in which A and B are uncorrelated and the CIE is set at 1%, 5%, and 10% of the median, respectively. We describe several patterns evident in the results:

First, as measurement error increases, the expected validity of the raw ratio drops quickly (solid red lines), relative to validity of the log-ratio (solid blue lines). This is true whether CDE (x-axis in all figures) or CIE (moving from Fig. 2 to Fig. 3 to Fig. 4) increases.

Second, as measurement error (both CDE and CIE) increases, the range of variation of the validity becomes wider (red bands around the solid red lines), meaning that the validity of raw hormone ratios can fluctuate wildly from one study to the next. Once again, this pattern contrasts with the pattern for log-ratios (blue bands around solid blue lines), where the range of validity across simulation runs is much smaller. Accordingly, log-ratios can be expected to yield more replicable findings all else being equal.

Third, the skewness of the hormone at the denominator (increasing from left to right in each figure) increases the validity of log-ratios but reduces that of raw ratios, while making the latter more variable; in doing so, it amplifies the discrepancy between the performance of the two indices. For example, even when the CIE is modest (5% of the median, Figure 3) and the skewness of the denominator is moderate (middle column of figures), the validity of raw hormone levels is comparatively low and highly variable, especially as the CDE increases. When the skewness of the denominator is high (right column of Figure 3), the validity of the raw hormone ratios is low and highly variable even at low levels of CDE. The skewness of the hormone at the numerator (increasing from top to bottom) increases the validity of both raw and log-ratios while making the former more variable. The net impact on the relative performance of the two indices is small compared with that of the skewness at the denominator. Also note that the validity of log-ratios is much less sensitive to changes in the skewness of the denominator hormone. Indeed, though this validity is obviously affected by measurement error, it remains substantial compared to that of the raw ratio, as well as much less variable across conditions.

[2] In three large samples (*N* from 419 to 2180), each with values on 2 to 5 hormones (estradiol, progesterone, testosterone, cortisol, and oxytocin), skewness ranged from 0.87 to 20.22, with a median (across 11 values) of 3.24; with the extreme on each end removed, range = 1.67 to 8.55. See Supplementary Material, Table S1.

[3] For example, in one data set of naturally cycling women (*N* = 708; Dinh et al., 2021), we found correlations of .38, .37, and .26 between levels of estradiol and progesterone, testosterone and cortisol, and testosterone and estradiol, respectively. In another data set (*N* = 2180; Jones et al., 2018), these correlations were .27, .15, and .30, respectively.

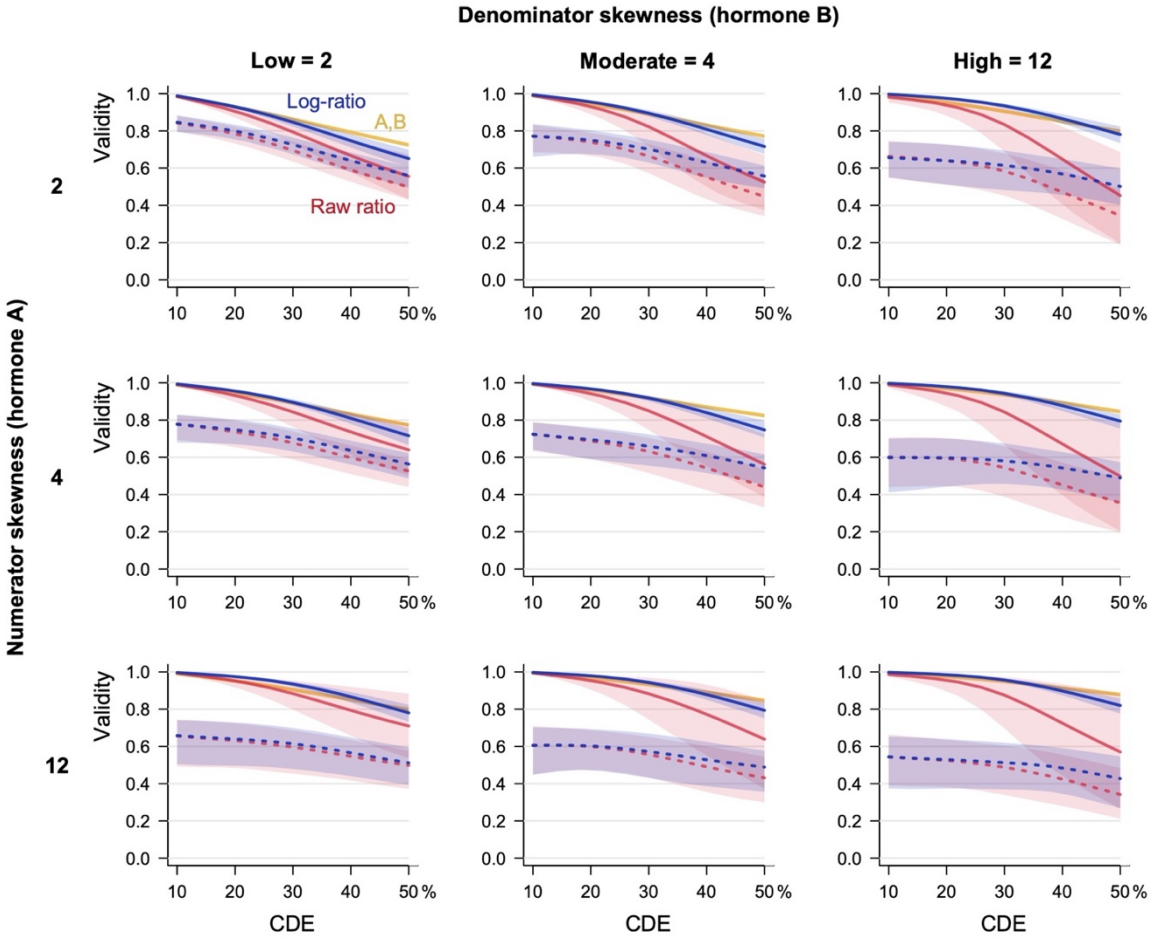

Fig. 2. Simulated validities of raw hormone ratios and log-ratios when the concentration-independent error (CIE) is set at 1% of the median and the levels of the two hormones are uncorrelated. Validities are shown as a function of concentration-dependent error (CDE), skewness of numerator hormone levels (A), and skewness of the denominator hormone levels (B). Values of hormones A and B were sampled from log-normal distributions. Solid red line: Correlation between measured raw hormone ratios and underlying raw hormone ratios. Solid blue line: Correlation between measured log-ratios and underlying log-ratios. Dashed red line: Correlation between measured raw hormone ratios and underlying log-ratios. Dashed blue line: Correlation between measured log-ratios and underlying raw hormone ratios. When the dashed blue line is higher than the solid red line, the measured log-ratio has greater validity for measuring the underlying raw ratio than does the measured raw ratio itself. Red and blue bands show the 5th and 95th percentiles of the validities across 100 simulated studies (N = 500 each). Solid yellow line: Mean correlation of the measured hormone levels with the underlying raw levels of A and B, shown for comparison purposes.

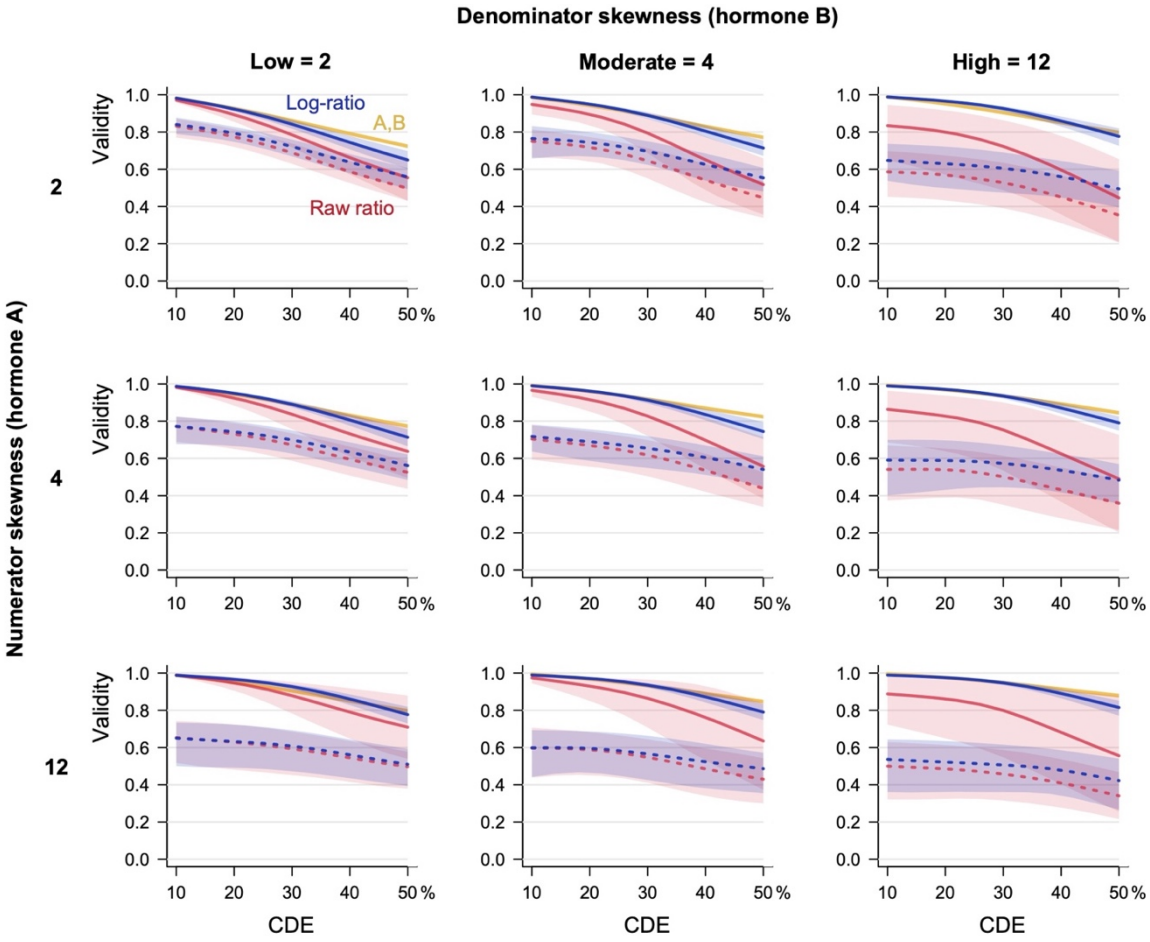

Fig. 3. Simulated validities of raw hormone ratios and log-ratios when the concentration-independent error (CIE) is set at 5% of the median and the levels of the two hormones are uncorrelated. Validities are shown as a function of concentration-dependent error (CDE), skewness of numerator hormone levels (A), and skewness of the denominator hormone levels (B). Values of hormones A and B were sampled from log-normal distributions. Solid red line: Correlation between measured raw hormone ratios and underlying raw hormone ratios. Solid blue line: Correlation between measured log-ratios and underlying log-ratios. Dashed red line: Correlation between measured raw hormone ratios and underlying log-ratios. Dashed blue line: Correlation between measured log-ratios and underlying raw hormone ratios. When the dashed blue line is higher than the solid red line, the measured log-ratio has greater validity for measuring the underlying raw ratio than does the measured raw ratio itself. Red and blue bands show the 5th and 95th percentiles of the validities across 100 simulated studies (N = 500 each). Solid yellow line: Mean correlation of the measured hormone levels with the underlying raw levels of A and B, shown for comparison purposes.

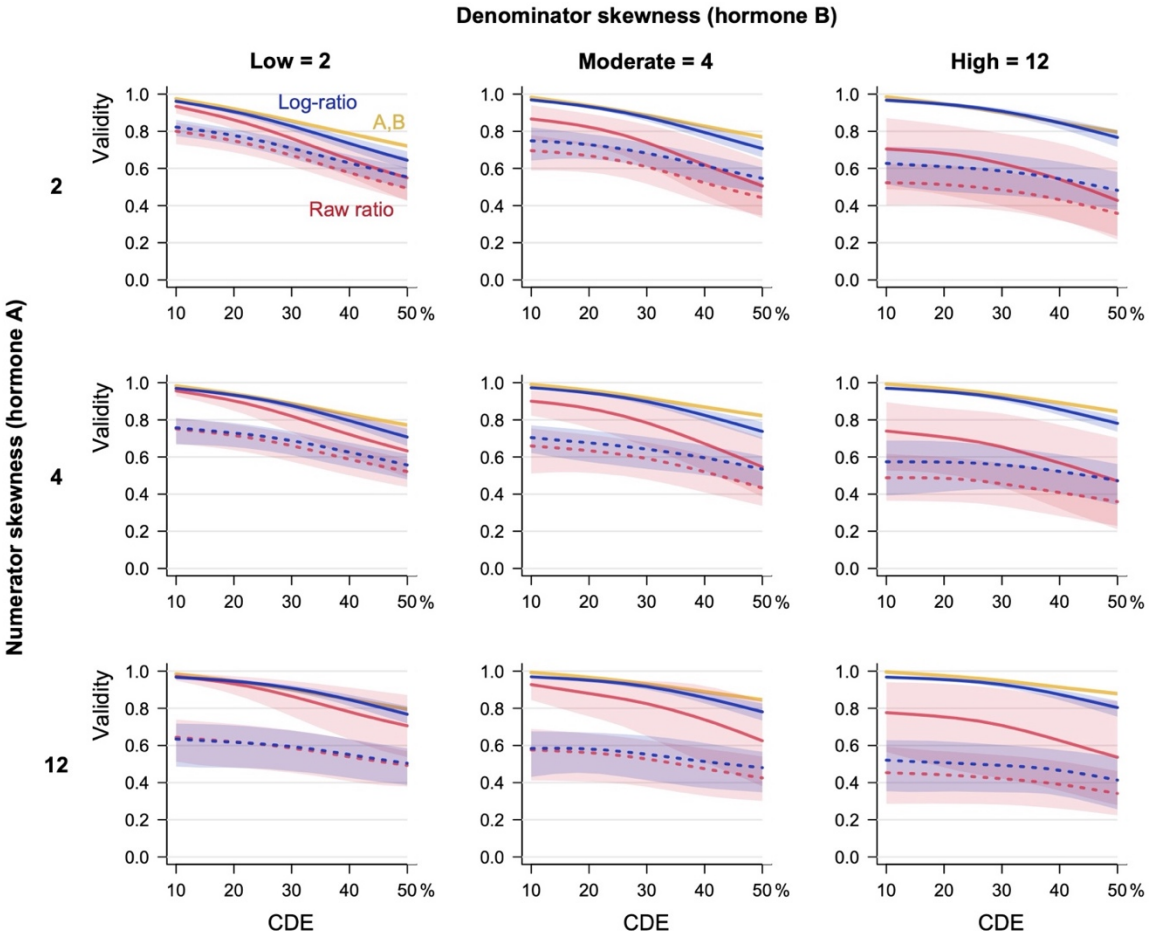

Fig. 4. Simulated validities of raw hormone ratios and log-ratios when the concentration-independent error (CIE) is set at 10% of the median and the levels of the two hormones are uncorrelated. Validities are shown as a function of concentration-dependent error (CDE), skewness of numerator hormone levels (A), and skewness of the denominator hormone levels (B). Values of A and B were sampled from log-normal distributions. Solid red line: Correlation between measured raw hormone ratios and underlying raw hormone ratios. Solid blue line: Correlation between measured log-ratios and underlying log-ratios. Dashed red line: Correlation between measured raw hormone ratios and underlying log-ratios. Dashed blue line: Correlation between measured log-ratios and underlying raw hormone ratios. When the dashed blue line is higher than the solid red line, the measured log-ratio has greater validity for measuring the underlying raw ratio than does the measured raw ratio itself. Red and blue bands show the 5th and 95th percentiles of the validities across 100 simulated studies (N = 500 each). Solid yellow line: Average correlation of the measured hormone levels of A and B with the underlying raw levels, shown for comparison purposes.

Fourth and relatedly, Figures 2-4 reveal an interesting and counterintuitive pattern: Under certain conditions, the measured log-ratio is not just a valid indicator of the underlying log-ratio, but becomes a better indicator of the underlying *raw* ratio than the measured raw ratio itself. This occurs in these figures whenever the dashed blue line (validity of the log-ratio in measuring the true raw ratio) crosses above the solid red line (validity of the raw ratio in measuring the true raw ratio). (Note that the reverse does not happen—the dashed red line and the solid blue line never cross, meaning the raw ratio never measures the log-ratio better than does the log-ratio itself.) This means that, even in cases where the underlying raw hormone ratio is physiologically more meaningful (see e.g., Roney, 2019), the measured log-ratio may still be the metric of choice, owing to the comparatively lower validity of the measured raw ratio.

Figures 5 shows simulation results for the scenario in which A and B are positively correlated. The CIE is set to 5% of the median, as in Figure 3. The results for CIE = 1% and 10% can be found in the Supplementary Material (Fig. S1 and S2), as can be results when A and B are negatively correlated (Fig. S3-S5). When the two hormones are positively correlated (Figure 5), the validity of both ratios decreases, but the drop is especially steep for the raw ratio. At the same time, the "crossed" validity of the log-ratio (dashed blue lines) increases. As a result, the conditions under which the measured log-ratio is a better indicator of the underlying raw ratio than the measured raw ratio itself expand. Hence, when skewness of the denominator is moderate or high (middle and right-hand columns), even a modest about of CDE error leads the measured log-ratio to measure the true underlying raw ratio as well or better than the measured raw ratio itself. This observation may be important in light of the fact that correlations between hormone pairs for which ratios have been commonly used (estradiol and progesterone; testosterone and cortisol; testosterone and estradiol) are often positive (see Footnote 2). When A and B are negatively correlated (Supplementary Material, Fig. S3-S5), the validity of both ratios increases and their crossed validity decreases. Although the log-ratio maintains the highest validity, it does not outperform the raw ratio as an indicator of the underlying raw ratio (at least within the range of measurement error considered in the simulation).

3.2 Simulation 2: Resampled E/P data

In the first set of simulations, hormone values were generated by sampling from log-normal distributions. This idealized approach should provide a reasonable approximation of real-world hormonal data; however, it is reasonable to wonder how the results would look if we had used a more realistic model of the underlying distributions. To address this question, we carried out another set of simulation in which hormone values were sampled from three empirical datasets of estradiol and progesterone. Dataset 1 was from Dinh et al. (2021) and included $N = 708$ complete pairs of E and P measurements. The correlation between the two hormones was .38; skewness values were 3.24 for E and 8.55 for P. Dataset 2 was from Jones et al. (2018) and included $N = 2{,}180$ complete pairs (correlation: .27). The skewness was 3.30 for E and 2.84 for P. Dataset 3 was from Stern et al. (2021) and included $N = 868$ complete pairs (correlation: .02). The skewness was 2.22 for E and 5.71 for P.

For each combination of parameters, we simulated 100 replicate studies by sampling $N = 500$ pairs of measurements from the empirical dataset (with replacement). Measurement error was added to these values in the same way as before. This procedure employs measured hormone values from the empirical datasets as underlying hormone values in the simulation. Because the former already include measurement error, the distributions of measured hormones in the simulation will be somewhat less skewed and less strongly correlated with one another than it would be the case in a perfectly accurate simulation. Despite this limitation, sampling from real-world datasets provides a more realistic and fine-grained model of hormone distributions in the specific case of E/P ratios, as contrasted with the generic model of the first set of simulations.

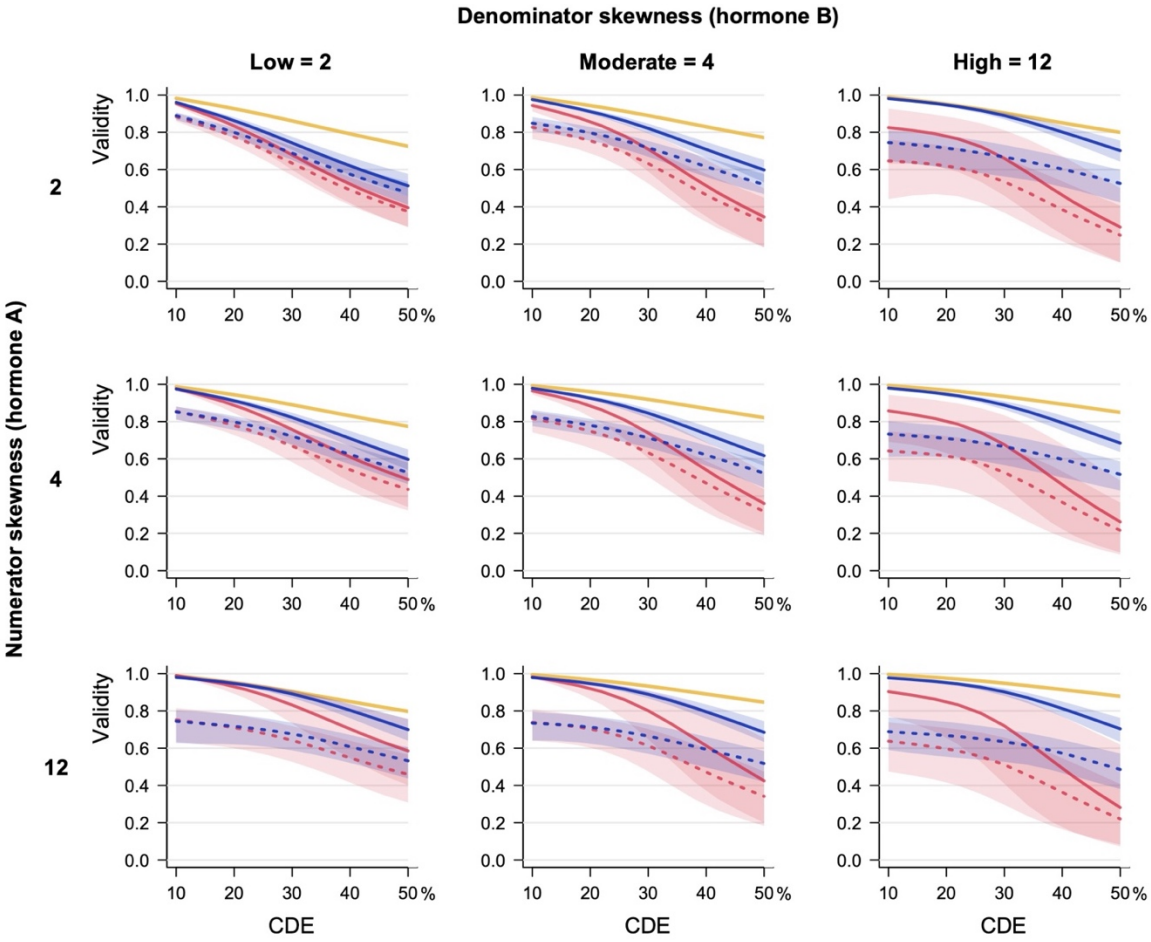

Fig. 5. Simulated validities of raw hormone ratios and log-ratios when the concentration-independent error (CIE) is set at 5% of the median and the levels of the two hormones are positively correlated (about .25 to .45 between measured levels). Validities are shown as a function of concentration-dependent error (CDE), skewness of numerator hormone levels (A), and skewness of the denominator hormone levels (B). Values of A and B were sampled from log-normal distributions. Solid red line: Correlation between measured raw hormone ratios and underlying raw hormone ratios. Solid blue line: Correlation between measured log-ratios and underlying log-ratios. Dashed red line: Correlation between measured raw hormone ratios and underlying log-ratios. Dashed blue line: Correlation between measured log-ratios and underlying raw hormone ratios. When the dashed blue line is higher than the solid red line, the measured log-ratio has greater validity for measuring the underlying raw ratio than does the measured raw ratio itself. Red and blue bands show the 5th and 95th percentiles of the validities across 100 simulated studies (N = 500 each). Solid yellow line: Average correlation of the measured hormone levels of A and B with the underlying raw levels, shown for comparison purposes.

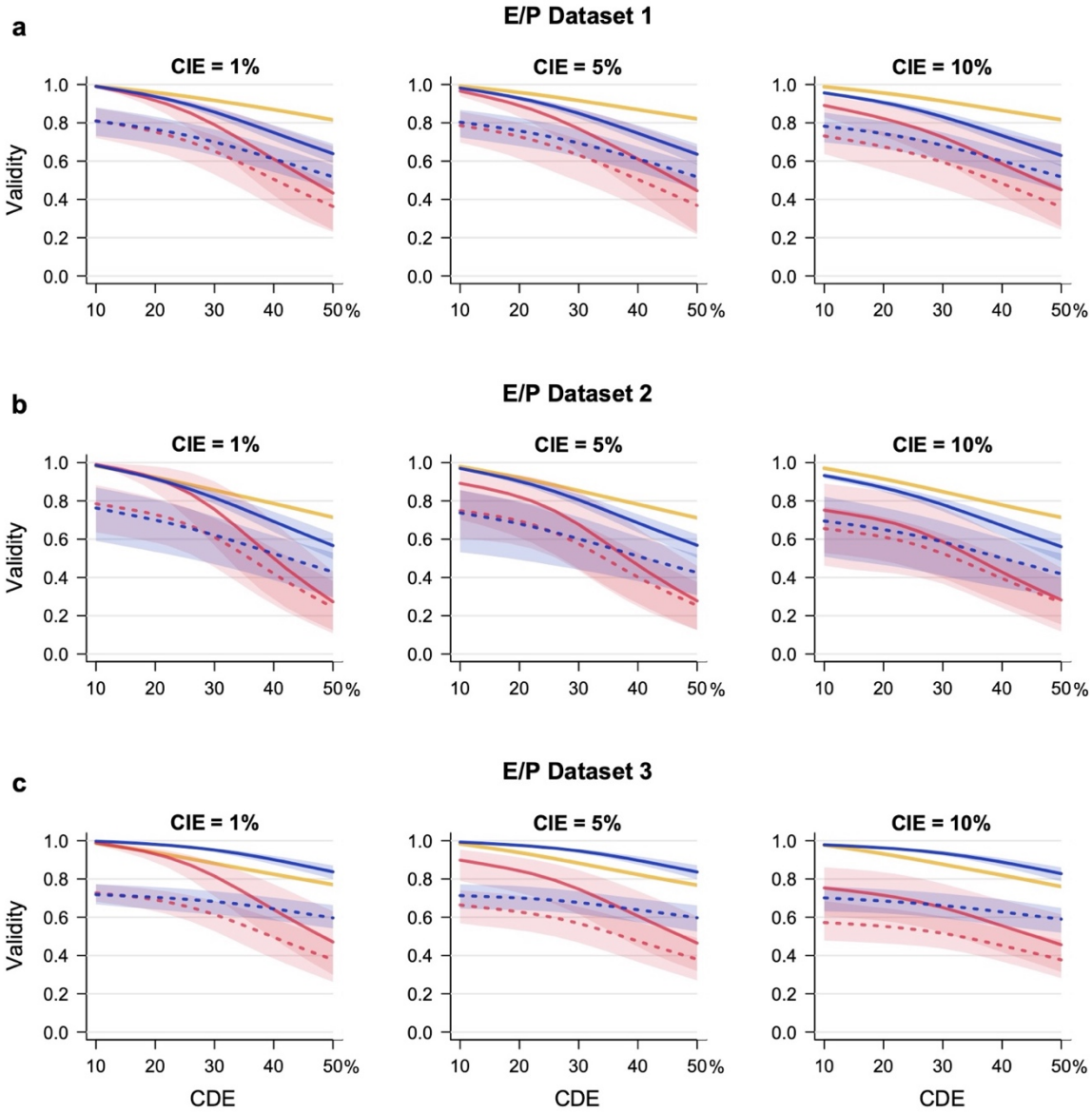

Fig. 6. Simulated validities of raw estradiol/progesterone (E/P) ratios and log-ratios, with hormone values sampled from three empirical datasets. (a) Dataset 1 (Dinh et al., 2021); (b) Dataset 2 (Jones et al., 2018); and (c) Dataset 3 (Stern et al., 2021; see the main text for more details). Validities are shown as a function of concentration-independent error (CIE) and concentration-dependent error (CDE). Solid red line: Correlation between measured raw hormone ratios and underlying raw hormone ratios. Solid blue line: Correlation between measured log-ratios and underlying log-ratios. Dashed red line: Correlation between measured raw hormone ratios and underlying log-ratios. Dashed blue line: Correlation between measured log-ratios and underlying raw hormone ratios. When the dashed blue line is higher than the solid red line, the measured log-ratio has greater validity for measuring the underlying raw ratio than does the measured raw ratio itself. Red and blue bands show the 5th and 95th percentiles of the validities across 100 simulated studies (N = 500 each). Solid yellow line: Average correlation of the measured hormone levels of E and P with the underlying raw levels, shown for comparison purposes.

The results of these simulations are shown in Figure 6. For all three datasets, the results are very close to what one would expect given the correlations between E and P and the skewness of their distributions. These results confirm that the patterns we described above are robust, and do not depend on the fine details of the hormone distributions. They also suggest that, in the specific case of E and P, the log-ratio ln(E/P) may easily prove a better indicator of the underlying E/P ratio than the measured ratio itself. This might be relevant to the fact that, even though Roney (2019) found that the [mean E/mean P] ratio tracked conception status more closely than ln[mean E/mean P], the log-ratio ln(E/P) was a better predictor of sexual desire than the raw E/P in the same dataset (Gangestad et al., 2019, p. 541).

3.3 Summary of the Simulation Results

As expected from the mathematical properties of ratios, simulations showed that raw hormone ratios quickly lose validity in presence of measurement error. In addition to reducing validity on average, measurement error broadens its range of variation across samples; the implication is that using raw ratios in presence of imperfect measurement will tend to yield unreplicable results from one study to the next. The validity of log-ratios was substantially less affected by error, and remained considerably more stable across samples. Counterintuitively, under some conditions the measured log-ratio turned out to be a better indicator of the underlying *raw* ratio than the measured raw ratio itself. The divergence between raw and log-ratios became especially pronounced when (a) the two hormones were positively correlated, and (b) the distribution of the hormone at the denominator was more strongly skewed. Increasing the realism of the simulations by sampling from empirical distributions of estradiol and progesterone did not change the results; this reinforces the idea that our conclusions can be generalized to real-world scenarios.

4 Discussion

Hormone ratios possess a number of statistical properties that, in many circumstances, are undesirable (Sollberger & Ehlert, 2016). Yet, despite concerns raised, many published studies continue to implement them. Here, we draw attention to yet another undesirable property: the validity of hormone ratios can be greatly affected by even modest levels of measurement error. This is particularly true when values of the hormone at the denominator are at least moderately skewed and when hormone levels are positively correlated, two conditions that are commonly observed in the literature.

A key implication of this lack of robustness is that reductions in validity can substantially reduce the power to detect true effects. Hence, for instance, consider the situation in which the skewness of both hormones is moderate, hormone values positively covary, and concentration independent error is modest (middle panel of Figure 5). With moderately high levels of concentration dependent error (40%), the validity of individual hormone values remains high, close to .90. The amount of error is individual hormone levels is not unrealistic, yet the validity of the raw hormone ratio barely exceeds .50. To illustrate: if we assume a true underlying correlation of .20 between a hormone measure and an outcome of interest (assumed to be measured without error for simplicity) in a simple between-subjects study, a sample size of $N = 250$ yields 82% power if validity is .90. If validity is .50, by contrast, the power of the study drops to 35%, and a sample size of about $N = 820$ is needed to achieve the same power (calculations in G*Power 3.1.9.4; Faul et al., 2017). Power is similarly compromised in other kinds of study designs (e.g., within-subject designs).

Our simulations also show that the validity of raw ratios may vary widely across samples. Due to sampling variability, then, in some instances validity may be substantially higher than average, leading to greater power. Yet validity is just as likely to be substantially lower than average, suppressing power. Across replications, then, positive findings can be expected to be spotty—and therefore discounted—even when a true effect exists.

Naturally, researchers rarely know how much error exists in their data. (As we discuss above, reported intra-assay and inter-assay CVs reflect only a portion of measurement error of interest—in many cases, a small portion.) Moreover, they rarely know, a priori, what true underlying hormone variable relates most strongly to an outcome of interest. Does the underlying raw hormone ratio best capture true hormonal effects? Does the log of the underlying hormone ratio (equivalent to the difference between the logs of each hormone level) capture true effects better (e.g., Dinh et al., 2021)? Are effects best represented by the simple additive effects of the hormones, whether treated as raw untransformed values or log-transformed levels? Do hormone levels interact to affect outcomes (e.g., Sollberger & Ehlert, 2016)? In the face of these uncertainties, using raw hormone ratios is a particularly risky choice. Even in cases where the underlying raw ratio best captures the joint effect of the two hormones, the measured raw ratio may perform worse than the log-ratio, as shown in our simulations. By contrast, if the joint hormonal effect is best captured by the underlying log-ratio, the measured log-ratio will uniformly perform better than the measured raw ratio—in most realistic circumstances, much better. In other words, utilization of log-ratios typically reduces risk in the face of uncertainties about underlying realities, relative to utilization of raw ratios.

Our simulations, then, offer a clear recommendation against utilization of raw hormone ratios. Some exceptions could potentially exist: e.g., when hormone levels are minimally skewed and measurement error is known to be very small (with respect to the effective hormone levels of interest, not simply because of small CVs "in the tube"), measured hormone ratios may be reasonably valid. But we suspect that these circumstances are rare. Furthermore, raw ratios suffer from the other problems already discussed by Sollberger and Ehlert (2016): highly skewed distributions; asymmetry between ratios of the same two hormones with numerator and denominator reversed; and lack of interpretability of effects involving a ratio. In our view, researchers should not report results of analyses using raw hormone ratios in absence of clearly stated strong justification, including discussion of reasons why their validity is *not* problematically affected by measurement errors.

It may be worth stressing that our simulations do not necessarily recommend that researchers use log-transformed hormone ratios instead. If researchers want to explore whether hormone ratios capture joint effects of hormones in particular instances, log-ratios are clearly preferable to raw ratios, as they *are* much more robust to measurement error and do not share other limitations of raw ratios (e.g., extreme skewness, asymmetry between A/B and B/A; see Sollberger & Ehlert, 2016). Furthermore, a log-ratio *may* capture the joint additive effects of log-transformed hormones with a single index such that, if hormones have opposing additive effects, use of the log-ratio may offer greater statistical power to detect those effects (relative to separate entry of each hormone as a predictor; Sollberger & Ehlert, 2016). Still, log-ratios impose constraints on these opposing effects: The two hormones, log-transformed, are constrained to have equal and opposite effects. ($\text{Ln}(A/B) = \ln(A) - \ln(B)$. If, in a regression equation, $\ln(A/B)$ receives a weight of b_1, then the regression equation is equivalent to one where (b_1)log-ratio is replaced by $(b_1)\ln(A) + (-b_1)\ln(B)$, that is, one in which ln(A) and ln(B) are separately entered but constrained to have equal weights opposite in sign. Because individual hormone levels may capture associations not fully embodied in log-ratios, we recommend that—even when log-ratio effects are examined—researchers perform and report follow-up analyses that probe the contributions of individual hormone levels, whether raw or log-transformed or (perhaps preferably) both, with each hormone entered in a prediction equation separately (with their weights freely allowed to vary). Furthermore, as noted earlier, log-ratios do not capture interaction effects. Simple linear × linear interaction effects may be examined by entering the interaction between hormones along with main effects, whether transformed or not: e.g., ln(A), ln(B), ln(A) × ln(B). Researchers should not confuse the use of a log-ratio with examination of statistical interactions. For example, Sollberger et al. (2016) reported a statistical interaction between testosterone and cortisol levels in predicting pro-environmental attitudes (consistent with the dual hormone hypothesis regarding testosterone and cortisol; e.g., Mehta & Prasad, 2015), but no effect of the logged testosterone/cortisol ratio. As the latter does not tap interaction effects, there is no inconsistency in this pattern of effects.

In conclusion, we hope that these results will inform data analytic practices by alerting researchers to the severe limitations of hormone ratios. Increasing the validity of hormonal measures will contribute to enhance the power and replicability of future research in this field.

Acknowledgments

We thank Oliver Schultheiss for his perceptive comments and suggestions.

References

Baskaran, C., Plessow, F., Silva, L., Asanza, E., Marengi, D., Eddy, K.T., Sluss, P.M., Johnson, M.L., Misra, M., Lawson, E.A., 2017. Oxytocin secretion is pulsatile in men and is related to social-emotional functioning. Psychoneuroendocrinol. 85, 28-34. https://doi.org/10.1016/j.psyneuen.2017.07.486

Beaven, C.M., Ingram, J.R., Gill, N.D., Hopkins, W.G., 2010. Ultradian rhythmicity and induced changes in salivary testosterone. Eur. J. Appl. Physiol. 110, 405-413. https://doi.org/10.1007/s00421-010-1518-3

Crespi, B.J., 2016. Oxytocin, testosterone, and human social cognition. Biol. Rev. 91, 390-408. https://doi.org/10.1111/brv.12175

Dinh, T., Emery Thompson, M., Gangestad, S. W., 2021. Ovarian hormones in relation to naturally cycling women's conception risk: Empirical evidence and implications for behavioral endocrinology. Unpublished manuscript, submitted for review.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Meth. 39, 175–191. https://doi.org/10.3758/BF03193146

Filicori, M., Butler, J. P., & Crowley, W. F. (1984). Neuroendocrine regulation of the corpus luteum in the human. Evidence for pulsatile progesterone secretion. *Journal of Clinical Investigation, 73,* 1638-1647. https://doi.org/10.1172/JCI111370

Fragkaki, I., Cima, M., Granic, I., 2018. The role of trauma in the hormonal interplay of cortisol, testosterone, and oxytocin in adolescent aggression. Psychoneuroendocrinol. 88, 24–37. https://doi.org/10.1016/j.psyneuen.2017.11.005

Gelman, A., 2019. Why you should usually log-tranform your positive data. Online blog at https://statmodeling.stat.columbia.edu/2019/08/21/you-should-usually-log-transform-your-positive-data/

Gong, Y., Xiao, H., Li, C., Bai, J., Cheng, X., Jin, M., Sun, B., Lu, Y., Shao, Y., Tian, H., 2013. Elevated t/e2 ratio is associated with an increased risk of cerebrovascular disease in elderly men. PloS one 8, e61598. https://doi.org/10.1371/journal.pone.0061598

Granger, D.A., Shirtcliff, E.A., Booth. A., Kivlighan, K.T., Schwartz, E.B. 2004. The "trouble" with salivary testosterone. Psychoneuroendocrinol. 29, 1229-1240. https://doi.org/10.1016/j.psyneuen.2004.02.005

Johnson, E.O., Kamilaris, T.C., Chrousos, G.P., Gold, P.W., 1992. Mechanisms of stress: A dynamic overview of hormonal and behavioral homeostasis. Neurosci. Biobehav. Rev. 16, 115-130. https://doi.org/10.1016/S0149-7634(05)80175-7

Jones, B.C., Hahn, A.C., Fisher, C.I., Wang, H., Kandrik, M., DeBruine, L.M., 2018. General sexual desire, but not desire for uncommitted sexual relationships, tracks changes in women's hormonal status. Psychoendocrinol. 88, 153-157. https://doi.org/10.1016/j.psyneuen.2017.12.015

Ketterson, E. D., & Nolan, V., Jr. (1999) Adaptation, exaptation, and constraint: a hormonal perspective. Am. Nat. *125*, S4-S25. https://doi.org/10.1086/303280

Kletzky, O.A., Nakamura, R.M., Thorneycroft, I.H., Mishell, D.R., 1975). Log normal distribution of gonadotropins and ovarian steroid values in the normal menstrual cycle. Am. J. Obstet. Gynecol. 121, 688-694. https://doi.org/10.1016/0002-9378(75)90474-3

Licinio, J., Negrão, A. B., Mantzoros, C., Kaklamani, V., Wong, M. L., Bongiorno, P. B., ... & Gold, P. W. (1998). Synchronicity of frequently sampled, 24-h concentrations of circulating leptin, luteinizing hormone, and estradiol in healthy women. *Proceedings of the National Academy of Sciences USA, 95,* 2541-2546. https://doi.org/10.1073/pnas.95.5.2541

Mehta, P.H., Prasad, S. 2015. The dual-hormone hypothesis: A brief review and future research agenda. Curr. Opin. Behav. Sci. 3, 163–168. http://doi.org/10.1016/j.cobeha.2015.04.008

Prasad, S., Lassetter, B., Welker, K.M., Mehta, P. H. 2019. Unstable correspondence between salivary testosterone measured with enzyme immunoassays and tandem mass spectrometry. Psychoneuroendocrinol. 109, 104373. https://doi.org/10.1016/j.psyneuen.2019.104373

Righetti, F., Tybur, J., Van Lange, P.A.M., Echelmeyer, L., van Esveld, S., Kroese, J., van Brecht, J., Gangestad, S.W., 2020. How reproductive hormonal changes affect relationship dynamics for women AND men: A 15-day diary study. Biol. Psychol. 149, 1-8. https://doi.org/10.1016/j.biopsycho.2019.107784

Roney, J.R., 2019. On the use of log transformations when testing hormonal predictors of cycle phase shifts: commentary on Gangestad, Dinh, Grebe, Del Giudice and Emery Thompson (2019). Evol. Hum. Behav. 40, 526–530. https://doi.org/10.1016/j.evolhumbehav.2019.08.006

Roney, J.R., Simmons, Z.L., 2013. Hormonal predictors of women's sexual desire in normal menstrual cycles. Horm. Behav. 63, 636-645. https://doi.org/10.1016/j.yhbeh.2013.02.013

Rossmanith, W.G., Laughlin, G.A., Mortola, J.F., Johnson, M.L., Veldhuis, J.D., Yen, S.S. Pulsatile cosecretion of estradiol and progesterone by the midluteal phase corpus luteum: temporal link to luteinizing hormone pulses. J. Clin. Endocrinol. Metab. 70, 990-995. doi: 10.1210/jcem-70-4-990. PMID: 2318954. https://doi.org/10.1016/j.psyneuen.2015.09.031

Schultheiss, O.C., Dlugash, G., Mehta, P. H. 2019. Hormone measurement in social neuroendocrinology: A comparison of immunoassay and mass spectrometry methods. In Routledge international handbook of social neuroendocrinology (pp. 26-40). Routledge. https://doi.org/10.4324/9781315200439-3

Selcer, K.W., Leavitt, W.W., 1988. Progesterone down-regulation of nuclear estrogen receptor: a fundamental mechanism in birds and mammals. Gen. Comp. Endocrinol. 72, 443-52. https://doi.org/10.1016/0016-6480(88)90167-0

Shirtcliff, E.A., Granger, D.A., Schwartz, E.B., Curran, M.J., Booth, A., Overman, W.H. 2000. Assessing estradiol in biobehavioral studies using saliva and blood spots: Simple radioimmunoassay protocols, reliability, and comparative validity. Horm. Behav. 38, 137-147. https://doi/org/10.1006/hbeh.2000.1614

Sollberger, S., Bernauer, T., Ehlert, U. 2016. Salivary testosterone and cortisol are jointly related to pro-environmental behavior in men, Soc. Neurosci. 11, 553-566. http://doi.org/10.1080/17470919.2015.1117987

Sollberger, S., Ehlert, U., 2016. How to use and interpret hormone ratios. Psychoneuroendocrinol. 63, 285-297. https://doi.org/10.1016/j.psyneuen.2015.09.031

Stamatiades, G.A., Kaiser, U.B., 2018. Gonadotropin regulation by pulsatile GnRH: Signaling and gene expression. Molecul. Cell. Endocrinol. 463, 131–141. https://doi.org/10.1016/j.mce.2017.10.015

Stern, J., Arslan, R.C., Gerlach, T.M., Penke, L. 2019. No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad et al. (2019). Evol. Hum. Behav. 40, 517-525. https://doi.org/10.1016/j.evolhumbehav.2019.08.005

Stern, J., Kordsmeyer, T.L., Penke, L., 2021. A longitudinal evaluation of ovulatory cycle shifts in women's mate attraction and preferences. Horm. Behav. 128, 104916. https://doi.org/10.1016/j.yhbeh.2020.104916

Terburg , D., Morgan, B., van Honk, J., 2009. The testosterone-cortisol ratio: A hormonal marker for proneness to social aggression. Int. J. Law Psychiat. 32, 216-223. https://doi.org/10.1016/j.ijlp.2009.04.008 .

Tilbrook, A. J., Turner, A. I., & Clarke, I. J. (2000). Effects of stress on reproduction in non-rodent mammals: the role of glucocorticoids and sex differences. *Reviews of Reproduction, 5,* 105-113. https://doi.org/10.1530/ror.0.0050105

Viau, V., 2002. Functional cross-talk between the hypothalamic-pituitary-gonadal and-adrenal axes. J Neuroendocrinol. 14, 506-513. https://doi.org/10.1046/j.1365-2826.2002.00798.x

Welker, K.M., Lassetter, B., Brandes, C.M., Prasad, S., Koop, D.R., Mehta, P.H. 2016. A comparison of salivary testosterone measurement using immunoassays and tandem mass spectrometry. Psychoneuroendocrinol. 71, 180-188. https://doi.org/10.1016/j.psyneuen.2016.05.022

Woolley, C.S., McEwen, B.S., 1993. Roles of estradiol and progesterone in regulation of hippocampal dendritic spine density during the estrous cycle in the rat. J. Comp. Neurol. 336, 293–306. https://doi.org/10.1002/cne.903360210

Young, E.A., Abelson, J., Lightman, S.L., 2004. Cortisol pulsatility and its role in stress regulation and health. Front. Neuroendocrinol. 25, 69-76. https://doi.org/10.1016/j.yfrne.2004.07.001