Accepted Manuscript

Title: Unstable correspondence between salivary testosterone measured with enzyme immunoassays and tandem mass spectrometry

Authors: Smrithi Prasad, Bethany Lassetter, Keith M. Welker, Pranjal H. Mehta

PII: S0306-4530(18)31144-2
DOI: https://doi.org/10.1016/j.psyneuen.2019.104373
Article Number: 104373

Reference: PNEC 104373

To appear in:

Received date: 13 November 2018
Revised date: 19 February 2019
Accepted date: 3 July 2019

Please cite this article as: Prasad S, Lassetter B, Welker KM, Mehta PH, Unstable correspondence between salivary testosterone measured with enzyme immunoassays and tandem mass spectrometry, *Psychoneuroendocrinology* (2019), https://doi.org/10.1016/j.psyneuen.2019.104373

Unstable correspondence between salivary testosterone measured with enzyme immunoassays and tandem mass spectrometry

Smrithi Prasad[1,2], Bethany Lassetter[3], Keith M. Welker[4], & Pranjal H. Mehta[1,5]

[1] Department of Psychology, University of Oregon, United States of America
[2] Marshall School of Business, University of Southern California, United States of America
[3] Department of Psychology, University of Toronto, Canada
[4] College of Liberal Arts, University of Massachusetts Boston, United States of America
[5] Department of Experimental Psychology, University College London, United Kingdom

Conflict of interest and financial disclosures: The authors report no financial interests or potential conflicts of interest. This research was supported by NSF Grants #1451848 and #1303743 awarded to PHM.

Acknowledgements: We thank Cassandra M. Brandes, Dennis R. Koop, Colton Christian, and Erik L. Knight. We also thank Nicholas Grebe and Victor Shiramizu for their feedback on the manuscript's pre-print.

Correspondence to: Smrithi Prasad, Marshall School of Business, University of Southern California, smrithip@marshall.usc.edu

Abstract: 248 words
Main document: 6249 words

Highlights

- Salimetrics EIAs have unstable correspondence with MS for testosterone measurement
  - Antibody performance and matrix interference may explain unstable correspondence
  - An updated composite Salimetrics EIA-MS correspondence is provided
  - Recommendations for future research in psychoneuroendocrinology are discussed

Abstract

Although some studies reveal that saliva handling and storage practices may influence salivary testosterone concentrations measured with immunoassays, the effect of these method factors on the validity of testosterone immunoassays remains unknown. The validity of immunoassays can be assessed by comparing hormone concentrations measured with immunoassays to a standard reference method: liquid chromatography tandem mass spectrometry (MS). We previously reported the correspondence between salivary testosterone measured with enzyme immunoassays (EIAs) and with MS when there was less variance in (or more control over) method factors related to saliva handling and storage across measurement methods (Welker et al., 2016). In the present study, we expanded the original dataset and compared the correspondence between Salimetrics EIAs and MS when there was greater variance in (or less control over) method factors across EIAs and MS (high method variance), to when there was less variance in these factors (low method variance). If variance in these method factors impacts the validity of testosterone measurement, then the EIA-MS correspondence should be stronger when method variance is low compared to when it is high. Our results contradicted this hypothesis: Salimetrics EIA-MS correspondence was *stronger* when variance in method factors was high compared to when it was low. The composite average correlation across both method variance comparisons provides an updated estimate of Salimetrics EIA-MS correspondence, but the instability in this correspondence may pose challenges to the reproducibility of psychoneuroendocrinology research. We discuss possible explanations for the surprising pattern of results and provide recommendations for future research.

Keywords: salivary testosterone; immunoassays; liquid chromatography tandem mass spectrometry; method variance

1. Introduction

Over the last decade, there has been a substantial increase in empirical work examining salivary testosterone's association with psychosocial processes and behaviors such as competition, relationships and mating behaviors, and social and economic decision-making (for reviews see Casto and Edwards, 2016; Eisenegger et al., 2011; Roney and Gettler, 2015; Stanton, 2016). To date, many studies in the field of psychoneuroendocrinology have adopted immunoassays (IAs) to measure salivary testosterone due to their technical ease of use, cost effectiveness, and accessibility to many researchers and labs (Taylor et al., 2015). This research, however, assumes that IAs provide reliable and valid estimates of testosterone concentrations. Surprisingly, little work in the field of psychoneuroendocrinology has empirically tested this assumption (but see Welker et al., 2016; Yasuda et al., 2008).

In a recent comprehensive study, Welker et al. (2016) assessed the validity of IAs for salivary testosterone measurement by comparing testosterone concentrations measured with three commercially manufactured enzyme immunoassays (EIAs; from Salimetrics, DRG, and IBL International) to a highly accurate reference method: liquid chromatography tandem mass spectrometry (LC-MS/MS; referred to as MS in the current research; see also Yasuda et al., 2008).[1] The results revealed that: (i) compared to MS, EIAs inflated estimates of salivary testosterone, especially at lower concentrations; and (ii) the strength of the EIA-MS linear association was small to moderate in statistical magnitude within sexes (per the benchmarks of Cohen, 1988; Overall: r=.47-.57; Males: r=.10-.17; Females: r=-.17-.22). These findings call into question the validity of salivary testosterone estimates provided by EIAs, an ongoing concern in the field of clinical chemistry that was first highlighted over two decades ago (Fitzgerald and Herold, 1996; Handelsman and Wartofsky, 2013). The findings from Welker et al. (2016) are also consistent with the growing clinical chemistry

[1]An EIA is a type of IA that measures hormone levels via an enzyme-triggered color change reaction. MS, as reported in Welker et al. (2016; p. 181), "is a sensitive reference measure, allowing for both the identification and quantification of compounds by combining the physical separation capacity of liquid chromatography with the mass analysis capability of mass spectrometry." Moreover, MS demonstrates greater specificity and sensitivity in hormone measurement and is free of some of the limitations found in EIAs (Hoofnagle and Wener, 2009; Soldin & Soldin, 2009), thereby making it a highly accurate reference method.

literature that has illustrated potential sources of inaccuracy in IAs, such as cross-reactivity and matrix effects (Keefe et al., 2014; Stanczyk et al., 2003; Taieb et al., 2003; Tate and Ward, 2004).[2]

Although the findings from Welker et al. (2016) provide insight into the strength of the EIA-MS correspondence of salivary testosterone measurement, two open questions remain. First, Welker et al. (2016) was primarily conducted to assess cross-manufacturer comparisons, and therefore provided only a *single* estimate of the EIA-MS relationship for each manufacturer. Whether the EIA-MS correspondence will maintain stability across repeated measurement within the same EIA manufacturer remains unknown. Second, because different laboratories adopt different procedures for saliva handling and storage (Chiappin et al., 2007; Dabbs, 1991; Granger et al., 2004), it is unclear whether variance in method procedures influences the correspondence between EIA and MS testosterone measurements. We address these two questions with the current research.

Variance in freeze-thaw cycles, centrifugation, and storage time influences mean estimates of salivary testosterone measured with IAs (Chiappin et al., 2007; Granger et al., 2004). More specifically, repeated freeze-thaw cycles increase bacterial contamination and exacerbate hormone degradation (Chiappin et al., 2007; Whembolua et al., 2006), more centrifugation decreases hormone concentrations (Durdiaková et al., 2013), and longer freezer storage periods alter salivary testosterone concentrations, albeit inconsistently (inflation: Toone et al., 2013; no change: Durdiaková et al., 2013; Granger et al., 2004). Despite these insights, many of these studies rely upon relatively small samples and samples that are predominantly male, therefore limiting generalizability, especially across sexes and hormone concentration ranges. Further, these studies examine the effects of these saliva storage and handling factors on *mean hormone concentrations*. To date, no study has examined the influence of these methodological factors on the *validity* of IAs compared to a reference method, MS, for salivary testosterone measurement.

[2] Cross-reactivity refers to the erroneous measurement of compounds structurally similar to the hormone of interest and can be caused by the lack of specificity of antibodies used in IAs. The sample's matrix composition (e.g., pH of saliva and the presence of mucins) can also interfere in antibody binding thereby resulting in inaccurate measurement (Tate and Ward, 2004). For a detailed discussion of factors that impact IA measurement, see Selby (1999).

The impact of factors related to saliva handling and storage on the validity of IAs can be determined by comparing IA-MS correspondence when there is *less* variance in (or greater control over) these method factors across measurement methods compared to when there is *greater* variance in (or less control over) method factors. In a previous study, we reported EIA-MS correspondences when there was low method variance across the measurement methods of EIAs and MS (Welker et al., 2016). In the present research, we expand the original dataset (see Figure 1) and compare correspondence between Salimetrics EIAs and MS when there is greater method variance across measurement methods (i.e., high method variance; new data reported in this paper) to when there is lower method variance across the two methods (i.e., low method variance; data from Welker et al., 2016). We predicted that if variance in method factors does indeed impact the validity of testosterone measurement, then the EIA-MS correspondence should be stronger when method variance is low compared to when it is high. However, if these factors do not systematically influence the validity of testosterone measurement, then we should see no difference in the EIA-MS correspondence across low and high method variance comparisons.

2. Methods

One hundred saliva samples were randomly selected from a set of samples collected for a different study (Prasad et al., *in prep*). These samples were obtained via passive drool using standard collection procedures (Schultheiss and Stanton, 2009). Immediately after collection, samples were stored at -80°C until they were assayed. Prior to hormone analysis, samples were aliquoted into smaller volumes so that the same samples could be used across multiple methods (i.e., EIAs and MS) while controlling variability in sample preparation. EIAs were conducted in duplicate using Salimetrics EIA kits in our in-house laboratory at the University of Oregon following standard protocols and specifications provided by the manufacturer (Salimetrics, LLC, 2014). MS was conducted at Oregon Health and Science University's Bioanalytical Shared Resource/Pharmacokinetics Core labs (see Welker et al., 2016 for more information about MS specifications[3]). The reported lower limit of quantification (LOQ) for Salimetrics EIAs was <1pg/mL

[3] MS procedure adhered to standard practices in the literature (Star-Weinstock et al., 2012; Keevil, 2013; Turpeinen, et al., 2012). However, we note in Welker et al. (2016) that future research could include additional

(Salimetrics, LLC, 2014). The LOQ for MS was 1pg/mL with a signal to noise (S/N) of 5:1. No samples in the current dataset fell below the LOQ for EIA and MS analyses.

To test if method factors associated with saliva handling and storage influenced correspondence between testosterone concentrations obtained from Salimetrics EIAs and MS, we compared the Salimetrics EIA-MS correspondence when variance in these factors was low compared when it was high (see Figure 1).

2.1. Low method variance comparison

In this comparison, factors related to saliva handling and storage were tightly controlled across Salimetrics EIAs and MS. All samples underwent three freeze-thaw cycles and three rounds of centrifugation (each round for 10 minutes at 3500 rpm), prior to *both* Salimetrics EIAs and MS. Samples were stored at -80°C until they were analyzed using EIAs and MS. Samples were stored for an additional five months before MS compared to EIAs. Average intra- and inter-assay coefficients of variation (CVs) for EIAs were 6.96% and 8.54%, respectively. For more details about methodological specifications for analyses conducted in the low method variance comparison, see Welker et al. (2016).

2.2. High method variance comparison

In this comparison, there was greater variance in the methodological practices related to saliva handling and storage between Salimetrics EIAs and MS. Whereas prior to MS (as noted above) samples underwent three freeze-thaw cycles and three rounds of centrifugation, prior to EIAs samples underwent two freeze-thaw cycles and two rounds of centrifugation. Samples were stored at -80°C for 18.18 months (*SD*=1.79) longer before MS compared to EIAs. Average intra- and inter-assay CVs for EIAs in this comparison were 6.59% and 15.79% respectively.[4]

quality control measures for cross-method comparisons. Researchers could verify the calibration consistency of MS analysis by adding certified reference materials (i.e., samples with predetermined concentrations) to their analyses, and analyze controls and samples used to create the IA assay's standard curve with MS. Finally, future research could assay samples in duplicate across both methods to obtain more accurate metrics of reliability.

[4] We note a few other differences across method variance comparisons. First, fewer assay batches were used under conditions of low method variance (2 batches) relative to the high method variance (4 batches). Second, to maintain consistency in assayer technique, only one lab personnel conducted EIAs in the low method variance comparison, relative to three lab personnel in the high method variance comparison.

Although we reported testosterone obtained from both Salimetrics EIAs and MS under conditions of low method variance in Welker et al. (2016), we did not examine concentrations from Salimetrics EIAs measured with high method variance. The reason for only including samples with low method variance in this past paper was to minimize and control for apparent methodological variability in freeze-thaw cycles, centrifugation, and storage times. However, after publishing this work (i.e., Welker et al., 2016), additional analyses of these data and feedback from the scientific community led us to recognize the additional insights to be gained from examining samples prepared under high method variance.[5]

2.3. Statistical analysis strategy

To test if variance in method practices between measurement techniques of EIAs and MS influenced the stability of EIA-MS correspondence, we compared how conditions of high and low method variance influenced: (i) mean testosterone concentrations from EIAs, (ii) mean differences in testosterone concentrations from Salimetrics EIAs relative to MS (or fixed bias analysis), (iii) the Salimetrics EIA-MS correlations using Fisher's *r* to *z* transformations for dependent samples (Steiger, 1980), and (iv) the degree of convergence of testosterone from Salimetrics EIAs with the reference method (MS) using Deming regression (Martin, 2000).

Importantly, the validity of testosterone measurement with IAs is found to differ at lower and higher concentrations, with greater quantification errors in populations with lower concentrations (e.g., in females: Schultheiss and Stanton, 2009, Welker et al., 2016; or older males: Mazur and Clifton, 2018). Therefore, we examined the stability of the EIA-MS correspondence in four sample sub-groups that represent higher concentrations (i.e., males and the upper 50% of the sampling distribution) or lower concentrations (i.e., females and the lower 50% of the sampling distribution). The cutoff for the upper and lower 50% sample sub-groups was determined by concentrations from

[5] Despite having previously assayed our samples using Salimetrics EIAs (high method variance comparison), we decided *a priori* to re-assay the samples for Welker et al. (2016) using the same EIA kit manufacturer (low method variance comparison) for greater consistency between measurement techniques. Further, Welker et al. (2016) compared the validity of three commercially manufactured EIAs (i.e., DRG, IBL International, and Salimetrics). Data from DRG or IBL are not included in the current paper because we did not conduct those assays under conditions of high method variance. Therefore, we were unable to test the effect of method variance on DRG and IBL EIA-MS relationships. Instead, a supplementary correlation table across all assay kits and MS is reported (Table S1).

MS. Finally, in supplementary analysis, we conducted Bland-Altman plot analysis to test how method variance altered magnitude of inflation across the sampling distribution (or proportional bias) in testosterone measurement (Bland and Altman, 1986).

2.4. Missing data and outliers

Three female participants (two in the high variance comparison, and one in the low variance comparison) were excluded from all main inferential analyses.[6] A fourth female with inadequate saliva volume for MS was also excluded from the main analyses. Because we excluded four females across all main analyses, we acknowledge discrepancies between the results we report here and those reported in Welker et al. (2016). We excluded these participants to permit appropriate statistical comparisons across both variance comparisons.

2.5. Transparent reporting and pre-printing

Data are available on the Open Science Framework (osf.io/3wn5p). Some of these data have been published previously (Welker et al., 2016, but see Section 2.4 for outlier exclusion criteria applied to the current study). This paper was also pre-printed prior to submission for the peer-review process (osf.io/5rcq7).

3. Results

See Table 1 for descriptive statistics (means and SDs) and correlations of testosterone concentrations from Salimetrics EIAs and MS, split by method variance comparisons, males and females, and the upper and lower 50% of concentrations in the sample.

3.1. Mean differences in testosterone concentrations from EIAs.

Overall, testosterone concentrations measured with EIAs did not differ significantly across conditions of high and low method variance (p>.48). Further, we found no differences in testosterone concentrations measured with EIAs across method variance comparisons when examining the sample separately by sex and at the upper and lower 50% of the distribution (ps>.23).

3.2. Fixed bias comparisons

[6] The criterion for excluding outliers was set at 3 SDs above the mean within males and females. The three females who were excluded from the analyses had testosterone values at 3.39 and 3.02 SDs above the mean in the high method variance comparison and 3.48 SDs above the mean in the low method variance comparison.

We examined whether the magnitude of fixed bias differed across the two method variance comparisons (see Table 2 for fixed bias estimates and inferential statistics for each comparison). In both high and low method variance comparisons, Salimetrics EIAs inflated testosterone concentrations compared to MS (high method variance: $F(1,95)=186.90$, $p<.001$, $\eta_p^2=.66$; low method variance: $F(1,95)=66.83$, $p<.001$, $\eta_p^2=.41$). Importantly, the magnitude of fixed bias did not statistically differ across comparisons ($p>.48$), suggesting that Salimetrics EIAs inflated levels of testosterone compared to MS and differences in method variance had no significant effects on the magnitude of inflation.

Analyses in the sample sub-groups revealed that Salimetrics EIAs significantly inflated levels of testosterone compared to MS in males, females, and in the upper and lower 50% of the sampling distribution across both the high and low method variance comparisons (see Table 2). Further, the magnitude of fixed bias in the four sample sub-groups was not significantly different across high and low method variance comparisons ($ps>.22$).

In summary, we found a similar pattern of fixed bias with Salimetrics EIAs inflating levels of testosterone compared to MS across the entire sample, in both sexes, and in the upper and lower 50% of the distribution in both method variance comparisons.

3.3. Correlational analyses

Next, we tested for differences in correlations between testosterone concentrations estimated via EIAs and MS, across high and low method variance comparisons. In the high method variance comparison, testosterone from Salimetrics EIAs strongly correlated with MS ($r(94)=.80$, $p<.001$), and this correlation was statistically greater in magnitude than that found in the low method variance comparison ($r(94)=.55$, $p<.001$; $z=5.77$, $p<.001$).

Similarly, the Salimetrics EIA-MS correlation (see Table 2) was significantly stronger in the high compared to low method variance comparison for males ($z=3.16$, $p=.002$), females[7] ($z=6.14$,

[7] Out of the three female outliers excluded from the analyses, one had a testosterone concentration 3.02 SDs above the mean within females. Because this participant was close to the cut off for outliers, we re-ran the EIA-MS correlation in the high method variance comparison including this female participant. Upon inclusion, the EIA-MS correlation in females in the high method variance comparison was $r(53)=.61$, $p<.001$.

p<.001), and in the upper 50% (*z*=3.76, *p*<.001) and lower 50% (*z*=3.83, *p*<.001) of the sampling distribution.

Surprisingly, these findings were in the opposite of the predicted direction: the Salimetrics EIA-MS correlation was stronger in the high rather than low method variance comparison across the entire sample, in both sexes, and in the upper and lower 50% of the distribution. Collectively, these findings provide evidence for instability in Salimetrics EIA-MS correspondence.

3.4. Deming regression

Next, Deming regression analysis was conducted to assess a direct one-to-one correspondence between methods. Deming regression determines how closely the relationship between two methods conforms to an identity line which assumes equality between methods (intercept= 0, slope= 1; see Table 3). Figure 2 depicts a scatterplot for testosterone measured with Salimetrics EIAs under comparisons of high and low method variance in relation to MS (left and right panels, respectively), the line of best fit from the Deming regression, and the identity line. We found that for Salimetrics EIAs, both the high and low method variance comparisons differed from the line of identity (i.e., the 95% CI for the slope did not include 1), indicating a lack of direct one-to-one correspondence between Salimetrics EIAs and MS (high method variance: 1.26, *95%CI*[1.07, 1.45]; low method variance: 1.63, *95%CI*[1.12, 2.13]). However, the slope in the high method variance comparison more closely approximated the line of identity: the slope was closer to 1, the 95% CI for the slope was narrower, and the CI did not include the higher slope value found in the low method variance comparison. These findings are consistent with the EIA-MS correlations reported above.

Breaking down these patterns by sample sub-group, males demonstrated convergence with the line of identity (i.e., the 95% CI for the slope included 1) in the high but not low method variance comparison (see Table 3 and Figure S1). Although in the upper 50% of the distribution we found convergence with the line of identity across both method variance comparisons (see Table 3 and Figure S3), the line of identity was approximated better in the high relative to the low method variance comparison. However, females and the lower 50% of the sampling distribution demonstrated poor approximation of the line of identity across both comparisons, with descriptively poorer

convergence in the low method variance comparison (the 95% CIs for the slope included 0). For more details, see Table 3, and Figure S2 (for females) and Figure S4 (for the lower 50%).

Overall, the Deming regression analyses were consistent with previous correlational analyses suggesting better one-to-one correspondence between Salimetrics EIAs and MS in the high method variance comparison. However, we note generally good convergence with the line of identity at higher concentrations of testosterone – in males and in the upper 50% of the sample distribution – and a lack of convergence with the line of identity at lower concentrations – in females and in the lower 50% of the sampling distribution.

3.5. Supplementary analysis: Bland-Altman plots

We used Bland-Altman plots to assess whether method variance influenced inflation in testosterone from Salimetrics EIAs compared to MS across the sample distribution (i.e., proportional bias; see Supplement). We note that unlike the low method variance comparison wherein we found greater inflation of testosterone at lower concentrations of the sample distribution, there was no proportional bias under conditions of high method variance (Figure S5).

3.6. Average analysis

Overall, we found that Salimetrics EIA-MS correspondence was stronger when variance in method factors was high compared to when it was low. We also note that the testosterone concentrations from EIAs across both variance comparisons were well-correlated (r=.79, p<.001), and therefore averaged them together to provide a composite estimate of Salimetrics EIA-MS correspondence across all available data. Using the averaged EIA testosterone concentrations, we conducted correlational analyses to estimate a composite Salimetrics EIA-MS correspondence (overall sample: $r(94)$=.71, p<.001; males: $r(40)$=.36, p=.018; females: $r(52)$=.34, p=.012; upper 50% of testosterone distribution: $r(46)$=.49, p<.001; lower 50% of testosterone distribution: $r(46)$=.43, p=.002). These composite EIA-MS correspondences are descriptively higher than those reported in Welker et al. (2016).

We also conducted Deming regression and Bland-Altman plot analyses using averaged testosterone levels. Deming regression analysis suggests that although the average testosterone values did not demonstrate direct one-to-one correspondence with MS concentrations, they more closely

approximated the line of identity compared to those reported in Welker et al. (2016) (see Table 3 and Figures S10-S12). Bland-Altman plot analysis indicated that despite marginally more inflation at lower levels of testosterone, the magnitude of bias was descriptively lower than that reported in Welker et al. (2016).

4. Discussion

We examined whether variance in method factors related to saliva handling and storage influences the validity of EIAs in measuring salivary testosterone. To test the primary research question, we used archival data to compare testosterone concentrations measured with Salimetrics EIAs to a standard reference method: MS. We predicted that if variance in method factors does indeed influence the validity of EIAs, then the Salimetrics EIA-MS correspondence will be stronger when there is less variance in method factors across EIAs and MS (low method variance comparison) compared to when there is greater variance in these factors (high method variance comparison). Contrary to this prediction, we found a surprising pattern of results: Salimetrics EIA-MS correspondence was stronger in the high compared to the low method variance comparison.

We offer two potential explanations for this surprising pattern of results. First, instability in EIA-MS correspondence may have been caused by *differences in the performance of antibodies* used in the manufacturing of IA kits. Commercially manufactured IAs use antibodies to isolate the substrate of interest (in the current research, testosterone). The specificity of antibodies in detecting the substrate, as opposed to other compounds with chemical structures similar to the substrate (i.e., cross-reactivity), may serve as a marker of the kit's performance. Recent papers suggest that antibodies may vary across different manufactured batches of IA kits thereby resulting in batch-to-batch variability in their performance (Baker, 2015; Tate and Ward, 2004). Different IAs use different classes of antibodies (e.g., monoclonal vs. polyclonal). Polyclonal antibodies may be especially prone to cross-reactivity (Frank, 2002), and several scientists have argued that the use of polyclonal antibodies should be phased out of research entirely (Bradbury and Plückthun, 2015). Because Salimetrics EIAs use polyclonal antibodies, they may demonstrate increased variability in performance (i.e. unstable validity). The current study's stronger EIA-MS correspondence in the high method variance comparison therefore could be explained by differences in antibody performance.

More specifically, the antibodies from the EIA kits in the high method variance comparison may have performed better, and more accurately estimated testosterone concentrations. If this explanation is accurate, it suggests that a method factor largely outside of a psychoneuroendocrinology researcher's control (i.e., antibody performance) is producing unstable validity in salivary testosterone measured using IAs. The current study was not designed to directly quantify performance of the antibodies by measuring cross-reactivity. A direct test of this potential explanation will require additional research and new study protocols (see Krasowski et al., 2014; Valdes and Jortani, 2002).

Second, the instability in EIA-MS correspondence may have been caused by *differential sensitivity of IAs and MS to matrix interference* from saliva handling and storage factors (i.e., storage times, freeze-thaw cycles, and centrifugation). Because IAs are more sensitive to matrix interference than MS (Tate and Ward, 2004), method factors that directly impact the saliva matrix may interfere more in hormone measurement using IAs compared to MS. For example, changes in the constitution of saliva linked to methods in saliva handling and storage (e.g., bacterial growth and hormone degradation from longer storage times or more freeze-thaw cycles) may impact hormone measurement via IAs more so than via MS. In the current study, samples in the low method variance comparison were stored for approximately 13 months longer, underwent one additional freeze-thaw cycle, and one additional round of centrifugation than those in the high method variance comparison. Matrix interference therefore may have been more pronounced in hormone measurement with EIAs in the low (vs. high) variance comparison, thereby resulting in a weaker correspondence with MS in the low methods comparison. If this explanation is accurate, it suggests that methods related to saliva handling and storage, which are largely within a researcher's control, may affect the validity of salivary testosterone measured with EIAs. This explanation indicates that researchers could improve the validity of salivary testosterone measurement by changing their practices related to saliva handling and storage. However, because the two method variance comparisons in the current study differed across multiple method factors, it was not possible to discern which (if any) of the method factors directly caused more or less matrix interference in testosterone measurement using EIAs in comparison to MS. Future research should systematically vary each method factor and test its independent influence on testosterone obtained from IAs and MS. Doing so will allow researchers to

examine whether differential sensitivity to matrix interference across measurement methods is indeed an explanation for the surprising pattern of results.

Salimetrics EIA-MS instability also did not differ across different subsets of the sample (i.e., males and females; the lower and upper 50% of sampling distribution): Salimetrics EIA concentrations correlated more strongly with MS concentrations in the high (compared to low) method variance comparison. However, additional analyses of these sample subsets using Deming regression and Bland-Altman plots revealed greater discrepancies and error in testosterone measurement for lower concentration samples. With Deming regression for example, males demonstrated better one-to-one correspondence between EIAs and MS in the high compared to the low method variance comparison, but females lacked convergence across EIAs and MS in both comparisons. Further, analysis of proportional bias with Bland-Altman plots using average testosterone concentrations indicated marginally greater inflation of testosterone at lower concentrations. This inflation of testosterone at lower concentrations is also apparent when descriptively comparing testosterone concentrations obtained via EIAs relative to MS *between* the two sexes: On average, EIAs (compared to MS) inflated testosterone by nearly five-fold in females vs. two-fold in males (see Table 1). This pattern of disproportionate inflation of testosterone concentrations in females vs. males is consistent with past research (see Schultheiss et al., 2018). Collectively, our results are consistent with prior research highlighting increased error in using EIAs for testosterone measurement in samples with lower concentrations (Herold and Fitzgerald, 2003). Future research should continue assessing the validity and stability of testosterone measurement, especially at lower testosterone concentrations.

Even though Salimetrics EIA-MS correspondence demonstrated instability, similar magnitudes of fixed bias (i.e., inflation of testosterone levels) in both variance comparisons emerged. Previous research suggests that the magnitude of fixed bias for IAs could serve as a crude proxy for IA validity with lower fixed bias suggesting more valid measurement (testosterone: Welker et al., 2016; cortisol: Miller et al., 2013). The present results, however, suggest that these fixed bias estimates are not entirely foolproof heuristics of IA validity. If the magnitudes of inflation are indeed good heuristics, then we should have found a smaller magnitude of fixed bias and lower mean

concentrations (from which fixed bias estimated are calculated) in the high compared to the low method variance comparison – a pattern of results that aligns with the stronger EIA-MS correspondence or better validity of EIAs in the high method variance comparison. Because we did not find this pattern, we do not recommend that researchers rely purely on fixed bias estimates or IA means to provide definitive evidence for IA validity. Instead, researchers should conduct direct tests of validity by comparing hormone concentrations from IAs to MS.

Despite instability in EIA-MS correspondence, we averaged testosterone concentrations measured with Salimetrics EIAs across both method variance comparisons and report an average Salimetrics EIA-MS correspondence across all available data. This composite Salimetrics EIA-MS correlation is descriptively stronger than what we originally reported in Welker et al. (2016; see Section 3.6). However, researchers should consider the following two points when interpreting these average results. First, these average estimates potentially mask instability in EIA-MS correspondence. We propose that this instability may be due to variability in antibody performance or matrix interference due to methodological factors, but additional research is needed to identify the exact sources of the instability. Second, fine-grained analyses with the averaged concentrations still reveal measurement limitations of Salimetrics EIAs. For instance, Deming regression analysis on the average testosterone concentrations indicated a lack of direct one-to-one correspondence with MS concentrations (see also the Supplement for evidence using Bland-Altman plot analysis), suggesting that researchers should continue to be concerned about the validity of EIAs for the measurement of salivary testosterone.

4.1. Theoretical and methodological implications

The primary result revealing instability in EIA-MS correspondence poses theoretical challenges to the interpretation of hormone-behavior relationships. Prior social endocrinology research has documented inconsistent testosterone-behavior associations, with some studies revealing strong effects and other studies revealing weak or null effects. Although researchers have attributed these inconsistencies to contextual and individual difference moderators (e.g., Mehta and Prasad, 2015; Norman et al., 2015; Zilioli et al., 2014), the current results suggest that instability in the validity of the salivary testosterone measures across studies may also explain some of the

inconsistencies. Because the field of psychoneuroendocrinology relies heavily on IAs for the measurement of hormone concentrations, this potential instability in IA validity across studies poses a direct and serious challenge to the reproducibility of research in the field (see Baker, 2015; Roy et al., 2019). Follow-up research will be needed to identify the exact causes for instability in EIA-MS correspondence, such as variance in antibody performance or matrix interference, to improve measurement validity and in turn enhance reproducibility. These concerns about reproducibility echo concerns raised by other scientists, who have argued that batch-to-batch variability in antibodies may be creating a "reproducibility crisis" across the biological sciences (Baker, 2015).

The observed instability in EIA-MS correspondence also has methodological implications. Below, we discuss implications related to both proposed explanations of the current results: (1) variance in antibody performance, and (2) method-linked sensitivity to matrix interference.

If instability in EIA-MS correspondence is attributable to variability in antibody performance, then methods should be adopted to improve the validity and stability of antibody performance. One potential solution is to phase out the use polyclonal antibodies, which may be especially prone to variable performance due to cross-reactivity (Bradbury and Plückthun, 2015). A complementary solution is to improve quality control in the supply of antibodies, such as the establishment of an independent body to certify commercial antibodies prior to their use (Baker, 2015; Bradbury and Plückthun, 2015). In the absence of such an organization, researchers can develop new protocols to specifically quantify antibody performance via the measurement of cross-reactivity. However, a test of antibody performance may lie outside the expertise of several psychoneuroendocrinology researchers and require collaboration with experts capable of assessing cross-reactivity. What may lie within researcher control however is the prioritization of methodological validation studies that examine IA-MS correspondence and its stability. The current study suggests that drawing conclusions from a single estimate is insufficient because IA performance may be unstable from study to study due, in part, to differences in antibody performance. Therefore, future research should incorporate repeated testing of the IA-MS correspondence to arrive at more accurate conclusions about IA validity.

If the current study's results are due to differences between IAs and MS in their sensitivity to matrix interference, then researchers may be able to reduce matrix interference and improve measurement validity by changing their practices related to saliva handling and storage. However, we do not know which if any of the saliva handling and storage factors may have affected matrix interference and in turn EIA-MS correspondence in our study. Thus, we recommend tightly controlled follow-up studies to test how variance in method factors may have affected the validity of salivary hormone measurement. The present research was not equipped to directly test this question given the archival nature of the data and focus on testing, in an ecologically-valid manner, the influence of variance in multiple method factors (i.e., storage times, freeze-thaw cycles, and centrifugation) on the validity of EIAs. Therefore, to assess the unique and independent impact of methods factors on the validity of hormone measurement via IAs, follow-up methods work may systematically vary the magnitude of each method factor (e.g., one versus two versus three freeze-thaw cycles; shorter versus longer storage times), examine each factor's independent effect on hormone measurement using IAs and MS, and ultimately make concrete recommendations for future studies (e.g., use fewer freeze-thaw cycles; use shorter storage times).

4.2. Conclusions and Recommendations

In the current research, we tested how variance in saliva handling and storage practices affects the validity of EIAs relative to MS in estimating testosterone concentrations. Contrary to predictions, we found that the correspondence between Salimetrics EIAs and MS was stronger when there was *greater* variance in method factors compared to when there was *less* variance. These findings highlight instability in Salimetrics EIA-MS correspondence. We provide two possible explanations for this pattern of results: (1) variability in antibody performance across the two method comparisons, and (2) differential sensitivity of EIAs and MS to matrix interference. Upon averaging testosterone concentrations from Salimetrics EIAs across both variance comparisons, we report a composite Salimetrics EIA-MS correlation that is descriptively stronger than what was originally reported in Welker et al. (2016; see Section 3.6). This average result can be interpreted alongside the main result showing instability in EIA-MS correspondence.

In light of the present findings, Welker et al. (2016), and other convergent evidence (e.g., Taieb et al., 2003), we make four recommendations for improved measurement of salivary hormones.

4.2.1. Acknowledge the limitations of IAs. First, we recommend that researchers who have previously analyzed their samples with IAs and are in the process of publishing those data acknowledge the limitations of IAs highlighted in the past and current research (e.g., Taieb et al., 2003; Welker et al., 2016), including the potential for instability in the validity of IAs as suggested by the current study. This practice will strengthen the foundation of psychoneuroendocrinology research, encourage open and continued dialogues about open science within the field as well as the broader scientific community, and finally, motivate future research to adopt increasingly improved methods.

4.2.2. Report detailed methodological information in papers. Second, we recommend that researchers report methodological details that may potentially affect IA measurement validity via both matrix interference (e.g., freeze-thaw cycles, rounds of centrifugation, storage length and temperature) and antibody performance (e.g., cross-reactivity information reported by IA manufacturers, type of antibody used, number of batches of assays kits used across assays).

4.2.3. Independently conduct method studies to validate IAs. Third, we recommend that researchers independently test the validity of IAs they currently use and plan to use. In the current research we assessed the capability of just one EIA kit: the Salimetrics Testosterone EIA. This decision was not made a priori but was a result of the archival dataset on which we relied. We encourage investigations with other commercially manufactured EIA kits (e.g., DRG and IBL), other types of IAs (e.g., radioimmunoassays and chemiluminescence IAs), and other sex hormones (e.g., estradiol; Gao et al., 2015; Keefe et al., 2014; Stanczyk et al., 2003). While doing so, researchers should keep in mind that the stability of IA-MS correspondence can be influenced by variance in antibody performance and/or by matrix interference from method factors. To account for variability in antibody performance, researchers can test IA-MS correspondence across multiple measurement instances and provide composite effect sizes of IA-MS correspondence as well as estimates of heterogeneity in effect sizes. To account for matrix interference, researchers can test the independent influence of method factors on hormone measurement to provide insights into the impact of these factors on IA validity and also yield recommendations for future research. We encourage researchers

to disseminate the results of their method studies to support reproducibility of methodological research.

Given the many benefits of IAs (e.g., cost-effectiveness, relative ease of use, and accessibility to researchers) we hope that communication between researchers (e.g., via methodological validation studies) and IA manufacturers will encourage IA manufacturing companies to continue assessing and improving quality control measures that address the impact of antibody performance and matrix interference on the validity of their assays (Tate and Ward, 2004; Baker, 2015).

4.2.4. Prioritize the use of MS for hormone measurement. Fourth and lastly, in line with suggestions from others (Hoofnagle and Wener, 2009; Schultheiss et al., 2018) we recommend the use of MS for salivary hormone measurement given its advantages over IAs including (i) better analytical specificity that is not contingent on antibody performance and is less sensitive to matrix interference, (ii) greater sensitivity at lower concentration ranges, and (iii) capability of measuring multiple hormones using a single test panel. For nearly two decades, clinical endocrinologists have expressed concerns about the validity of IAs, especially in the low range of measurement for salivary hormones (Herold and Fitzgerald, 2003; Stanczyk et al., 2003; Taieb et al., 2002). The *Journal of Clinical Endocrinology & Metabolism* has even explicitly advised against the use of IAs in submitted manuscripts and has suggested that MS methods be adopted instead (Handelsman and Wartofsky, 2013; cf. Wierman et al., 2014). Despite these long-standing recommendations, the field of psychoneuroendocrinology continues to rely on IAs to measure salivary hormones, perhaps because MS requires high instrument and technician costs, and is often inaccessible to researchers and labs (see Taylor et al., 2015). Researchers may consider including costs for MS analyses in future grants and collaborating with clinical chemistry experts and labs that regularly conduct MS. Even after circumventing feasibility concerns, hormone analyses using MS can be error-free only if conducted in labs that have necessary technical expertise, calibration techniques, and adequate quality control protocols. Despite the challenges associated with MS, using this measurement technique will allow researchers to conduct more valid and replicable science in the growing field of psychoneuroendocrinology.

References

Baker, M., 2015. Blame it on the Antibodies. Nature 521, 274–275. doi:10.1038/521274a

Bland, J.M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327, 307–310.

Bradbury, A., Plückthun, A., 2015. Reproducibility: Standardize antibodies used in research. Nature. https://doi.org/10.1038/518027a

Casto, K. V., Edwards, D.A., 2016 Testosterone, cortisol, and human competition. Horm Behav 82, 21–37.

Chiappin, S., Antonelli, G., Gatti, R., De Palo, E.F., 2007. Saliva specimen: A new laboratory tool for diagnostic and basic investigation. Clin. Chim. Acta. doi:10.1016/j.cca.2007.04.011

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci. doi:10.1234/12345678

Dabbs, J.M., 1991. Salivary testosterone measurements: Collecting, storing, and mailing saliva samples. Physiol. Behav. 49, 815–817. doi:10.1016/0031-9384(91)90323-G

Durdiaková, J., Fábryová, H., Koborová, I., Ostatníková, D., Celec, P., 2013. The effects of saliva collection, handling and storage on salivary testosterone measurement. Steroids 78, 1325–1331. doi:10.1016/j.steroids.2013.09.002

Eisenegger, C., Haushofer, J., Fehr, E., 2011. The role of testosterone in social interaction. Trends Cogn. Sci. doi:10.1016/j.tics.2011.04.008

Fitzgerald, R.L., Herold, D.A., 1996. Serum total testosterone: Immunoassay compared with negative chemical ionization gas chromatography-mass spectrometry. Clin. Chem.

Frank, S., 2002. Specificity and Cross-Reactivity, in: Immunology and Evolution of Infectious Disease. https://doi.org/10.1038/420741b

Gao, W., Stalder, T., Kirschbaum, C., 2015. Quantitative analysis of estradiol and six other steroid hormones in human saliva using a high throughput liquid chromatography–tandem mass spectrometry assay. Talanta 143, 353–358. doi:10.1016/j.talanta.2015.05.004

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

Handelsman, D.J., Wartofsky, L., 2013. Requirement for mass spectrometry sex steroid assays in the journal of clinical endocrinology and metabolism. J. Clin. Endocrinol. Metab. doi:10.1210/jc.2013-3375

Herold, D. A., & Fitzgerald, R. L. 2003. Immunoassays for testosterone in women: better than a guess? Clin. Chem. 49, 1250-1251.

Hoofnagle, A.N., Wener, M.H., 2009. The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. J. Immunol. Methods. doi:10.1016/j.jim.2009.06.003

Keefe, C.C., Goldman, M.M., Zhang, K., Clarke, N., Reitz, R.E., Welt, C.K. 2014. Simultaneous Measurement of Thirteen Steroid Hormones in Women with Polycystic Ovary Syndrome and Control Women Using Liquid Chromatography-Tandem Mass Spectrometry. PLoS One 9, e93805. doi:10.1371/journal.pone.0093805

Keevil, B.G., 2013. Novel liquid chromatography tandem mass spectrometry (LC-MS/MS) methods for measuring steroids. Best Pract. Res. Clin. Endocrinol. Metab. 27, 663–674. doi:10.1016/j.beem.2013.05.015

Krasowski, M.D., Drees, D., Morris, C.S., Maakestad, J., Blau, J.L., Ekins, S., 2014. Cross-reactivity of steroid hormone immunoassays: clinical significance and two-dimensional molecular similarity prediction. BMC Clin. Pathol. 14, 33. doi:10.1186/1472-6890-14-33

Martin, R.F., 2000. General deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clin. Chem. 46, 100–104.

Mazur, A., Clifton, S., 2018. Enzyme immunoassay may be inadequate for measuring salivary testosterone in older men. Aging Male. https://doi.org/10.1080/13685538.2018.1509206

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

Miller, R., Plessow, F., Rauh, M., Gröschl, M., Kirschbaum, C., 2013. Comparison of salivary cortisol as measured by different immunoassays and tandem mass spectrometry. Psychoneuroendocrinology 38, 50–57. doi:10.1016/j.psyneuen.2012.04.019

Norman, R.E., Moreau, B.J.P., Welker, K.M., Carré, J.M., 2015. Trait Anxiety Moderates the Relationship Between Testosterone Responses to Competition and Aggressive Behavior. Adapt. Hum. Behav. Physiol. 1, 312–324. doi:10.1007/s40750-014-0016-y

Prasad, S., Knight, E. L., Sarkar, A., Welker, K. M., Lassetter, B., & Mehta, P. H. (in prep). Testosterone fluctuations in response to a democratic election predict partisan attitudes toward the elected leader.

Roney, J.R., Gettler, L.T. 2015. The role of testosterone in human romantic relationships. Curr Opin Psychol 1, 81–86.

Roy, A.R.K., Cook, T., Carré, J.M., Welker, K.M., 2019. Dual-hormone regulation of psychopathy: Evidence from mass spectrometry. Psychoneuroendocrinology 99, 243–250. doi:10.1016/j.psyneuen.2018.09.006

Salimetrics, LLC, 2014. Salivary Testosterone: Enzyme Immunoassay Kit Salimetrics. LLC., State College, PA, pp. 1–25.

Schultheiss, O.C., Dlugash, G., Mehta, P.H. 2018. Hormone measurement in social endocrinology: A comparison of immunoassay and mass spectrometry methods. In: O.C Schultheiss and P.H. Mehta (Eds.). The International Handbook of Social Neuroendocrinology. Routledge Press.

Schultheiss, O.C., Stanton, S.J., 2009. Assessment of salivary hormones. In: Harmon-Jones, E., Beer, J.S. (Eds.), Methods in Social Neuroscience. Guilford Press, New York, pp. 17–44.

Selby, C. 1999. Interference in Immunoassay. Annals of Clinical Biochemistry 36, 704-721. doi: 10.1177/000456329903600603

Soldin, S.J., Soldin, O.P., 2009. Steroid Hormone Analysis by Tandem Mass Spectrometry. Clin. Chem. 55, 1061–1066. doi:10.1373/clinchem.2007.100008

Stanczyk, F.Z., Cho, M.M., Endres, D.B., Morrison, J.L., Patel, S., Paulson, R.J., 2003. Limitations of direct estradiol and testosterone immunoassay kits. Steroids 68, 1173–1178. doi:10.1016/j.steroids.2003.08.01

Stanton, S. J. 2016. The role of testosterone and estrogen in consumer behavior and social & economic decision making: A review. Horm and Beh. 92, 155-163.

Star-Weinstock, M., Williamson, B.L., Dey, S., Pillai, S., Purkayastha, S., 2012. LC-ESI-MS/MS analysis of testosterone at sub-picogram levels using a novel derivatization reagent. Anal. Chem. 84, 9310–9317. doi:10.1021/ac302036r

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. Psych. Bull. 87, 245-251.

Taieb, J., Benattar, C., Birr, A.S., Lindenbaum, A., 2002. Limitations of steroid determination by direct immunoassay. Clin. Chem. 48, 583–585.

Taieb, J., Mathian, B., Millot, F., Patricot, M.C., Mathieu, E., Queyrel, N., Lacroix, I., Somma-Delpero, C., Boudou, P., 2003. Testosterone measured by 10 immunoassays and by isotope-dilution gas chromatography-mass spectrometry in sera from 116 men, women, and children. Clin. Chem. 49, 1381–1395. doi:10.1373/49.8.1381

Tate, J.,Ward, G. 2004. Interferences in immunoassay. Clin. Biochem. Rev. 32, 105-120.

Taylor, A.E., Keevil, B., Huhtaniemi, I.T., 2015. Mass spectrometry and immunoassay: How to measure steroid hormones today and tomorrow. Eur. J. Endocrinol. 173, D1-12. doi:10.1530/EJE-15-0338

Toone, R.J., Peacock, O.J., Smith, A.A., Thompson, D., Drawer, S., Cook, C., Stokes, K.A., 2013. Measurement of steroid hormones in saliva: Effects of sample storage condition. Scand. J. Clin. Lab. Investig. 73, 615–621. doi:10.3109/00365513.2013.835862

Turpeinen, U., Hämäläinen, E., Haanpää, M., Dunkel, L., 2012. Determination of salivary testosterone and androstendione by liquid chromatography-tandem mass spectrometry. Clin. Chim. Acta 413, 594–599. doi:10.1016/j.cca.2011.11.029

Valdes, R., Jortani, S.A., 2002. Unexpected suppression of immunoassay results by cross-reactivity: Now a demonstrated cause for concern. Clin. Chem. 48, 405–406.

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. Psychoneuroendocrinology 71, 180–188. doi:10.1016/j.psyneuen.2016.05.022

Whembolua, G.L.S., Granger, D.A., Singer, S., Kivlighan, K.T., Marguin, J.A., 2006. Bacteria in the oral mucosa and its effects on the measurement of cortisol, dehydroepiandrosterone, and testosterone in saliva. Horm. Behav. 49, 478–483. doi:10.1016/j.yhbeh.2005.10.005

Wierman, M.E., Auchus, R.J., Haisenleder, D.J., Hall, J.E., Handelsman, D., Hankinson, S., Rosner, W., Singh, R.J., Sluss, P.M., Stanczyk, F.Z., 2014. Editorial: The new instructions to authors for the reporting of steroid hormone measurements. Endocr. Rev. 35, 849. doi:10.1210/er.2014-1067

Yasuda, M., Honma, S., Furuya, K., Yoshii, T., Kamiyama, Y., Ide, H., Muto, S., Horie, S., 2008. Diagnostic significance of salivary testosterone measurement revisited: using liquid chromatography/mass spectrometry and enzyme-linked immunosorbent assay. J. Mens. Health 5, 56–63. doi:10.1016/j.jomh.2007.12.00

Zilioli, S., Mehta, P.H., Watson, N. V., 2014. Losing the battle but winning the war: Uncertain outcomes reverse the usual effect of winning on testosterone. Biol. Psychol. 103, 54–62. doi:10.1016/j.biopsycho.2014.07.02

Tables

DESCRIPTIVE STATISTICS AND CORRELATIONS

Overall	***M (pg/mL)***	***SD***	**MS**	**Salimetrics EIA High Meth Var**	**Salimetrics EIA Low Meth Var**
MS	48.49	53.15			
Salimetrics EIA High Meth Var	101.78	64.11	.80** [.72, .86]; *p*<.001		
Salimetrics EIA Low Meth Var	98.66	70.00	.55** [.40, .68]; *p*<.001	.79** [.70, .86]; *p*<.001	
Salimetrics EIA Average	100.22	63.47	.71** [.59, .80]; *p*<.001	.94** [.91, .96]; *p*<.001	.95** [.93, .97]; *p*<.001
Males					
MS	96.52	47.21			
Salimetrics EIA High Meth Var	154.48	60.39	.54** [.28, .72]; *p*<.001		
Salimetrics EIA Low Meth Var	145.95	77.23	.17 [-.14, .45]; *p*=.286	.66** [.44, .80]; *p*<.001	
Salimetrics EIA Average	150.21	62.73	.36* [.07, .60]; *p*=.018	.89** [.80, .94]; *p*<.001	.93** [.88, .96]; *p*<.001
Females					
MS	11.14	9.42			
Salimetrics EIA High Meth Var	60.79	25.15	.66** [.47, .79]; *p*<.001		
Salimetrics EIA Low Meth Var	61.88	31.91	.03 [-.24, .29]; *p*=.838	.61** [.40, .75]; *p*<.001	
Salimetrics EIA Average	61.34	25.61	.34* [.08, .56]; *p*=.012	.87** [.78, .92]; *p*<.001	.92** [.87, .95]; *p*<.001
Upper 50%					
MS	88.43	49.28			
Salimetrics EIA High Meth Var	144.08	63.10	.65** [.44, .79]; *p*<.001		
Salimetrics EIA Low Meth Var	134.41	78.84	.31* [.03, .55]; *p*=.032	.72** [.55, .83]; *p*<.001	
Salimetrics EIA Average	139.24	65.87	.49** [.25, .68]; *p*<.001	.91** [.84, .95]; *p*<.001	.94** [.90, .97]; *p*<.001
Lower 50%					
MS	8.56	4.81			
Salimetrics EIA High Meth Var	59.49	25.92	.62**[.41, .77]; *p*<.001		
Salimetrics EIA Low Meth Var	62.91	32.84	.19[-.10, .45]; *p*=.186	.60** [.38, .76]; *p*<.001	
Salimetrics EIA Average	61.20	26.32	.43**[.16, .63]; *p*=002	.87** [.77, .92]; *p*<.001	.92** [.86, .95]; *p*<.001

* indicates $p < .05$. ** indicates $p < .01$.

Table 1. *M* (in pg/mL) and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. Key: MS- Liquid chromatography tandem mass spectrometry; Salimetrics EIA High Meth Var- Salimetrics EIAs in the high method variance comparison; Salimetrics EIA Low Meth Var - Salimetrics EIAs in the low method variance comparison; Salimetrics EIA Average- Testosterone concentrations from Salimetrics EIAs averaged across both comparisons. Note: Welker et al. (2016) reported descriptive statistics and correlations for Salimetrics EIAs in the low method variance comparison and MS in Table 1 across all 99 participants in the study. However, we report statistics after the exclusion of outliers across all measurement methods and measurement instances (Overall n=96; Males= 42; Females= 54; Upper 50%= 48; Lower 50%= 48).

FIXED BIAS ESTIMATES

	Salimetrics EIA High Meth Var		Salimetrics EIA Low Meth Var		Salimetrics EIA Average	
	Fixed bias estimates	**EIAs vs. MS**	**Fixed bias estimates**	**EIAs vs. MS**	**Fixed bias estimates**	**EIAs vs. MS**
Overall	53.29 [45.55, 61.03]	$F(1, 95)=186.90$, $p<.001$, $\eta_p^2=.66$	50.17 [37.98, 62.35]	$F(1, 95)=66.83$, $p<.001$, $\eta_p^2=.41$	51.73 [42.53, 60.92]	$F(1, 95)=124.70$, $p<.001$, $\eta_p^2=0.57$
Males	57.96 [41.44, 74.47]	$F(1, 41)=50.22$, $p<.001$, $\eta_p^2=0.55$	49.42 [23.42, 75.43]	$F(1, 41)=14.73$, $p<.001$, $\eta_p^2=0.26$	53.69 [33.94, 73.44]	$F(1, 41)=30.14$, $p<.001$, $\eta_p^2=0.42$
Females	49.66 [44.13, 55.19]	$F(1, 53)=324.20$, $p<.001$, $\eta_p^2=0.86$	50.74 [41.73, 59.75]	$F(1, 53)=127.50$, $p<.001$, $\eta_p^2=0.71$	50.20 [43.62, 56.78]	$F(1, 53)=234.40$, $p<.001$, $\eta_p^2=0.82$
Upper 50%	55.65 [41.43, 69.86]	$F(1, 47)=62.01$, $p<.001$, $\eta_p^2=0.57$	45.98 [23.06, 68.90]	$F(1, 47)=16.29$, $p<.001$, $\eta_p^2=0.26$	50.81 [33.50, 68.13]	$F(1, 47)=34.86$, $p<.001$, $\eta_p^2=0.43$
Lower 50%	50.93 [44.18, 57.68]	$F(1, 47)=230.60$, $p<.001$, $\eta_p^2=0.83$	54.35 [44.99, 63.72]	$F(1, 47)=136.30$, $p<.001$, $\eta_p^2=0.74$	52.64 [45.48, 59.80]	$F(1, 95)=218.80$, $p<.001$, $\eta_p^2=0.82$

Table 2. Fixed bias (or magnitude of inflation) in testosterone concentrations obtained from Salimetrics EIAs in the high method variance comparison (Salimetrics EIA High Meth Var), Salimetrics EIAs in the low method variance comparison (Salimetrics EIA Low Meth Var), and testosterone concentrations from Salimetrics EIAs that were averaged across both comparisons (Salimetrics EIA Average), relative testosterone from liquid chromatography tandem mass spectrometry (MS). We also report inferential statistics comparing testosterone concentrations from EIAs with MS concentrations. Values in square brackets indicate the 95% confidence interval for each fixed bias estimate.

DEMING REGRESSION

	Salimetrics EIA High Meth Var	Salimetrics EIA Low Meth Var	Salimetrics EIA Average
Deming Regression: Overall			
Intercept	40.59 [26.86, 54.32]	19.84 [-16.21, 55.88]	38.04 [19.37, 56.70]
Slope	1.26 [1.07, 1.45]	1.63 [1.12, 2.13]	1.28 [1.02, 1.54]
Deming Regression: Males			
Intercept	3.47 [-80.55, 87.49]	-457.05 [-1709.95, 795.85]	-49.81 [-232.32, 132.70]
Slope	1.56 [0.78, 2.35]	6.25 [-5.44, 17.93]	2.07 [0.37, 3.77]
Deming Regression: Females			
Intercept	18.85 [1.29, 36.41]	-1144.02 [-16473.92, 14185.88]	-17.28 [-96.33, 61.78]
Slope	3.77 [2.56, 4.97]	108.28 [-946.58, 1163.14]	7.06 [1.62, 12.50]
Deming Regression: Upper 50%			
Intercept	15.03 [-36.75, 66.81]	-168.96 [-483.65, 145.74]	-16.26 [-108.80, 76.29]
Slope	1.46 [0.95, 1.97]	3.43 [0.31, 6.55]	1.76 [0.84, 2.67]
Deming Regression: Lower 50%			
Intercept	-13.07 [-44.09, 17.95]	-231.34 [-735.78, 273.10]	-45.31 [-121.81, 31.19]
Slope	8.48 [5.31, 11.65]	34.38 [-17.12, 85.88]	12.44 [4.63, 20.26]

Table 3. Deming regression comparing testosterone concentrations obtained from Salimetrics EIAs in the high method variance comparison (Salimetrics EIA High Meth Var) and Salimetrics EIAs conducted in the low method variance comparison (Salimetrics EIA Low Meth Var) with liquid chromatography tandem mass spectrometry (MS). For the Deming Regression, we compared the identity line (intercept = 0, slope = 1) with the data from Salimetrics EIAs across the two comparisons. We also report Deming regression for testosterone concentrations from Salimetrics EIAs that were averaged across both comparisons (Salimetrics EIA Average). Values in square brackets indicate the 95% confidence interval for the slopes and intercepts.

Figures

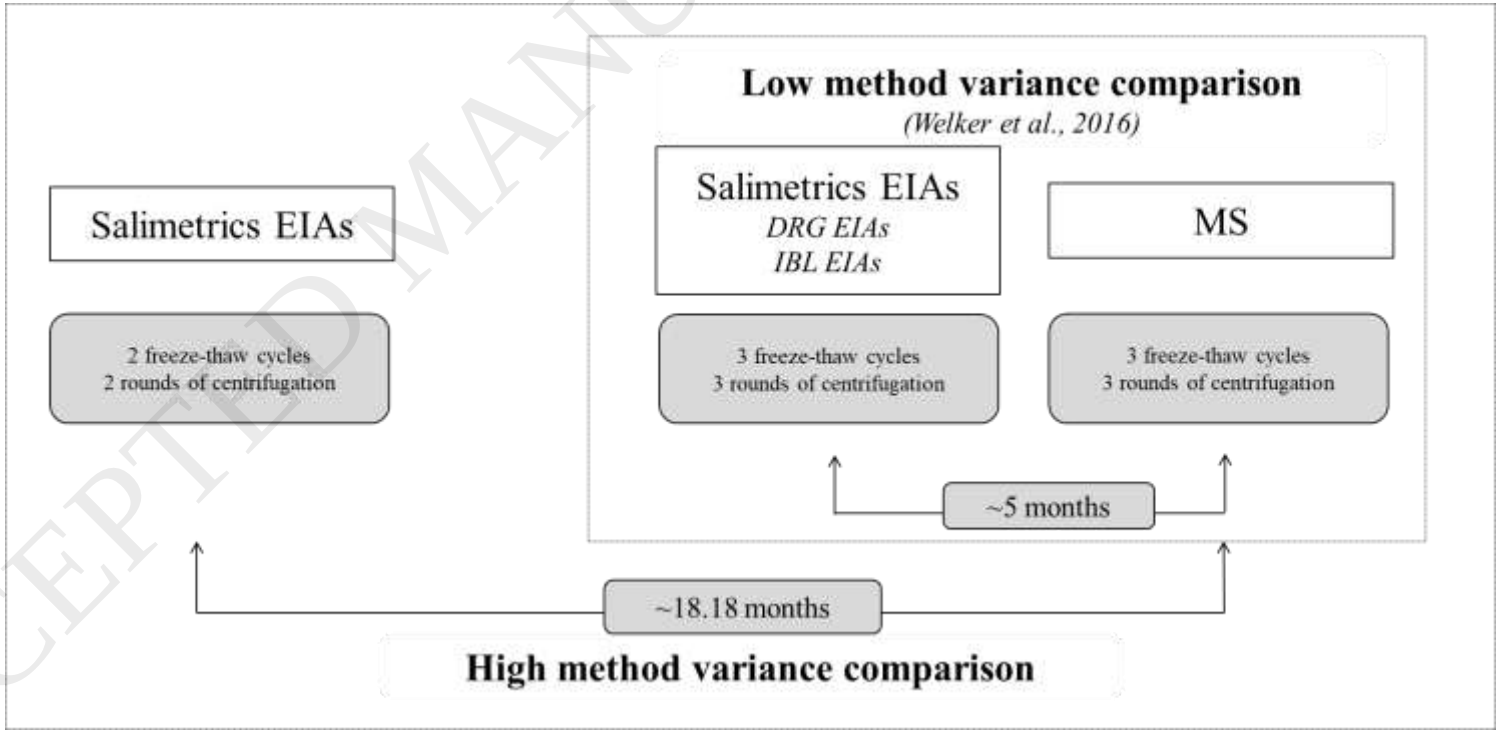

Figure 1. Salimetrics EIAs and MS conducted across high and low method variance comparisons. The interim storage time of 13.18 (*SD*=1.79) months between both sets of EIAs was calculated as the average difference between the dates of assay for each sample across both Salimetrics EIAs. Samples were stored at -80°C prior to being processed at each measurement instance. Key: EIA- Enzyme Immunoassays; MS- liquid chromatography tandem mass spectrometry.

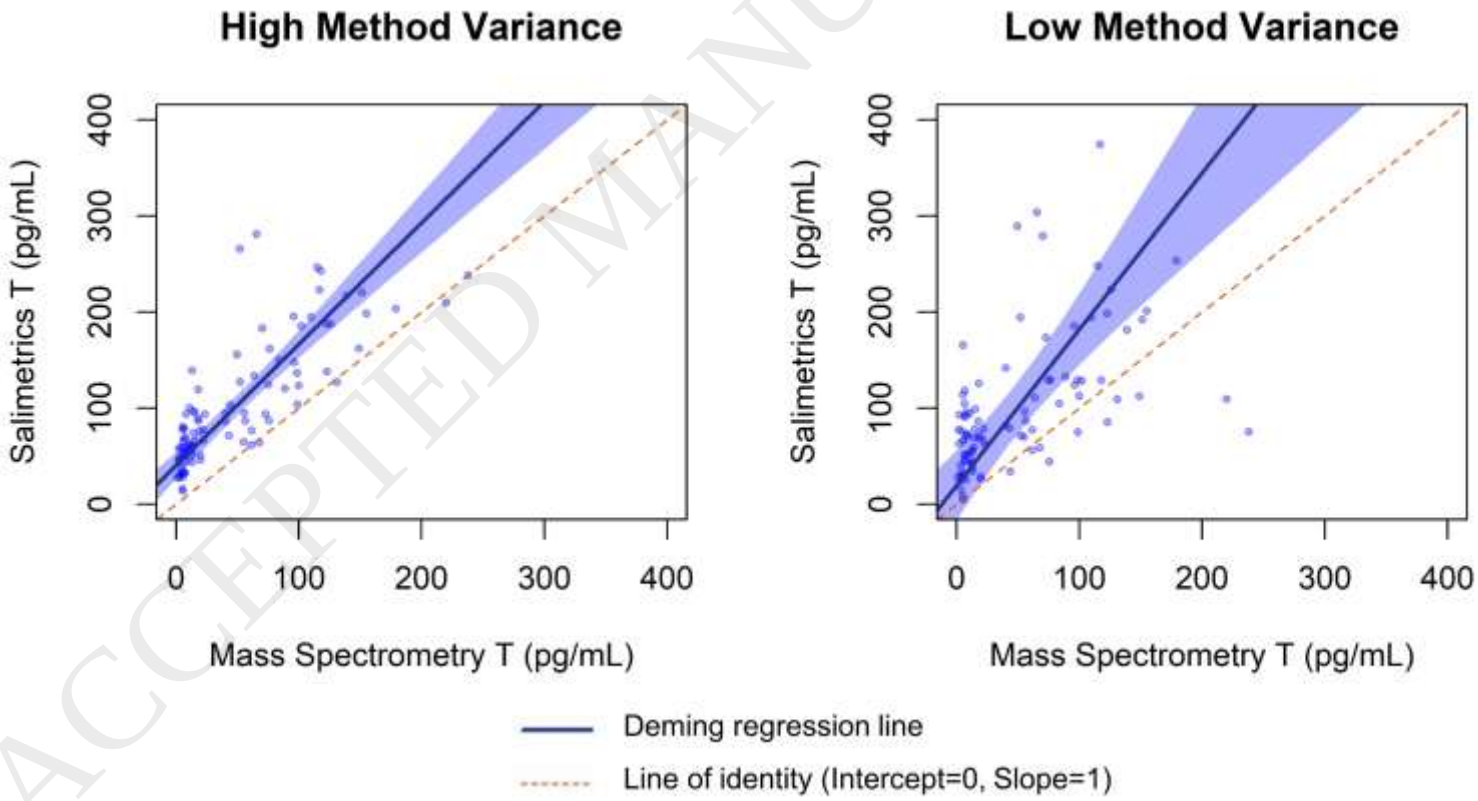

Figure 2. Deming regression between testosterone (T) concentrations from Salimetrics EIAs and liquid chromatography tandem mass spectrometry (MS) in the high method variance comparison (Left Panel) and low method variance comparison (Right Panel). Note: The dashed line represents the line of identity (if Salimetrics EIAs and MS testosterone concentrations were equivalent), whereas the blue line represents the Deming regression line.