Hormone measurement in social neuroendocrinology:

A comparison of immunoassay and mass spectrometry methods

Oliver C. Schultheiss, Gelena Dlugash & Pranjal H. Mehta

Please cite as:

Schultheiss, O. C., Dlugash, G., & Mehta, P. H. (in press). Hormone measurement in social neuroendocrinology: A comparison of immunoassay and mass spectroscopy methods. In O. C. Schultheiss & P. H. Mehta (Eds.), *Routledge international handbook of social neuroendocrinology*. Abingdon, UK: Routledge.

1 A primer on hormones

As a field, social neuroendocrinology almost always involves the measurement of hormones in some manner, be it as a predictor or correlate of physiology, affect, cognition, and behavior, or be it as an outcome of experimental manipulations. This applies regardless of whether the research is conducted on non-human animals or humans. Hormones differ regarding the ease and validity with which they can be measured due to their chemical structure, the properties that result from it, and their concentrations. Three large groups of hormones can be distinguished: (1) hormones that are derived from single amino acids, such as adrenaline, noradrenaline, dopamine, or the thyroid hormones; (2) hormones that are peptides and proteins, such as oxytocin or growth hormone; and (3) steroid hormones, derived from cholesterol, such as cortisol, progesterone, or testosterone (Ojeda & Kovacs, 2012).

While the first two groups are comprised of hormones that due to either their large molecular structure or their polarity cannot permeate the barrier of cell membranes, steroids, the third group, can do that. This means that steroid hormones, but not amines or peptides and proteins, can travel to and affect the brain (e.g., Bäckström, Carstensen, & Södergard, 1976), even though steroid hormones are released by glands in the periphery of the body (outside the central nervous system). To some extent, this eliminates the issue of whether central levels of steroids correspond to peripheral levels, which is a thorny problem in research on non-steroid hormones (e.g., Leng & Ludwig, 2016; McCullough, Churchland, & Mendez, 2013). Another crucial difference between these three groups of hormones is that amines, peptides, and proteins are broken down by enzymes and are therefore difficult to measure unambiguously (e.g., McCullough et al., 2013). In contrast, steroids represent rather robust hormones that need to be cleared from the bloodstream by the liver. This structural stability increases their attractiveness for researchers. But it also comes at the cost of a very small size that can pose its own set of challenges for measurement, as we will show below. Finally, because steroids can be found in comparable concentrations in virtually all body compartments and do not degrade quickly, they provide researchers with many substrates from which to assay these hormones. As a consequence, steroids are not only routinely measured in blood, but can also be assessed from urine (Campbell, 1994), feces (Palme, 2005), hair (Stalder & Kirschbaum, 2012), and saliva (Gröschl, 2008; Schultheiss & Stanton, 2009), with the latter showing good convergence with measurements in blood (e.g., Keevil, MacDonald, Macdowall, Lee, & Wu, 2014). For these reasons, they represent a highly attractive target for social neuroendocrinologists, particularly when researchers want to avoid the use of invasive sample collection methods such as venipuncture. For research with human participants, salivary steroid assessment has become the method of choice in many laboratories.

While salivary steroid assessment provides a rather accessible window into endocrine processes and their relationship to brain functions and behavior, considerable care and forethought need to go into all steps of the measurement process. For instance, measurement results can be biased by the devices used for stimulating saliva flow for easier sample collection (e.g., Gröschl & Rauh, 2006; Schultheiss, 2013), by improper storage (e.g., Gröschl, Wagner, Rauh, & Dörr, 2001; Whembolua, Granger, Singer, Kivlighan, & Marguin, 2006), and by the pre-processing of the samples themselves (e.g., Durdiakova, Fabryova, Koborova, Ostatnikova, & Celec, 2013). Schultheiss and Stanton (2009) and Schultheiss, Schiepe, and Rawolle (2012) provide detailed discussions and recommendations regarding these issues.

However, despite social neuroendocrinologists' awareness of these issues, the field as a whole has not paid as much attention as it should to another, more consequential problem associated with salivary steroid assessment. For quite some time, clinical endocrinologists, who are by profession particularly concerned about valid, reliable measurement from which medical diagnoses can be derived, have voiced concerns about the validity of immunoassay methods for the assessment of steroids, particularly in the low ranges that is typical of salivary steroids (e.g., Herold & Fitzgerald, 2003; Matsumoto & Bremner, 2004; Stanczyk et al., 2003; Taieb, Benattar, Birr, & Lindenbaum, 2002). By and large, these concerns have not received attention within the field of human social neuroendocrinology – many researchers continue to use steroid immunoassays, encouraged by the proliferation of specialized commercially available assay kits for salivary hormones and perhaps trusting that a technique that had worked well for other parts of endocrinology for so long could not be the wrong approach for their own area of research. Social neuroendocrinologists were largely unaware that they increasingly fell out of step with the developments in clinical endocrinology, which by 2013 even led to a de-facto ban by the *Journal of Clinical Endocrinology & Metabolism*

on papers based on immunoassays and the stipulation that mass spectrometry should be used instead (Handelsman & Wartofsky, 2013; but see Wierman et al., 2014). To explain how this situation arose and what practical consequences it may have for the measurement and interpretation of salivary steroids, we will first provide some background on immunoassay and mass spectrometry approaches to hormone assessment and then compare the results that these methods give for various steroids, with an emphasis on testosterone for illustrative purposes. We will also discuss the convergence between steroid measurements made with both methods and draw some conclusions about the future of steroid hormone assessment in social neuroendocrinology.

2 Immunoassays

Immunoassays are based on organisms' immune responses to exogenous molecules and compounds entering the body. The body's immune system analyzes the exogenous matter (also called antigen) and manufactures antibodies that fit the specific molecular surface of the antigen well enough to selectively attach to it and incapacitate it or mark it as a target for other components of the immune system. Antibodies can be raised in and harvested from animals (usually rabbits), resulting in *polyclonal* antibodies – that is, antibodies that are produced by different bone marrow cells and therefore represent a mix of heterogeneous antibodies, targeting different surface sites of the antigen. In contrast, *monoclonal* antibodies are raised in specific hybrid cell lines that can produce an unlimited and highly uniform supply of a homogenous antibody, targeting a specific antigen site (Kane & Banks, 2000).

Antibodies can be raised against specific target analytes, such as testosterone or oxytocin, that one would like to measure. This is the basic idea behind the immunoassay approach. However, most mammals do not show a strong immune system response against lipids or very small molecules (steroids are both) or substances that their own body produces

in exactly the same way, too. This is why for immunization purposes the targeted analyte is usually conjugated to a protein to increase its immunogenicity, that is, its capacity to elicit an immune system response leading to the production of antibodies (Kane & Banks, 2000). For small molecules like steroid hormones, this means that some of the few binding sites that they have are occupied by the conjugated protein and are therefore not available for detection by the immune system. Because steroid hormones and their various metabolites differ only in very minute details, conjugation to a binding site that differentiates between steroid variants can therefore be one reason for an antibody's lack of specificity. And because antibodies are tailored to recognize the compound of target and conjugated protein, they may also have a binding affinity to just the protein that was used and other proteins with similar partial binding properties. This is important to keep in mind when considering the validity limits of immunoassays in endocrinology (see also Gosling, 2000, pp. 14 & 15).

Having an antibody that specifically binds to a certain substance is per se insufficient for measuring the substance. Instead of one thing that is invisible and unweighable (the targeted substance), now we have two such things – the target and the antibody. Yalow and Berson (1960) solved this problem in the following way: They used a fixed amount of antibody and added a fixed amount of antigen, with a label attached that could be measured, plus a sample with an unknown amount of the unlabeled antigen. This resulted in a competition of the labeled antigen and the unlabeled antigen for antibody binding. The greater the relative amount of unlabeled antigen, the less of a signal the labeled antigen could emit after the competition was stopped. Conversely, the smaller the relative amount of unlabeled antigen, the greater the signal emitted by the labeled antigen. The meaning of variations in the signal that different amounts of labeled antigen emit, reflecting different gradations of competitive displacement by unlabeled antigen, can be derived by including

samples with known concentrations of unlabeled antigen – the so-called standards or calibrators of an assay.

The majority of immunoassays used in social neuroendocrinology these days are based on this competitive binding principle originally introduced by Yalow and Berson (1960). Yalow and Berson used a radioisotope as label and a gamma counter to measure its signal, thus creating the *radio*immunoassay (RIA) technique that is still in use today. Later, researchers developed *enzymatic* immunoassays (EIA) that replaced the radioactive label with one whose color (or sometimes luminescence) indicated the degree of antigen present in the assay. Here, the measurement is done through photometers detecting light at a wavelength corresponding to the enzyme's color. The deeper the coloration, the stronger is the labeled signal and hence the less unlabeled antigen is present. So whether it is a RIA or an EIA, immunoassays do not measure the naturally occurring hormone itself, but its effect on the concentration of the labeled hormone of a fixed quantity that it competes with.

The quality and performance characteristics of an immunoassay can be gauged from several indicators. Its *validity* is usually assessed from its ability to measure the concentration of samples with known concentrations as accurately as possible (*recovery*). For instance, if a sample with known concentrations (e.g., by creating the sample through first weighing and then dissolving the analyte in a liquid matrix) of 1 and 5 ng/mL of cortisol gives readings of 1.05 and 4.85 ng/mL by cortisol immunoassay, then recovery is 105% for the former and 97% for the latter sample and thus very good. Recovery thus represents an instance of a causal link between values given by the measure and systematic variations of the measured target, a key concept in validity theory (Borsboom, Mellenbergh, & van Heerden, 2004).

An immunoassay's validity can also be gauged from its *specificity* for the targeted analyte; that is, how well it can discriminate the target from other analytes with a similar

structure. For this purpose, samples are created that not only contain known amounts of the target analyte (e.g., cortisol), but also of analytes that the antibody might bind to, too, because they are structurally similar to the target analyte (e.g., other steroids). While recovery should be close to 100% for the target analyte, it should be close to 0% for all other analytes. Determining specificity usually requires a lot of work and biochemical knowhow and intuition about which substances might cross-react with the antigen. Assay users rarely do their own specificity checks and instead rely on relevant information by the manufacturer. However, it is often unclear whether manufacturer-supplied information encompasses all substances that could conceivably bind to the antibody. This means that although the list may give the assay a clean bill of specificity (e.g., no cross-reactivity with other steroids), this may not be the whole story, as other substances also contained in a given sample type (e.g., saliva) may substantially cross-react with the antibody. Thus, for highly complex biological fluids like serum or saliva, lists of substances for which cross-reactivity has been determined can hardly ever be exhaustive. And one should also keep in mind that the information supplied by the manufacturer (a) may try to present a flawless picture of the assay and (b) may be based on antibodies or assay procedures at one stage on assay manufacturing and may have changed with subsequent alterations to assay ingredients and manufacturing. In the terms of psychological validity theory, specificity can be viewed as an instance of discriminant validity (e.g., Campbell & Fiske, 1959).

A third aspect of an assay's validity is its *sensitivity*, indicating the minimal concentration it can discern from a zero-concentration measurement. Most frequently, sensitivity is operationalized via the limit of detection (LOD), defined as the mean signal for a zero-concentration measurement, plus 3 times its SD, or the limit of quantification (LOQ), derived in the same way, but with 10 times the SD. Thus, if a cortisol assay has a SD of 0.002 ng/mL for a zero-concentration standard sample, its LOD would be 3 x 0.002 ng/mL = 0.006

ng/mL and its LOQ would be 10 x 0.002 ng/mL = 0.020 ng/mL. Note that this determination of sensitivity can be criticized for many of the same reasons that null-hypothesis testing has been criticized more generally (e.g., Nickerson, 2000). It does not positively establish a lower limit for the measurement of a non-zero-concentration sample. And its results depend on measurement error, whose magnitude is in turn influenced by the number of zero-concentration measurements that were done (10 measurements will give a lower SD and hence lower LOD and LOQ than 5 measurements).

An assay's reliability is gauged via its precision, assessed by measuring the same sample twice or more. The SD of these repeated measurements is then divided by their mean and multiplied by 100, yielding the coefficient of variation, or CV, in %. For instance, if the same saliva sample gives values of 1.03 ng/mL and 0.97 ng/mL in two successive measurements, the SD is 0.042 ng/mL and the mean is 1.00 ng/mL; hence (0.042/1.00) x 100 = 4.2%. To determine the CV of all samples measured in one and the same assay, all CVs are averaged to arrive at the *intra-assay CV*. However, an assay's precision may also vary across repeated performances of this assay, such as when one assay is run today and another next week. Will independently conducted assays give the same values for the same samples? To determine the *inter-assay CV*, the same samples are included in two or more assays that constitute an integrated measurement series (e.g., for salivary cortisol of all participants in a large panel study), and results are converted to a CV according to the same formula as the intraassay CV. Note that intra- and interassay CVs again depend on measurement error and thus on the total number of measurements. Including more measurements will yield lower CVs.

Note that some of these measures of validity and reliability also apply to MS. For instance, research using MS for the assessment of hormones frequently reports the limit of detection (or quantification), recovery, and intra- and interassay CV (e.g., Gao, Stalder, &

Kirschbaum, 2015). We will return to some of these issues in our next section, which describes the principles of the MS approach.

3 Mass spectrometry

(Ultra-)high pressure liquid chromatography (HPLC or simply LC), coupled with mass spectrometry (MS), has become the "gold standard" of hormone assessment. Depending on the measurement goals, components of the LC-MS system can be adjusted to better accommodate a particular analyte or matrix. In the following, we will focus only on the most important and common components of various LC-MS systems as they are applied to hormone assessment (see Gouveia et al., 2013; McDonald, Matthew, & Auchus, 2011).

The MS component is based on two physical properties of molecules: mass and charge. In the mass analyzer, the core of MS, ions are accelerated in a vacuum through an electromagnetic field towards the detector. This field can only be passed by charged molecules with a specific mass-to-charge-ratio, which can be set by the user. Compounds with a different ratio get thrown off on their way through the mass analyzer and therefore do not reach the detector. MS thus measures a target analyte directly and with high specificity.

Accurately measuring a given analyte in a given substance (often called "matrix") via MS is challenging for several reasons. First, complex matrices, like serum, saliva, or urine, contain various salts and different nonvolatile compounds (e.g., Chiu et al., 2010; Selby, 1999). In an MS system, these can interfere with measurement by suppressing ionization. As a second challenge, analytes with an identical mass-to-charge ratio (= isobars), like testosterone and epitestosterone or 17-α-estradiol and 17-β-estradiol, cannot be differentiated by MS itself. A third challenge is posed by the sheer number and variety of compounds present in saliva or serum. These would let the vacuum collapse upon entering the mass analyzer.

To resolve these issues, MS is coupled to a chromatography system (either liquid chromatography, LC, for nonvolatile compounds, or gas chromatography, GC, for volatile compounds, as when fluids are vaporized by heating), which (a) purifies and (b) separates analytes by their chemical and physical properties and lets them enter MS separately. This results in an enormous improvement of measurement sensitivity and specificity. Thus, the coupling of LC (GC) and MS is the reason why this measurement approach can detect analytes with unrivaled specificity.

Generally, a LC-MS or GC-MS system can be divided into four main components: chromatography, ion source, mass analyzer, and detector (Fig. 1).

<FIGURE 2.1 HERE>

The first part of the system is chromatography. GC features excellent resolution and separation of over 65 compounds in one single run and is therefore frequently used in steroid metabolomics of urine samples. However, because GC is only applicable to volatile and nonpolar compounds, neither of which steroids and their metabolites belong to, a time- and labor-consuming derivatization step must be executed first. *Derivatization* refers to a chemical reaction in which unwanted nonvolatile and polar properties of analytes are modified by conjunction of special derivates to the analyte, making it nonpolar and volatile. Additionally, derivatization can also enhance MS performance, resulting in lower limits of detection. Therefore, derivatization can also be used in LC assays, but only as an option.

Mostly, LC has become the chromatography method of choice. It is easy to conduct and does not generally require derivatization. It offers both a high throughput of samples and the detection of up to a dozen steroids in one run. Therefore, it is often used for analyzing steroids in blood and saliva samples. By developing efficient interfaces for transferring

analytes that have been separated by LC to the vacuumed gas phase of MS, LC can now easily be coupled to MS.

The ion source is the second essential part of this system, as only ionized and gaseous components can enter the vacuumed mass analyzer. The most common ion sources are electric spray ionization (ESI), atmospheric pressure chemical ionization (APCI), and atmospheric pressure photoionization (APPI). All of these techniques can be run in a positive or negative ion mode for detecting positively or negatively charged ions. Also, they all are suitable for measuring various steroids simultaneously.

Both ESI and APCI in positive mode are often used for measuring steroid profiles (Gao, Stalder, & Kirschbaum, 2015; Gaudl et al., 2016). However, because of its gentle ionization, ESI is the most frequently used ion source. It provides high measurement sensitivity with less analyte fragmentation, but is also vulnerable to whatever matrix effects remain after LC, which, as we have pointed out above, can cause ionization and suppress evaporation. Nevertheless, ESI is particularly suitable for assessing testosterone, because it comes with a very low LOD and can therefore detect testosterone even in populations that have low circulating levels of this hormone (e.g., women and children)

In contrast to ESI, APCI is less impaired by ion suppression caused by matrix effects, but generates more analyte fragmentation and can be therefore less sensitive. APPI is a newer ion source, which was designed for a soft ionization of nonpolar compounds and steroids in particular. Using this technique combined with a negative ion mode has resulted in a huge improvement of sensitivity in estradiol detection, a hormone that is notoriously difficult to assess due to its low levels.

Nowadays, modern mass spectrometers contain several ion sources, such as a combination of ESI and APCI, and use these sources simultaneously in one single run in

positive as well as in negative ion modes. Thus the mass spectrometer can be aligned to any type of analyte in any type of sample.

After ionization, analytes enter the vacuum of a mass analyzer. The most commonly used type of mass analyzer in steroid endocrinology is the triple quadrupole mass analyzer (Triple Quad or MS/MS), which is well suited for the quantification of small molecules. It consists of three quadrupoles, each of which consists of two pairs of electrically charged metal rods that select or fragment ions.

<FIGURE 2.2 HERE>

More specifically, the first quadrupole filters the analyte by its mass-to-charge ratio. Analytes that correspond to this ratio are the only ions that can travel on a stable path through the electromagnetic field towards second quadrupole, the so-called collision cell. Here, the previously selected ion (precursor ion), which frequently do not suffice for identifying the analyte, get fragmented in a gas. This results in characteristic, more identifiable fragments, which then travel to the third quadrupole, where they undergo further mass-to-charge-ratio selection. Finally, the ions hit a detector, usually an electron multiplier.

One special feature of all mass spectrometers is the option of stable isotope analysis by using an internal standard (IS). Typically, the IS is the same type of molecule as the target analyte. Crucially, though, the IS is labeled with a stable isotope like deuterium or $C^{13.}$ It therefore has the same chemical properties as the analyte, except for its molecular mass. Thus, both compounds can be detected and quantified by MS separately. The IS is added to all samples and calibrators before sample preparation and therefore before entering the LC-MS procedure. During sample preparation, ionization, and quantification, the IS undergoes

the same losses as the target analyte and is therefore an internal control *for every single sample*. Usually, MS quantification is computed by analyte-to-IS ratio.

4 Comparison of mass spectrometry with immunoassays

The key difference between MS and immunoassays lies in analyte detection. While MS exploits stable physical properties of molecules for measurement, immunoassays are based on more fickle biochemical reactions, which can be affected by any little fluctuation in external factors like enzyme purity or activity, temperature, ionic strength, pH, and so on (see Selby, 1999). These factors have a big impact on assay validity; and more so for enzymatic immunoassays than for radioimmunoassays, which use a physical characteristic of molecules (i.e., radiation) for detection.

Another major problem of all immunoassay techniques is a lack of specificity due to cross-reactions between antibody and matrix compounds that are structurally similar to the targeted analyte. These interfering bindings cause immunoassays to overestimate analytes, sometimes up to a factor of three. Welker et al. (2016) showed this effect for testosterone in saliva, comparing immunoassay performance to LC-MS. They observed that the overestimation was higher when testosterone was lower (e.g., in the female range). Bae et al. (2016) reported similar results for salivary cortisol. Relative to LC-MS, immunoassay overestimated cortisol in saliva. The authors attributed this effect to the different standardization of immunoassays relative to MS and pointed out that this issue can be resolved by standardizing immunoassay s against MS. While this is certainly true, they also reported the more problematic observation that at low cortisol concentrations (< 1.8 ng/mL), there was substantial cross-reactivity with cortisone (up to 254% recovery). The authors also observed interference by the protein α-amylase. The latter effect is particularly troubling, because α-amylase is proportionally the most important protein in saliva and it appears to affect the measurement of cortisol at levels that are in the typical salivary range. Most

importantly, these findings provide direct evidence that a protein can bias a steroid immunoassay, and particularly at levels that are typical of salivary steroids. RIAs of gonadal steroids also suffer from cross-reactivity to other substances and can yield higher estimates in comparison to MS (e.g., Hsing et al., 2007). However, it is important to keep in mind that there can be substantial differences between immunoassays in the extent to which they are susceptible to overestimation and cross-reactivity. A lot hinges on proper standardization and antibody design (Baker, 2015).

4.1 Testosterone assessment

To illustrate the issues involved in measuring a steroid in different matrices (blood, saliva) and with different methods (LC-MS, RIA, EIA), we have compiled data from studies that looked at testosterone in combined samples of women and men (see Table 1). This ensures that whatever processing steps and methods were used, they were held constant within a study and thus across sample gender. Although the data listed in Table 1 (by no means) represent a comprehensive overview of all published studies, some consistent observations can be made. First, the gender difference that is typical of testosterone (Stanton, 2011) is always highest for LC-MS, followed by RIA and then EIA. For serum, this difference is particularly dramatic, as LC-MS indicates an up to 22-fold difference, whereas for RIA and EIA the difference is only about 8- to 9-fold. For saliva, the difference is smaller, with LC-MS and RIA providing similar estimates of a 5- to 6-fold difference. This suggests that perhaps due to gender differences in protein binding of testosterone, the gender differences in testosterone in saliva, which contains only the free, unbound fraction of this hormone, is considerably less pronounced than in serum. But here, too, EIAs underestimate the difference as only about 2-fold. Second, for serum, which contains both protein-bound and unbound (or free) testosterone, male concentrations are very comparable across studies and methods. However, female testosterone is overestimated by a factor of two by all

immunoassays relative to LC-MS. For salivary EIAs we specifically looked at the Salimetrics testosterone assay, which is frequently used in research with human participants, and, for RIAs, the equally popular Coat-a-Count RIA by Siemens (previously manufactured by Diagnostic Products Corporation) and the Diagnostic Systems Laboratories RIA. There are two notable observations here. One is that RIA and LC-MS yield measurements of similar magnitude, which may reflect the good measurement convergence between RIA with MS methods (for the Coat-a-Count RIA, see Groenestege et al., 2012; Taieb et al., 2003; Wang et al., 2004). The second is the pronounced tendency of the Salimetrics assay to overestimate salivary testosterone and produce rather variable measurements across studies, despite testing similar populations. For women in particular, the average across-study testosterone level was more than 5-fold of what would be expected based on the average LC-MS results. For men, the same comparison yielded a less extreme, but still twofold higher average than LC-MS. The peculiarity of salivary testosterone values as assayed with EIA appears not to be specific to the Salimetrics assay, however. Welker et al. (2016), who assayed the same samples also with other EIAs in addition to the one by Salimetrics, obtained average concentrations of 82 pg/mL for women and 222 pg/mL for men for a testosterone EIA by IBL, and concentrations of 45 pg/mL and 107 pg/mL, respectively, for a testosterone EIA by DRG (see Table S3 in Welker et al.. 2016). Thus, both the reduced gender difference and the inflated concentrations appear to be a more general feature of testosterone EIAs. This conclusion is further underscored by the fact that these values come from the very same samples for which Welker et al. (2016) reported considerably lower testosterone concentrations for LC-MS, and with a much larger gender difference (see Table 1).

<TABLE 2.1 HERE>

So what can be concluded from the illustrative findings displayed in Table 1? If LC-MS with its ability to isolate and directly detect analytes is taken as a reference standard, then immunoassays yield similar concentration estimates for testosterone as long as there is a large signal-to-noise ratio, with "signal" denoting the targeted analyte (i.e., testosterone) and "noise" denoting anything that can cross-react with the antibodies. For serum samples taken from men, there seems to be little difference overall between LC-MS and immunoassay measurements. But even with serum samples obtained in women, the first deviations between immunoassay and LC-MS become apparent, presumably due to immunoassays' susceptibility to matrix effects (e.g., protein content) and steroid cross-reactivity. These susceptibility issues are amplified at the concentrations typically observed in salivary testosterone, particularly for the female range of concentrations and particularly for EIAs. However, although not quite as dramatic, the high concentrations for male saliva samples measured with EIA are also troubling. So can testosterone immunoassays be even trusted, particularly when used with saliva samples?

Another way to look at this issue is through the prism of convergent validity. Although absolute salivary testosterone levels may differ by assay type, there still might be good convergence between LC-MS and immunoassay. While such evidence would not persuade a clinician, who needs accurate absolute measurements for diagnosis, linear shifts in value ranges are less of a concern for researchers who are less interested in absolute levels and more in measurement specificity ("Is it really testosterone I am measuring?") and covariance ("Can I use measured concentrations to explore correlations with other variables or with treatments?"). Welker et al. (2016) have addressed this question by examining the convergence between LC-MS and EIA measurements of salivary testosterone in men and women. When correlating LC-MS measurements with EIA measurements across both genders, there is some evidence of convergence, with correlations ranging from .47 (IBL) to

.57 (DRG). However, these coefficients are not very meaningful, because they are mainly driven by the substantial gender difference in testosterone, which is detected more or less well by all assays (see Table 1). When Welker et al. (2016) looked at the more relevant within-gender correlations, results were disappointing. For men, correlations between all three EIAs and LC-MS did not exceed .17. For women, correlations ranged between -.17 (IBL) and .22 (DRG). These findings suggest that the EIAs tested by Welker et al. (2016) are not suitable to assess valid individual differences in salivary testosterone, particularly when looking at each gender separately. They thus reinforce the doubts already raised by the elevated measurement levels of salivary T as obtained via EIA and the curiously low gender difference associated with them.

What about RIA assessments of testosterone? Unfortunately, there are no published data yet on the degree of convergence between LC-MS and RIA measurements of salivary testosterone. Perhaps a study by Groenestege et al. (2012) can serve as an interim estimate. These researchers looked at the convergence between LC-MS and direct immunoassays specifically for serum samples with low testosterone concentrations (i.e., < 1,153 pg/mL). Convergence coefficients varied considerably across assays, from .59 to .92. Notably, the latter, highest value was reached by the Coat-A-Count RIA for testosterone that has also been frequently used for assessing testosterone in saliva. It is unclear, however, whether similarly high convergence coefficients would be found for samples in the actual, much lower salivary range of testosterone (i.e., from 1 to 250 pg/mL) and particularly whether substantial convergence could be reached in within-gender analyses. To some extent, such speculations are moot by now, because the Coat-a-Count RIA, like the DSL testosterone RIA, is no longer manufactured, reflecting an overall tendency of researchers to move away from RIAs and the radioactive waste they generate. Still, the illustrative data presented in Table 1 seem to suggest that some of the problems that testosterone EIAs are grappling with are absent or at

least greatly attenuated in RIAs, presumably because the label conjugated to the target hormone is a complex, biochemically active enzyme in the former case a simple and biochemically inert radioisotope in the latter. The mean values and gender differences RIAs produce appear to be very comparable to data generated by LC-MS.

4.2 Other steroids

In the previous paragraphs, we have focused on the illustrative case of salivary testosterone assessment. We would also like to briefly comment on how immunoassay assessment of other key steroids compares to MS assessment. Cortisol is the steroid with very high salivary levels (an order of magnitude higher than testosterone) and should therefore be easier to assess by immunoassay than most other salivary steroids. While this is generally true, we have already pointed out that immunoassays tend to overestimate cortisol, reflecting cross-reactivity and protein interference effects (Baid et al., 2007; Bae et al., 2016; Miller et al., 2013). Baid et al. (2007) report a convergence of .72 (Spearman) between RIA and LC-MS assessment of salivary cortisol; Miller et al. (2013) report convergence coefficients ranging from .90 to .97 for various EIAs with LC-MS. Welker et al. (2016) report a correlation coefficient of .80 between EIA and LC-MS cortisol.

Compared to salivary cortisol, the assessment of salivary progesterone and particularly of salivary estradiol represents much greater challenges due to the extremely low concentrations of these hormones. Even LC-MS faces a challenge when trying to detect salivary estradiol, for which measurements are typically in the low single-digit pg/mL range. For instance, Gao et al. (2015) report a limit of quantification of 1 pg/mL for salivary estradiol and of 5 pg/mL for salivary progesterone. Both boundaries are thus relatively close to the typical mean values of these hormones in the saliva of men and normally cycling women. However, LC-MS appears to be suitable for assaying salivary estradiol, as judged by its high convergence with serum estradiol (r = .82; Fiers et al., 2017). Although so far no

direct comparisons of immunoassay and MS assessment of salivary estradiol and progesterone appear to have been published, there are studies that compared various immunoassays with LC-MS for serum samples. Ray et al. (2015) report that although an EIA of progesterone showed good convergence with LC-MS (r = .92), the former method overestimated progesterone concentrations by a factor of 2.3 relative to the latter method. Similarly, Ketha et al. (2015) report that all of 14 tested EIAs overestimate estradiol, some up to threefold. RIAs, on the other hand, again appear to give values more similar to LC-MS methods (Rosner et al., 2013). In a concentration range of up to 330 pg/mL, Gaudl et al. (2016) observed excellent convergence between EIA and LC-MS (r = .96). However, Huhtaniemi et al. (2012) found that at the low serum estradiol concentrations (< 11 pg/mL) – that is, in the range of salivary estradiol concentrations -- immunoassay and LC-MS show unacceptably low convergence (r = .32). It is safe to expect that whatever problems of overestimation or method convergence exist for serum measurements will be exacerbated for assessment of estradiol and progesterone in saliva, with its much lower steroid concentrations relative to serum.

5 Implications

From the detailed discussion of salivary testosterone assessment and our cursory overview of the challenges associated with the assessment of other steroids that are of interest to social neuroendocrinologists, the following picture emerges. Immunoassay approaches to measuring steroids have come under increasing fire due to their lack of sensitivity and tendency to overestimate true concentrations (e.g., Harold & Fitzgerald, 2003; Matsumoto & Bremner, 2004; Taieb et al., 2002). To the extent that overestimation is due only to a misalignment of standardization between immunoassay and LC-MS, the issue would be one of scaling only. But it appears that it is also one of validity, because antibody cross-reactivity, non-specificity, and susceptibility to the general biochemical milieu in the assay have been

shown to contribute to measurement variance. Although there are some indications that these problems may be more severe in EIAs than in RIAs, there is no a-priori reason why some RIAs should not also be prone to cross-reactivity and non-specificity effects. As we have pointed out at the beginning, for the measurement of steroids in serum, this state of affairs has already led to a de-facto ban on immunoassay methods in at least one clinical endocrinology journal (Handelsman & Wartofsky, 2013). Others have taken a more measured approach, arguing that it is unrealistic to expect all laboratories to give up immunoassays and adopt LC-MS, as this would be tantamount to abandoning many assay techniques that work reasonably well and require a sizable investment in terms of equipment and training (Taylor, Keevil, & Huhtaniemi, 2015). Besides, a badly done LC-MS measurement can yield less valid results than a carefully validated and executed immunoassay with suitable quality controls; thus, LC-MS is not a sure-fire guarantee of better results. In the long run, however, LC-MS will become the norm not only in clinical endocrinology, but also in social neuroendocrinology, simply because scientific progress and the advancement of theory crucially depend on the most valid, accurate, and sensitive measurement available. In this regard, the sun is rising for LC-MS, while it is gradually setting for immunoassays.

In closing, we offer some more pragmatically oriented arguments that may aid researchers interested in measuring steroids. The benefits of immunoassays are clear: their performance is easy and they are relatively cheap. Researchers new to hormone assessment can readily learn how do run immunoassays, without advanced analytical knowledge. But every hormone must be assayed individually, which is time- and labor-consuming. Moreover, assays of several steroids also require larger sample volumes (50-400 µl per hormone). In comparison to immunoassays, LC-MS is an effective, high throughput method that requires only small sample volumes of 100 µl for measuring several steroids simultaneously (e.g., testosterone, estradiol, progesterone and cortisol). Despite all performance advantages of LC-

MS, this technique is very complex and requires well-advanced chemical and analytical knowledge. In comparison to immunoassays, newcomers to this technique have to be trained intensively. Due to its universal application, LC-MS/MS methods are now being developed that can be used for steroid analysis in saliva and serum as well as urine or hair . Further, several LC-MS/MS methods for quantification of estradiol in serum with a limit of quantification < 0.5 pg/mL have already been developed (Ketha et al, 2015; Fiers et al, 2017). Modern immunoassays are generally not capable of reaching "down" this far while also retaining good precision.

In summary, immunoassays show satisfying performance in quantifying steroids in larger concentrations, such as in serum. But when it comes to measurement of hormones at lower concentrations, like testosterone in women or gonadal steroids in saliva more generally, immunoassays' validity appears to be limited and findings should be evaluated with a healthy dose of skepticism. In the long run –that is, in the next 5 to 10 years—we anticipate that LC-MS methods will become easier to use and, through methodological advances, even more sensitive so that they can provide precise and valid measurements of steroids even in the single-picogram range. By that time, they will have become the new standard for accurate hormone measurement in social neuroendocrinology.

References

Bäckström, T., Carstensen, H., & Södergard, R. (1976). Concentration of estradiol, testosterone and progesterone in cerebrospinal fluid compared to plasma unbound and total concentrations. *Journal of Steroid Biochemistry, 7*(6), 469-472. doi: https://doi.org/10.1016/0022-4731(76)90114-X

Bae, Y. J., Gaudl, A., Jaeger, S., Stadelmann, S., Hiemisch, A., Kiess, W., . . . Kratzsch, J. (2016). Immunoassay or LC-MS/MS for the measurement of salivary cortisol in children? *Clin Chem Lab Med, 54*(5), 811-822. doi: 10.1515/cclm-2015-0412

Bae Yoon, J., Gaudl, A., Jaeger, S., Stadelmann, S., Hiemisch, A., Kiess, W., . . . Kratzsch, J. (2016). Immunoassay or LC-MS/MS for the measurement of salivary cortisol in children? *Clinical Chemistry and Laboratory Medicine (CCLM)* (Vol. 54, pp. 811).

Baid, S. K., Sinaii, N., Wade, M., Rubino, D., & Nieman, L. K. (2007). Radioimmunoassay and tandem mass spectrometry measurement of bedtime salivary cortisol levels: a comparison of assays to establish hypercortisolism. *Journal of Clinical Endocrinology and Metabolism, 92*(8), 3102-3107. doi: 10.1210/jc.2006-2861

Baker, M. (2015). Reproducibility crisis: Blame it on the antibodies. *Nature, 521*(7552), 274-276. doi: 10.1038/521274a

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. doi: 2004-19012-010 [pii]
10.1037/0033-295X.111.4.1061

Burkitt, J., Widman, D., & Saucier, D. M. (2007). Evidence for the influence of testosterone in the performance of spatial navigation in a virtual water maze in women but not in men. *Hormones and Behavior, 51*(5), 649-654. doi: 10.1016/j.yhbeh.2007.03.007

Buttler, R. M., Martens, F., Ackermans, M. T., Davison, A. S., van Herwaarden, A. E., Kortz, L., . . . Heijboer, A. C. (2016). Comparison of eight routine unpublished LC-MS/MS

methods for the simultaneous measurement of testosterone and androstenedione in serum. *Clin Chim Acta, 454*, 112-118. doi: 10.1016/j.cca.2016.01.002

Buttler, R. M., Martens, F., Fanelli, F., Pham, H. T., Kushnir, M. M., Janssen, M. J., . . . Heijboer, A. C. (2015). Comparison of 7 Published LC-MS/MS Methods for the Simultaneous Measurement of Testosterone, Androstenedione, and Dehydroepiandrosterone in Serum. *Clinical Chemistry, 61*(12), 1475-1483. doi: 10.1373/clinchem.2015.242859

Campbell, D. T., & Fiske, D. W. (1959). Covergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Campbell, K. L. (1994). Blood, urine, saliva and dip-sticks: Experiences in Africa, New Guinea, and Boston. *Annals of the New York Academy of Sciences, 709*, 312-330.

Ceglarek, U., Kortz, L., Leichtle, A., Fiedler, G. M., Kratzsch, J., & Thiery, J. (2009). Rapid quantification of steroid patterns in human serum by on-line solid phase extraction combined with liquid chromatography-triple quadrupole linear ion trap mass spectrometry. *Clin Chim Acta, 401*(1-2), 114-118. doi: 10.1016/j.cca.2008.11.022

Chiu, M. L., Lawi, W., Snyder, S. T., Wong, P. K., Liao, J. C., & Gau, V. (2010). Matrix Effects—A Challenge toward Automation of Molecular Analysis. *JALA: Journal of the Association for Laboratory Automation, 15*(3), 233-242. doi: 10.1016/j.jala.2010.02.001

Clifton, S., Macdowall, W., Copas, A. J., Tanton, C., Keevil, B. G., Lee, D. M., . . . Wu, F. C. W. (2016). Salivary Testosterone Levels and Health Status in Men and Women in the British General Population: Findings from the Third National Survey of Sexual Attitudes and Lifestyles (Natsal-3). *The Journal of Clinical Endocrinology & Metabolism, 101*(11), 3939-3951. doi: 10.1210/jc.2016-1669

Durdiakova, J., Fabryova, H., Koborova, I., Ostatnikova, D., & Celec, P. (2013). The effects of saliva collection, handling and storage on salivary testosterone measurement. *Steroids, 78*(14), 1325-1331. doi: 10.1016/j.steroids.2013.09.002

Evrin, P. E., Nilsson, S. E., Öberg, T., & Malmberg, B. (2005). Serum C-reactive protein in elderly men and women: Association with mortality, morbidity and various biochemical values. *Scandinavian Journal of Clinical and Laboratory Investigation, 65*(1), 23-31. doi: 10.1080/00365510510013505

Fiers, T., Dielen, C., Somers, S., Kaufman, J. M., & Gerris, J. (2017). Salivary estradiol as a surrogate marker for serum estradiol in assisted reproduction treatment. *Clin Biochem, 50*(3), 145-149. doi: 10.1016/j.clinbiochem.2016.09.016

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

Gaudl, A., Kratzsch, J., Bae, Y. J., Kiess, W., Thiery, J., & Ceglarek, U. (2016). Liquid chromatography quadrupole linear ion trap mass spectrometry for quantitative steroid hormone analysis in plasma, urine, saliva and hair. *J Chromatogr A, 1464*, 64-71. doi: 10.1016/j.chroma.2016.07.087

Gosling, J. P. (2000). Analysis by specific binding. In J. P. Gosling (Ed.), *Immunoassays: A practical approach* (pp. 1-17). Oxford: Oxford University Press.

Gouveia, M. J., Brindley, P. J., Santos, L. L., Correia da Costa, J. M., Gomes, P., & Vale, N. (2013). Mass spectrometry techniques in the survey of steroid metabolites as potential disease biomarkers: A review. *Metabolism, 62*(9), 1206-1217. doi: https://doi.org/10.1016/j.metabol.2013.04.003

Gouveia, M. J., Brindley, P. J., Santos, L. L., Correia da Costa, J. M., Gomes, P., & Vale, N. (2013). Mass spectrometry techniques in the survey of steroid metabolites as potential disease biomarkers: a review. *Metabolism, 62*(9), 1206-1217. doi: 10.1016/j.metabol.2013.04.003

Groenestege, W. M. T., Bui, H. N., ten Kate, J., Menheere, P. P., Oosterhuis, W. P., Vader, H. L., . . . Janssen, M. J. (2012). Accuracy of first and second generation testosterone assays and improvement through sample extraction. *Clinical Chemistry, 58*(7), 1154-1156. doi: 10.1373/clinchem.2011.181735

Gröschl, M. (2008). Current status of salivary hormone analysis. *Clinical Chemistry, 54*(11), 1759-1769. doi: clinchem.2008.108910 [pii]

10.1373/clinchem.2008.108910

Gröschl, M., & Rauh, M. (2006). Influence of commercial collection devices for saliva on the reliability of salivary steroids analysis. *Steroids, 71*(13-14), 1097-1100. doi: S0039-128X(06)00195-4 [pii]

10.1016/j.steroids.2006.09.007

Gröschl, M., Wagner, R., Rauh, M., & Dorr, H. G. (2001). Stability of salivary steroids: the influences of storage, food and dental care. *Steroids, 66*(10), 737-741. doi: S0039-128X(01)00111-8 [pii]

Ha, Y. W., Moon, J. Y., Jung, H. J., Chung, B. C., & Choi, M. H. (2009). Evaluation of plasma enzyme activities using gas chromatography-mass spectrometry based steroid signatures. *J Chromatogr B Analyt Technol Biomed Life Sci, 877*(32), 4125-4132. doi: 10.1016/j.jchromb.2009.11.010

Hakkinen, K., Pakarinen, A., Kraemer, W. J., Newton, R. U., & Alen, M. (2000). Basal concentrations and acute responses of serum hormones and strength development

during heavy resistance training in middle-aged and elderly men and women. *J Gerontol A Biol Sci Med Sci, 55*(2), B95-105.

Handelsman, D. J., & Wartofsky, L. (2013). Requirement for Mass Spectrometry Sex Steroid Assays in the Journal of Clinical Endocrinology and Metabolism. *The Journal of Clinical Endocrinology & Metabolism, 98*(10), 3971-3973. doi: 10.1210/jc.2013-3375

Herold, D. A., & Fitzgerald, R. L. (2003). Immunoassays for testosterone in women: better than a guess? *Clinical Chemistry, 49*(8), 1250-1251.

Hsing, A. W., Stanczyk, F. Z., Belanger, A., Schroeder, P., Chang, L., Falk, R. T., & Fears, T. R. (2007). Reproducibility of serum sex steroid assays in men by RIA and mass spectrometry. *Cancer Epidemiol Biomarkers Prev, 16*(5), 1004-1008. doi: 10.1158/1055-9965.EPI-06-0792

Huhtaniemi, I. T., Tajar, A., Lee, D. M., O'Neill, T. W., Finn, J. D., Bartfai, G., . . . Group, E. (2012). Comparison of serum testosterone and estradiol measurements in 3174 European men using platform immunoassay and mass spectrometry; relevance for the diagnostics in aging men. *Eur J Endocrinol, 166*(6), 983-991. doi: 10.1530/EJE-11-1051

Kane, M. M., & Banks, J. N. (2000). Raising antibodies. In J. P. Gosling (Ed.), *Immunoassays: A practical approach* (pp. 19-58). Oxford: Oxford University Press.

Keevil, B. G., MacDonald, P., Macdowall, W., Lee, D. M., Wu, F. C., & Team, N. (2014). Salivary testosterone measurement by liquid chromatography tandem mass spectrometry in adult males and females. *Ann Clin Biochem, 51*(Pt 3), 368-378. doi: 10.1177/0004563213506412

Ketha, H., Girtman, A., & Singh, R. J. (2015). Estradiol assays--The path ahead. *Steroids, 99*(Pt A), 39-44. doi: 10.1016/j.steroids.2014.08.009

Khosla, S., Arrighi, H. M., Melton, I., L. J., Atkinson, E. J., O'Fallon, W. M., Dunstan, C., & Riggs, B. L. (2002). Correlates of Osteoprotegerin Levels in Women and Men. *Osteoporosis International, 13*(5), 394-399. doi: 10.1007/s001980200045

L1, B.-Z. Comparison of methods: Passing and Bablok regression. *Biochem Med (Zagreb). 2011*.

Leng, G., & Ludwig, M. (2016). Intranasal Oxytocin: Myths and Delusions. *Biological Psychiatry, 79*(3), 243-250. doi: https://doi.org/10.1016/j.biopsych.2015.05.003

Matsumoto, A. M., & Bremner, W. J. (2004). Serum Testosterone Assays—Accuracy Matters. *The Journal of Clinical Endocrinology & Metabolism, 89*(2), 520-524. doi: 10.1210/jc.2003-032175

McCullough, M. E., Churchland, P. S., & Mendez, A. J. (2013). Problems with measuring peripheral oxytocin: Can the data on oxytocin and human behavior be trusted? *Neuroscience & Biobehavioral Reviews, 37*(8), 1485-1492. doi: https://doi.org/10.1016/j.neubiorev.2013.04.018

McDonald, J. G., Matthew, S., & Auchus, R. J. (2011). Steroid Profiling by Gas Chromatography–Mass Spectrometry and High Performance Liquid Chromatography–Mass Spectrometry for Adrenal Diseases. *Hormones and Cancer, 2*(6), 324-332. doi: 10.1007/s12672-011-0099-x

Mehta, P. H., Jones, A. C., & Josephs, R. A. (2008). The social endocrinology of dominance: Basal testosterone predicts cortisol changes and behavior following victory and defeat. *Journal of Personality and Social Psychology, 94*(6), 1078-1093. doi: 2008-06135-011 [pii]

10.1037/0022-3514.94.6.1078

Mehta, P. H., & Josephs, R. A. (2010). Testosterone and cortisol jointly regulate dominance: Evidence for a dual-hormone hypothesis. *Hormones and Behavior, 58,* 898–906. doi: S0018-506X(10)00241-2 [pii]

10.1016/j.yhbeh.2010.08.020

Miller, R., Plessow, F., Rauh, M., Groschl, M., & Kirschbaum, C. (2013). Comparison of salivary cortisol as measured by different immunoassays and tandem mass spectrometry. *Psychoneuroendocrinology, 38*(1), 50-57. doi: 10.1016/j.psyneuen.2012.04.019

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods, 5*(2), 241-301.

Ojeda, S. R., & Kovacs, W. J. (2012). Organization of the endocrine system. In W. J. Kovacs & S. R. Ojeda (Eds.), *Textbook of endocrine physiology* (6 ed., pp. 3-20). New York, NY: Oxford University Press.

Oxford, J. K., Tiedtke, J. M., Ossmann, A., Özbe, D., & Schultheiss, O. C. (2017). Endocrine and aggressive responses to competition are moderated by contest outcome, gender, individual versus team competition, and implicit motives. *PLoS One, 12*(7), e0181610. doi: 10.1371/journal.pone.0181610

Palme, R. (2005). Measuring Fecal Steroids: Guidelines for Practical Application. *Annals of the New York Academy of Sciences, 1046*(1), 75-80. doi: 10.1196/annals.1343.007

Ray, J. A., Kushnir, M. M., Yost, R. A., Rockwood, A. L., & Wayne Meikle, A. (2015). Performance enhancement in the measurement of 5 endogenous steroids by LC-MS/MS combined with differential ion mobility spectrometry. *Clin Chim Acta, 438*, 330-336. doi: 10.1016/j.cca.2014.07.036

Rosner, W., Hankinson, S. E., Sluss, P. M., Vesper, H. W., & Wierman, M. E. (2013). Challenges to the measurement of estradiol: an endocrine society position statement.

Journal of Clinical Endocrinology and Metabolism, 98(4), 1376-1387. doi: 10.1210/jc.2012-3780

Schultheiss, O. C. (2013). Effects of sugarless chewing gum as a stimulant on progesterone, cortisol, and testosterone concentrations assessed in saliva. *International Journal of Psychophysiology, 87*, 111-114. doi: 10.1016/j.ijpsycho.2012.11.012

Schultheiss, O. C., Schiepe, A., & Rawolle, M. (2012). Hormone assays. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *Handbook of Research Methods in Psychology* (Vol. 1: Foundations, planning, measures, and psychometrics, pp. 489-500). Washington DC: American Psychological Association.

Schultheiss, O. C., & Stanton, S. J. (2009). Assessment of salivary hormones. In E. Harmon-Jones & J. S. Beer (Eds.), *Methods in social neuroscience* (pp. 17-44). New York, NY: Guilford.

Schultheiss, O. C., Wirth, M. M., Torges, C. M., Pang, J. S., Villacorta, M. A., & Welsh, K. M. (2005). Effects of implicit power motivation on men's and women's implicit learning and testosterone changes after social victory or defeat. *Journal of Personality and Social Psychology*, *88*(1), 174–188.

Selby, C. (1999). Interference in Immunoassay. *Annals of Clinical Biochemistry, 36*(6), 704-721. doi: 10.1177/000456329903600603

Söderberg, S., Olsson, T., Eliasson, M., Johnson, O., Brismar, K., Carlström, K., & Ahren, B. (2001). A strong association between biologically active testosterone and leptin in non-obese men and women is lost with increasing (central) adiposity. *International journal of obesity*, *25*(1), 98.

Stalder, T., & Kirschbaum, C. (2012). Analysis of cortisol in hair – State of the art and future directions. *Brain, Behavior, and Immunity*, *26*(7), 1019-1029. doi: https://doi.org/10.1016/j.bbi.2012.02.002

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*(14), 1173-1178.

Stanton, S. J. (2011). The essential implications of gender in human behavioral endocrinology studies. *Front Behav Neurosci, 5*, 9. doi: 10.3389/fnbeh.2011.00009

Taieb, J., Benattar, C., Birr, A. S., & Lindenbaum, A. (2002). Limitations of steroid determination by direct immunoassay. *Clinical Chemistry, 48*(3), 583-585.

Taieb, J., Mathian, B., Millot, F., Patricot, M. C., Mathieu, E., Queyrel, N., . . . Boudou, P. (2003). Testosterone measured by 10 immunoassays and by isotope-dilution gas chromatography-mass spectrometry in sera from 116 men, women, and children. *Clinical Chemistry, 49*(8), 1381-1395.

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*(2), D1-12. doi: 10.1530/EJE-15-0338

Torjesen, P. A., & Sandnes, L. (2004). Serum testosterone in women as measured by an automated immunoassay and a RIA. *Clinical Chemistry, 50*(3), 678; author reply 678-679. doi: 10.1373/clinchem.2003.027565

Turpeinen, U., Hämäläinen, E., Haanpää, M., & Dunkel, L. (2012). Determination of salivary testosterone and androstendione by liquid chromatography–tandem mass spectrometry. *Clinica Chimica Acta, 413*(5), 594-599. doi: https://doi.org/10.1016/j.cca.2011.11.029

Wang, C., Catlin, D. H., Demers, L. M., Starcevic, B., & Swerdloff, R. S. (2004). Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry. *The Journal of Clinical Endocrinology & Metabolism, 89*(2), 534-543.

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

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: http://dx.doi.org/10.1016/j.psyneuen.2016.05.022

Whembolua, G. L., 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. *Hormones and Behavior, 49*(4), 478-483. doi: S0018-506X(05)00236-9 [pii]

10.1016/j.yhbeh.2005.10.005

Wierman, M. E., Auchus, R. J., Haisenleder, D. J., Hall, J. E., Handelsman, D., Hankinson, S., . . . Stanczyk, F. Z. (2014). Editorial: The new instructions to authors for the reporting of steroid hormone measurements. *Journal of Clinical Endocrinology and Metabolism, 99*(12), 4375. doi: 10.1210/jc.2014-3424

Yalow, R. S., & Berson, S. A. (1960). Immunoassay of endogenous plasma insulin in man. *J Clin Invest, 39*, 1157-1175.

Zilioli, S., Caldbick, E., & Watson, N. V. (2014). Testosterone reactivity to facial display of emotions in men and women. *Hormones and Behavior, 65*(5), 461-468. doi: 10.1016/j.yhbeh.2014.04.006

Table 1.

Illustrative findings for testosterone in saliva and serum, as assessed with liquid chromatography-mass spectrometry (LC-MS), enzyme-linked immunosorbent assays (ELISA) or enzymatic immunoassays (EIA), and radioimmunoassays (RIA)

LC-MS				ELISA/EIA				RIA			
Study	N (♀/♂)	♀	♂	Study	N (♀/♂)	♀	♂	Study	N (♀/♂)	♀	♂
Saliva											
Welker et al (2016)	56/42	11.6	63.6	Welker et al (2016)	57/42	63.6	146.0	Stanton (2011)	262/296	13.3	77.5
Keevil et al (2014)	86/104	4.6	63.7	Zilioli et al (2014)	79/85	63.4	121.8	Oxford et al (2017)	53/152	14.6	79.0
Turpeinen et al (2012)	47/36	5.5	49.3	Mehta et al (2008)	91/93	38.9	119.4	Mehta & Josephs (2010)	50/50	21.7	99.9
Clifton et al (2016)	2123/1599	10.7	64.4	Burkitt et al (2007)	39/36	96.8	183.5	Schultheiss et al (2005)	86/125	18.0	125.0
Weighted mean	2312/1781	10.4	64.0	Weighted mean	266/256	60.0	133.6	Weighted mean	451/543	15.3	87.4
♂/♀ ratio		6.16		♂/♀ ratio		2.23		♂/♀ ratio		5.72	
Serum (total testosterone)											
Keevil et al (2014)	91/94	231	5192	Taieb et al (2003)	55/50	1470	6311	Taieb et al (2003)	51/50	807	5965
Büttler et al (2015)	23/22	228	5228	Torjesen & Sandnes (2003)	2057/1447	605	4813	Torjesen & Sandnes (2003)	2180/1505	490	4006
Wang et al (2014)	75/75	230	5209	Häkkinen et al (2000)	11/11	490	4842	Khosla et al (2002)	152/173	400	5370
Büttler et al (2016)	31/29	190	3135	Evrin et al (2005)	197/189	288	4611	Söderberg et al (2001)	45/85	317	5504
Weighted mean	220/220	225	4930	Weighted mean	2320/1697	598	4834	Weighted mean	2428/1813	487	4260
♂/♀ ratio		21.95		♂/♀ ratio		8.08		♂/♀ ratio		8.73	

Note. For Häkkinen et al (2000), data are reported for middle-aged participants only. For Khosla et al (2002), data are reported for premenopausal individuals and individuals up to an age of 50 years.

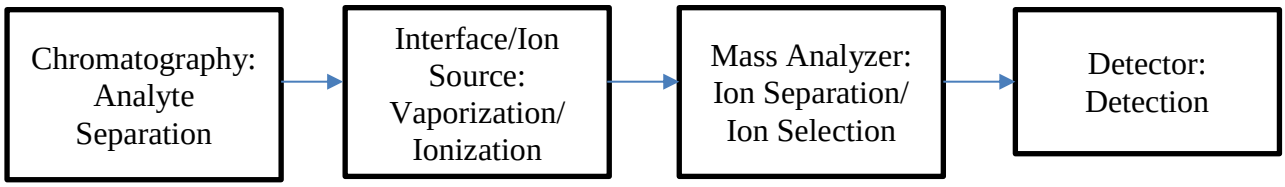

Figure 1. Individual components of mass spectrometry and their functions

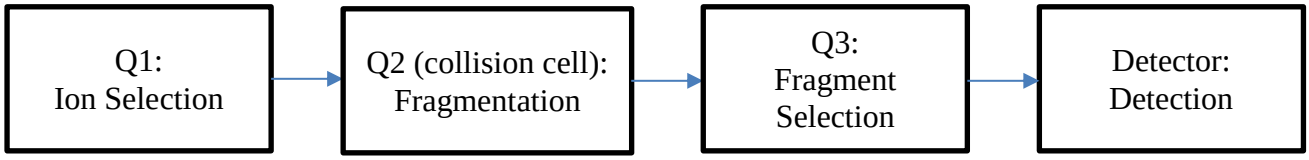

Figure 2. Triple quadrupole mass spectrometer (tandem mass spectrometer; Q: quadrupole)