Published in final edited form as:

J Res Adolesc. 2012 June 1; 22(2): 381–391. doi:10.1111/j.1532-7795.2011.00773.x.

Correspondence Between Gonadal Steroid Hormone Concentrations and Secondary Sexual Characteristics Assessed by Clinicians, Adolescents, and Parents

Bin Huang, PhD,

Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center;
Department of Pediatrics, University of Cincinnati College of Medicine

Jennifer Hillman, MD,

Division of Adolescent Medicine, Cincinnati Children's Hospital Medical Center, Department of
Pediatrics, University of Cincinnati College of Medicine

Frank M. Biro, MD,

Division of Adolescent Medicine, Cincinnati Children's Hospital Medical Center, Department of
Pediatrics, University of Cincinnati College of Medicine

Lili Ding, MS,

Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center

Lorah D. Dorn, PhD, and

Division of Adolescent Medicine, Cincinnati Children's Hospital Medical Center, Department of
Pediatrics, University of Cincinnati College of Medicine

Elizabeth J. Susman, PhD

Department of Biobehavioral Health, The Pennsylvania State University

Abstract

Adolescent sexual maturation is staged using Tanner criteria assessed by clinicians, parents, or adolescents. The physiology of sexual maturation is driven by gonadal hormones. We investigate Tanner stage progression as a function of increasing gonadal hormone concentration and compare performances of different raters. Fifty-six boys (mean age, 12.7 ± 1.3 years) and 52 girls (mean age, 12.0 ± 1.6 years) were seen at baseline, 6, and 12 months. Estradiol and testosterone concentrations were determined from 3 morning serum samples and Tanner stage by three different raters (clinician, parent, adolescent). Results confirm that Tanner criteria reflect gonadal hormone concentrations, and clinician rating provides optimal assessment. Detailed insight about the strengths and limitations of different raters is provided, augmenting the scientific understanding of pubertal development.

Keywords

gonadal hormones; puberty; sexual maturation; adolescents; Tanner criteria

Puberty is the stage in human development during which an adolescent's body undergoes physical and sexual maturation. During puberty, the reproductive organs mature, body

composition changes, and linear growth reaches a peak velocity. Characterization and measurement of these physical changes are important for clinical purposes and for research regarding puberty. Tanner staging is extensively used to assess the advancement of sexual maturity (Marshall & Tanner, 1969, 1970), identify the timing of pubertal onset, mark pubertal progress, and comprehend unique aspects of the adolescent's developmental process.

Since the development of Tanner criteria in the 1960s, a significant amount of research has evaluated its usefulness. Some of the research has focused on inter-rater reliability and agreement between Tanner ratings made by clinicians and other methods, such as adolescent- or parent-report. Only a very small number of studies have focused on the critical question of how well Tanner criteria relate to gonadal steroid hormone concentrations (Akslaede, Sørensen, Petersen, Skakkebaek, & Juul, 2009), the physiological driving force underlying sexual maturation. Previous studies have noted that increased concentrations of testosterone and estradiol are associated with more advanced maturity ratings by clinicians (Biro, Lucky, Huster, & Morrison, 1995; Garcés et al., 2008; Nottelmann et al., 1987), and significant differences in hormone concentrations have been demonstrated between noncontiguous (Nottelmann, et al., 1987) and some contiguous Tanner stages when rated by clinicians (Biro et al., 1995; Garcés, et al., 2008). One recent study (Shirtcliff, Dahl, & Pollak, 2009) evaluated the correlations between sexual maturity rating by Tanner criteria and gonadal steroid hormone concentration and compared clinician- and adolescent-reported assessments of sexual maturity ratings. No study has investigated how Tanner stage progresses as a function of increasing gonadal hormone concentrations and how the correspondence may vary using different methods of applying Tanner criteria.

To assess pubertal maturation, clinicians perform a visual and physical assessment of the Tanner criteria. This assessment includes evaluation of the development of breasts (whole breast and areolar changes) and pubic hair in girls and development of genitals (scrotum appearance, testicular size, and penis enlargement) and pubic hair in boys. With these observations, clinicians then assign a Tanner stage rating, with stage I representing prepubertal development; stage II, the onset of pubertal development; and stage V, full sexual maturity. In males, testicular enlargement is often the first sign of puberty. Assessment of testicular volume, however, is not included in the original Tanner criteria, although this measurement is used commonly in clinical practice to determine onset and advancement of puberty.

One area of clinical and methodological concern regarding Tanner staging involves the person making the assessment, or the "rater." Although clinician assessment via physical examination is uniformly considered the "gold standard" (Marshall & Tanner, 1969, 1970), adolescent- and parent-reports are often used in lieu of the physical examination, especially in epidemiological or community studies where conducting extensive physical examinations may not be feasible. Agreement between the adolescent- or parent-report and the physical examination varies greatly and ranges from 0.09 to 0.88 (Dorn, Dahl, Woodward, & Biro, 2006). This lack of agreement between raters is concerning because of the potential impact on accurate evaluation of the individual's timing, tempo, and synchrony of pubertal processes, which are critical in understanding puberty-related psychosocial development and growth (Golub et al., 2008). Studies have demonstrated associations between timing of puberty and development of affective disorders and risk-taking behaviors (Graber, Lewinsohn, Seeley, & Brooks-Gunn, 1997; Kaltiala-Heino, Marttunen, Rantanen, & Rimpelä, 2003; Williams & Dunlop, 1999), as well as development of later adult disease (Golub, et al., 2008).

Given the importance of using Tanner criteria to understand pubertal development in clinical and research settings, there is a need to understand the limitations and biases associated with different approaches to assess maturation. With this goal, the current study is designed to understand how Tanner stages progress as a function of advancement in gonadal hormone concentrations and to compare ratings among three different types of raters. Using unique statistical modeling, the study reveals a number of inherent strengths and limitations of Tanner staging as assessed by different raters. Such results add a novel and critical layer to the scientific understanding of puberty and assessment of pubertal development.

The overall aims of the current study are to (1) determine whether gonadal steroid hormone concentrations (in boys and girls) and testicular volume (in boys) differ by Tanner stage as determined by three types of raters (clinician, adolescent, and parent); (2) evaluate how Tanner stage may progress as a function of advancement in hormone concentration (in boys and girls) and testicular volume (in boys) as determined by three types of raters (clinician, adolescent, and parent); and (3) test whether estradiol concentration (in girls) and testosterone concentration and testicular volume (in boys) demonstrate high sensitivity and specificity in differentiating between prepubertal and early pubertal stages of development as determined by the clinician rater.

We hypothesize that (1) gonadal steroid hormone concentrations (in boys and girls) and testicular volume (in boys) will differ by Tanner stage of sexual maturity and that such associations with gonadal steroid hormone concentrations and testicular volume will be stronger when sexual maturity is assessed by clinicians; (2) the likelihood of determining a higher Tanner stage will increase as hormone concentrations increase (for boys and girls), as well as when testicular volume increases (for boys), and that such findings will be stronger when sexual maturity is assessed by clinicians compared to adolescent or parent raters; and (3) estradiol concentrations (in girls) and testicular volume and testosterone concentrations (in boys) will demonstrate high sensitivity and specificity in differentiating between prepubertal and early pubertal stages of development as determined by the clinician rater.

Method

Participants

This is a secondary data analysis of a longitudinal study of girls (aged 9–14 years; $n = 52$) and boys (aged 10–15 years; $n = 56$) from a mid-Atlantic region of the United States who were seen at baseline, 6 months, and 12 months. The original study was designed to evaluate associations between puberty, growth, hormone status, and psychological development during the pubertal transition (Nottelmann, et al., 1987; Susman et al., 1985). Exclusion criteria included any chronic illness or major mental health problems. The mean age $\pm$ SD of the participants was 12.0 ± 1.6 years for girls and 12.7 ± 1.3 years for boys. Participants were recruited to have approximately equal representation of the five maturity stages ($n = 10, 10, 13, 10$, and 9 for girls; $n = 12, 14, 10, 7$, and 13 for boys). Eighteen girls were post-menarcheal at baseline. Almost all participants were Caucasian except for three African American boys. The sample was predominantly middle to upper-middle class (Hollingshead, 1975), and the majority of participants had BMIs < 85 th percentile. Table 1 summarizes the baseline characteristics of study participants. Informed consent was obtained from parents, as well as assent from adolescents. Of the 108 participants, only two boys and two girls (4%) dropped out of the study. The parent study was approved by the Institutional Review Board of the intramural programs of the NICHD and NIMH. The current secondary data analyses were approved by the Institutional Review Board of Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio.

Measures and Procedures

Hormonal assessment—Adolescents, accompanied by a parent, were scheduled for three study visits six months apart at Time 1, Time 2, and Time 3. Three serial serum samples were collected at each visit, with 20 minutes between each sample. All visits occurred between 0800 and 1000 hours, and therefore all samples were collected during these morning hours. Radioimmunoassays were performed for total testosterone and estradiol concentrations. For testosterone, the inter-assay and intra-assay coefficients of variation (CVs) were 18.2% and 6.3%, respectively, with a detection limit of 10.0 ng/dL. For estradiol, the inter-assay and intra-assay CVs were 12.2% and 4.2%, respectively, with a detection limit of 8.0 pg/mL. The inter-assay and intra-assay CVs for both testosterone and estradiol are within desired ranges. The mean value of the three samples from each visit was used in the analyses.

Maturity stage—At each visit, an experienced and trained pediatric research nurse explained the five stages of pubertal maturation and showed the adolescent and parent photos commonly used to assist with Tanner staging, while describing the physical changes by stage. Using the photos but without examination, the parent and adolescent then independently determined the adolescent's sexual maturity rating. Finally, the pediatric research nurse conducted a physical examination to determine a sexual maturity rating. A randomly selected subsample of four girls and four boys were also evaluated by a female pediatric research nurse or male pediatrician. Inter-rater agreement between clinicians was 1.0 for breast stage in girls, 0.99 for genital stage in boys, and 1.0 for pubic hair stage in both boys and girls. For boys, the pediatric research nurse measured testicular volume of both left and right testes using a Prader orchidometer during the clinical examination. The intra-cluster correlation coefficient of left and right testicular volumes was 0.98; thus, the mean value of the left and right testicles was used in the analyses.

Covariates—To assess adiposity, during the same physical examination the research nurse obtained three repeated measures of height on a stadiometer and weight on a balance scale. Participants were dressed in street clothes without shoes, sweaters, or jackets. The mean height and weight were used to calculate body mass index (BMI) as weight (kg) divided by height (meters squared). BMI Z-score and the corresponding BMI percentile were calculated using Centers for Disease Control and Prevention age- and sex-specific standards (Kuczmarski et al., 2000). To assess socioeconomic status, the Hollingshead Measure of Social Status (Hollingshead, 1975) was administered to the parent.

Statistical analysis

Box-and-whisker plots were used to examine distributions of hormone concentrations and testicular volume, which were stratified by breast/genital Tanner stage as reported by the three raters. Hormone concentrations were compared using the nonparametric Wilcoxon rank sum test, a robust method to compare maturity stages for each rater. Goodness-of-fit tests suggested a normal distribution fit of the log-transformed hormone values.

For the first hypothesis, we postulated that gonadal steroid hormone concentrations (in boys and girls) and testicular volume (in boys) would differ at successive Tanner stages and that the findings would be stronger when the sexual maturity rating was assessed by clinicians. To address this hypothesis, we first compared hormone concentrations at different Tanner stages as determined by each type of rater (parent, adolescent, or clinician). Repeated measures analyses of variance (ANOVAs) were conducted using the log-transformed data, in which the repeated measures were data from the three visits of each participant and the main independent factor was the Tanner stage. Adiposity (BMI) was included as a covariate initially; due to lack of statistical significance, BMI was not retained in the final models.

Least square means (LSM) and the corresponding 95% confidence intervals (CIs) were estimated, and differences between adjacent maturity stages were compared. Multiple comparisons were adjusted using the Bonferroni step-down procedure (Hochberg, 1988). For overall model fit, Akaike's Information Criteria (AIC) were reported and compared for the repeated measures ANOVAs. When comparing two models, a smaller AIC value suggests a better model fit; a difference in AIC > 10 suggests a substantially better fit of the model (Burnham & Anderson, 2004).

For the second hypothesis, stereotype logistic regression analyses (Anderson, 1984) were conducted to evaluate the likelihood of each type of rater to determine each Tanner stage at a given hormone concentration and testicular volume. Stereotype logistic regression is a more general modeling technique than conventional ordinal logistic regression because it does not assume proportional odds. The results from the stereotype logistic regression were presented by predicted probability plots, depicting the probability that a given type of rater will rate the adolescent's maturity level at a particular Tanner stage given a particular hormone concentration (or testicular volume). Predicted probabilities were summarized by plotting mean scores of the Tanner stage, presenting the expected stage corresponding to the hormone concentration or testicular volume as recommended for logistic modeling for ordinal outcomes (Hannah & Quigley, 1996).

For the final aim, subset logistic analyses were conducted to assess the ability of testicular volume, testosterone, and estradiol to distinguish prepubertal (stage I) from early pubertal maturation (stages II & III) as determined by clinician examination. The area under the receiver operating characteristic (ROC) curve was calculated. The ROC is a commonly used statistical method of evaluating the accuracy of a diagnostic test; the area under the ROC curve, or AUC, is a statistic that summarizes sensitivity and specificity of the test. All analyses were conducted using SAS v9 (SAS Institute Inc., Cary, North Carolina). Analyses with testosterone concentrations and testicular volumes were conducted only for boys; analyses with estradiol concentrations were conducted only for girls.

Results

Distributions of Hormone Concentrations and Testicular Volume

Box-and-whisker plots for hormone concentrations and testicular volume by Tanner stage and by rater are presented in Figure 1, stratified by sex. There is a greater inter-quartile range within Tanner stages I-IV for estradiol (girls), testosterone (boys), and testicular volume (boys) when rated by the adolescent or parent compared to the clinician. The only exception occurs in the testosterone concentration in Tanner stage I. Here the apparent smaller variance in adolescent-reported Tanner stage may be related to the small number of boys (8 by adolescent-report vs. 12 by parent- and clinician-report) who were classified as Tanner stage I.

Wilcoxon rank sum tests compared the distributions of hormone concentrations/ testicular volume across the Tanner stages as determined by all three raters. All comparisons were significantly different ($p < .0001$). In boys, overall Chi-square statistics ($df = 4$) were 66.64, 59.13, and 42.52 for testosterone and 84.02, 40.21, and 60.65 for testicular volume by clinician-, adolescent-, and parent-report, respectively. A higher Chi-square statistic indicates a stronger association between Tanner stage and testosterone concentration and testicular volume. We also examined correlation coefficients and found significantly higher correlations with hormones and Tanner stages when rated by the clinician. For testosterone, Spearman's $\rho = 0.86, 0.73$, and 0.60 for clinician-, adolescent-, and parent-report, respectively; $p = 0.055$ and 0.003 when contrasting adolescent- and parent-report vs. clinician examination, respectively. For testicular volume, $\rho = 0.92, 0.83$, and 0.67 for

clinician-, adolescent-, and parent-report, respectively; $p = 0.03$ and 0.0001 when contrasting adolescent- and parent-report vs. clinician examination, respectively. In girls, we examined the association between estradiol concentration and Tanner stage. The overall Chi-square statistics in girls ($df = 4$) from the Wilcoxon rank sum test were 39.81, 43.84, and 41.29 for clinician-, adolescent-, and parent-report, respectively. No significant differences in the correlation coefficients for estradiol concentration and Tanner stage were found when contrasting the three raters ($\rho = 0.56, 0.51$, and 0.55 for clinician-, adolescent-, and parent-report, respectively).

Hormone Concentrations and Testicular Volume by Maturity Stage

Table 2 summarizes the LSM and corresponding 95% CIs of gonadal steroid hormone concentrations and testicular volume across maturity stages estimated from repeated measures ANOVA. Figure 2 presents the corresponding LSM differences and multiple comparison-adjusted p -values, comparing consecutive pairs of Tanner stages for each type of rater. In boys, testosterone concentrations and testicular volume are significantly higher from one Tanner stage to the next (in the direction from Tanner stages I to V) when assessed by the clinician. When assessed by adolescents and parents, however, testosterone concentrations differ significantly only between stages II and III; testicular volume differs between stages II and III and stages III and IV. In girls, estradiol concentrations are significantly different between stages II and III and stages III and IV, and are marginally different between stages I and II, when assessed by the clinician. Estradiol concentrations are only distinguishable between stages II and III for parent-report and between stages II and III and stages III and IV for adolescent-report. With one exception, all AICs improve by > 10 when comparing clinician examination with adolescent- or parent-report, indicating clinician examination has a substantially better model fit.

Likelihood of Tanner Stage at a Given Hormone Concentration/Testicular Volume According to Rater

Figure 3 presents the probability of each type of rater to determine a particular Tanner stage at a given hormone concentration or testicular volume. The figure shows five probability curves for each Tanner stage. The rise or decline of a probability curve indicates an increased or decreased chance, respectively, of being rated at a specific Tanner stage when given a hormone concentration or testicular volume, with a peak indicating the hormone/testicular level at which a Tanner stage is most likely to be assigned. In general, probability curves peak in an orderly fashion from stages I to V from low to high hormone concentration or testicular volume, suggesting an increased chance of progressing to the next Tanner stage as the hormone/testicular values increase. The main pattern of results is described below.

Testosterone concentration and Tanner stage in boys—Regardless of testosterone concentration, parents preferred to rate a boy at stage III rather than stage IV, while clinician-report showed a stronger preference for either stages III or V over stage IV. Of note, given a sufficiently high testosterone concentration (> 500 ng/dL), the probability of being rated at Tanner stage V was at least 80% by clinician-report, contrasted with a probability of only 20% by adolescent-report and 18% by parent-report. By overlaying the mean score plots for the three types of raters, it is evident that the clinician curve monotonically increases and eventually reaches stage V (Figure 4). In contrast, the parent and adolescent curves increase sharply in the early Tanner stages (compared to the clinician curve) when testosterone concentrations are very low (< 5 ng/dL), reach a plateau quickly, and never progress to stage V.

Testicular volume and Tanner stage in boys—At a testicular volume of 30 mL, 40% of adolescent-report and 30% of parent-report placed boys in a Tanner stage less than stage V, or full sexual maturity, whereas at the same testicular volume, stage V was almost always (> 99%) the rating chosen by clinician-report. Given low testicular volumes (approximately 2 mL), however, one-third of adolescent- and parent-report placed boys in Tanner stage II compared to < 3% by clinician-report (i.e., > 97% of clinician-report was labeled stage I).

Estradiol concentration and breast stage in girls—At higher estradiol concentrations (110 pg/mL), approximately one-fourth of adolescent- and parent-reports placed girls at Tanner stage V compared to > 50% of clinician-reports. Given very low estradiol concentrations (2 pg/mL), the chance of being rated at stage II is much greater with adolescent- and parent-report compared to clinician examination; clinician-report was more likely to be labeled stage I. Clinician-report was more likely to place a girl at stage V than at stage IV, while adolescent- and parent-report were more likely to place girls at stage IV than at stage V, independent of estradiol concentration.

Distinguishing Prepubertal (Stage I) from Early Pubertal (Stages II and III) Maturation

Clinician examination of sexual maturity rating was used for these analyses. The number of observations is represented by *n*. Because three measures were obtained for each patient, some patients contributed more than one observation. Statistical analyses used the generalized estimation equation to account for lack of independence in observations. For testosterone, the repeated binary logistic regression analyzing prepubertal stage (Tanner stage I, *n* = 24) vs. early pubertal stage (Tanner stages II & III, *n* = 59) estimated the AUC as 0.67 (95% CI, 0.65–0.88), suggesting testosterone concentration has an acceptable ability to distinguish prepubertal and early pubertal maturation in males. For testicular volume (Tanner stage I, *n* = 21 vs. stages II & III, *n* = 60), the AUC was 0.86 (95% CI, 0.78–0.93), suggesting testicular volume has a good ability to distinguish prepubertal from early pubertal maturation. Finally, for estradiol (Tanner stage I, *n* = 23 vs. stages II & III, *n* = 53), AUC was 0.94 (95% CI, 0.89–0.99), suggesting estradiol concentration has an excellent ability to distinguish prepubertal from early pubertal maturation.

Discussion

The current study investigated the critical question of how Tanner stage progresses in relation to advancement in gonadal steroid hormone concentrations as assessed by different raters. Aside from confirming results from earlier studies, the current study generated a number of new findings that offer insights into the inherent strengths and limitations when three different types of raters apply Tanner criteria.

Consistent with previous literature, our results confirm that Tanner staging determined by a clinician's physical examination corresponds well with both the physiological changes in gonadal steroid hormone concentrations that drive pubertal maturation as well as with testicular volume changes occurring in boys during puberty. Furthermore, clinician ratings of Tanner stage have the strongest correlation with these indicators of maturation compared to adolescent and parent ratings. This is the first study to confirm via serum hormone concentrations that clinicians are more accurate raters of sexual maturation, especially among boys; the results are presented graphically using probability plots.

A new contribution of the current study regards the differentiation between pre-puberty and early puberty (i.e., when the onset of puberty occurs). We found that Tanner staging via clinician examination does effectively distinguish between consecutive Tanner stages in boys based on correspondence with gonadal steroid hormone concentrations. Adolescent- and parent-report of Tanner staging are less effective, however, at making such distinctions,

particularly between stages I and II. In addition, our results suggest that testicular volume and testosterone and estradiol concentrations provide good utility in differentiating pre-pubertal from early pubertal maturation by the clinician. These findings are of great relevance to clinicians and researchers by validating the use of the Tanner criteria and Tanner staging, a method that has been relied on for decades to assess degrees of pubertal development. A unique aspect of the current study is that serum as opposed to salivary hormone levels were used to differentiate between stages. Blood is the substance of choice by endocrinologists and other health care providers to rule out pathology in pubertal-age youth.

Another new contribution is that our study demonstrates some inherent limitations of the Tanner method based on who is making the assessment. A major limitation of the adolescent-and parent-report is that both raters underestimate the maturity stage at higher hormone concentrations and testicular volumes, and they overestimate the maturity stage at lower hormone concentrations and testicular volumes. It is possible that there exists a social bias or desire by parents or teens to be in the middle rather than at the extreme ends of sexual maturation. In other words, being perceived at either extreme end (i.e., Stage I or V) of pubertal maturation is less desirable for parents and adolescents. If a researcher was particularly interested in capturing the early or late stages of pubertal development, it would be paramount to understand that adolescent- and parent-report are less reliable at these stages than at other maturity stages. For the clinician examination, we noted a lower likelihood of reporting stage IV when assessing male Tanner (genital) stage; stage IV was used for a narrow range of testicular volume (13–16 mL). For girls, we also noted a stronger preference for the clinician to rate at Tanner stage V vs. stage IV, regardless of estradiol concentration, which may reflect difficulty in assessing either breast development or, possibly, fluctuations in estradiol concentrations across the menstrual cycle. Given these limitations, it bears merit to consider a four-stage rating system for sexual maturity, even for clinicians. Our findings indicate that collapsing Tanner stages IV and V into one category might be a more efficient strategy but one that needs validation.

Finally, this study found that the correspondence between estradiol concentration and Tanner stage for girls was not as strong as the correspondence between testosterone concentration and Tanner stage for boys. These findings are likely explained by estradiol's monthly cycling (Bulun, 2011), which leads to increased error in the measurement of estradiol concentrations. Additionally, assessment of Tanner staging in girls can be complicated by adipose tissue in the area of the breast, which may be confused with breast tissue and thereby introduce an opportunity for misclassification. Based on the ROC analyses (AUC, 0.94), estradiol may be more helpful in discriminating prepubertal vs. early pubertal maturation as opposed to later stages of sexual maturity.

Few studies have compared measures of sexual maturity by mapping these methods against gonadal steroid hormone concentrations. To our knowledge, only one such study using saliva has been published (Shirtcliff et al., 2009). Both our study and the Shirtcliff et al. study suggest that clinician examination performs better than adolescent- and parent-report when mapping Tanner stages against testosterone concentrations for boys. For girls, both our study and the Shirtcliff et al. study suggest only a weak correlation between Tanner stage and estradiol concentrations and that clinician rating is not significantly different from adolescent- or parent-report. For boys, both studies found a strong correlation between testosterone concentration and Tanner stage. Further, both studies suggest that adolescents underestimate their maturity at higher stages and overestimate it at lower stages. Our primary departure from the Shirtcliff et al. study is that we found statistically significant differences when comparing clinician examination to adolescent- and parent-report in boys, whereas the Shirtcliff et al. study detected a difference without statistical significance.

Similarities between the two studies are very important in that our report validates the findings of the previous report by strengthening insights that carry clinical significance but that lack published research support. Our additional finding also bears important clinical significance: for determining the onset of puberty or pubertal stage in boys, it will be important to defer to a clinician's evaluation for greatest accuracy.

Some methodological differences between our study and the Shirtcliff et al. study are worth elaborating, as they can affect the comparability of some data. First, the current study analyzed serum hormone concentrations, whereas Shirtcliff et al. analyzed salivary hormone concentrations. Salivary and serum hormone concentrations cannot be directly compared; the former reflects free hormone concentration while the latter reflects total hormone concentration (Albrecht & Styne, 2007; Stanczyk, 2006). Also, the Shirtcliff et al. study used a parallel linear structural equation modeling approach; our study utilized repeated measures ANOVA and stereotype logistic regression to elicit a more detailed understanding of the association between Tanner stage and hormone concentrations while requiring fewer assumptions. Of note, unlike linear structural equation modeling, repeated measure ANOVA modeling does not impose a linear correlation between Tanner stage and hormone concentrations or testicular volume.

Our study does contain limitations. Although pubertal maturation inevitably advanced for adolescents in this study during the course of one year, only six girls and 15 boys advanced two or more stages, corresponding to changes in hormone concentrations longitudinally; this progress limited our ability to model transitions. Future studies might follow participants for an additional time point or introduce a longer interval between time points. Additionally, radioimmunoassay for estradiol may be less reliable at lower concentrations, and estradiol can be affected by the phase of the menstrual cycle. Regardless of our acquired information about the menstrual cycles of the girls in our study, adolescent cycles are inherently irregular, especially at these early gynecological ages, and only 18 girls were menarcheal at Time 1. Future studies should continue collecting data that capture the particular day in the menstrual cycle in addition to undertaking a purposeful study of girls at a particular time in their menstrual cycle (e.g., days 5–9) when possible. The vast majority of our participants were white adolescents of upper-middle socioeconomic class, and thus our findings may not necessarily extend to other populations. Results for testicular volume might be somewhat overestimated because the same examiner assessed Tanner stage and testicular volume. Because knowledge of the boys' testicular volume may have influenced the assessment of Tanner stage and vice versa, a separate examiner may be preferable to complete the assessment of testicular volume independent of the Tanner staging examiner. Finally, the small sample size may have limited the power of the study; however, repeated measures analyses utilized data from three visits for each participant, providing a substantial increase in the amount of available data.

In conclusion, using the Tanner criteria to determine a sexual maturity rating, or Tanner stage, via clinician examination effectively reflects the advancement of hormone concentrations in boys and girls and testicular volume in boys during pubertal maturation. Furthermore, clinician examination most clearly approximates the underlying endocrine changes driving pubertal maturation compared with adolescent- and parent-report. When choosing an approach to assess sexual maturity, researchers and clinicians should determine the importance of measuring maturity relative to the hormonal driving force vs. practical constraints and ease of measurement. When the onset of puberty is of primary interest, a clinician examination is necessary. Previous studies offer practical suggestions for overcoming some perceived constraints in conducting a clinician examination (Dorn, et al., 2006). When clinician examination is not possible, researchers should be mindful of potential misclassification of pubertal maturity when interpreting findings; sensitivity

analyses might be helpful in these interpretations. Identifying reasons why adolescent- and parent-report are less accurate, as well as methods for improving their accuracy, will be important goals for future research.

Acknowledgments

Authors are very grateful to Jan Clavey for her administrative, technical, and editorial support. We acknowledge funding support from 1R01DA01965, NIDA, NIH to this study, and to intramural program between NICHD and NIMH that supported the parent study.

Funding Sources: 1R01DA01965, NIDA, NIH

References

- Aksglaede L, Sørensen K, Petersen JH, Skakkebaek NE, Juul A. Recent decline in age at breast development: The Copenhagen Puberty Study. *Pediatrics*. 2009; 123(5):e932–e939.10.1542/peds.2008-2491 [PubMed: 19403485]
- Albrecht L, Styne D. Laboratory testing of gonadal steroids in children. *Pediatric Endocrinology Reviews*. 2007; 5(suppl 1):599–607. [PubMed: 18167469]
- Anderson JA. Regression and ordered categorical variables. *Journal of the Royal Statistical Society Series B (Methodological)*. 1984; 46(1):1–30.10.2307/2345457
- Biro FM, Lucky AW, Huster GA, Morrison JA. Pubertal staging in boys. *The Journal of Pediatrics*. 1995; 127(1):100–102.10.1016/S0022-3476(95)70265-2 [PubMed: 7608791]
- Bulun, SE. Physiology and pathology of the female reproductive axis. In: Melmed, S.; Polonsky, KS.; Larsen, PR.; Kronenberg, HM., editors. *Williams textbook of endocrinology*. 12. Philadelphia, PA: Elsevier Saunders; 2011. p. 581
- Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*. 2004; 33(2):261–304.10.1177/0049124104268644
- Dorn LD, Dahl RE, Woodward HR, Biro F. Defining the boundaries of early adolescence: A user's guide to assessing pubertal status and pubertal timing in research with adolescents. *Applied Developmental Science*. 2006; 10(1):30–56.10.1207/s1532480xads1001_3
- Garcés C, de Oya I, López-Simón L, Cano B, Schoppen S, Gil A, de Oya M. Hormone levels in 12- to 15-year-old boys and girls in Spain and their relationship with anthropometric variables. *Clinical Biochemistry*. 2008; 41(7–8):621–624.10.1016/j.clinbiochem.2008.01.003 [PubMed: 18237556]
- Golub MS, Collman GW, Foster PM, Kimmel CA, Rajpert-De Meyts E, Reiter EO, Toppari J. Public health implications of altered puberty timing. *Pediatrics*. 2008; 121(suppl 3):S218–S230.10.1542/peds.2007-1813G [PubMed: 18245514]
- Graber JA, Lewinsohn PM, Seeley JR, Brooks-Gunn J. Is psychopathology associated with the timing of pubertal development? *Journal of the American Academy of Child & Adolescent Psychiatry*. 1997; 36(12):1768–1776.10.1097/00004583-199712000-00026 [PubMed: 9401339]
- Hannah M, Quigley P. Presentation of ordinal regression analysis on the original scale. *Biometrics*. 1996; 52(2):771–775.
- Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75(4):800–802.10.1093/biomet/75.4.800
- Hollingshead, AB. Four-factor index of social status. New Haven, CT: Yale University Press; 1975.
- Kaltiala-Heino R, Marttunen M, Rantanen P, Rimpelä M. Early puberty is associated with mental health problems in middle adolescence. *Social Science & Medicine*. 2003; 57(6):1055–1064.10.1016/S0277-9536(02)00480-X [PubMed: 12878105]
- Kuczmarski RJ, Ogden CL, Grummer-Strawn LM, Flegal KM, Guo SS, Wei R, Johnson CL. CDC growth charts: United States. Advance data from vital and health statistics of the National Center for Health Statistics. 2000; 8(314):1–27. [PubMed: 11183293]
- Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. *Archives of Disease in Childhood*. 1969; 44(235):291–303. [PubMed: 5785179]

- Marshall WA, Tanner JM. Variations in the pattern of pubertal change in boys. *Archives of Disease in Childhood*. 1970; 45(239):13–23.10.1136/adc.45.239.13 [PubMed: 5440182]
- Nottelmann ED, Susman EJ, Dorn LD, Inoff-Germain G, Loriaux DL, Cutler GB Jr, Chrousos GP. Developmental processes in early adolescence: Relations among chronologic age, pubertal stage, height, weight, and serum levels of gonadotropins, sex steroids, and adrenal androgens. *Journal of Adolescent Health*. 1987; 8(3):246–260.
- Shirtcliff EA, Dahl RE, Pollak SD. Pubertal development: Correspondence between hormonal and physical development. *Child Development*. 2009; 80(2):327–337.10.1111/j.1467-8624.2009.01263.x [PubMed: 19466995]
- Stanczyk FZ. Measurement of androgens in women. *Seminars in Reproductive Medicine*. 2006; 24(2): 78–85.10.1055/s-2006-939566 [PubMed: 16633981]
- Susman EJ, Nottelmann ED, Inoff-Germain GE, Dorn LD, Cutler GB, Loriaux DL, Chrousos GP. The relation of relative hormonal levels and physical development and social-emotional behavior in young adolescents. *Journal of Youth and Adolescence*. 1985; 14(3):245–264.10.1007/bf02090322
- Williams JM, Dunlop LC. Pubertal timing and self-reported delinquency among male adolescents. *Journal of Adolescence*. 1999; 22(1):157–171.10.1006/jado.1998.0208 [PubMed: 10066339]

Figure 1.
Box-and-Whisker Plots Displaying Hormone Concentrations and Testicular Volume at Maturity Stages I-V Determined by Clinician Examination (PE), Adolescent-Report (AR), and Parent-Report (PR)

Figure 2.

Repeated Measures ANOVA Least Square Mean Differences, the Corresponding 95% Confidence Intervals, and Adjusted p Values (Bonferoni Step-down) of Gonadal Steroid Hormone Concentrations and Testicular Volumes across Adjacent Tanner Stages as Assessed by Clinician Examination (PE), Adolescent-Report (AR), and Parent-Report (PR)

Figure 3. Probability Plots of Tanner Stage as a Function of Gonadal Steroid Hormone Concentration or Testicular Volume. Probability plots depict the probability that a given type of rater will rate the adolescent's maturity level at a particular Tanner stage (corresponding to line legends) given a particular hormone concentration or testicular volume.

Figure 4.

Mean Score Plots of Tanner Stage as a Function of Gonadal Steroid Hormone Concentration or Testicular Volume by Clinician Examination (PE), Adolescent-Report (AR), and Parent-Report (PR)

Table 1

Descriptive and Demographic Characteristics of the Participants

Baseline Characteristics	Boys (N = 56)	Girls (N = 52)
	Mean ± SD	Mean ± SD
Age (years)	12.7 ± 1.30	12.0 ± 1.60
Age Menarche (years) ^a	-	12.1 ± 0.92
Socioeconomic Status ^b	49.2 ± 11.65	48.6 ± 14.16
Estradiol (pg/mL)	12.8 ± 8.20	45.5 ± 49.22
Testosterone (ng/dL)	214.7 ± 204.30	19.3 ± 9.24
Testicular Volume (ml)	11.7 ± 7.14	-
Body Mass Index (BMI)	19.5 ± 3.07	18.6 ± 2.74

	N (%)	N (%)
Race		
Caucasian	53 (94.6%)	52 (100%)
African American	3 (5.4%)	0 (0%)
Tanner Stage (Clinician)		
Stage I	12 (21.4%)	10 (19.2%)
Stage II	14 (25.0%)	10 (19.2%)
Stage III	10 (17.9%)	13 (25.0%)
Stage IV	7 (12.5%)	10 (19.2%)
Stage V	13 (23.2%)	9 (17.3%)
Menarche Present	-	18 (34.6%)
BMI > 85 th percentile	13 (23.2%)	7 (13.5%)

^aData are only available for *n* = 18 girls who were post-menarcheal at baseline.

^bThe Hollingshead four-factor index of social status is based on the parent's sex, education, occupation, and marital status; score ranges between 8 – 66, with higher score indicating higher socioeconomic status.

Table 2

Repeated Measures ANOVA Least Square Means (LSM) and the Corresponding 95% Confidence Limit (95% CL) of Maturation Hormone Concentration and Testicular Volume by Maturity Stage as Assessed by Clinician Examination and Adolescent- and Parent-Report

	<u>Clinician Examination</u>	<u>Adolescent-Report</u>	<u>Parent-Report</u>
Maturity Stage	LSM [95% CL]	LSM [95% CL]	LSM [95% CL]
Testosterone (ng/dL) in Boys (N = 56)			
I	26.9 [18.95,38.03]	34.0 [19.71,58.76]	56.2 [35.40, 89.17]
II	52.5 [39.87, 69.13]	95.6 [70.95, 129.79]	86.0 [60.21,122.79]
III	173.2 [134.37, 223.18]	203.3 [150.11, 275.42]	215.7 [156.95, 296.52]
IV	264.9 [198.64, 353.22]	340.0 [230.95, 500.50]	251.5 [168.04, 376.34]
V	380.9 [299.20, 485.02]	379.4 [233.74, 615.79]	303.7 [179.47, 513.83]
	AIC = 259.4	AIC = 703.1	AIC = 723.9
Testicular Volume (ng/dL) in Boys (N = 56)			
I	3.96 [2.18, 7.73]	6.6 [3.74, 9.47]	9.6 [7.13, 12.08]
II	6.8 [5.43, 8.18]	10.2 [8.52, 11.83]	11.3 [9.27, 13.38]
III	11.0 [9.67, 12.25]	16.5 [14.79, 18.13]	14.9 [12.94, 16.84]
IV	16.1 [14.63, 17.61]	19.8 [17.72, 21.89]	17.4 [15.25, 19.57]
V	22.2 [21.07, 23.24]	22.1 [19.43, 24.74]	18.4 [15.74, 20.95]
	AIC = 626.9	AIC = 703.1	AIC = 723.9
Estradiol (ng/dL) in Girls (N = 52)			
I	9.0 [6.29, 12.75]	10.9 [7.56, 15.74]	10.6 [7.34, 15.33]
II	14.9 [10.58, 220.88]	14.8 [10.86, 20.07]	12.8 [9.33, 17.52]
III	28.5 [21.35, 37.97]	36.3 [27.47, 48.01]	37.2 [28.00, 49.37]
IV	56.4 [40.63, 78.35]	58.6 [44.25, 77.58]	59.5 [43.05, 82.30]
V	59.3 [45.79, 76.77]	56.3 [38.65, 81.97]	55.6 [38.68, 79.81]
	AIC = 289.5	AIC = 298.3	AIC = 295.3

Note. Lower AIC corresponds to better fit. AIC = Akaike's Information Criteria; ANOVA = analysis of variance.