Pubertal Development: Correspondence Between Hormonal and Physical Development

Elizabeth A. Shirtcliff
University of New Orleans
Ronald E. Dahl
University of Pittsburgh School of Medicine
Seth D. Pollak
University of Wisconsin-Madison

Puberty is advanced by sex hormones, yet it is not clear how it is best measured. The interrelation of multiple indices of puberty was examined, including the Pubertal Development Scale (PDS), a picture-based interview about puberty (PBIP), and a physical exam. These physical pubertal measures were then associated with basal hormones responsible for advancing puberty. Participants included 160 early adolescents (82 boys). Puberty indices were highly correlated with each other. The physical exam stages correlated well with boys’ and girls’ testosterone and dehydroepiandrosterone and less so with girls’ estradiol. The PDS and PBIP were similarly related to basal hormones. Self-report may be adequate when precise agreement is unnecessary. Multiple measures of puberty are viable options, each with respective strengths.

Adolescence constitutes a transition between childhood and adulthood whose onset includes pubertal maturation. Puberty has important implications for the development of regulatory competence and many aspects of physical, emotional, cognitive, and social development, including decision making and mental health (Steinberg et al., 2006). For these reasons, biobehavioral researchers increasingly seek to examine measures of puberty to clarify studies of emotion-related neural circuitry (Nelson, Leibenluft, McClure, & Pine, 2005; Sisk & Foster, 2004), psychopathology (Angold & Worthman, 1993; Cyranowski, Frank, Young, & Shear, 2000), cognition (Steinberg, 2005), and behavioral changes (Carskadon, Acebo, Jenni, Dahl, & Spear, 2004; Steinberg, 2000). Yet, it is not clear how to best evaluate pubertal development (Brooks-Gunn, Warren, Rosso, & Gargiulo, 1987). Here, we compare several measures of pubertal maturation, including hormonal indices. The main hormones responsible for advancing secondary sexual characteristics were captured by measuring testosterone and dehydroepiandrosterone (DHEA), two androgens that facilitate masculine development, and estradiol, an estrogen that facilitates feminine development. We evaluated agreement between physical exam and different methods of self-report, the associations between hormones and the physical exam, and the extent to which self-report methods led to parallel relations with hormonal measures as did the physical exam.

Physical Measures of Puberty

Nearly five decades ago, Tanner (1962) described five stages of puberty, ranging from 1 (no development) to 5 (adult development). These stages capture visible secondary sexual characteristics such as breast/genital development and pubic hair growth. Since its introduction, the gold standard for measuring pubertal status has been a physical exam conducted by a clinician employing Tanner’s...
methods (Dorn, Dahl, Woodward, & Biro, 2006). Yet, researchers often find it difficult to integrate physical exams into nonclinical settings, and the Tanner stages measure only one dimension of development—external signs of physical development. To address these problems, the Pubertal Development Scale (PDS) asks adolescents to answer less invasive questions about puberty without mapping directly onto Tanner stages (Petersen, Crockett, Richards, & Boxer, 1988). The kappa (κ) concordance between the physical exam and PDS is only .24, however (Brooks-Gunn et al., 1987).

An alternative self-report method maps directly onto Tanner stages. Adolescents examine photographs or line drawings of models at each Tanner stage and indicate which image they most closely resemble (Morris & Udry, 1980). Although it is easy to administer, there is only moderate agreement between the physical exam and various versions of self-reported Tanner stage, with an average κ around .50 (see review by Coleman & Coleman, 2002; as well as more recent work by Desmangles, Lappe, Lipaczewski, & Haynatzki, 2006; Hergenreeder, Hill, Wong, Sangi-Haghpeykar, & Taylor, 1999; Schmitz et al., 2004). A few frequently cited studies, however, report excellent agreement (κs above .70; Boas, Falsetti, Murphy, & Orenstein, 1995; Carskadon et al., 1980; Norris & Richter, 2005). Agreement between self-reported PDS and self-reported Tanner stage was also moderate, κ = .50 (Bond et al., 2006). In addition to examining the agreement between multiple puberty measures, we then also explored whether certain factors predicted which adolescents had low agreement between puberty measures.

Hormonal Measures of Puberty

Although steroid hormones advance pubertal maturation, there is an imperfect match of hormones with physical measures of puberty for several reasons. Hormone levels change across the day, there are individual differences in hormone concentrations necessary to advance puberty, and there is overlap in hormone levels across each pubertal stage (Dawes et al., 1999). Nevertheless, knowledge about underlying hormonal processes provides information about pubertal maturation not available from overt physical measures alone.

The earliest peripheral sign of puberty occurs when androgens begin to be released gradually from the adrenal gland (Palmert et al., 2001). DHEA and other adrenal androgens cause pubic hair growth, body odor, acne, and prepubertal growth (Havelock, Auchus, & Rainey, 2004; Lucky, Biro, Simbartl, Morrison, & Sorg, 1997). Adrenal androgens increase twofold in boys from when they show no pubertal development to when they reach adultlike development (Biro, Lucky, Huster, & Morrison, 1995). Puberty shows moderate correlations with DHEA in both sexes (Shirtcliff, Zahn-Waxler, Klimes-Dougan, & Slattery, 2007), although a few frequently cited studies report excellent agreement (κs above .70; Boas, Falsetti, Murphy, & Orenstein, 1995; Carskadon et al., 1980; Norris & Richter, 2005). Although these studies have found moderate to high agreement, additional research is needed to fully understand the relationship between DHEA and pubertal development.

In contrast, estradiol is the primary estrogen released from the gonads and other peripheral tissues (Fernandez-Garcia et al., 2002). Estradiol causes breast development, encourages female-typical fat distributions and long bone fusion during growth spurts, and helps stimulate ovulation and menstruation (Frank, 2003; MacGillivray, Morishima, Conte, Grumbach, & Smith, 1998). Girls with delayed puberty show advancing pubertal maturation when administered estradiol (Finkelstein et al., 1999; Rosenfield et al., 2005). Estradiol is 4–9 times higher in late adolescent girls as compared to childhood (Ikegami et al., 2001). Although more hormonal signals are involved, measuring these three hormones (testosterone, DHEA, and estradiol) should provide converging information about gonadal and adrenal hormonal signals of puberty, two distinct components of maturation in early adolescents. The goal of the current study was to examine how well different physical pubertal stages corresponded with each other and, in addition, how these measures captured basal hormones.

Method

Participants

Participants were 82 boys and 78 girls recruited from the community through an existing laboratory registry and local advertisements. Adolescents ranged from 9 through 14 years of age.
(M = 11.2 years), capturing early adolescence when pubertal stage is most variable. Exclusion criteria included use of allergy or asthma medication. Participants were from diverse backgrounds; Hollingshead scores spanned the full socioeconomic gradient (M = 41.9, SD = 15.6, range = 5–66). Forty-eight percent of participants were White, 26% were Black, and 26% were Asian, Hispanic, Mixed, or unspecified. Body mass index (BMI) was 21.8 on average (SD = 9.3), with 21% of the sample above the 95th percentile of BMI for age (overweight) and 2% below the 5th percentile (underweight).

**Procedures**

Adolescents and their parent(s) provided informed assent and consent, respectively. Adolescents completed the self-report measures, and then a pediatric nurse practitioner (PNP) conducted the physical exam. Participants provided saliva throughout the laboratory day and were sent home with supplies for additional saliva collection. Procedures were approved by the University of Wisconsin Institutional Review Board.

**Measures**

**PDS.** Adolescents completed the five PDS questions about physical development, scored from 1 (no) to 4 (development seems complete; Petersen et al., 1988). Reliability of the PDS was high (α = .77 for boys, α = .81 for girls). Few (3%) adolescents had missing PDS scores. We developed a coding system to convert the PDS to a 5-point scale in order to parallel the physical exam Tanner stages (available upon request). Although interrelated, puberty is not a single event. Therefore, our coding system differentially captured gonadal and adrenal hormonal signals of physical development. In girls, growth spurt, breast development, and menarche are associated with gonadal hormonal signals. In boys, growth spurt, deepening of voice, and facial hair growth are associated with gonadal hormones. For both sexes, pubic/body hair and skin changes are associated with adrenal hormones.

**Picture-Based Interview About Puberty (PBIP).** In a comfortable room, a research assistant spoke with adolescents about “changes that happen when you grow up” with the assistance of a script and photographs (Dorn & Susman, 2002). Following this discussion, researchers left the room while adolescents reported their assessment of pubertal stage. Female researchers interviewed girls and male researchers interviewed half of the boys. There was no difference in accuracy of staging based on the sex of the interviewer, p > .29.

**Physical exam.** Adolescents were given the option to wear a hospital gown or loose clothing for the exam. Height and weight were measured and later used to calculate BMI (age-corrected using Centers for Disease Control and Prevention CDC guidelines). Experienced PNPs were trained to conduct exams for research purposes by the second author. PNPs inspected breast development with brief palpation for girls and visually examined pubic hair. An orchidometer measured testicular size in boys (Genentech, 1997) along with visual inspection of genitals and pubic hair. Interobserver reliability (N = 10 exams, 6.3%) was good, κ = .88. Thirteen percent of participants refused the exam, but assented to self-report measures. Those who refused the exam did not differ in age, race, BMI, or stage; there was a trend for boys (N = 14) to refuse more than girls (N = 6), χ²(1) = 3.2, p = .07.

**Hormonal measures.** Adolescents provided eight saliva samples across the laboratory day, beginning immediately after providing informed consent (M = 9:38, SD = 1:27 hr) through bedtime (M = 21:04, SD = 1:32 hr). Saliva was collected by passive drool (Shirtcliff, Granger, Schwartz, & Curran, 2001). At the time of each sample, participants completed a short diary, which has been used previously (Granger et al., 2003). Samples were immediately frozen at –80°C until they were aliquotted to minimize freeze/thaw cycles. Participants were also sent home with supplies for home collection. To capture the full day, participants collected six samples on each of 4 days at prespecified times between waking (M = 7:45, SD = 1:19 hr) and bedtime (M = 21:17, SD = 1:13 hr), prior to mealtimes. To ensure compliance, cryuvials were stored in time-locked caps (Aardex, Zug, Switzerland), which recorded collection time and date. Samples were stored in home freezers until all samples were collected, then the batch was shipped overnight on ice and stored at –80°C (Dabbs, 1991). All 32 samples were assayed for testosterone and DHEA. One morning sample from each day of saliva collection was assayed for estradiol in girls (M = 9:39, SD = 0:51 hr). Salivary estradiol is not valid in boys (Shirtcliff et al., 2000).

**Hormone determination.** Enzymeimmunoassays were completed by Madison Biodiagnostics (Madison, WI) using Salimetrics’ kits (State College, PA). Samples were measured in duplicate; duplicates that varied by more than 7% were repeat tested. For DHEA, the range of sensitivity was from 5 to 1,000 pg/ml. The average intra-assay coefficient of
variation (CV) was 5.6% and the average interassay CV was 8.2%. For testosterone, the range of sensitivity was from 1 to 600 pg/ml. Average intra- and interassay CVs were 4.6% and 8.3%, respectively. For estradiol, the range of sensitivity was 1–32 pg/ml. Average intra- and interassay CVs were 7.1% and 7.5%, respectively.

A basal hormone measure was calculated using hierarchical linear modeling, which separates within-the-day and day-to-day variation in hormone levels (N = 3,704) from individual basal levels (N = 160), thereby allowing removal of the effects of several important control variables and accurate aggregation across repeated measures of each hormone to a single basal level. At the within-the-day level, we controlled for linear and quadratic time since waking (in minutes) and time of day to account for the individual’s intrinsic and extrinsic rhythm. These values were allowed to vary so that each individual had his or her own rhythm removed from the basal estimate. An empirical Bayes estimate of each log-transformed hormone was extracted after accounting for additional control variables at both within-the-day and day-to-day levels (e.g., flow rate, response to awakening, location, medication usage, exercise, emotion, who the child was with). Basal DHEA comprised 73.8% of the total variation in DHEA, p < .0001. Basal testosterone comprised 80.0% of the variation in testosterone, p < .0001. For estradiol, an average across the five samples/individual was calculated as analyses revealed no day-to-day predictors of estradiol (including controls above as well as menstrual cycle day-count, cycle regularity, menarchal status). Basal estradiol comprised 36% of the variation in estradiol, p < .0001. Less variation in estradiol was basal than the other hormones, perhaps due to the reduced number of samples.

Statistical analyses. Kappas and percent accuracy described precise agreement between the three puberty measures. Pearson correlations examined whether measures were associated, without necessitating precision. To examine whether adolescents were inaccurate informants, we calculated the discrepancy between the exam and the PDS and PBIP, respectively, using a difference score. We assessed whether gender, age, stage, BMI, or race influenced accuracy of adolescents’ self-report using linear regression. Race was coded as White, Black, or Other. Structural equation modeling (SEM) simultaneously examined how the physical exam was associated with steroid hormones, with separate models for boys and girls (because estradiol was measured in girls only). The physical exam was first modeled with basal hormones, removing nonsignificant coefficients. Poor model fit was indicated by significant $\chi^2$ values, comparative fit index (CFI) less than .95, or root mean square error of approximation (RMSEA) greater than .10. Next, parallel models were fit substituting the respective self-report measures. To test whether models were parallel to the exam, we fixed coefficients to be identical to the physical exam and examined the reduction in model fit compared to when coefficients were unconstrained. If the indices of practical fit were too high/low or the $\chi^2$ was significant (indicating models were not parallel), we removed constraints on coefficients which resulted in the greatest model improvements.

Results and Discussion

How Did Self-Report PDS Map Onto the Physical Exam?

Correlations between the physical exam and the PDS are presented in Table 1. The concordance between the physical exam and the PDS gonadal stage was modest, $\kappa = .36, \chi^2(16) = 93.0, p < .0001$ (Table 2). Accuracy was defined as self-report of the same stage as the physical exam. Fifty-two percent of adolescents’ gonadal scores were accurate (54% boys, 47% girls), whereas 18% underestimated (15% boys, 27% girls) and 30% underestimated stage (31% boys, 27% girls) compared to the exam. The concordance between the physical exam and the PDS adrenal stage was also modest, $\kappa = .36, \chi^2(16) = 90.6, p < .0001$ (Table 3). Fifty percent of adolescents were accurate (60% boys, 44% girls), whereas 29% underestimated (26% boys, 34% girls) and 21% overestimated pubic hair (14% boys, 23% girls).

How Did the PBIP Map Onto the Physical Exam?

The concordance between the physical exam and PBIP breast/genital stage was modest, $\kappa = .36, \chi^2(16) = 120.9, p < .0001$ (Tables 1 and 2). Forty-nine percent of adolescents reported the same breast/genital stage as the exam (41% boys, 57% girls), whereas 26% overestimated (35% boys, 17% girls) and 25% underestimated stage (24% boys, 17% girls). The parallel concordance for pubic hair was good, $\kappa = .43, \chi^2(16) = 137.2, p < .0001$ (Table 3). Fifty-six percent of adolescents reported the same pubic stage as the exam (54% boys, 58% girls), whereas 24% overestimated (26% boys, 21% girls)
Table 1
Intercorrelations of the Physical Exam, Picture-Based Interview About Puberty (PBIP), and Pubertal Development Scale (PDS) With Girls Above (in Italics) and Boys Below (in Bold)

<table>
<thead>
<tr>
<th></th>
<th>Physical Exam</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Breast/genital</td>
<td>Pubic hair</td>
<td>Breast/genital</td>
<td>Pubic hair</td>
<td>Gonadal</td>
<td>Adrenal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast/genital</td>
<td>.93</td>
<td>.85</td>
<td>.83</td>
<td>.76</td>
<td>.65</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pubic hair</td>
<td></td>
<td>.73</td>
<td>.75</td>
<td>.88</td>
<td>.69</td>
<td>.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast/genital</td>
<td>.60</td>
<td>.60</td>
<td></td>
<td>.79</td>
<td>.77</td>
<td>.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pubic hair</td>
<td></td>
<td>.69</td>
<td>.71</td>
<td>.71</td>
<td>.73</td>
<td>.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonadal</td>
<td>.65</td>
<td>.63</td>
<td>.59</td>
<td>.68</td>
<td></td>
<td>.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adrenal</td>
<td>.63</td>
<td>.68</td>
<td>.70</td>
<td>.70</td>
<td></td>
<td></td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>Mean (SD) stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td>2.4 (1.3)</td>
<td>2.3 (1.3)</td>
<td>2.7 (1.2)</td>
<td>2.3 (1.2)</td>
<td>2.0 (1.1)</td>
<td>2.3 (1.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girls</td>
<td>2.9 (1.5)</td>
<td>2.8 (1.5)</td>
<td>2.9 (1.2)</td>
<td>2.7 (1.4)</td>
<td>2.5 (1.3)</td>
<td>2.9 (1.3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. All intercorrelations have ps < .001. There are no mean differences in staging between the physical exam and PBIP, ps > .2, or PDS, ps > .06.

Table 2
Concordance of the Physical Exam Breast/Genital Stage, Pubertal Development Scale (PDS), and the Picture-Based Interview About Puberty (PBIP)

<table>
<thead>
<tr>
<th>A. PDS gonadal score</th>
<th>Physical exam breast/genital stagea</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Totalb</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>61.1</td>
<td>28.1</td>
<td>10.3</td>
<td>5.0</td>
<td>5.3</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>II.</td>
<td>19.4</td>
<td>53.1</td>
<td>20.7</td>
<td>15.0</td>
<td>5.3</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>III.</td>
<td>16.7</td>
<td>18.8</td>
<td>51.7</td>
<td>30.0</td>
<td>21.1</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>IV.</td>
<td>2.8</td>
<td>6.9</td>
<td>30.0</td>
<td>26.3</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>V.</td>
<td>10.3</td>
<td>20.0</td>
<td>42.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (N)b</td>
<td>36</td>
<td>32</td>
<td>29</td>
<td>20</td>
<td>19</td>
<td>136</td>
<td></td>
</tr>
</tbody>
</table>

| B. PBIP              | Physical exam breast/genital stagea | I: No development | 54.1| 18.2| 3.4 | 4.8 | 28 |
|                      |                                   | II: Breast bud     | 24.3| 48.5| 20.7| 14.3| 34 |
|                      |                                   | II: Testes started to grow | 18.9| 30.3| 48.3| 19.0| 4.8 | 36 |
|                      |                                   | III: Breast tissue beyond areola | 2.7 | 3.0 | 17.2| 61.9| 15 |
|                      |                                   | IV: Penis growth in length | 2.7 | 3.0 | 17.2| 61.9| 33 |
|                      |                                   | IV: Penis growth in width and length | 10.3 | 33.3 | 32.6| 53.3| 36 |
|                      |                                   | Total (N)b         | 37 | 33 | 29 | 21 | 141 |

<table>
<thead>
<tr>
<th>C. PBIP</th>
<th>PDS gonadal scorea</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Totalb</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>51.3</td>
<td>18.4</td>
<td>4.3</td>
<td></td>
<td></td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>II.</td>
<td>33.3</td>
<td>44.7</td>
<td>13.0</td>
<td>13.3</td>
<td>5.9</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>III.</td>
<td>12.8</td>
<td>28.9</td>
<td>41.3</td>
<td>26.7</td>
<td>11.8</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>IV.</td>
<td>2.6</td>
<td>7.9</td>
<td>32.6</td>
<td>53.3</td>
<td>52.9</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>V.</td>
<td>8.7</td>
<td>6.7</td>
<td>29.4</td>
<td></td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total (N)b</td>
<td>39</td>
<td>38</td>
<td>46</td>
<td>15</td>
<td>17</td>
<td>155</td>
<td></td>
</tr>
</tbody>
</table>

aColumn percentages. bNumber of participants.
and 20% underestimated stage (19% boys, 21% girls).

How Did Self-Report PDS Map Onto the PBIP?

Correlations between the two self-report measures are reported in Table 1. The concordance between the PDS gonadal stage and the breast/genital PBIP stage was low, $\kappa = .29$, $\chi^2(16) = 98.4$, $p < .0001$ (Table 2). Forty-five percent of adolescents reported the same breast/genital PBIP stage as the PDS gonadal score (37% boys, 52% girls). The parallel concordance for pubic hair was moderate, $\kappa = .37$, $\chi^2(16) = 152.1$, $p < .0001$ (Table 3). Fifty-two percent of adolescents reported the same pubic stage on the PBIP as the PDS (47% boys, 57% girls).

The physical exam is well suited for a wide range of behavioral endocrinology-oriented questions or when an objective measure of physical development is desirable (Dorn et al., 2006). Though precise agreement was modest, the two self-report measures were correlated with the physical exam, suggesting they mutually capture underlying pubertal processes. If precision is not necessary, adolescents are relatively good observers. We should note that 13% of adolescents refused the physical exam, but none refused the PBIP. Researchers conducting an exam might consider supplementation with a self-report measure to capture this subset.

Which Adolescents Were Inaccurate Informants?

Using linear regression, where the discrepancy between the physical exam and PDS was the outcome, we found that neither sex nor BMI influenced accuracy, $ps > .09$. Stage (based on the exam) qualified an effect of age ($\beta = .32$, $p < .003$ for age; $\beta = -.68$, $p < .001$ for stage). In general, adolescents overestimated pubertal maturation when they were at lower stages of development relative to their peers and underestimated development when they were at higher stages than their peers. As expected, this distortion of staging was age specific. For example, 11- and 12-year-olds were accurate at Stage 3 but overestimated Stage 2 and underestimated Stages 4+; 13-; and 14-year-olds were
accurate at Stage 4, but tended to overestimate Stages 2 and 3 and underestimate Stage 5. This may reflect the desirability of adolescents to appear like the developmental stage that is most typical for their age. White adolescents overestimated stage more often than non-Caucasian adolescents ($\beta = .21$, $p < .02$).

Analyses of the discrepancies between the physical exam and PBIP yielded similar findings. Sex and BMI did not influence accuracy, $ps > .15$. Stage qualified the effect of age, ($\beta = .32$, $p < .003$ for age; $\beta = -.68$, $p < .001$ for stage), such that adolescents sometimes overestimated development when at lower stages and underestimated development when at higher stages. Again, adolescents tended to report stages that were most typical of their age. White or Black adolescents overestimated stage more often than other adolescents ($\beta = .19$, $p = .03$).

Young adolescents may not be able to self-report an exact stage—particularly if they are maturing earlier or later than their peers. Measurement problems may primarily affect studies that employ a cut-score to describe adolescents as pre- or postpubertal rather than as a continuous process; this may be especially problematic in research designed to isolate early and late maturing adolescents.

**Was the Physical Exam Stage Associated With Hormones?**

Table 4 presents hormone values across Tanner stages for boys and girls. Boys’ basal testosterone and DHEA were predicted by the physical exam, with one exception. Genital development was not associated with DHEA; dropping this coefficient did not change model fit, $\chi^2(1) = 0.5$, $p = .46$, CFI > .999, RMSEA < .0001 (Figure 1A). For girls, breast development was not associated with testosterone or DHEA, and pubic hair was not associated with estradiol (Figure 2A). These three coefficients were dropped without reducing the goodness of fit, $\chi^2(3) = 1.75$, $p = .63$, CFI > .999, RMSEA < .0001. In sum, the physical exam captured basal testosterone and DHEA well in both sexes, though estradiol was modestly related to the exam.

We were surprised that breast development explained such a small amount of variability in estradiol and await replication in a different (and perhaps older) sample or in which more of the variability was basal. Pubic hair was particularly good at capturing basal hormones in both sexes. The endocrine signaling of pubic hair generally begins earlier (between ages 6 and 9) and may be more established than breast/genital development in this early age range. When interested in comparing boys and girls, assessments of pubic hair may be emphasized, as they performed well in both sexes.

**Was the Self-Report PDS Associated With Hormones Like the Physical Exam?**

An identical model substituting the PDS stages for boys demonstrated marginal model fit, $\chi^2(4) = 9.15$, $p = .06$, CFI = .97, RMSEA = .13, but no single coefficient differed from the physical
exam model, ps > .15. Like the exam, gonadal development measured using the PDS did not predict DHEA, p = .58, but all other coefficients were significant. The PDS basically led to parallel relations with boys’ basal hormones, as did the physical exam (Figure 1B).

For girls, an identical model substituting PDS stages for the exam resulted in marginal model fit, $\chi^2(6) = 10.7$, p = .10, CFI = .96, RMSEA = .10. Three coefficients were substantially different from the physical exam, accounting for most of the model misspecification, $\chi^2(3) = 8.4$, p = .04, CFI = .96, RMSEA = .15. Unlike the physical exam, the PDS gonadal score was associated with testosterone and DHEA, and the PDS adrenal score was not as highly related to DHEA as in the physical exam model. Constraining the remaining coefficients to be parallel to the physical exam did not reduce model fit, $\chi^2(3) = 4.16$, p = .25, CFI = .99, RMSEA = .07, indicating that the PDS gonadal score was more broadly related to basal hormones than the physical exam, whereas the adrenal score was less predictive of girls’ basal DHEA (Figure 2B).

Figure 1. Structural equation model for boys.
Note. (A) Standardized $\beta$ coefficients when testosterone and dehydroepiandrosterone (DHEA) are predicted by the physical exam genital and pubic hair stage. (B) Model parallel to (A) with testosterone and DHEA predicted by the Pubertal Development Scale stages (in italics) and Picture-Based Interview About Puberty stages (in bold).
*p < .05, **p < .01.

Figure 2. Structural equation model for girls.
Note. (A) Standardized $\beta$ coefficients when testosterone, dehydroepiandrosterone (DHEA), and estradiol are predicted by the physical exam breast and pubic hair stage. (B) Model parallel to (A) with basal hormones predicted by the Pubertal Development Scale (in italics) and Picture-Based Interview About Puberty (in bold).
*p < .05, **p < .01.

Given that the PDS is a common measure and is easily employed in a variety of settings (e.g., schools, screening mailers), it should be welcome news that this self-report measure captured basal hormones in parallel with the physical exam in boys; in girls, the gonadal score performed slightly better than the exam. That girls’ adrenal score was not associated with basal hormones is perplexing because the adrenal score included items, such as body/pubic hair growth and skin changes, that are related to hormones like DHEA (Grumbach, 2002).

Because the basal hormones were complicated, all SEM analyses were recalculated using a simple average. These models fit similarly—the RMSEA was on average .05 different, the $\chi^2$ never exceeded 2.3, and the percent of variance in each hormone explained by pubertal status differed on average by only 3.7%.
Was the PBIP Associated With Hormones Like the Physical Exam?

An identical model that substituted PBIP stages for the physical exam for boys demonstrated poor model fit, \( \chi^2(4) = 13.4, p < .01, \text{CFI} = .94, \text{RMSEA} = .17 \), indicating that PBIP was not parallel with the physical exam (Figure 1B). We allowed genital development to be related to DHEA and found it significantly predicted DHEA, \( p = .009 \). When the other three coefficients were constrained to be identical to the exam model, goodness of fit was excellent, \( \chi^2(3) = 0.87, p = .8, \text{CFI} > .999, \text{RMSEA} < .0001 \), suggesting that PBIP led to parallel relations with basal hormones as did the physical exam and, in addition, that the PBIP genital stage predicted DHEA better than the physical exam.

For girls, an identical model that substituted PBIP for the physical exam fit well, \( \chi^2(6) = 3.96, p = .7, \text{CFI} > .999, \text{RMSEA} < .0001 \), indicating that PBIP captured basal hormones in a similar manner as the physical exam (Figure 2B). In sum, the PDS and PBIP were related to basal hormones in parallel or occasionally better than the physical exam.

That the PBIP mapped onto basal hormones in parallel to the physical exam (or slightly better for boys’ prediction of basal DHEA) is an additional advantage of the PBIP for potentially addressing hormone-related research questions. Use of the PBIP is most viable (a) when high correlations with an objective measure like the physical exam are sought after, (b) when the Tanner metric is desirable, and (c) when basal hormones (particularly in boys) are outcomes of interest or are proximally associated with outcomes of interest. Nevertheless, even the best measure of external pubertal status captured less than half of the variability in basal hormones. Directly measuring hormones is often feasible.

Although we were agnostic about which measure would be optimal, we were surprised that self-reported PDS and PBIP scores were occasionally better correlates with basal hormones than the exam. This may be due to the unique perspectives of clinicians and adolescents. Although clinicians have a range of knowledge comparing one adolescent to another, rarely do they observe the same adolescent across time. In contrast, adolescents have little experience with other individuals, yet they have daily insights into their own pubertal changes. The adolescents’ perspective may be optimal for noticing changes in their bodies across months and years. Basal hormones likewise capture a gradual, continuous developmental process. Adolescents may generally be more attuned to the confluence of this internal process with external developmental changes. Choosing measures that encompass the subjective adolescent experiences may be suitable for many biopsychosocial research questions.

Limitations

Several limitations should be considered. First, although our study is ethnically and socioeconomically diverse, the sample size limited the extent to which we could explore individual differences, such as mechanisms behind racial differences in accuracy of self-report. Second, other hormones involved in pubertal maturation (e.g., DHEA-sulfate, androstenedione, and progesterone) could yield different associations with exam and self-report measures. Third, estradiol varies across the menstrual cycle. This limitation is noticeable because estradiol was weakly related to pubertal development. Girls were not recruited to come to the laboratory during a particular phase of their menstrual cycle because (a) estradiol begins to cycle years before girls’ first menstruation (menarche), so this would reduce cycle effects in menarcheal but not premenarcheal girls; (b) after menarche, cycles are often irregular (48% of our girls), so it would be difficult to schedule by day-count; and (c) days in which girls were likely accurate (i.e., during menstruation) are when estradiol is at its nadir and least likely to differentiate early from late puberty (Dawes et al., 1999). More frequent or systematic estradiol measurement may reveal stronger associations with puberty than we found.

Conclusions

Here we reported different ways to measure pubertal development in early adolescence. Our broad goal was to understand the association between these measures so that researchers can be informed about which measure(s) best addresses particular research questions. Because puberty encompasses a suite of changes and is not a single process, different measures may best capture different aspects of adolescent development. The answer about which measure(s) is best may depend on which aspects of puberty are of interest for a particular study or research question. Inclusion of pubertal measures provides essential information about developmental changes in adolescence.
References


