1 Methodological concerns in puberty-related research

Chris Hayward

Defining puberty

Puberty is not a single event, but rather a complex metamorphosis. It is a cascade of changes that result in adult appearance, adult physiology, and altered identity. Although sexual dimorphism, differences in form and structure between males and females, are initiated at conception, some of the most salient biological differences between males and females emerge during pubertal transition. However, identifying exactly when puberty begins has been difficult. It is easier to know that puberty has already started than to pinpoint its exact onset, since the initiation of puberty is not completely understood.

As described by Patricia Fechner in chapter 2, puberty consists of both adrenarche and gonadarche. Adrenarche occurs when the adrenal gland begins to increase production of androgen in both males and females, and is responsible for the development of pubic and axillary hair. This begins much earlier than what is typically thought of as the age of onset of puberty, beginning normatively as early as 6 years of age and typically having started by 8 years of age. Gonadarche is characterized by the development of the gonads, with increased release of estrogen in females and testosterone in males, which results in breast development in girls and testicular enlargement in boys.

As puberty is a process and not an event, its definition partly depends on the purpose for which the definition is being used. It is not necessary to measure hormones to define puberty if the purpose of the definition is to determine rate of growth. On the other hand, if an understanding of the interplay between different aspects of puberty is desired then the definition and measurement need to be more complex. In determining the source of the decrement in body image that many girls experience at puberty, to take one specific example, it may be best to measure multiple characteristics of puberty (increase in body fat, breast development, hormones, etc.), as well as the contextual factors in which these biological changes occur (degree of weight-related teasing by peers, media-induced

culture of the thin ideal, parent preoccupation with body weight and shape, etc.). Arguably, both the individual's pubertal changes and the context in which these changes occur constitute the best definition of puberty for understanding issues such as body image change. In fact, it can be argued that a full understanding of most psychological aspects of puberty requires measuring both the individual pubertal changes and the environmental factors that give these changes meaning. In this view, the definition of puberty is "purpose dependent" and in its more complex form includes interrelated biological, psychological, and social factors.

Measuring puberty

Having argued that the definition of puberty can either be narrow (e.g., Tanner stage) or broad (e.g., Tanner stage, hormones, growth, social context, etc.), depending on the purpose for which the definition is to be used, then it follows that the appropriate measurement of puberty is also "purpose dependent." Different biological systems are developing at different rates and times and may have variable downstream effects (e.g., estrogen's effect on serotonin), intrapsychic meaning, and elicit different external responses. Although the measurement of puberty using different markers may yield highly correlated indicators, they are not equivalent. For example, puberty may be measured by assessing secondary sexual characteristics (e.g., Tanner staging either by physical exam or self-report), bone age, growth spurt, menarche, or hormonal indicators (estrogen, testosterone, or adrenal androgen, etc.). None of these represent a "gold standard," as each captures a different aspect of the pubertal process. Each indicator may be more or less an imperfect proxy for another. If the purpose of the measurement is to determine general categories (e.g., prepubertal or not), then any of these indicators may suffice. On the other hand, if the purpose is to determine any "direct effect" an indicator might have on an outcome (versus one indicator being a proxy for another), multiple indicators must be measured (see below).

Thus, the selection of the appropriate indicator of puberty is best based on the desired purpose, but in practice (clinical and research) it is also determined by convenience, feasibility, and cost. It is important to note, therefore, the limitations of various pubertal indicators. In early adolescent girls self-reported onset of first menses may be difficult to measure reliably (Petersen, 1983; Hayward, *et al.*, 1997). For example, a girl may have her first period followed by several months of being amenorrheic. On the other hand, in older adolescents and adults, menarche is a reliable measure of puberty (Petersen, 1983; Dubas, Graber, and Petersen, 1991; Brooks-Gunn and Warren, 1985). Menarche is also the most commonly

used measure in psychological research, as it is easily collected. Unfortunately, there is no equivalently validated convenient measure of puberty for boys.

Self-reported Tanner stage can be measured in both sexes and has fairly good agreement with physician examination, but the validity of self-ratings may vary by ethnicity and degree of body image disturbance in girls (Litt, 1999; Hick and Katzman, 1999). Also, Tanner self-staging requires showing diagrams of genitalia. This can be problematic in non-clinical studies. For this reason, self-ratings that use a questionnaire index (e.g., the Petersen Development Scale) may be preferable (Petersen, *et al.*, 1988) and can be given to both sexes. Physician visual inspection versus physical examination may confound puberty and obesity (Kaplowitz, *et al.*, 2001). Measurements of hormonal indicators have methodological problems as well. Diurnal, menstrual cycle, and pubertal variations make cross-sectional measurements of sex hormone levels difficult to interpret. Measurements at the same time of day, at the same stage of the menstrual cycle in girls would be ideal. Longitudinal hormonal measurements are often more informative, allowing for estimates of rate of change and direction of change over time. Finally, because of the variability in the tempo of various aspects of puberty (e.g., female increase in body fat occurs later than height spurt), relationships between different indicators vary by pubertal stage (see chapter 8 below). There may be individual asynchronies in the sequencing of pubertal changes (e.g., delayed height spurt), which can have significant psychological effects (Eichorn, 1975). Ideally, multiple indicators of puberty measured over time provide the best characterization of the pubertal process. Short of this, qualifying inferences from measurements that are inevitably less than ideal continues to be the best protection against unwarranted conclusions.

Differentiating different pubertal effects

As I have previously stated, different indicators of puberty may be more or less correlated. For example, teasing apart the effects of adiposity from timing of menarche (Striegel-Moore, *et al.*, 2001) or the effect of Tanner stage from estrogen levels at puberty (Angold, *et al.*, 1999) can be challenging. In studying how different aspects of puberty are related to outcomes, how can their effects be differentiated? Any apparent association between puberty and an outcome is going to be dependent on which pubertal process is measured. If body image worsens in most females at puberty this change will likely be associated with increases in estrogen, BMI, Tanner stage, height, and so forth. Which, if any, of these different components of puberty is most critical for understanding the

development of body image disturbance in girls at puberty? The most common statistical method used to parcel effects is multiple regression, the results of which are partially dependent on the measurement characteristics of each variable and the degree of colinearity between variables. Highly correlated independent variables, such as different indicators of puberty, may yield unstable results. Dimensional variables and variables with a metric that has a broad distribution and low measurement error yield larger effect sizes. For example, BMI is frequently observed to be a more powerful predictor than self-reported pubertal stage in multiple regression analyses. Yet, BMI and Tanner stage are highly correlated. Which is more important? For the purposes of multiple regression, BMI has better measurement characteristics, as it offers a continuous measure usually with a broad distribution, whereas the measure of pubertal stage is ordinal and frequently the samples used are truncated at one of the extremes of the five Tanner stages. Self-reported Tanner stage is also less likely to be reliable than direct measurements of height and weight. By virtue of the different measurement characteristics and all other things being equal, BMI would be expected to have a better chance of showing more of an association than Tanner stage. Techniques such as centering and rescaling can address some of the differences in measurement characteristics of the different indicators of puberty, although not measurement unreliability. The problem of parceling effects from colinear variables is more insidious.

If the goal is to have an overall marker of puberty, then strategies to deal with colinearity can include creating an index (i.e., combining different indicators into one index) or factor analysis that produces a set of truly independent variables. However, if the intent is to determine the relative contribution of different (but correlated) aspects of puberty, then stratifying the sample on those factors of interest may be preferable. For example, examining the effects of BMI within Tanner stage groups on a particular outcome would allow for differentiating effects attributable to increasing BMI while holding pubertal stage constant. Similarly, examining effects of Tanner stage within different BMI levels allows an estimate of pubertal stage effects while holding BMI constant. Because stratifying a representative sample by two highly correlated variables will yield smaller numbers at the "corners" (e.g., low BMI at Tanner stage 5 and high BMI for Tanner 1), sampling stratification may be necessary to provide adequate power.

Finally, BMI and pubertal stage may interact in their effect on an outcome. Although including interaction terms is the preferred method for testing for interactive effects, negative findings may be subject to type II error, as more statistical power is required to observe significant

interaction effects compared to main effects. This raises the practical problem of adequate sample sizes for teasing apart the main and interactive effects of correlated indicators of puberty; sample sizes of less than 100 subjects are rarely adequate and typically samples need to be quite large (e.g., 500–1000 subjects).

Differentiating pubertal status and pubertal timing effects

Differentiating age from pubertal status effects is important in determining if outcomes occurring in early adolescence are part of "getting older" or are linked specifically to puberty (Angold and Worthman, 1998). Examining pubertal status effects within the age groups where variation in pubertal status is expected provides information about the relative contribution of pubertal status at different ages and vice versa. The age range in which this can be accomplished is limited and differs between the genders (later in boys). Figure 1.1 shows hypothetical data demonstrating age effects and not pubertal stage effects, while figure 1.2 shows the reverse. Figure 1.3 shows additive effects of age and pubertal status and figure 1.4 demonstrates an interaction between age and pubertal status. Interestingly, interactive effects between age and pubertal status suggest a pubertal timing effect. These two features of puberty, pubertal status and pubertal timing, are sometimes confused (Steinberg, 1987). Pubertal

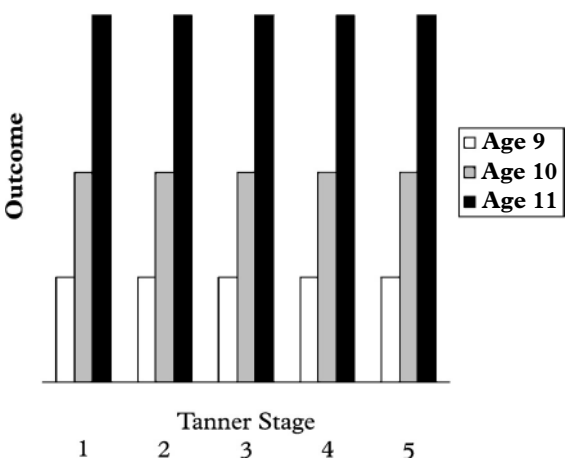

Figure 1.1 Hypothetical outcome data showing stratification by age and Tanner Stage. This figure shows an age effect but no pubertal stage effect.

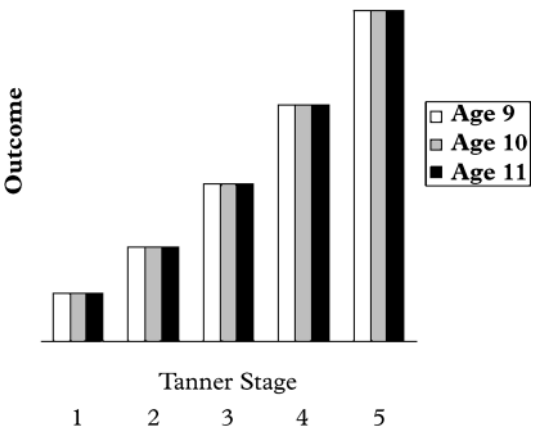

Figure 1.2 Hypothetical outcome data showing stratification by age and Tanner Stage. This figure shows a pubertal stage effect but no age effect.

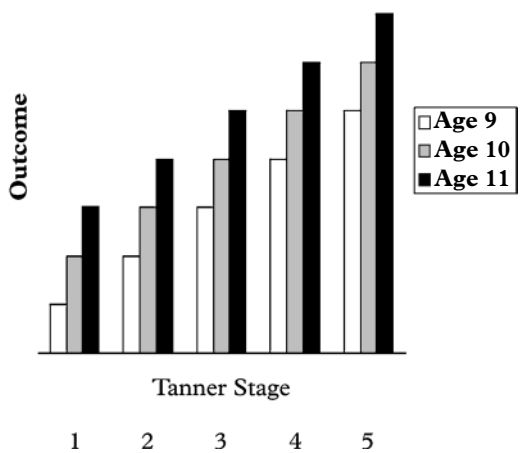

Figure 1.3 Hypothetical outcome data showing stratification by age and Tanner Stage. This figure shows an additive age and pubertal stage effect.

status refers to the level or stage of pubertal development, while pubertal timing refers to the age of a pubertal event and is often categorized early, on time, or late in comparison to a defined reference group. Measuring pubertal status effects requires a sufficient distribution of subjects in different pubertal stages. Obviously, pubertal status effects cannot be measured prior to puberty or after its completion.

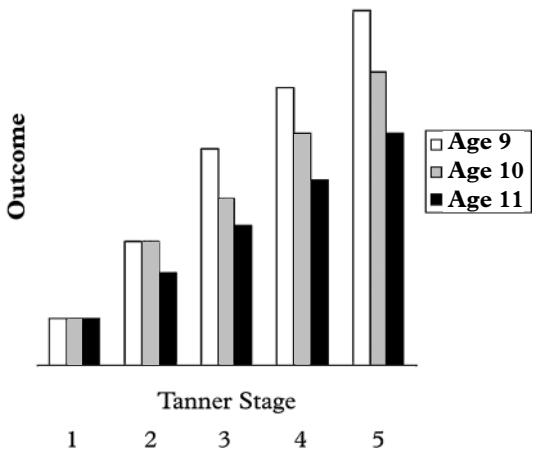

Figure 1.4 Hypothetical outcome data showing stratification by age and Tanner Stage. This figure shows an interaction effect between age and pubertal stage. The interaction in this figure represents an early pubertal timing effect.

Pubertal timing effects are, however, often confounded with pubertal status effects in cross-sectional studies limited to one age or grade. If more sexually mature fifth grade girls have higher depression scores, it is difficult to know if this is a status effect or a timing effect. The less sexually mature girls may or may not "catch up" when they proceed through puberty. Longitudinal studies or studies with sufficient age distributions across all levels of pubertal development can help differentiate status effects from timing effects (Angold and Worthman, 1998; Ge, Conger, and Elder, 2001). Also, both pubertal status and timing effects may be important. In other words, there may be a main effect for pubertal status and an interaction effect between age and status (i.e., a timing effect). This is graphically shown in figure 1.4.

Untangling short-term and long-term pubertal effects can also be difficult. For example, the sexually mature sixth grader may have more depression than the eighth grader at the same level of sexual maturation, but both may be similar by tenth grade. Cross-sectional studies in the peripubertal age range cannot differentiate short-term pubertal timing effects from long-term timing effects that persist after all subjects have completed puberty. Longitudinal studies that continue past the time when most subjects have completed puberty (Stattin and Magnussen, 1990) or studies of postpubertal subjects who retrospectively report their pubertal timing can both yield results that provide information about long-term pubertal

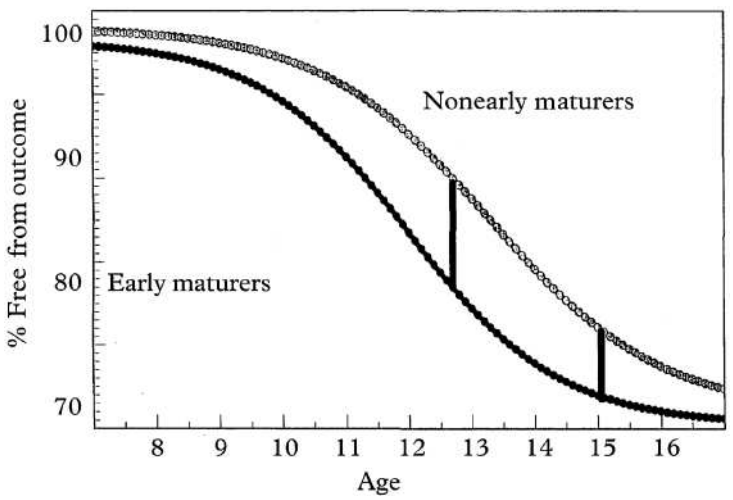

Figure 1.5 Survival curves using hypothetical outcome data comparing those with early pubertal timing and those with nonearly pubertal timing. This figure demonstrates a short-term early pubertal timing effect.

timing effects (Graber, *et al.*, 1997). However, the retrospective report of pubertal timing may be subject to recall bias. Figures 1.5, 1.6 and 1.7 show survival curves from hypothetical data to demonstrate short-term, long-term, and no pubertal timing effects.

It is also difficult to know the degree to which even purported long-term pubertal timing effects are related to the length of time between the onset of puberty and the measurement of the outcome. For example, if increasing levels of estrogen at puberty are found to be related to depression in girls with early onset of puberty, is this due to problems of being an "early bloomer" or to the effects of a longer exposure to estrogen? Measuring the outcome in all subjects at the same time interval from the onset of puberty while controlling for age may help. For example, if depression is measured at age 16 in subjects with pubertal onset at age 10, a comparable test would be rates of depression in 18-year-olds who had onset of puberty at age 12, adjusting for age effects. Statistically controlling for the number of years since pubertal onset might accomplish the same end.

In summary, evaluating status effects requires dividing samples into different levels of pubertal development during the peri-pubertal time period. Observing short-term pubertal timing effects requires knowing the age most subjects start puberty and for long-term pubertal timing

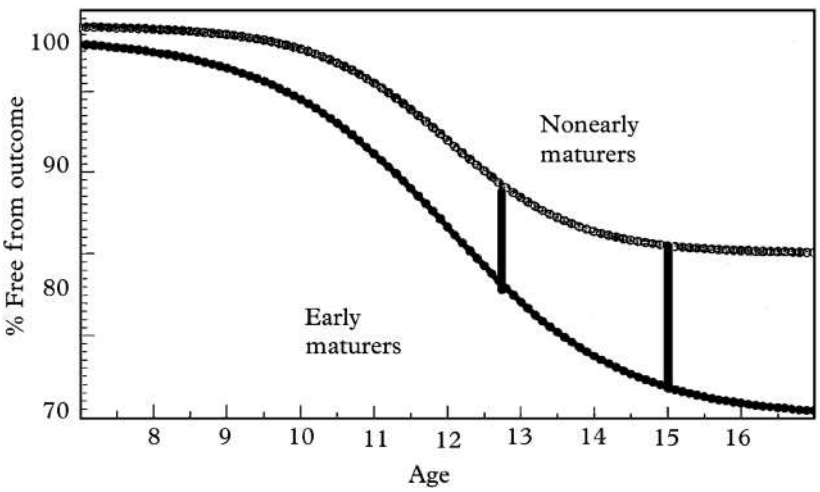

Figure 1.6 Survival curves using hypothetical outcome data comparing those with early pubertal timing and those with nonearly pubertal timing. This figure demonstrates a long-term early pubertal timing effect.

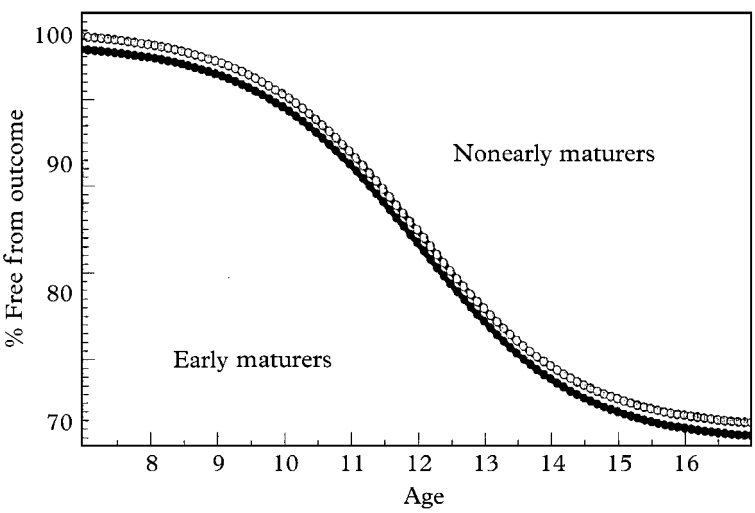

Figure 1.7 Survival curves using hypothetical outcome data comparing those with early pubertal timing and those with nonearly pubertal timing. This figure demonstrates no early pubertal timing effect.

effects most subjects must have completed puberty. Controlling for the length of time "exposed" to puberty may help separate effects that are related to the amount of time since pubertal onset versus effects of being an early maturer. Finally, most of the comments offered in reference to early maturation apply equally well to late maturation.

Recent secular trend in the onset of secondary sexual characteristics but not menarche

One of the most puzzling recent findings in puberty research is the apparent decrease in the age of onset of secondary sex characteristics by approximately one year in Caucasians and African American girls in the United States over the last two decades, while the mean age of menarche has remained unchanged (Herman-Giddons, *et al.*, 1997). Is pubertal onset earlier but the tempo (duration of puberty) slower? This reported finding has received considerable attention in the media and has alarmed those who are concerned about the psychological well-being of girls who enter into puberty at a younger age. In this report the mean age of beginning breast development was 10 for Caucasian girls and just under 9 for African American girls. Previously, Tanner reported mean age for breast development onset to be 11.2 years of age (Marshall and Tanner, 1969). Similar secular trends were observed for pubic hair growth in girls. No decrease in age of onset at puberty was observed for boys.

The finding that the onset of secondary sex characteristics is occurring earlier in Caucasian and African American girls comes from one study, the Pediatric Research in Office Settings Network (Herman-Giddens, *et al.*, 1997). Consisting of trained pediatricians, this network reported staged breast and pubic hair development in over 17,000 children between the ages of 3 and 12. There are some notable limitations to this study. The sample, although large, is drawn from pediatric office visits and is not representative of the general population. Parents who were concerned about early breast development may have been more likely to bring their daughters to the pediatrician (although not necessarily state that as the reason for the visit) than those parents without this concern. This would bias the sample in favor of early maturers.

Another concern was that breast development staging was performed by visual inspection for 60 percent of the sample. In obese girls, increased fat can be mistaken for breast tissue. A follow-up study using the same sample addressed the role of BMI in explaining the reported findings (Kaplowitz, *et al.*, 2001). When the reanalysis included only the subsample (40%) that received breast examinations, the observation of earlier age of breast development persisted. Given the trend for increasing levels