Statistical Methods in Medical Research
http://smm.sagepub.com/

Growth charts of human development

Stef van Buuren

Stat Methods Med Res published online 12 March 2013
DOI: 10.1177/0962280212473300

The online version of this article can be found at:
http://smm.sagepub.com/content/early/2013/02/22/0962280212473300

Published by:
SAGE
http://www.sagepublications.com

Original Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–23

DOI: 10.1177/0962280212473300
smm.sagepub.com
SAGE

Growth charts of human development

Stef van Buuren[1,2]

Abstract

This article reviews and compares two types of growth charts for tracking human development over age. Both charts assume the existence of a continuous latent variable, but relate to the observed data in different ways. The *D-score diagram* summarizes developmental indicators into a single aggregate score measuring global development. The relations between the indicators should be consistent with the Rasch model. If true, the *D*-score is a measure with interval scale properties, and allows for the calculation of meaningful differences both within and across age. The *stage line diagram* describes the natural development of ordinal indicators. The method models the transition probabilities between successive stages of the indicator as smoothly varying functions of age. The location of each stage is quantified by the mid-*P*-value. Both types of diagrams assist in identifying early and delayed development, as well as finding differences in tempo. The relevant techniques are illustrated to track global development during infancy and early childhood (0–2 years) and Tanner pubertal stages (8–21 years). New reference values for both applications are provided.

Keywords

Bayley scale, Tanner stages, references, D-score diagram, stage line diagram, age continuity, gain score

1 Introduction

Growth charts are widely being used to track growth in children. Conventional growth diagrams portray the distribution of continuous measures (e.g. height, weight) against age. Growth charts aid in detecting and monitoring growth-related diseases in children. Methods for fitting growth charts to continuous measures have been well developed during the last two decades. The most popular approach for fitting growth references is the LMS method,[1] but many other models exist.[2,3] Specialized diagnostic methods have been developed for evaluating model fit,[4,5] many of which have been implemented in GAMLSS.[6]

It has been long recognized that clinical practice should assess both growth and development.[7] Whereas the appraisal of growth depends on continuous anthropometric quantities, evaluating development during infancy, childhood and adolescence is more difficult because of the lack of

[1]Netherlands Organisation for Applied Scientific Research TNO, Leiden, The Netherlands
[2]Department of Methodology and Statistics, FSS, University of Utrecht, The Netherlands

Corresponding author:
Stef van Buuren, TNO, PO Box 2215, 2301 CE Leiden, The Netherlands.
Email: stef.vanbuuren@tno.nl

precise objective measures. Development is typically classified in phases and stages, e.g. phases of cognitive development,[8] milestones (e.g. as in the Bayley Scales of Infant Development[9]), stages of pubertal development,[10,11] or stages of moral development.[12]

The quantitative methodology for stage measurements is less well developed than for quantitative measures. In practice, one often relates the child's score to a distribution from an age-specific norm group. This can be done for separate indicators, as well as for aggregates. The comparison provides an idea of the position of the child relative to the norm group at that age. Alternatively, one could compare the current age of the child to the age at which 10% or 90% of the norm population achieves the milestone. If the child's age is below the 10th centile, it can be classified as early, whereas children whose age is beyond the 90th centile are classified at late.

These approaches, however, do not yield quantitative measures of development that can be compared across time. Unlike continuous anthropometry, it is not possible to calculate a meaningful difference between two developmental scores obtained at different ages. Major shortcomings of current methods are:

- Outcomes are relative to a specific population, the norm group,
- There is no common metric to compare outcomes. Difference scores are not meaningful because there is no underlying quantitative scale,
- The exact meaning of the same score may differ across age; it is not possible to quantify a child's progress in time in terms of a gain in developmental units.

There is no equivalent to 'height gained' or 'weight gained' for developmental measures, which complicates tracking individual development over time. This article presents and discusses two novel growth charts that enable individual tracking of development both within and across time.

2 Development as a continuous latent variable

2.1 Latent variable theory

Borrowing from the social sciences, we will distinguish between *manifest* and *latent* variables. A manifest variable is a variable that can be measured directly. An example of a manifest variable is an indicator to code whether or not the patient has a particular symptom. Another example is observed blood pressure in mmHg. Latent variables, by contrast, cannot be observed directly. The values on the latent variable are inferred through a mathematical model from manifest variables. Examples of latent variables include the true (but unknown) disease status, and the true (but unknown) blood pressure. Both manifest and latent variables can be of continuous or categorical nature.

The new growth charts both assume the existence of a continuous latent variable on which the 'true' developmental score of a person can be placed. The person's location on the latent variable changes over time as development progresses. The person's location is inferred from the person's measurements on the manifest variables. These measured variables can be observed scores on developmental indicators, milestones, stages, and so on. The mathematical model allows us to estimate the true but unobservable status on the latent continuum from the data. We use these estimates to chart changes in developmental status over time. Thus, this study deals with the situation where the measurements are categorical and where the latent variables are continuous, or 'dimension-like'.[13] The methodology for connecting discrete measurements to a latent continuum is known as latent trait analysis, or item response theory.

2.2 Related work

One of the basic measurement assumptions of all latent variable models is *longitudinal measurement equivalence*; that is, the same unidimensional attribute is measured on the same persons with the same scale of measurement at every occasion.[14] In practice, equivalence is hard to achieve. McArdle et al. distinguished five general strategies to solve the problem of obtaining scale equivalence:

(1) *Absolute scaling*. Scales are constructed such that growth is linear in both the mean and the standard deviations.[15] Such linear assumption are often unrealistic, so this approach is almost never used.
(2) *Over time prediction*. This approaches predicts later scores from earlier scores.[16] Such prediction models, however, do not attempt to construct a common scale, or do not directly estimate change over time at the individual level.
(3) *Within-occasion rescoring*. All scores are standardized relative to an age-dependent norm.[17] This enables analyses of the relative positions, but does does not allow the estimation of change over time, as there is no common unit that is invariant of time.
(4) *SEM with convergent factor patterns*. For multiple scales, the factor pattern is assumed to be invariant over time formated as a structural equations model (SEM). This allows us to estimate changes in terms of a latent growth model of the second order.[18] This approach requires identification of the parameters, which may become problematic in real-life applications.
(5) *IRT linkage of common items*. The item response theory (IRT) approach postulates a single factor model for different measures across age. Unless there is enough overlap over time, it will not be possible to test the assumed measurement invariances. The IRT model is often estimated simultaneously with the longitudinal model.[19]

The next sections describe two novel methods for measuring development on a common scale with a unit that is invariant of time. The *D*-score method for infant development described in section 3 falls into category 5. The method for pubertal method can be categorized into category 3.

3 Tracking development by the *D*-score

3.1 SMOCC Data

The Social Medical Survey of Children attending Child Health Clinics (SMOCC) project[20] collected data on 2151 infants born between April 1988 and October 1989. Data were obtained at nine occasions between birth and 30 months of age. A total of 57 developmental indicators was sampled. At each occasion, a doctor or a trained nurse assessed whether a child could perform a set of developmental behaviors and tasks, and assigned a pass/fail score to each child for each indicator. The difficulty of indicators is matched to the infant's age so that approximately 90% of the children will achieve a pass. A fail score is a signal of a potential delayed development, and a reason for the youth health care physician to consider further investigation of the child. The set of 57 indicator is known as *Van Wiechenschema*, and forms an integral part of routine care in preventive Child Health Care Centers in The Netherlands. For details on the interpretation of each indicator, see the handbook by Laurent de Angulo.[21]

Each record in the data corresponds to a visit. Records without valid ages or with missing scores on all developmental indicators were eliminated. The total number of available records was 16,538, pertaining to 2038 infants. The total number of measurements made on the 57 indicators was equal to 164,885. Table 1 is a breakdown of the number of indicators actually measured per occasion. The table shows that per child usually 7 or 11–13 measurements were taken during the visit. A unique

Table 1. Number of actual indicators measured per child per visit (IPV) on 57 developmental milestones. SMOCC data

IPV	Frequency	IPV	Frequency
1	33	8	388
2	910	9	401
3	101	10	636
4	86	11	2305
5	613	12	2818
6	682	13	3970
7	2708	14	887

aspect of the design was that the more difficult set of indicator belonging to the next occasion was also sampled. This allows linking of the indicators across time. The variation in the number of measurements resulted from the design of the study, and was not related to the actual outcomes.

It is common practice to evaluate each indicator separately. Since 10% of the infants will fail on a given indicator, and since there are on average six indicators per occasion, the number of infants with at least one fail present could become quite large. A better and more reliable estimate developmental of delay can be attained by a composite score that combines the scores on multiple indicators.

3.2 Model

The Rasch model[22] assumes the existence of one continuous latent variable θ. In the present application we interpret θ as global development. Suppose that the sample contains n children that are observed at ages $t > 0$ (in days). The position of a child on the latent variable is denoted by θ_t, a number indicating the child's maturation at age t. If all is well, θ_t increases with t as the child matures with age.

Let there be $j = 1, \ldots, m$ indicators. The jth indicator is characterized by a number β_j on the same latent variable. The parameter β_j is the difficulty of the indicator, with higher values of β_j being associated with more difficult indicators. We assume that the difficulty β_j is fixed and does not depend on age.

The Rasch model stipulates that the probability of passing an indicator depends on only two parameters: the child's development status (θ_t) and the difficulty of the indicator (β_j). More precisely, suppose that Y_{jt} is the score a child on indicator j at age t, where $Y_{jt} = 0$ if the child fails and where $Y_{jt} = 1$ if the child passes. The Rasch model describes the probability that the child passes item j at age t as

$$P(Y_{jt} = 1|\theta_t, \beta_j) = \frac{\exp(\theta_t - \beta_j)}{1 + \exp(\theta_t - \beta_j)} \quad (1)$$

which corresponds to the logistic model based on the difference between θ_t and β_j. If developmental status equals the difficulty of the indicator, i.e. if $\theta_t = \beta_j$, then $P(Y_{jt} = 1|\theta_t, \beta_j) = \exp(0)/(1 + \exp(0)) = 1/(1 + 1) = 0.5$. If $\theta_t > \beta_j$ then the probability of passing exceeds 0.5.

For $m = 2$ there are four possible response vectors of (Y_{1t}, Y_{2t}): (0, 0), (0, 1), (1, 0) and (1, 1). The Rasch model expresses the probability of observing each of these vectors simply as the product of the separate indicator probabilities, i.e.

$$P(Y_{1t} \cap Y_{2t}|\theta_t, \beta_1, \beta_2) = P(Y_{1t}|\theta_t, \beta_1)P(Y_{2t}|\theta_t, \beta_2). \quad (2)$$

The generalization to $m > 2$ will be obvious. Likewise, suppose that the child responds to indicator j at ages t_1 and t_2. The probability of the four response vectors is then equal to

$$P(Y_{jt_1} \cap Y_{jt_2}|\theta_{t_1}, \theta_{t_2}, \beta_j) = P(Y_{jt_1}|\theta_{t_1}, \beta_j)P(Y_{jt_2}|\theta_{t_2}, \beta_j), \quad (3)$$

which illustrates that the Rasch model is able to predict response vector probabilities for any combination of θ_t and β_j.

An important and unique property of the Rasch model is the principle of invariant comparison. Rasch summarized the principle of invariant comparison as follows (p. 332)[23]:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion.

If the Rasch model holds, differences between β_j and $\beta_{j'}$ with $j \neq j'$ are identical no matter what individuals we have in the sample. Vice versa, differences between θ_t and $\theta_{t'}$ for $t \neq t'$ are identical no matter which indicators we use for comparison. The principle of invariant comparison is an extremely powerful concept that enables generalization of comparisons across different measures and samples. Under the Rasch model, the intervals between indicators and individuals remain invariant under addition and multiplication, so θ is effectively an interval scale.

3.3 Estimation and model fit

Both θ_t and β_j parameters are unknown and must be estimated from the data. We use the RUMM2020 software to estimate both set of parameters from the SMOCC data. Estimates $\hat{\theta}_t$ and $\hat{\beta}_j$ are calculated such that the total probability of obtaining the observed data under the Rasch model is maximal. RUMM2020 implements the pairwise conditional method using principal components.[24]

Observe that the data consist of multiple observations for each subject, so the data rows are not independent. The model does not attempt to model the longitudinal character of the data, and the fitting process treats all observations as independent. Jacobusse et al.[25] observed that this causes the standard errors to be too small by a factor of three relative to a sample size of 2151. In addition, the rows dependency influences the reference distribution of the RUMM residuals (called OUTFIT ZSTD in the Winsteps software for Rasch analysis), resulting in estimated RUMM residuals that are too extreme. For more details on the data fitting process, see Jacobusse et al.[25]

3.4 External scale anchoring

The mean and the variance of the scale are arbitrary and must be set by user. In our previous work, we standardized the scale to a sample mean of 50 and a sample standard deviation of 10.[25,26]

This choice has been criticized as 'meaningless' because internal anchors are sample dependent and cannot be reproduced by others by objective means.[27] Since we find shifted θ-estimates from a sample with different ability,[28] this criticism is justified. In this article, we will therefore use an alternative approach based on external scale anchors. There are many instances in science where external anchoring has proven useful. In 1742, Celcius fixed his temperature scale by setting two anchors at easily determined and objective values, the freezing and boiling points of water at an air pressure of 1 bar. The two anchors fixed his temperature unit and made it easily reproducible.

We have chosen to anchor the scale relative to two indicators. The lower anchor is 'Lifts head to 45° in prone position', which we equate to D=20. Thus, at D=20 the probability of passing this item is 50%. The upper anchor is 'Sits in stable position, without support', which we equate to D=40. Specifying these two anchors fixes the location and the unit of the D-score scale. There are several good arguments for these settings. The milestones associated with the lower and upper anchors fitted the model well, they are widely used in different instruments, and they are easy to measure. The lower and upper anchors are located at approximately one third and two third of the scale for 0–2 years infants. The two indicators have very different difficulties so the estimates of the measurement unit can well be made. Around the age of 1 month, D-scores will start approximately at zero (though negative values may occur). In the present set of 57 indicators, every unit increase corresponds to approximately passing one additional indicator. Since the scale has no end points, it is straightforward to extend the scale to lower and upper D-scores that will appear outside the age range 0–2 years.

Table 2 contains the estimated difficulty $\hat{\beta}_j$ of the 57 indicators in the D-score scale. For binary indicators, the difficulty level $\hat{\beta}_j$ can be interpreted as the point on the scale where the probability of passing indicator j is exactly 50%.

To wrap things up, there are now three different scales: the scale θ_{RUMM} constructed by RUMM to calculate the estimates, the scale D_{internal} with internals anchors of mean 50 and standard deviation 10,[25,26] and the scale D_{external} with two external anchors as defined above. These scales can be transformed into each other in the following way:

$$
\begin{aligned}
D_{\text{internal}} &= 49.273 + 1.1981\ \theta_{\text{RUMM}} \\
D_{\text{external}} &= 38.906 + 2.1044\ \theta_{\text{RUMM}} \\
D_{\text{external}} &= -47.63945 + 1.756448\ D_{\text{internal}}
\end{aligned}
$$

The same transformation applies to item difficulties and ability estimates. The advice is to work in the scale with external anchors, D_{external}.

3.5 *D*-score estimation

Estimates of person ability, $\hat{\theta}_t$, are called developmental scores, or D-scores. In general, the more milestones the infant passes, the higher his or her D-score. Suppose that we measure the entire set of 57 indicators for a given infant at day t. In that case, we can simply calculate the proportion p of indicators that the child passes, transform the result onto the logit scale by $\text{logit}(p) = \log(p) - \log(1-p)$, and apply a linear transformation to logit(p) to obtain $\hat{\theta}_t = D_t$ (p. 133)[29]. In practice, however, we often have scores on a limited subset of indicators. For identical set of indicators, we can use the above approach with linear transformations depending on the average difficulty of the set. There should be at least 5 dichotomous items of appropriate difficulty for this to work well. See Jacobusse and Van Buuren's Table II for an example.[26]

Table 2. Estimated item difficulties ($\hat{\beta}_j$) of 57 indicators sorted according to difficulty

Variable	Item	Label	Difficulty
v1432	52	Moves arms equally well	−2.2
v1434	53	Moves legs equally well	−1.9
v1431	29	Reacts when spoken to	1.7
v1436	56	Lifts chin off table for a moment	5.2
v1430	1	Eyes fixate	5.4
v1437	30	Smiles in response	11.3
v1438	2	Follows with eyes and head $30^\circ < 0 > 30^\circ$	14.5
v1443	31	Vocalizes in response	14.5
v1444	54	Stays suspended when lifted under the armpits	15.8
v1440	3	Hands occasionally open	16.5
v1445	57	Lifts head to 45° in prone position (*anchor*)	20.0
v1442	4	Watches own hands	20.7
v1452	59	Flexes or stomps legs while being swung	25.7
v1449	55	No head lag if pulled to sitting	26.0
v1454	58	Looks around to side with angle face-table 90°	27.8
v1446	5	Plays with hands in midline	28.2
v1447	6	Supine position: grasps object within reach	29.9
v1450		Turns head to sound	31.1
v1460	61	Balances head well while sitting	32.5
v1457	9	Plays with both feet	33.2
v1459	60	Rolls over, back and forth	34.7
v1461	62	Sits on buttocks while legs stretched	34.9
v1455	7	Passes cube from hand to hand	36.0
v1462	33	Says 'dada', 'baba' or 'gaga'	36.0
v1456	8	Holds cube, grasps another one with other hand	36.5
v1463	63	Sits in stable position, without support (*anchor*)	40.0
v1469	34	Babbles while playing	40.9
v1464	10	Picks up pellet between thumb and index finger	43.1
v1466	64	Crawls forward, abdomen on the floor	43.1
v1468	36	Waves 'bye-bye'	43.1
v1467	65	Pulls up to standing position	44.3
v1475	35	Reacts to verbal request	45.7
v1470	11	Puts cube in cup on command	46.0
v1473	66	Crawls, abdomen off the floor	46.1
v1474	67	Walks along	46.1
v1472	12	Plays 'give and take'	46.5
v1514	14	Explores environment energetically	46.9
v1476	37	Says 2 'sound-words' with comprehension	50.1
v1517	68	Walks alone, few steps	51.9
v1522	16	Imitates everyday activities	52.3
v1515	39	Says three 'words'	53.2
v1526	70	Picks up object from floor without falling	55.3
v1516		Identifies two named objects	55.4
v1527		Walks well	55.5
v1518	69	Throws ball without falling	56.0
v1512	13	Builds tower of two cubes	56.4

(continued)

Table 2. Continued

Variable	Item	Label	Difficulty
v1525	40	Understands 'play' orders	57.8
v1523		Drinks from cup	58.5
v1531		Eats with spoon without help	58.5
v1520	15	Builds tower of three cubes	59.2
v1524	41	Says 'sentences' of 2 words	60.2
v1529	18	Places round block	60.3
v1530	19	Takes off shoes and socks	60.6
v1532	43	Refers to self using 'me' or 'I'	61.7
v1533	44	Points at 5 pictures in the book	62.2
v1528	17	Builds tower of six cubes	62.6
v1534	71	Kicks ball away	64.2

Note: Items 1–28: Fine motor behavior and personal/social behavior; Items 29–51: Communication; Items 52–75 Gross motor activity. Unnumbered items are not part of the 2005 classification of the items of the Van Wiechenschema.[21]

In many realistic settings the composition of the indicator sets differs between children. For example, if a child fails a particular indicator, this indicator is added to the set to be measured on the next occasion. Also, missing data may occur if the child ceases to cooperate. Since we cannot compare sum scores based on different numbers of indicators, we need to resort to methods that explicitly take the difficulty per indicator into account.

One popular method is conditional maximum likelihood (CML). The conditional likelihood for a given data set can be maximized over θ. The CML estimator is unbiased, efficient and has normally distributed errors. However, CML may not work very well for a small number of items.[29] The CML estimator is incapable of providing estimates for perfect response profiles, and may give rise to local minima for short scales.

Alternatives to the CML estimator include the maximum a posteriori (MAP) estimator and expected a posteriori (EAP) estimator.[29,30] The estimators can be used even for $k=1$. Both generally yield similar results, but the EAP is noniterative and faster. The EAP estimator works as follows. Suppose that $P(\theta_{j-1})$ denotes the probability density of the proficiency of the infant at age t after seeing $j-1$ scores. Let $y_j=1$ and $y_j=0$ denotes the infant's pass and the fail scores. Then the posterior density $P(\theta_j|Y_j=y_j)$ after seeing j items can be calculated by Bayes theorem as

$$P(\theta_j \,|\, Y_j = y_j) = \frac{P(Y_j = 1|\theta_{j-1})P(\theta_{j-1})}{P(Y_j = 0|\theta_{j-1})P(\theta_{j-1}) + P(Y_j = 1|\theta_{j-1})P(\theta_{j-1})} \tag{4}$$

This equation works for one indicator at a time. We apply it successively to all $j=1,\ldots,k$ indicators by setting the prior for $P(\theta_{j-1})$ equal to the posterior $P(\theta_j|Y_j)$, and rerun the formula for $j=j+1$. The sequence in which the indicators are entered is irrelevant to the end result. The EAP estimator $\hat{\theta}_j = \mathrm{E}[\mathrm{P}(\theta_\mathrm{j}\,|\mathrm{Y}_1 = \mathrm{y}_1, \ldots, \mathrm{Y}_\mathrm{k} = \mathrm{y}_\mathrm{k})]$ is the mean of the posterior distribution after processing k items. For charting application, we are generally not interested in intermediate $\hat{\theta}_j$ for $j<k$, but they are available. The integers between -10 and 80 are taken as the quadrature points for θ. The set of points easily cover the range of D implied by the anchors. A total of 91 quadrature points is on the safe side.[31] The procedure is repeated for all children and ages.

One thing remains to be specified, the starting prior $P(\theta_0)$. This choice is critical if the number of indicators k is low, e.g. one or two. Usual uninformative global priors fail because these pull the

D-scores too much towards to global mean for low k. After careful experimentation, we decided to use the age-dependent normal prior $N(\mu_t, 5)$, where μ_t is the mean of the *D*-score distribution at age t, and where the standard deviation of 5 is almost twice the within-day variation of the *D*-score. Note that the use of an age-specific prior implies that identical scores at distinct ages are mapped into (slightly) different *D*-scores. The discrepancies rapidly vanish as k grows.

3.6 Number of indicators per *D*-score

One potential worry in *D*-score estimation is that the *D*-score could be sensitive to the size of the indicator set. In order to study this issue, we divided the records into 14 groups according to set size (c.f. Table 1), and calculated the mean *D*-score per group after correcting for age and sex. This was done for two outcomes: the Maximum Likelihood estimator and the EAP estimate with the age-dependent prior.

Table 3 contains the parameter estimates of the two regressions. The proportion of explained variance of both regression models is very high (0.97). The residual deviation $\hat{\sigma}$ around the regression line is about 3 *D*-score units, which is small relative to the range of the *D*-score. Age was entered as a third-order polynomial. The regression weights for different set size are expressed as differences with respect to the last category (14 observations per occasion). Girls develop slightly

Table 3. Regression weights for predicting *D*-scores estimated by maximum likelihood (ML) and expected a posterior (EAP) methods. The labels '1-14' correspond to dummy variables of the number of indicators used to estimate the *D*-score. SMOCC data ($n = 2038$)

	ML		EAP	
Term	Estimate	Std. Error	Estimate	Std. Error
Intercept	39.42	0.15	38.96	0.16
Age(1)	2082.97	4.73	2094.22	4.85
Age(2)	−566.80	4.43	−561.71	4.57
Age(3)	116.94	3.67	141.54	3.78
Sex	0.39	0.05	0.38	0.05
1			−0.27	0.56
2	−6.49	0.27	0.03	0.19
3	−4.89	0.70	1.10	0.35
4	−2.29	0.35	0.34	0.36
5	−4.12	0.19	−0.46	0.20
6	−2.43	0.18	−0.58	0.18
7	−2.43	0.18	−1.41	0.17
8	−0.55	0.20	0.81	0.21
9	−0.23	0.20	0.67	0.21
10	−0.09	0.18	0.64	0.18
11	0.09	0.15	0.66	0.15
12	−0.51	0.15	−0.09	0.15
13	0.06	0.14	0.31	0.15
14 (ref)	0	0	0	0
r^2	0.97		0.97	

faster than boys. The difference is about 0.4 *D*-score units. Observe that the regression weights for ML are relatively large in magnitude. For example, if there are only two indicators observed at a specific time point, then the ML estimate is about 6.5 *D*-score points lower compared to a set size of 14. Since this difference is large (about 2 standard deviations in the *D*-score scale), the ML estimate appears biased downward from set sizes smaller than seven indicators. Also observe that ML cannot deal with set size of one. In contrast, the EAP estimator with the age-dependent prior is insensitive to set size. This means that we can use the EAP to validly calculate $\hat{\theta}$ from any number of indicators. Of course, the results will be more accurate (i.e. less biased and more precise) if we use more indicators.

3.7 Reference values

We calculated age-conditional references of the *D*-score for all boys and girls combined by the LMS method.[1] The LMS method assumes that the outcome has a normal distribution after a Box–Cox transformation. The reference distribution has three parameters, which model respectively the location (μ), the spread (σ), and the skewness (λ) of the distribution. Each of the three parameters can vary smoothly with age. Let μ_t, σ_t and λ_t be the parameter values at age t. The transformation

$$Z = \frac{(D_t/\mu_t)^{\lambda_t} - 1}{\lambda_t \sigma_t} \tag{5}$$

converts measurement D_t into its normal equivalent deviate Z. If λ_t is close to zero, we use

$$Z = \frac{\ln(D_t/\mu_t)}{\sigma_t} \tag{6}$$

The parameters are estimated by GAMLSS[6] using cubic splines smoothers. We used the worm plot[5] and *Q*-statistics[4] to determine the optimal degrees of freedom of the smoothers. The final solution used a log-transformed age scale and fitted the model with $\mathrm{df}(\mu) = 2$, $\mathrm{df}(\sigma) = 2$, $\mathrm{df}(\lambda) = 1$. Table 4 gives the LMS estimates that define normal reference values of the *D*-score.

Any required centile curve can be derived from Table 4. First, choose Z_α as the *Z*-score below which 100α percent of the distribution is located, for example, $Z_{0.05} = -1.64$. The *D*-score that defines the 100α centile is equal to[32]

$$D_t(\alpha) = \mu_t(1 + \lambda_t \sigma_t Z_\alpha)^{1/\lambda_t} \tag{7}$$

If λ_t is close to zero, we can use

$$D_t(\alpha) = \mu_t \exp(\sigma_t Z_\alpha) \tag{8}$$

Figure 1 is the reference diagram of the *D*-score of Dutch infants. The gray area between the −2SD and +2SD lines delineates the *D*-score expected if development is normal. Note that the shape of the reference is quite similar to that of weight and height, with rapid growth occurring in the first few months. The Pearson correlation between age and *D*-score (0.944) is comparable to the that of age–height (0.955) and age–weight (0.916). Thus, the *D*-score is extremely sensitive to detect and monitor age-related changes. Figure 1 illustrates this by showing the *D*-score trajectory of two infants, one with normal development and one with severely delayed development. Any measures and referral

Table 4. Reference distribution for *D*-scores, boys and girls combined. The table lists age-dependent values of the median μ, the coefficient of variation σ, and the skewness parameter λ of LMS reference distribution of *D*-scores

Week	μ	σ	λ	Week	μ	σ	λ
2	8.81	0.3126	1.3917	48	47.16	0.0647	1.4778
3	10.59	0.2801	1.4418	52	48.84	0.0627	1.4676
4	12.27	0.2526	1.4891	56	50.41	0.0608	1.4605
5	13.87	0.2291	1.5331	60	51.89	0.0591	1.4561
6	15.39	0.2089	1.5722	64	53.27	0.0574	1.4538
7	16.83	0.1916	1.6049	68	54.58	0.0559	1.4533
8	18.20	0.1767	1.6304	72	55.81	0.0544	1.4539
9	19.50	0.1640	1.6487	76	56.97	0.0530	1.4555
10	20.75	0.1531	1.6607	80	58.06	0.0517	1.4580
12	23.07	0.1354	1.6706	84	59.11	0.0505	1.4612
14	25.21	0.1220	1.6698	88	60.11	0.0494	1.4649
16	27.17	0.1117	1.6636	92	61.06	0.0483	1.4692
18	28.99	0.1035	1.6533	96	61.97	0.0474	1.4740
20	30.70	0.0970	1.6403	100	62.85	0.0465	1.4791
22	32.29	0.0917	1.6255	104	63.70	0.0457	1.4846
24	33.79	0.0873	1.6100	108	64.52	0.0449	1.4904
26	35.21	0.0837	1.5946	112	65.31	0.0441	1.4964
28	36.55	0.0807	1.5797	116	66.08	0.0434	1.5024
32	39.04	0.0759	1.5523	120	66.82	0.0428	1.5084
36	41.32	0.0723	1.5284	124	67.54	0.0421	1.5142
40	43.42	0.0693	1.5081	128	68.24	0.0415	1.5199
44	45.36	0.0669	1.4913	132	68.92	0.0410	1.5254

rules that have been developed for continuous measures[33] can be applied to the *D*-score. This opens up new possibilities for objectively identifying developmental delay.

This is the first time that age-conditional references have been created for development. There are some open issues. Inclusion of the standard error would theoretically be appropriate. On the other hand, it would complicate model fitting, and make application of the model in practice more difficult. The added value of incorporating standard errors still needs to be studied. Note that the references are purely cross-sectional. The correlation structure over time is not taken into account. For prediction purposes, it is useful to extend the modeling to include velocities and change scores, but this has not yet been done. Use of the Box–Cox transformation is the de facto standard in the construction of growth references for anthropometric data. It could be that other, perhaps more flexible distributions are needed to account for the typical features (e.g. severe skewness, ceiling effect) found in developmental data.

3.8 Related work

The *D*-score is an attempt to solve the problem of changing scales of measurement. The term '*D*-score' has been used previously for a similar purpose by Bayley.[17]

The *D*-score as proposed here uses a two-stage estimation procedure. The Rasch model is fitted first, followed by a calculation of change scores. Many applications in the social sciences

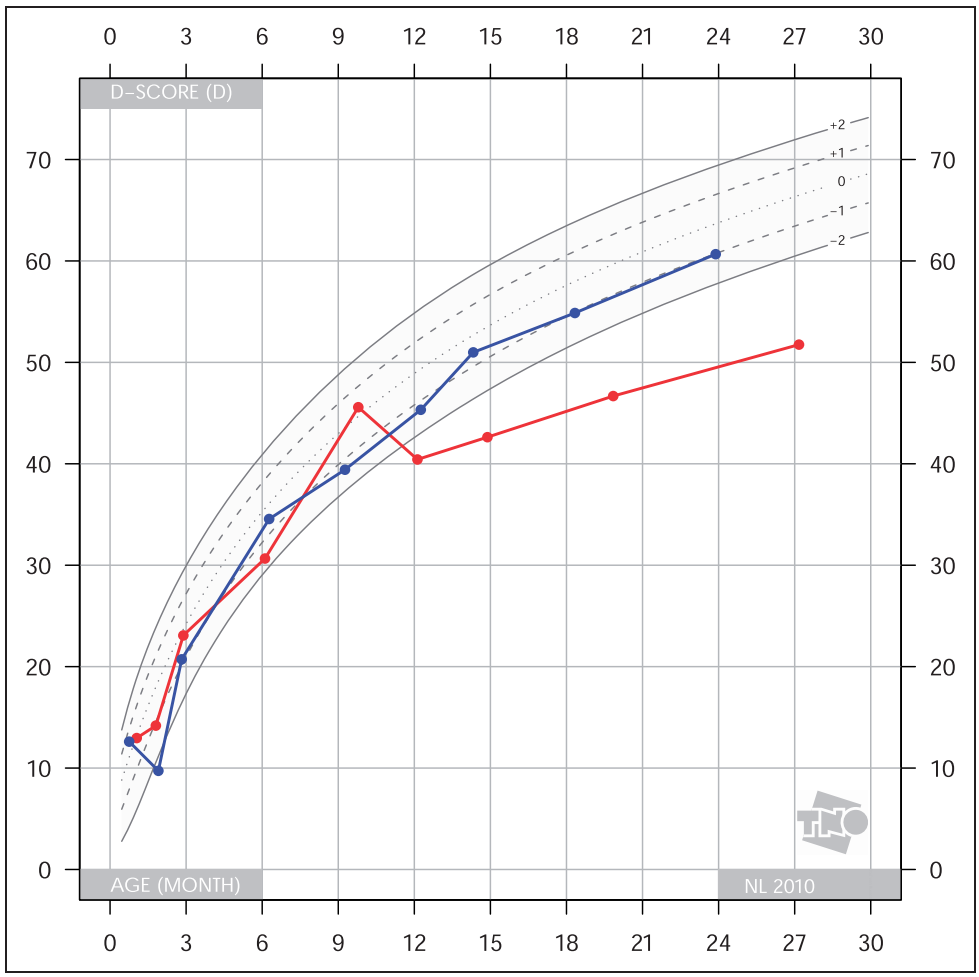

Figure 1. *D*-score reference chart, 0–30 months, with SD curves −2SD, −1SD, 0SD (median), +1SD and +2SD. Two child trajectories are superposed. The infant with the blue curve has a normal development around −1SD. Maturation of the infant with the red curve is severely delayed from the age of 12 months onwards.

rely on simultaneous estimation. McArdle et al.[14] considered two-stage estimation 'less optimal owing to the longitudinal dependencies within person' (p. 142). They acknowledge, however, that two-stage estimation substantially cuts down on the computational complexity and simplifies modeling.

To this, let us add another argument in favor or two-stage estimation. The two stages address conceptually very distinct problems. The first problem is to construct the measurement scale, which is done by fitting the Rasch model. The second problem is to estimate change over time on that measurement scale, which is done by calculating differences or by performing ANOVA. Solving both estimation problems simultaneously is undesirable since it confounds both problems. In the simultaneous procedure, the scale itself depends on the observed changes in the calibration sample. Unless the crossover between both stages is small, the scale produced by simultaneously estimation cannot be used across different samples. Moreover, the amount of change estimated from the

calibration sample is larger than the amount of change under two-stage estimation. The reason is that the simultaneous fitting process will emphasize features in the measurements that will bring out the estimated change most clearly. In sequential estimation, the amount of change plays no role in the way in which the scale is constructed, and hence the resulting change estimate will be equal or smaller. So both for simplicity and conceptually, we advocate two-stage estimation over simultaneous estimation.

The two-stage analysis is simple to do and yields generalizable results through the properties of the Rasch model. The model is now being accepted for modeling developmental data. Applications in human development have been published by Dawson,[34] Jacobusse et al.,[25,26] Draney,[35] and Boom et al.[36,37] Presenting human development as a growth diagram is natural and may enhance the understanding of the developmental phenomena under study.

The Rasch model is a very strict model and may not fit the data. Extensions such as the 2PL model and 3PL models[29] could be used instead, at the expense of losing the attractive invariance properties of the Rasch model. For applications in measuring cognitive development, it interesting alternative is the so-called Saltus model.[38] The Saltus model is an extension of the Rasch model that allows modeling leaps in development. The model can capture discontinuities as predicted by the theories of cognitive development of Gagné, van Hiele and Siegler by assuming that indicator difficulties vary between groups.

4 Tracking stages of development

4.1 Pubertal stages

Puberty is an important phase of life that connects childhood to adolescence. The timing and speed of pubertal maturation varies between individuals. A widely accepted measurement of pubertal maturation are the so-called Tanner stages.[10,11] This system classifies secondary sexual characteristics into a number of distinct stages. For boys, there are three types of measures: genital development (5 stages G1–G5), pubic hair (5 stages PH1–PH5), and testis size (12 stages T1–T25 corresponding to volumes of 1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25 mL). For girls, the Tanner system measures includes: breast development (5 stages B1–B5), pubic hair (6 stages PH1–PH5) and menarche (2 stages no/yes). Pubertal stages were determined by visual inspection, using Tanner's criteria according to the high-resolution photographs.[39] In boys testicular volume was assessed using the Prader orchidometer.

We use cross-sectional data collected within the Fourth Dutch Growth Study.[40] The original data were collected using an additional category for pubic hair (PH6). In the sequel, we combine stage PH6 with PH5 in order to conform to the original scoring system. The study contained data on 5436 children (2377 boys, 3059 girls) aged between 7 and 22 years that had one or more Tanner stages observed. This is a subset of a larger sample of approximately 50%, with more nonresponse occurring in the higher age groups. In this article we present the estimates always conditional on age, so this skewed age distribution will not affect the results.[41,42] The composition of the puberty sample was comparable with the sample of a national survey with regard to region and level of education.

4.2 Age continuity

The usual estimates of interest for pubertal development include

(1) the age at which some part of the population (e.g. 50%) reaches a stage,[39,40,43]
(2) the mean age at which a stage is reached,[10]

(3) the mean age of all children that are in a particular stage,[43]
(4) the mean duration of being in a particular stage.[10]

Note that these statistics all have age as their outcome and are conditional on stage. Estimation and interpretation of references with age as outcome is generally difficult as it requires us to take any censoring and selective drop out processes into account. In practice, such problems can easily occur since nonresponse can strongly depend on age.[41,42] Fortunately, we can estimate statistics of type 1 and 2 by solving the reverse problem. It is straightforward to determine the probability of reaching a stage at a given age, for example by probit or logistic regression. Given the estimated probability curve, we can find the ages at which 10%, 50% and 90% of the population reaches the stage. It is also possible to calculate the mean age at which the stage is reached. Age-conditional references are widely accepted for continuous measures like height and weight. In the sequel, we will concentrate on deriving truly age-conditional references of pubertal development.

We assume that pubertal development is a continuous process, even though the observations are always discrete. This induces a form of continuity that we call age-continuity. Suppose that two children, one young and one old, are in the same developmental stage. We can say that, on average, the younger child matures earlier than the older child. To see why this is the case, consider the fact that the younger child still has the opportunity to move into the next stage before his/her age reaches that of the older child, whereas the older child does not have this opportunity anymore. Thus, depending on age, the same stage is associated with different degrees of maturation. Moreover, the difference in maturation grows with the age gap.

In the previous section we aggregated different indicators of infant development into an overall summary, the *D*-score. Can we do the same for the three pubertal measures? The answer is 'no' for two reasons. First, the three items fail to fit the polytomous Rasch model as developed by Andrich.[44] In particular, the model does not cope well with the different numbers of categories (5, 5 and 2 for girls; 5, 5 and 12 for boys). For example, in the model a change in menarche status weights as much as a change from B3 to B4. In practice a change in menarche is considered much more important. A solution for this would be to calculate the model separately for the observation with and without menarche (Andrich, personal communication, 2004). Second, each of the measures is of clinical interest in its own right. Marshall and Tanner (p. 301)[10] state: 'It is important to recognize that the relation between the different events of puberty is a more significant index of normality than the chronological age at which they occur.' It is known that some diseases have disparate effect on different events. For these reasons, we model each measure separately.

4.3 Reference values

Let Y be an ordered stochastic variable whose values $Y \in \{1, \ldots, m\}$ correspond to stages 1 to m, and let X represent decimal age. The probability of achieving stage c at age X is written as $P(Y \geq c|X)$ where $c = 2, \ldots, m$. Let $P(Y < c|X) = 1 - P(Y \geq c|X)$. For each transition from stage $c-1$ to c, we model a reference curve conditional on X by an generalized additive model.[45] Let $\Phi(Z)$ be the cumulative distribution function of the normal distribution $N(0, 1)$, and let its inverse $\Phi^{-1}(P)$ be the probit transformation. We assume that we can model the probability of being in stage c or higher as a function of age as

$$\text{probit}(P(Y \geq c|X)) = \alpha_c + f_c(X), \quad c = 2, \ldots, m \tag{9}$$

Table 5. Age-conditional reference curves for pubertal development in boys. The table lists the probability*10,000 of achieving a stage at a given age between 8 and 21 years

	Genitalia				Pubic hair				Testicular volume (mL)										
Age	G2	G3	G4	G5	PH2	PH3	PH4	PH5	T2	T3	T4	T5	T6	T8	T10	T12	T15	T20	T25
8.00	1152	112	0	0	324	0	0	0	7152	3103	719	322	276	183	175	8	0	0	0
8.25	1317	128	0	0	418	1	0	0	7467	3336	838	369	310	211	192	12	0	0	0
8.50	1494	147	0	0	532	2	0	0	7763	3565	971	422	348	243	212	17	0	0	0
8.75	1681	167	0	0	669	3	0	0	8039	3784	1117	480	389	278	232	24	0	0	0
9.00	1875	190	0	0	831	7	0	0	8293	3988	1277	545	434	318	255	34	0	0	0
9.25	2079	216	0	0	1021	13	0	0	8524	4176	1449	620	485	362	279	47	0	0	0
9.50	2306	246	0	0	1240	24	0	0	8734	4370	1641	713	545	412	305	65	1	0	0
9.75	2565	284	1	0	1488	42	0	0	8922	4599	1878	840	620	471	338	88	2	1	0
10.00	2853	333	2	0	1774	73	0	0	9089	4907	2190	1015	715	544	378	118	4	2	0
10.25	3163	399	4	0	2102	120	1	1	9236	5295	2577	1246	837	633	430	157	8	3	0
10.50	3487	489	8	0	2475	188	3	2	9364	5744	3011	1528	993	745	499	208	17	6	0
10.75	3824	612	17	0	2890	286	6	5	9475	6237	3480	1863	1200	890	590	277	33	10	0
11.00	4198	779	33	0	3350	425	15	9	9570	6770	4000	2272	1482	1085	714	370	63	18	1
11.25	4643	1009	64	1	3861	624	34	16	9650	7318	4601	2798	1869	1350	886	498	114	31	2
11.50	5200	1324	119	4	4438	911	74	28	9718	7856	5318	3490	2392	1709	1123	671	197	51	4
11.75	5887	1744	214	9	5090	1321	153	48	9774	8352	6122	4325	3057	2181	1442	903	322	82	8
12.00	6645	2286	371	21	5801	1887	305	84	9821	8769	6903	5197	3833	2766	1853	1206	501	128	14
12.25	7400	2947	616	45	6533	2618	574	142	9859	9092	7578	6025	4665	3456	2359	1588	745	196	25
12.50	8076	3714	972	93	7241	3485	997	237	9890	9335	8134	6788	5493	4225	2955	2053	1059	292	43
12.75	8624	4552	1451	179	7874	4417	1584	386	9915	9515	8568	7461	6273	5032	3620	2592	1448	426	71
13.00	9045	5411	2050	323	8406	5340	2314	605	9935	9646	8899	8036	6987	5838	4326	3191	1908	610	114
13.25	9355	6240	2755	540	8833	6197	3150	908	9950	9742	9160	8521	7627	6612	5043	3839	2430	855	175
13.50	9576	6997	3545	842	9163	6963	4052	1300	9962	9814	9381	8938	8187	7328	5752	4523	3000	1164	257
13.75	9727	7654	4386	1232	9410	7629	4972	1774	9972	9868	9571	9291	8660	7963	6437	5218	3597	1535	365
14.00	9826	8196	5232	1703	9588	8188	5871	2321	9979	9909	9724	9563	9036	8489	7078	5905	4211	1963	498
14.25	9890	8631	6041	2247	9712	8637	6715	2934	9985	9940	9835	9750	9318	8900	7663	6568	4836	2438	655
14.50	9929	8973	6774	2853	9796	8988	7475	3606	9989	9962	9908	9868	9525	9207	8184	7190	5467	2953	834
14.75	9954	9239	7415	3500	9853	9259	8127	4326	9992	9977	9952	9935	9673	9432	8629	7753	6099	3498	1038
15.00	9970	9443	7961	4170	9892	9466	8649	5076	9994	9986	9976	9970	9778	9593	8992	8234	6711	4052	1267
15.25	9980	9595	8416	4835	9920	9621	9038	5817	9996	9992	9989	9987	9850	9707	9271	8623	7272	4590	1519
15.50	9987	9707	8784	5461	9939	9734	9314	6505	9997	9996	9995	9995	9898	9788	9477	8923	7755	5085	1785
15.75	9992	9789	9074	6009	9954	9814	9508	7101	9998	9998	9998	9998	9930	9846	9626	9146	8147	5520	2053
16.00	9995	9848	9294	6461	9964	9871	9644	7590	9999	9999	9999	9999	9951	9889	9730	9309	8456	5887	2312
16.25	9997	9890	9458	6815	9972	9911	9737	7978	*	*	*	*	9965	9919	9802	9427	8693	6182	2551
16.50	9998	9919	9578	7082	9978	9938	9801	8285	*	*	*	*	9975	9941	9852	9511	8865	6409	2758
16.75	9999	9940	9664	7284	9983	9958	9846	8533	*	*	*	*	9981	9958	9888	9570	8983	6573	2932
17.00	9999	9955	9728	7446	9986	9971	9879	8742	*	*	*	*	9986	9969	9913	9611	9059	6689	3076
17.25	*	9965	9777	7585	9989	9981	9907	8922	*	*	*	*	9990	9978	9931	9641	9105	6766	3194
17.50	*	9973	9814	7721	9992	9988	9931	9085	*	*	*	*	9993	9985	9945	9663	9134	6820	3297
17.75	*	9978	9845	7872	9994	9992	9950	9230	*	*	*	*	9995	9989	9956	9682	9155	6863	3397
18.00	*	9982	9869	8039	9995	9995	9965	9354	*	*	*	*	9996	9993	9964	9699	9172	6910	3500
18.25	*	9985	9888	8216	9997	9997	9976	9458	*	*	*	*	9997	9995	9971	9715	9189	6964	3603
18.50	*	9988	9904	8393	9998	9998	9984	9545	*	*	*	*	9998	9997	9976	9733	9207	7024	3696
18.75	*	9990	9916	8566	9999	9999	9990	9617	*	*	*	*	9999	9998	9980	9752	9226	7095	3786
19.00	*	9992	9927	8729	9999	9999	9993	9677	*	*	*	*	9999	9999	9984	9772	9244	7179	3892
19.25	*	9993	9935	8881	*	*	9996	9726	*	*	*	*	9999	9999	9987	9792	9263	7275	4023

(continued)

Table 5. Continued

	Genitalia				Pubic hair				Testicular volume (mL)										
Age	G2	G3	G4	G5	PH2	PH3	PH4	PH5	T2	T3	T4	T5	T6	T8	T10	T12	T15	T20	T25
19.50	*	9994	9943	9019	*	*	9997	9766	*	*	*	*	*	9999	9990	9812	9283	7383	4177
19.75	*	9995	9950	9146	*	*	9998	9800	*	*	*	*	*	*	9992	9832	9304	7504	4351
20.00	*	9996	9956	9261	*	*	9999	9828	*	*	*	*	*	*	9993	9850	9326	7635	4541
20.25	*	9997	9961	9365	*	*	9999	9854	*	*	*	*	*	*	9995	9866	9348	7771	4727
20.50	*	9998	9966	9458	*	*	*	9876	*	*	*	*	*	*	9996	9881	9370	7905	4903
20.75	*	9998	9970	9539	*	*	*	9895	*	*	*	*	*	*	9997	9895	9391	8034	5075
21.00	*	9998	9974	9611	*	*	*	9911	*	*	*	*	*	*	9998	9907	9411	8160	5246

*denotes a value of 10,000.

where $f_c(X)$ is a smooth univariate function of age. The calculations were done by the `gam` package in `R`. We applied smoothing splines to find the shape $f_c(X)$, and used analysis of deviance by `anova.gam()` to find the optimal degrees of freedom of the smoothing spline. In most cases, this resulted in the default smoothing parameter as calculated by the `gam()` function.

Tables 5 and 6 contain the fitted reference curves for successive stage transitions of puberty. One can determine the ages at which 10%, 50% and 90% of the reference population achieve a stage by linear interpolation. For example, for the transition B1-B2 we find 8.98, 10.72 and 12.17 years, respectively. Note: Due to recent advances in fitting methodology and software, slight differences may occur with the previously tabulated values (here 9.01, 10.72 and 12.16 years).[40]

4.4 Maturation scores

Tables 5 and 6 contain a complete summary of the reference distribution, but they are unsuitable to track individual development over time. The problem is that the measurements are stages, but that the reference values apply to stage *transitions*. For example, if we observe stage B3 at age t, we only know that the transition from B2 to B3 must have occurred at or before age t. We do not know when the transition occurred. The only exception to this is perhaps menarche, a clear event for which we sometimes do have an exact calendar date. Specialized methods for handling this case have been developed.[46] For the other measures, we are always unsure about the timing of transition. Consequently, the reference values as presented in Tables 5 and 6 are of limited value to monitor development.

The stage line diagram[47] remedies this problem. The essential idea is to convert the $m-1$ transition probabilities into m maturity scores, one per stage. The probability of observing stage c at age X is equal to the distance between two curves, i.e.

$$P(Y = c+1|X) = P(Y < c+1) - P(Y < c|X), \quad c = 1, \ldots, m \tag{10}$$

where $P(Y<1|X) \equiv 0$ and $P(Y<m+1|X) \equiv 1$, so that $\sum_c^m P(Y=c|X) = 1$ for all X. The maturity score $\pi_c|X$ corresponding to stage c at age X is defined as

$$\pi_c|X = P(Y < c|X) + P(Y = c|X)/2 \tag{11}$$

$$= \frac{P(Y < c|X) + P(Y < c+1|X)}{2} \tag{12}$$

Table 6. Age-conditional reference curves for pubertal development in girls. The table lists the probability*10,000 of achieving a stage at a given age between 8 and 21 years

	Breast				Pubic hair				Menarche
Age	B2	B3	B4	B5	PH2	PH3	PH4	PH5	Yes
8.00	213	4	0	0	175	1	0	0	46
8.25	328	7	0	0	251	2	0	0	53
8.50	491	15	0	0	354	4	1	0	62
8.75	719	29	0	0	493	9	2	0	71
9.00	1027	54	1	0	682	21	4	0	83
9.25	1424	97	2	0	934	44	9	1	95
9.50	1886	164	5	1	1250	88	17	2	110
9.75	2392	265	12	2	1628	166	33	5	129
10.00	2947	416	27	4	2082	298	61	10	155
10.25	3575	641	59	9	2644	510	109	20	190
10.50	4299	969	120	19	3335	834	190	38	239
10.75	5115	1419	231	37	4129	1299	320	70	310
11.00	5972	1998	415	69	4955	1914	523	124	414
11.25	6789	2705	698	124	5757	2666	823	211	567
11.50	7525	3514	1096	211	6529	3531	1247	348	787
11.75	8169	4409	1616	343	7270	4471	1820	552	1100
12.00	8711	5383	2256	531	7967	5442	2544	839	1528
12.25	9133	6385	3005	786	8569	6386	3386	1221	2085
12.50	9434	7318	3832	1114	9038	7247	4295	1701	2766
12.75	9635	8118	4695	1514	9371	7987	5211	2269	3547
13.00	9763	8749	5555	1988	9592	8580	6080	2907	4390
13.25	9843	9199	6375	2525	9734	9024	6865	3585	5245
13.50	9893	9499	7120	3106	9823	9340	7550	4265	6068
13.75	9924	9689	7762	3701	9880	9557	8122	4910	6821
14.00	9944	9806	8284	4273	9917	9704	8583	5493	7480
14.25	9957	9878	8686	4791	9940	9800	8941	6001	8033
14.50	9967	9922	8985	5244	9956	9864	9214	6437	8484
14.75	9973	9949	9205	5640	9967	9904	9420	6811	8843
15.00	9978	9966	9364	5985	9975	9931	9571	7136	9125
15.25	9982	9977	9481	6285	9981	9948	9680	7422	9343
15.50	9985	9984	9569	6547	9985	9959	9759	7676	9508
15.75	9989	9989	9638	6782	9988	9967	9815	7902	9630
16.00	9993	9993	9693	6995	9991	9972	9856	8100	9721
16.25	9995	9995	9739	7185	9993	9976	9884	8273	9786
16.50	9997	9997	9777	7355	9995	9978	9904	8425	9834
16.75	9998	9998	9807	7510	9996	9980	9918	8555	9869
17.00	9999	9999	9832	7653	9997	9981	9928	8666	9894
17.25	9999	9999	9851	7786	9998	9982	9935	8759	9913
17.50	*	*	9867	7910	9999	9982	9940	8838	9926
17.75	*	*	9880	8033	9999	9982	9943	8908	9936
18.00	*	*	9891	8161	9999	9982	9946	8973	9944
18.25	*	*	9901	8298	9999	9982	9948	9034	9950

(continued)

Table 6. Continued

	Breast				Pubic hair				Menarche
Age	B2	B3	B4	B5	PH2	PH3	PH4	PH5	Yes
18.50	*	*	9910	8444	*	9982	9950	9096	9955
18.75	*	*	9919	8599	*	9982	9952	9159	9959
19.00	*	*	9928	8759	*	9983	9955	9226	9963
19.25	*	*	9936	8918	*	9983	9957	9295	9966
19.50	*	*	9944	9071	*	9984	9960	9365	9968
19.75	*	*	9951	9213	*	9984	9963	9435	9971
20.00	*	*	9958	9340	*	9985	9965	9501	9973
20.25	*	*	9963	9451	*	9985	9968	9562	9976
20.50	*	*	9968	9547	*	9986	9970	9617	9978
20.75	*	*	9973	9629	*	9986	9972	9667	9980
21.00	*	*	9977	9699	*	9987	9974	9711	9982

*denotes a value of 10,000.

$$= 1 - \frac{P(Y \geq c|X) + P(Y \geq c+1|X)}{2} \tag{13}$$

for $c = 1, \ldots, m$.

Figure 2 illustrates the calculation of $\pi_c|X$. The diagram plots the probability $P(Y < 2|X)$ and $P(Y < 3|X)$ taken from Table 6 against age as two lines. The distance between the lines is equal to the absolute probability of observation stage 2 at age X. The most obvious place for defining a maturity score for stage 2 is exactly half way this distance. If this is done for all ages and points are connected, then we obtain the Y-coordinate of the stage line $\pi_2|X$. The same calculation applies to other stages.

The statistic $\pi_c|X$ is known as the mid-P-value, and was proposed as a correction for continuity in statistical tests. Using the mid-P-value to quantify categorical data seems to be novel. To see that the mid-P-value is a reasonable value for stage c imagine that the observable stage $Y = c$ is a coarse version of a continuous latent variable $\tilde{Y}$ that has a uniform distribution $\tilde{Y} \sim \mathrm{U}[0, 1]$. The link between $\tilde{Y}$ and the observed data Y is

$$Y = c \quad \text{if } P(Y < c|X) \leq \tilde{Y} < P(Y < c+1|X) \quad c = 1, \ldots, m \tag{14}$$

The distribution $\tilde{Y} \sim \mathrm{U}[0, 1]$ will be reasonable if the model adequately fits the reference data, which can be evaluated through worm plots or Q-statistics.[4,5] The interval $[P(Y < c|X),\ P(Y < c+1|X)]$ is also uniformly distributed, so the mid-P-value $\pi_c|X$ is the best single summary measure.

As before in section 3.7, it is possible to calculate Z-scores for the observed data. Suppose we observe $Y = c$ at age $X = t$. First, linearly interpolate $P(Y \geq c|X = t)$ and $P(Y \geq c+1|X = t)$ from the surrounding tabulated ages t_1 and t_2 by $P(Y \geq c|X = t) = hP(Y \geq c|X = t_1) + (1-h)P(Y \geq c|X = t_2)$ where $h = (t_2 - t)/(t_2 - t_1)$. The Z-score is calculated as

$$Z = \Phi^{-1}\left(1 - \frac{P(Y \geq c|X) + P(Y \geq c+1|X)}{2}\right) \tag{15}$$

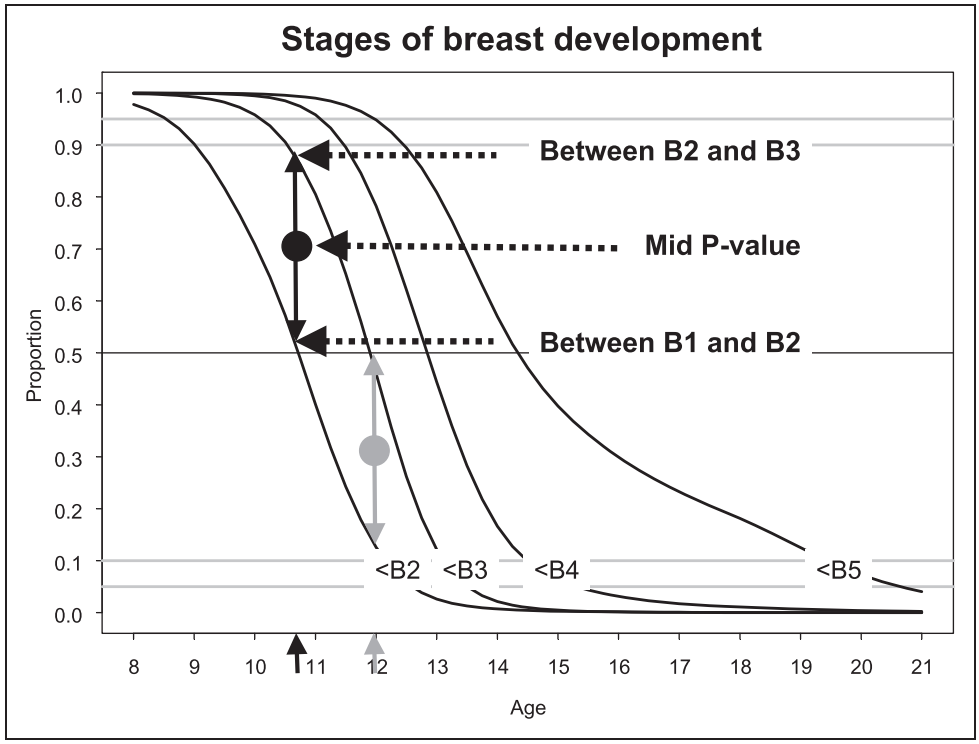

Figure 2. Calculation of stage lines positions from the reference curves by the mid-*P*-value. The mid-*P*-value is calculated for all ages and all stages.

The *Z*-score allows us to draw a diagram that relates the concept of early/late maturation to the proportion of children in the reference sample that achieves some stage. In the reference population, the mean of the *Z*-scores is equal to zero. The standard deviation is smaller than 1, due to rounding, floor and ceiling effects.[47]

4.5 Stage line diagram

Figure 3 contains the stage line diagrams of genital development as calculated from Table 5. The horizontal axis represents age between 8 and 21 years. The vertical axis indicates maturation status. The scale of this uses the probit scale, so the values can be interpreted as *Z*-scores. The *Z*-score is useful for tracking development because it provides high resolution at the extremes, the areas of most clinical significance. Lower values indicate delayed development, and higher values signal early maturation. The diagram contains five stage lines. Each stage line corresponds to one of a developmental stage.

The user places a mark on the stage line corresponding to the observed stage at the child's age, and connects the mark to the previous measurement. A move to the next stage produces a jump in the curve. The exact age at which the child reaches the next stage is unknown, and can be anywhere between the two ages surrounding the jump. Steeper jumps occur for measurements that are closer in time. Jumps can span two or more stages.

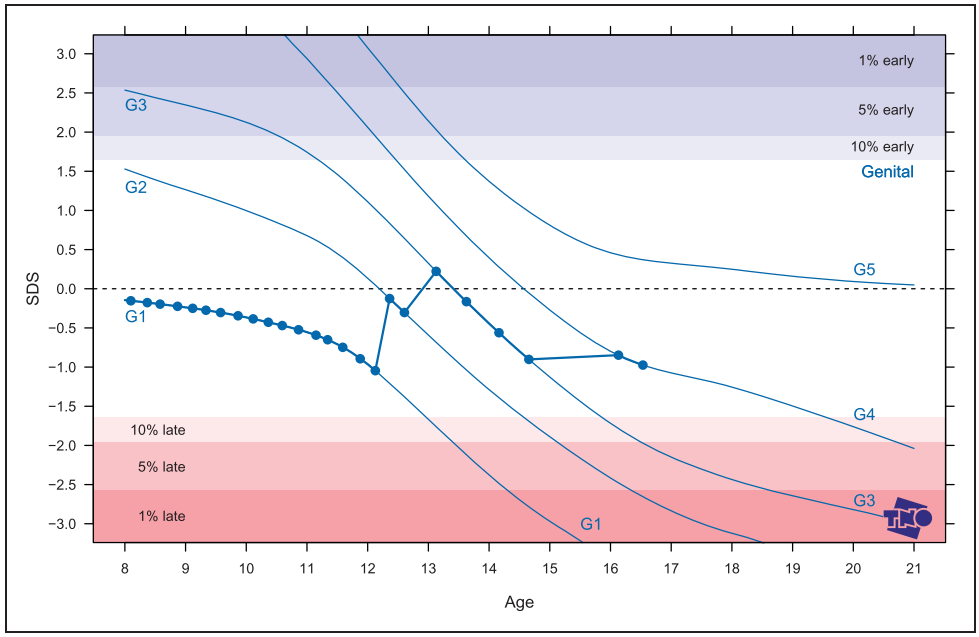

Figure 3. Stage line diagram for genital development in Tanner stages G1–G5.

The stage line diagram generalizes the format proposed by Sorva[48] to discrete data. Normal growth is shown as a horizontal line and any deviation from this indicates abnormal change in growth. Observations within the middle region, say between −2 SD and +2 SD lines, signal normal development. Early maturing children are placed near the top of the diagram, while children with developmental delay appear near the bottom. Regions signifying 10%, 5% and 1% extreme children are marked at both sides for easy reference. These regions can be used to set action levels and monitor treatment. The slope of the curve starting from the last B1-mark is a measure of developmental tempo in SDS/year.

We can read off the age interval [P10, P90] for a stage from the diagram as follows. First find the age at which the stage line crosses the 90% late region, i.e. P90(B1) = 12.17 years. Next, select the next higher stage line, and find the age at which the stage line crosses the 10% early region. For example, we find that P10(B2) = 8.98. Thus, the [P10, P90] interval for stage B1 is [8.98,12.17]. These values can be used conventionally for classifying children into 'early', 'normal', 'late'. Note that for stage lines, tail probabilities are doubled so $Z = -1.64$ matches the 10th percentile (and not the 5th).

In addition, we can compare maturation across different ages. Since all diagrams use the same Y-axis, we can plot multiple scores on the same diagram. This will visualize differences in status and tempo of different aspects of pubertal development within the same individual. Figure 4 shows pubertal development of a boy. Genital and testicular development are normal, but pubic hair is clearly ahead between ages 10 and 13 years. Since testicular volume has many stages (12), the resulting curves are generally less erratic. Note that how graph indicates that testicular volume was not measured at all occasions. This boy also had measurements before the age of 8 years. This is indicated by the line extending left from the first visible measurement. If desired, other measures (e.g. height SDS, BMI SDS) can be added.

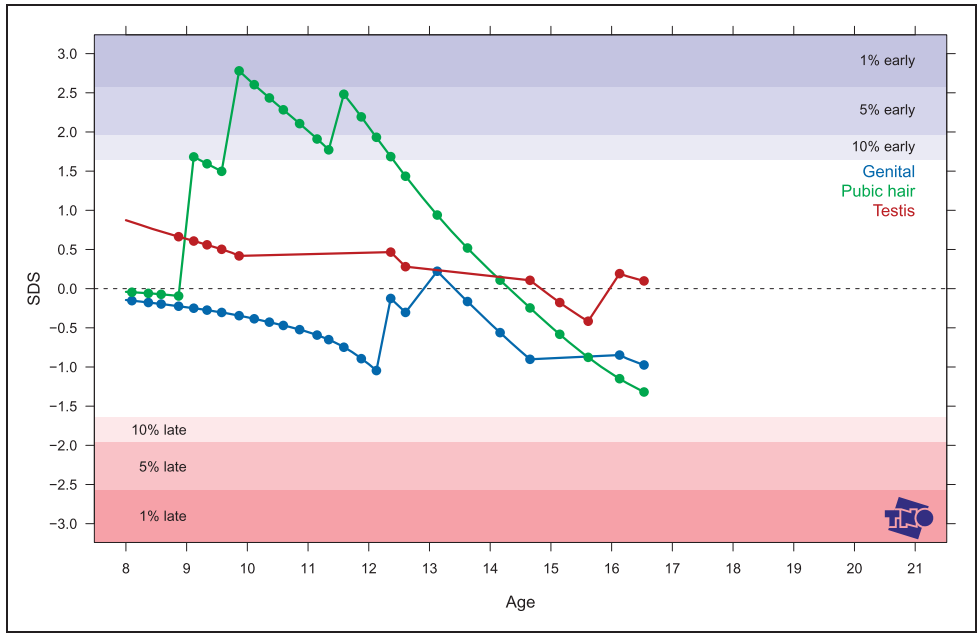

Figure 4. Stage line diagram for Tanner scores for genital, pubic hair and testis development combined.

4.6 Related work

Wade et al.[49] pioneered the construction of reference standard from ordinal data. Applications include visual acuity during childhood[50] and the recognition of emotions.[51] Royston extended the family of models to include multiple covariates and more liberal forms of dependencies between the outcome and the covariates.[52] The method is designed to model outcome variables with peculiar distributions (e.g. many zeroes) that are difficult to model in the conventional way.

Potential fields of applications of the stage line diagram include: dentistry (tooth eruption), oncology (tumor grading, cancer staging), virology (HIV infection and disease staging), psychology (stages of cognitive development), human development (pubertal stages) and chronic diseases (stages of dementia). The web site that implements the stage line diagram is located at http://vps.stefvanbuuren.nl/puberty. Readers interested in calculating maturation scores on their own data can do this via this web site.

5 Conclusion

Age-conditional growth charts aid in tracking development over time. The problem of creating reference diagrams for developmental data has not received proper attention in the past. This article presents and discusses two novel types of growth diagrams. The *D*-score diagram summarizes the information collected on multiple indicators into a single summary measure, the *D*-score. It is possible to calculate reference values and draw reference diagrams by well-tested techniques for continuous data. The stage line diagram preserves the discrete nature of the observed data, and estimates one parameter per stage that varies with age. Both allow us to spot abnormal development and to gauge differences in tempo for individuals.

Both diagrams assume the existence of a continuous latent variable on which the 'true' developmental score of a person can be placed. The person's location on the latent variable changes over time as development progresses. The primary difficulty is to construct appropriate models that translate the observed data into a location of the latent continuum. Latent variable models differ in the way in which they connect the latent variables to the data. The primary vehicle of the *D*-score diagram is the Rasch model. The EAP estimator turns out to be a good way of calculating *D*-scores for infant development. The stage line diagram assumes a latent variable model that represents the hypothetical 'true' developmental score prior to discretization by the measurement process. This is similar to the models considered by Scott Long.[53]

As a by-product, we obtain quantified versions of the discrete data. These maturation scores are interesting in their own right, and often much easier to analyze that the original measurements. For example, it is straightforward to calculate the Pearson correlation between breast development and pubic hair development (it is equal to 0.59). Alternatively, we can use maturation as a predictor in a risk model to predict developmental delay. In one of our recent applications, the *D*-score at the age of 2 years was found to be a highly discriminatory predictor for developmental disability at the age of 5–10 years.[54] We expect that creative researchers will find novel ways to put these new measures to work.

Acknowledgments

The studies were performed in cooperation with the Well Baby Clinics and Municipal Health Services. I thank the participating schools and universities, the Koninklijke Landmacht, and the Evangelische Omroep. Pieter Herngreen, Thea Reerink and Miranda Fredriks put great efforts into the collection of the data. I thank Elise Dusseldorp and two anonymous reviewers for their suggestions for improvement.

References

1. Cole TJ and Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat Med* 1992; **11**(10): 1305–1319.
2. Borghi E, de Onis M, Garza C, et al. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Stat Med* 2006; **25**(2): 247–265. (Available at: http://www.stefvanbuuren.nl/publications/Construction%20WHO%20-%20Stat%20Med%202006.pdf).
3. Van Buuren S. Growth references. In: Kelnar C, Savage M, Saenger P and Cowell C (eds) *Growth disorders*, 2nd ed. London: Hodder Arnold, 2007, pp.165–181.
4. Royston P and Wright EM. Goodness-of-fit statistics for age-specific reference intervals. *Stat Med* 2000; **19**: 2943–2962.
5. Van Buuren S and Fredriks AM. Worm plot: a simple diagnostic device for modelling growth reference curves. *Stat Med* 2001; **20**(8): 1259–1277.
6. Stasinopoulos DM and Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw* 2007; **23**(7): 1–46. (Available at: http://www.jstatsoft.org/v23/i07).
7. Tanner JM and Whitehouse RH. *Growth and development reference charts* (Tanner-Whitehouse Standards). Hertford, UK: Castlemead Publications, 1984.
8. Inhelder B and Piaget J. *The growth of logical thinking from childhood to adolescence*. New York: Basic Books, 1958.
9. Bayley N. *Bayley scales of infant development*, 2nd ed. San Antonio, TX: Psychological Corp, 1993.
10. Marshall WA and Tanner JM. Variations in pattern of pubertal changes in girls. *Arch Dis Child* 1969; **44**: 291–303.
11. Marshall WA and Tanner JM. Variations in pattern of pubertal changes in boys. *Arch Dis Child* 1970; **45**: 13–23.
12. Kohlberg L. *The psychology of moral development: the nature and validity of moral stages*. Vol 2, San Francisco: Harpen & Row, 1984.
13. De Boeck P, Wilson M and Scott Acton G. A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychol Rev* 2005; **112**(1): 129–158.
14. McArdle JJ, Grimm KJ, Hamagami F, et al. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychol Meth* 2009; **14**(2): 126–149.
15. Thurstone LL. The absolute zero in intelligence measurement. *Psychol Rev* 1928; **35**: 175–197.
16. Jones MC, Bayley N, McFarlane JW, et al. *The course of human development*. Selected Papers from the Longitudinal Studies. Institute of Human Development, the University of California, Berkeley. Waltham, MA: Xerox, 1971.
17. Bayley N. Individual patterns of development. *Child Develop* 1956; **27**: 45–74.
18. Sayer AG and Cumsille PE. Second-order latent growth models. In: Collins LM and Sayer AG (eds) *New methods for the analysis of change*. Washington DC: American Psychological Association, 2001, pp.179–200.
19. Fischer GH and Parzer P. An extention of the rating scale model with an application to the measurement of change. *Psychometrika* 1991; **56**: 637–651.

20. Herngreen WP, Reerink JD, van Noord-Zaadstra BM, et al. The SMOCC-study: design of a representative cohort of live-born infants in the Netherlands. *Eur J Public Health* 1992; **2**: 117–122.
21. Laurent de Angulo MS. *Ontwikkelingsonderzoek in de Jeugdgezondheidszorg*. Assen: Van Gorcum, 2008.
22. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.
23. Rasch G. On general laws and the meaning of measurement in psychology. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, IV*. Berkeley: University of California Press, 1961, pp.321–334.
24. Andrich D and Luo G. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *J Appl Meas* 2003; **4**(3): 205–221.
25. Jacobusse G, Van Buuren S and Verkerk PH. An interval scale for development of children aged 0-2 years. *Stat Med* 2006; **25**(13): 2272–2283.
26. Jacobusse G and Van Buuren S. Computerized adaptive testing for measuring development of young children. *Stat Med* 2007; **26**(13): 2629–2638.
27. Cheung YB, Gladstone M, Maleta K, et al. Comparison of four statistical approaches to score child development: a study of Malawian children. *Tropical Med Int Health* 2008; **8**: 987–993.
28. Vale CD. Linking item parameters onto a common scale. *Appl Psychol Meas* 1986; **10**: 333–344.
29. Embretsen SE and Reise SP. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum, 2000.
30. Bock DD and Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. *Appl Psychol Meas* 1982; **6**(4): 431–444.
31. Chen SK, Hou L and Dodd BG. A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educ Psychol Meas* 1998; **58**(4): 569–595.
32. Cole TJ, Freeman JV and Preece MA. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by mximum penalized likelihood. *Stat Med* 1998; **17**: 407–429.
33. van Dommelen P and van Buuren S. Evidence-based referral criteria in growth monitoring. *Stat Meth Med Res* 2012; (to appear).
34. Dawson TL. New tools, new insights: Kohlberg's moral judgement stages revisited. *Int J Behav Develop* 2002; **26**(2): 154–166.
35. Draney K. The saltus model applied to proportional reasoning data. *J Appl Meas* 2007; **8**(4): 438–455.
36. Boom J, Wouters H and Keller M. A cross-cultural validation of stage development: a Rasch re-analysis of longitudinal socio-moral reasoning data. *Cognit Develop* 2007; **22**: 213–229.
37. Boom J. Measuring moral development: stages as markers along a latent developmental dimension. In: Koops W, Brugman D, Ferguson TW and Sanders AF (eds) *The development and structure of conscience*. London: Psychology Press, 2010, pp.151–167.
38. Wilson M. Saltus: A psychometric model of discontinuity in cognitive development. *Psychol Bull* 1989; **105**(2): 276–289.
39. van Wieringen JC, Wafelbakker F, Verbrugge HP, et al. *Growth diagrams 1965 Netherlands*. Leiden: Nederlands Instituut Praeventieve Geneeskunde, 1971.
40. Fredriks AM, van Buuren S, Burgmeijer RJF, et al. Continuing positive secular growth change in The Netherlands 1955-1997. *Pediatr Res* 2000; **47**(3): 316–323.
41. Mul D, Fredriks AM, van Buuren S, et al. Pubertal development in the Netherlands 1965-1997. *Pediatr Res* 2001; **50**(4): 479–486.
42. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Meth Med Res* 2007; **16**(3): 219–242. (Available at: http://www.stefvanbuuren.nl/publications/MI%20by%20FCS%20-%20SMMR%202007.pdf).
43. Sun SS, Schubert MS, Chumlea WC, et al. National estimates of the timing of sexual maturation and racial differences among US children. *Pediatrics* 2002; **110**: 911–919.
44. Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43**: 561–573.
45. Hastie TJ and Tibshirani RJ. *Generalized additive models*. Vol 1, London: Chapman and Hall, 1990.
46. Atwood CL and Taube A. Estimating mean time to reach a milestone, using retrospective data. *Biometrics* 1976; **32**(1): 159–172.
47. van Buuren S and Ooms JCL. Stage line diagram: an age-conditional reference diagram for tracking development. *Stat Med* 2009; **28**(11): 1569–1579.
48. Sorva R, Perheentupa J and Tolppanen EM. A novel format for growth chart. *Acta Paediatrica* 1984; **73**(4): 527–529.
49. Wade AM, Ades AE, Salt AT, et al. Age-related standards for ordinal data: modelling the changes in visual acuity from 2 to 9 years of age. *Stat Med* 1995; **14**(3): 257–266.
50. Wade AM, Salt AT, Proffitt RV, et al. Likelihood-based modelling of age-related normal ranges for ordinal measurements: changes in visual acuity through early childhood. *Stat Med* 2004; **23**(23): 3623–3640.
51. Wade AM, Lawrence K, Mandy W, et al. Charting the development of emotion recognition from 6 years of age. *J Appl Stat* 2006; **33**(3): 297–315.
52. Royston P. A parametric model for ordinal response data, with application to estiming age-specific reference intervals. *Biostatistics* 2000; **1**(3): 263–277.
53. Scott Long J. *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage, 1997.
54. Boere-Boonekamp MM, Dusseldorp E, Hafkamp-de Groen E, et al. *Screening for developmental disability is possible* 2011; (submitted for publication).