Child Development with the D-score: Turning Milestones into Measurement

Authors: Stef van Buuren & Iris Eekhout

Contents

Frontmatter

Chapter 1

Short history

The measurement of child development has quite an extensive history. This section

- reviews definitions of child development (1.1)
- discusses concepts in the nature of child development (1.2)
- shows a classic example of motor measurements (1.3)
- summarizes typical questions whose answers need proper measurements (1.4)

1.1 What is child development?

In contrast to concepts like height or temperature, it is unclear what exactly constitutes child development. Shirley (1931) executed one of the first rigorous studies in the field with the explicit aim

> that the many aspects of development, anatomical, physical, motor, intellectual, and emotional, be studied simultaneously.

Shirley gave empirical definitions of each of these domains of development.

Certain domains advance through a fixed sequence. Figure 1.1 illustrates the various stages needed for going from a *fetal posture* to *walking alone*. The ages are indicative of when these events happen, but there is a considerable variation in timing between infants.

Gesell (1943) (p. 88) formulated the following definition of development:

> Development is a continuous process that proceeds stage by stage in an orderly sequence.

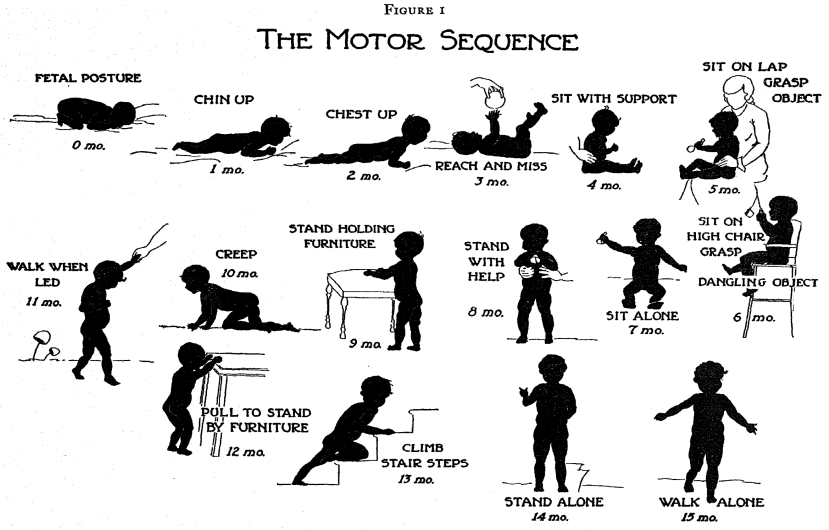

Figure 1.1: Gross motor development as a sequence of milestones. Source: Shirley (1933), with permission.

Gesell's definition emphasizes that development is a continuous process. The stages are useful as indicators to infer the level of maturity but are of limited interest by themselves.

Liebert, Poulos, and Strauss (1974) (p. 5) emphasized that development is not a phenomenon that unfolds in isolation.

> Development refers to a process in growth and capability over time, as a function of both maturation and interaction with the environment.

Cameron and Bogin (2012) (p. 11) defined an endpoint of development, as follows:

> "Growth" is defined as an increase in size, while "maturity" or "development" is an increase in functional ability... The endpoint of maturity is when a human is functionally able to procreate successfully ... not just biological maturity but also behavioural and perhaps social maturity.

Berk (2011) (p. 30) presented a dynamic systems perspective on child development as follows:

> Development cannot be characterized as a single line of change, and is more like a web of fibres branching out in many directions, each representing a different skill area that may undergo both continuous and stagewise transformation.

There are many more definitions of child development. The ones described here illustrate the main points of view in the field.

1.2 Theories of child development

The field of child development is vast and spans multiple academic disciplines. This short overview, therefore, cannot do justice to the enormous richness. Readers new to the field might orient themselves by browsing through an introductory academic titles (Santrock 2011; Berk 2011), or by searching for the topic of interest in an encyclopedia, e.g., Salkind (2002).

The introductions by Santrock (2011) and Berk (2011) both distinguish major theories in child development according to how each answer to following three questions:

1.2.1 Continuous or discontinuous?

Does development evolve gradually as a continuous process or are there qualitatively distinct stages, with jumps occurring from one step to another?

Many stage-based theories of human development have been proposed over the years: social and emotional development by psycho-sexual stages introduced by Freud and furthered by Erikson (Erikson 1963), Kohlberg's six stages of moral development (Kohlberg 1984) and Piaget's cognitive development theory (Piaget and Inhelder 1969). Piaget distinguishes four main periods throughout childhood. The first period, the *sensorimotor period* (approximately 0-2 years), is subdivided into six stages. When taken together, these six stages describe "the road to conceptual thought." Piaget's stages are qualitatively different and aim to unravel the mechanism involved in intellectual development.

On the other hand, Gesell and others emphasize development as a continuous process. Gesell (1943) (p. 88) says:

> A stage represents a degree or level of maturity in the cycle of development. A stage is simply a passing moment, while development, like time, keeps marching on.

1.2.2 One course or multiple parallel tracks?

Stage theorists assume that children progress sequentially through the same set of stages. This assumption is also explicit in the work of Gesell.

The ecological and dynamic systems theories view development as continuous, though not necessarily progressing in an orderly fashion, so there may be multiple, parallel ways to reach the same point. The developmental path taken by a given child will depend on the child's unique combination of personal and environmental circumstances, including cultural diversity in development.

1.2.3 Nature or nurture?

Figure 1.2 illustrates that children vary in appearance. Are genetic or environmental factors more important for influencing development? Most theories generally acknowledge the role of both but differ in emphasis. In practice, the debate centres on the question of how to explain individual differences.

Maturation is the process of becoming fully developed, much like the natural unfolding of a flower. The process depends on both genetic factors (species, breed) as well as environmental influences (sunlight, water, nutrition). Some theorists emphasize that differences in child development are innate and stable over time, although there may be differences in unfolding speed due to different

Figure 1.2: A group of culturally diverse children. Source: Shutterstock, under license.

environments. Others argue that environmental factors drive differences in development between children, and changing these factors could very well impact child development.

Our position in this debate has practical implications. If we believe that differences are natural and stable, then it may not make much sense trying to change the environment, as the impact on development is likely to be small. On the other hand, we may consider developmental potential as evenly distributed, with its expression governed by the environment. In the latter case, improving life circumstances may have substantial pay-offs in terms of better development.

1.3 Example of motor development

1.3.1 Shirley's motor data

For illustration, we use data on loco-motor development from a classic study on child development among 25 babies. Shirley (1931) collected measurements of the baby's walking ability, starting at ages around 13 weeks, in an ingenious way. The investigator lays out a white paper of twelve inches wide on the floor of the living room, and lightly greases the soles of the baby's feet with olive oil.

The baby was invited to "walk" on the sheet. Of course, very young infants need substantial assistance. Footprints left were later coloured by graphite and measured. Measurements during the first year were repeated every week or bi-weekly.

```
Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
R Markdown document.
```

Table 1.1: Age at beginning stages of walking (in weeks) for 21 babies. Source: Shirley (1931).

Name	Sex	Stepping	Standing	Walking with help	Walking alone
Martin	boy	15		21	50
Carol	girl	15	19	37	50
Max	boy	14		25	54
Virginia Ruth	girl		21	41	54
Sibyl	girl		22	37	58
David	boy	19	27	34	60
James D.	boy	19	30	45	60
Harvey	boy	14	27	42	62
Winnifred	girl	15	30	41	62
Quentin	boy	15	23	38	64
Maurice	boy	18	23	45	66
Judy	girl	18	29	45	66
Irene May	girl	19	34	45	66
Peter	boy	15	29	49	66
Walley	boy	18	33	54	68
Fred	boy	15	32	46	70
Donovan	boy		23	50	70
Patricia	girl	15	30	45	70
Torey	boy		21	72	74
Larry	boy	13	41	54	76

Name	Sex	Stepping	Standing	Walking with help	Walking alone
Doris	girl	23		44	

Table 1.1 (Shirley 1931, Appendix 8) lists the age (in weeks) of the 21 babies when they started, respectively, stepping, standing, walking with help, and walking alone. Blanks indicate missing data. A blank in the first column means that the baby was already stepping when the observation started (Virginia Ruth, Sibyl, Donovan, Torey and Doris). Max and Martin, who have blanks in the second column, skipped standing and went directly from stepping to walking with help. Doris has a blank in the last column because she passed away before she could walk alone.

1.3.2 Individual trajectories of motor development

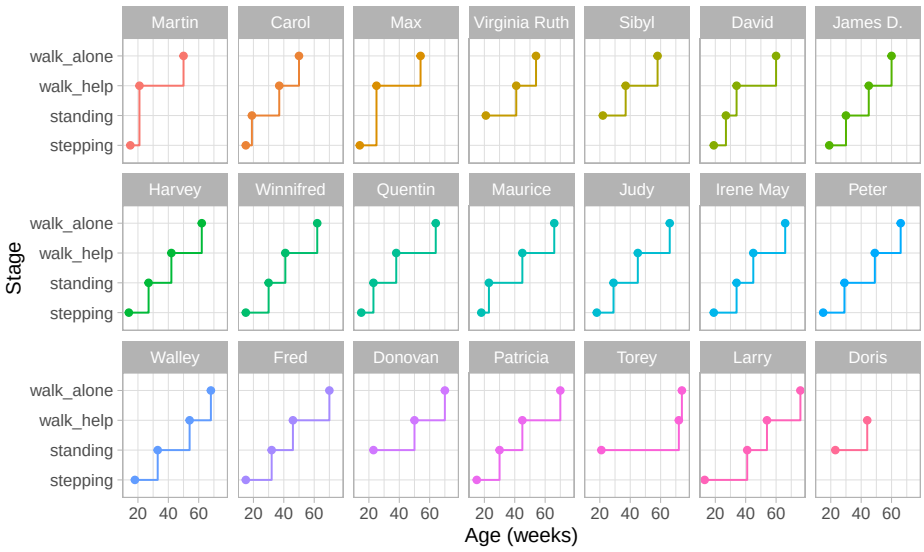

Figure 1.3: Staircase plot indicating the age at which each baby achieves a new milestone of gross-motor functioning.

Figure 1.3 is a visual representation of the information in Table 1.1. Each data point is the age of the first occurrence of the next stage. Before that age, we assume the baby is in the previous stage.

Figure 1.3 makes it easy to spot the quick walkers (Martin, Carol) and slow walkers (Patricia, Torey, Larry). Furthermore, we may also locate children who remain a long time in a particular stage (Torey, Larry) or who jump over stages (Martin, Max).

For ease of plotting, the categories on the vertical axis are equally spaced. The height of the jump from one stage to the next has no sensible interpretation. We might be inclined to think that the vertical distance portrays to how difficult it is to achieve the next stage, but this is inaccurate. Instead, the ability needed to set the next step corresponds to the *horizontal line length* between stages. For example, on average, the line for `stepping` is rather short in all plots, so going from `stepping` to `standing` is relatively easy.

Figure 1.3 presents data from only those visits where a jump occurred. The number of house visits made during the ages of 0-2 years was far higher. Shirley (1931) collected data from 1370 visits, whereas Figure 1.3 plot only the 76 occasions that showed a jump. Thus the data collection needs to be intense and costly to obtain individual curves. Fortunately, there are alternatives that are much more efficient.

1.4 Typical questions asked in child development

The emotional, social and physical development of the young child has a direct effect on the adult he or she will become. We may be interested in measuring child development for answering clinical, policy or public health questions.

```
Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
R Markdown document.
```

Table 1.2: Questions whose answers require quantitative measurements of child development.

Level	Question
Individual	What is the child's gain in development since the last visit?
Individual	What is the difference in development between the child and peers of the same age?
Individual	How does the child's development compare to a norm?
Group	What is the effect of this intervention on child development?
Group	What is the difference in child development between these two groups?
Population	What is the change in average child development since the last measurement?
Population	What was the effect of implementing this policy on child development?

Level	Question
Population	How does this country compare to other countries in terms of child development?

Table 1.2 lists typical questions whose answers require measuring child development. Note that all questions compare the amount of child development between groups or time points. A few questions compare development for the same child, group or population at different ages. Others compare development at the same age across different children, groups or populations.

Chapter 2

Quantifying child development

This section discusses four principles to quantify child development:

- Age-based measurement (2.1)
- Probability-based measurement (2.2)
- Score-based measurement (2.3)
- Unit-based measurement (2.4)

2.1 Age-based measurement of development

2.1.1 Motivation for age-based measurement

Milestones form the based building blocks for instruments to measure child development. Methods to quantify growth using separate milestones relate the milestone behaviour to the child's age. Gesell (1943) (p. 89) formulated this goal as follows:

> We think of behaviour in terms of age, and we think of age in terms of behaviour. For any selected age it is possible to sketch a portrait which delineates the behaviour characteristics typical of the age.

There is an extensive literature that quantifies development in terms of the ages at which the child is expected to show a specific behaviour. The oldest methods for quantifying child development calculate an *age equivalent* for achieving a milestone, and compare the child's age to this age equivalent.

2.1.2 Age equivalent and developmental age

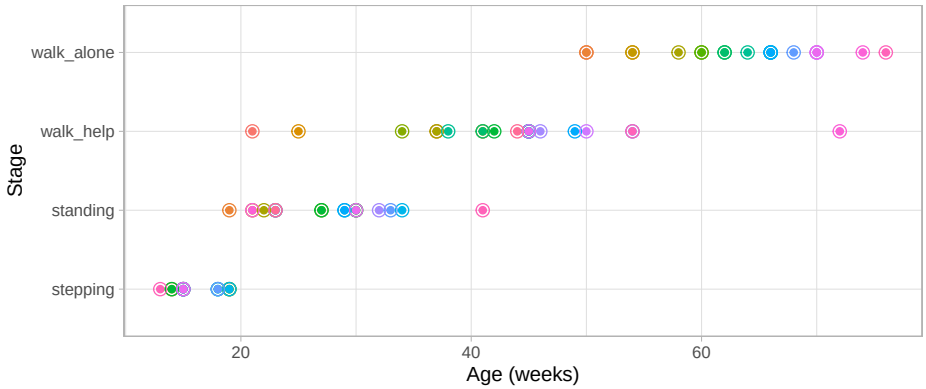

Figure 2.1: Ages at which 21 children achieve four motor development milestones.

Figure 2.1 graphs the ages at which each of the 21 children enter a given stage in Shirley's motor data of Table 1.1. Since `standing` follows `stepping`, children who can stand are older than the children who are stepping. Hence the ages for standing are located more to the right.

Since age and development are so intimately related, we can express the *difficulty* of a milestone as the *mean age* at which children achieve it. For example, Stott (1967) (p. 25) defines the *age equivalent* and its use for measurement, as follows:

> The age equivalent of a particular stage is simply the average age at which children reach that particular stage.

Figure 2.2 adds the mean age and the boxplot at which the children enter the four stages. The difficulty of these milestones can thus be expressed as age equivalents: 16.1 weeks for `stepping`, 27.2 weeks for `standing`, 43.3 weeks for `walking with help` and 63.3 weeks for `walking alone`.

Thus, a child that is stepping beyond the age of 16.1 weeks is considered later than average, whereas a child already stepping before 27.2 weeks earlier than average. We may also calculate age delta as the difference between the child's age and the norm age, and express it as "two weeks late" or "three weeks ahead." Summarizing age delta's over different milestones has led to concepts like *developmental age* as a measure of a child's development.

2.1.3 Limitations of age-based measurement

Age-based measurement is easy to understand, and widely used in the popular press, but not without pitfalls:

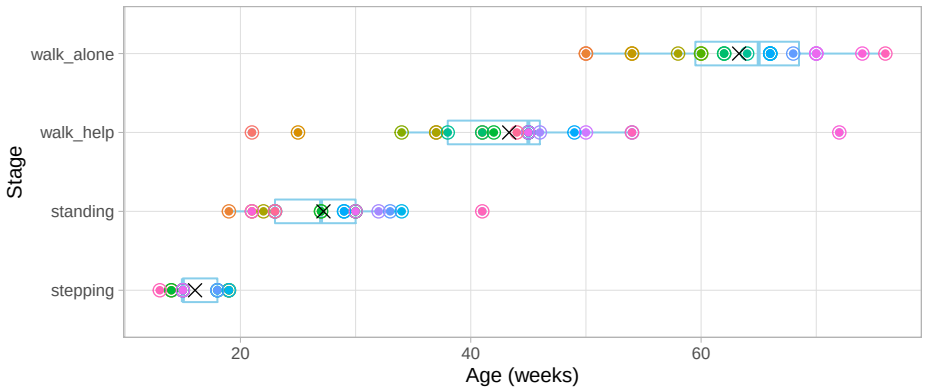

Figure 2.2: Mean (symbol `x`) and spread of the ages at which 21 children achieve four motor development milestones.

1. Age-based measurement requires us to know the ages at which the child entered a new stage. The mean age can be a biased estimate of item difficulty if visits are widely apart, irregular or missing.
2. Age-based measurement can inform us whether a child is achieving a given milestone early of late. However, it does not tell us what behaviours are characteristic for children of a given age.
3. Age-based measurement cannot exist without an age norm. When there are no norms, we cannot quantify development.
4. Age-based measurement works only at the item level. Although we may average age delta's over milestones, the choice of milestones is arbitrary.

2.2 Probability-based measurement

An alternative is to calculate the *probability* of achieving a milestone at a given age and compare the child's response to that probability.

The passing probability is an interpretable and relevant measure. An operational advantage of the approach is that the necessary calculations place fewer demands on the available data and can be done even for cross-sectional studies.

2.2.1 Example of probability-based measurement

Figure 2.3 plots the percentage of children achieving each of Shirley's motor stages against age. There are four cumulative curves, one for each milestone, that indicate the percentage of children that pass.

In analogy to the age equivalent introduced in Section 2.1.2 we can define the *difficulty* of the milestone as the age at which 50 per cent of the children pass.

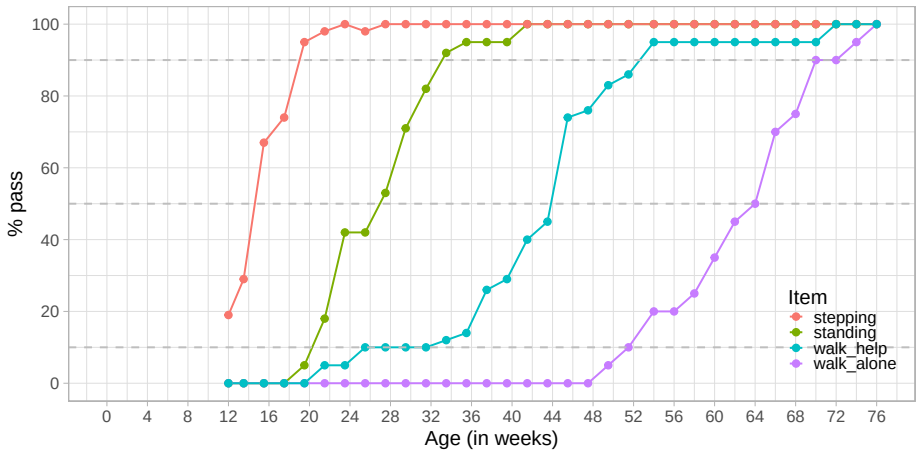

Figure 2.3: Probability of achieving four motor milestones against age.

In the Figure we see that the levels of difficulty are approximately 14.2 weeks (`stepping`), 27.0 weeks (`standing`), 43.8 weeks (`walking with help`) and 64.0 weeks (`walking alone`). Also, we may easily find the ages at which 10 per cent or 90 per cent of the children pass each milestone.

Observe there is a gradual decline in the steepness as we move from `stepping` to `walk_alone`. For example, we need an age interval of 13 weeks (33 - 20) to go from 10 to 90 per cent in `standing`, but need 19 weeks (71 - 52) to go from 10 to 90 per cent in `walking alone`. Thus, one step on the age axis corresponds to different increments in probability. The flattening pattern is typical for child development and represents evidence that evolution is faster at earlier ages.

2.2.2 Limitations of probability-based measurement

Probability-based measurement is a popular way to create instruments for screening on developmental delay. For example, each milestone in the Denver II (Frankenburg et al. 1992) has markers for the 25th, 50th, 75th and 90th age percentile.

1. The same age step corresponds to different probabilities.
2. The measurement cannot exist without some norm population. When norms differ, we cannot compare the measurements.
3. Interpretation is at the milestone level, sometimes supplemented by procedures for counting the number of delays. No aggregate takes all responses into account.

2.3 Score-based measurement of development

2.3.1 Motivation for score-based measurement

Score-based measurement takes the responses on multiple milestones and counts the total number of items passed as a measure of development. This approach takes all answers into account, hence leading to a more stable result.

One may order milestones in difficulty, and skip those that are too easy, and stop administration for those that are too difficult. In such cases, we cannot merely interpret the sum score of a measure of development. Instead, we need to correct for the subset of administered milestones. The usual working assumption is that the child would have passed all easier milestones and failed on all more difficult ones. We may repeat this procedure for different domains, e.g. motor, cognitive, and so on.

2.3.2 Example of score-based measurement

Figure 2.4 is a gross-motor score calculated as the number of milestones passed. It varies from 0 to 3.

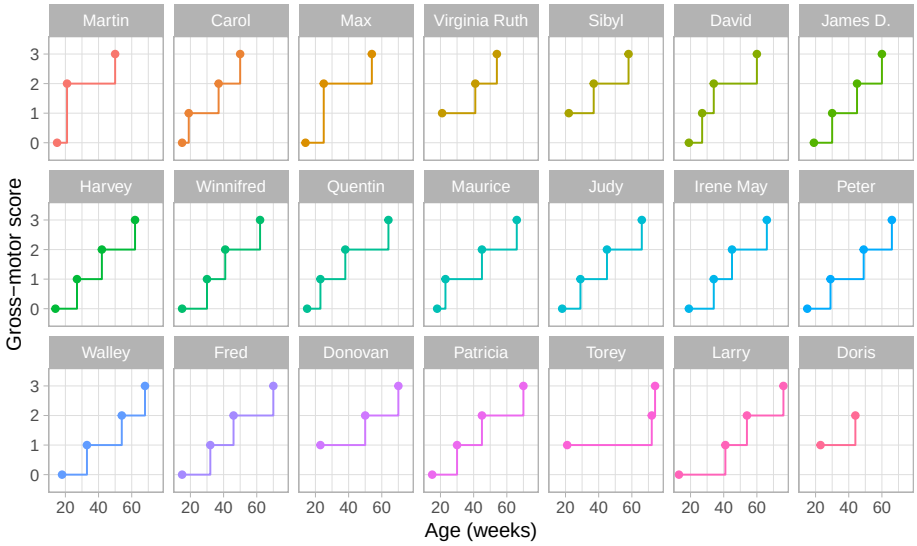

Figure 2.4: Same data as in Figure 1.3, but now with the vertical axis representing gross-motor score.

The plot suggests that the difference in development between scores 0 and 1 is the same as the difference between, say, scores 2 and 3. *This is not correct.*

For example, suppose that we express the difficulty of the milestone as an age-equivalent. From section 2.1.2 we see that the difference between stepping and standing is 27.2 - 16.1 = 11.1 weeks, whereas the difference between walking alone and walking with help is 63.3 - 43.3 = 20 weeks. Thus, according to age equivalents scores 0 and 1 should be closer to each other, and ratings 2 and 3 should be drawn more apart.

2.3.3 Limitations of score-based measurement

Score-based measurement is today's dominant approach, but is not without conceptual and logistical issues.

1. The total score depends not only on the actual developmental status of the child, but also on the set of milestones administered. If a milestone is skipped or added, the sum score cannot be interpreted anymore as a measure of developmental status. It might be possible to correct for starting and stopping rules under the assumptions described in Section 2.3.1, but such will be involved if intermediate milestones are missing.
2. It is not possible to compare the scores made by different instruments. Some instruments allow conversion to age-conditional scores. However, the sample used to derive such transformations pertain to that tool and does not generalise to others.
3. Domains are hard to separate. For example, some cognitive milestones tap into fine motor capabilities, and vice versa. There are different ways to define domains, so domain interpretation varies by instrument.
4. Administration of a full test may take substantial time. The materials are often proprietary and costly.

2.4 Unit-based measurement of development

2.4.1 Motivation for unit-based measurement

Unit-based measurement starts by defining ideal properties and derives a procedure to aggregate the responses on milestones into an overall score that will meet this ideal.

Section 1.4 highlighted questions for individuals, groups and populations. There are three questions:

- What is the difference in development over time for the same child, group or community?
- What is the difference in development between different children, groups or populations of the same age?

- How does child development compare to a norm?

In the ideal situation, we would like to have a continuous (latent) variable D (for development) that measures child development. The scale should allow us to quantify *ability* of persons, groups or populations from low to high. It should have a *constant unit* so that a given difference in ability refers to the same quantity across the entire scale. We find the same property in height, where a distance of 10 cm represents the same amount for molecules, people or galaxies. When are these conditions are met, we say that we measure on an *interval scale*.

If we succeed in creating an interval scale for child development, an enormous arsenal of techniques developed for quantitative variables opens up to measure, track and analyze child development. We may then evaluate the status of a child in terms of D points gained, create age-dependent diagrams (just like growth charts for height and weight), devise age-conditional measures for child development, and intelligent adaptive testing schemes. Promising studies on Dutch data van Buuren (2014) suggest that such benefits are well within reach.

2.4.2 Example of unit-based measurement

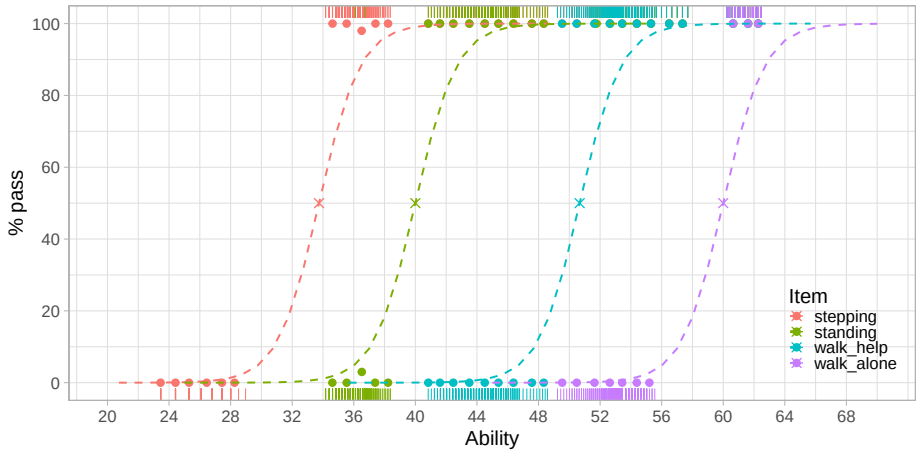

Figure 2.5: Modeled probability of achieving four motor milestones against the D-score.

Figure 2.5 is similar to Figure 2.3, but with `Age` replaced by `Ability`. Also, modelled curves have replaced empirical ones, but this is not essential.

We estimated the ability values on the horizontal axis from the data. The values correspond to the amount of development of each visit. Likewise, we calculated the logistic curves from the data. These reflect the probability of passing each milestone *at a given level of ability*.

Figure 2.5 shows that the probability of passing a milestone increases with ability. Items are sorted according to difficulty from left to right. Milestone `stepping` is the easiest and `walk_alone` is the most difficult. The point at which a logistic curve crosses the 50 per cent line (marked by a cross) is the *difficulty of the milestone*.

The increase in ability that is needed to go from 10 to 90 per cent is about five units here. Since all curves are parallel, the interval is constant for all scale locations. Thus, the scale is an *interval scale* with a *constant unit of measurement*, the type of measurement needed for answering the basic questions identified in Section 2.4.1.

2.4.3 Limitations of unit-based measurement

While unit-based measurement has many advantages, it cannot perform miracles.

1. An important assumption is that the milestones "measure the same thing," or put differently, are manifestations of a continuous latent variable that can be measured by empirical observations. Unit-based measurement won't work if there is no sensible latent scale.
2. The portrayed advantages hold only if the discrepancies between the data and the model are relatively small. Since the simplest and most powerful measurement models are strict, it is essential to obtain a good fit between the data and the model.
3. The construction of unit-based measurement requires psychometric expertise, specialized computer software and considerable sample sizes.

2.5 A unified framework

This section brings together the four approaches outlined in this section into a unified framework.

Figure 2.6: Placing milestones and children onto the same line reveals their positions.

Figure 2.6 shows the imaginary positions on a gross-motor continuum of three babies from Figure 1.1 at the age of 30 weeks. Both milestones and children are

ordered along the same continuum. Thus, standing is more difficult than stepping, and at week 30, Doris is ahead of Walley in terms of motor development.

More generally, measurement is the process of locating milestones and children on a line. This line represents a *latent variable*, a continuous construct that defines the different poles of the concept that we want to measure. A latent variable ranges from low to high.

The first part of measurement is to determine the location of the milestones on the latent variable. In many cases, the instrument maker has already done that. For example, each length marker on a ruler corresponds to a milestone for measuring length. The manufacturer of the ruler has already placed the marks at the appropriate places on the tool, and we take for granted that each marker has been calibrated correctly.

A milestone for child development is similar to a length marker, but

- we may not know how much development the milestone measures, so its location on the line is unknown, or uncertain;
- we may not know whether the milestone measures child development at all so that it may have no location on the line.

The second part of measurement is to find the location of each child on the line. For child height, this is easy: We place the horizontal headpiece on top of the child's head and read off the closest height marker. Since we lack a physical ruler for development, we must deduce the child's location on the line from the responses on a series of well-chosen milestones.

By definition, we cannot observe the values of a latent variable directly. However, we may be able to measure variables (milestones) that are related to the latent variable. For example, we may have scores on tasks like *standing* or *walking with help*.

The *measurement model* specifies the relations between the actual measurements and the latent variable. Under a given measurement model, we may estimate the locations of milestones and children on the line. Section 3.5 discusses measurement models in more detail.

2.6 Why unit-based measurement

```
Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
R Markdown document.
```

Table 2.1: Evaluation of four measurement approaches on seven criteria.

Criterion	Age	Probability	Score	Unit
Independent of age norm	No	No	Yes	Yes
Supports multiple milestones	No	No	Yes	Yes
Latent variable	No	No	Yes	Yes
Robust to milestone skipping	Yes	Yes	No	Yes
Comparable scores	Yes	Yes	No	Yes
Probability model	No	Yes	No	Yes
Defines measurement unit	No	No	No	Yes

This section distinguished four approaches to measure child development: *age-based*, *probability-based*, *score-based* and *unit-based* measurement. Table 2.1 summarizes how the approaches evaluate on nine criteria.

Age-based measurement expresses development in age equivalents, whose precise definition depends on the reference population. Age-based measurement does not support multiple milestones and does not use the concept of a latent variable.

Probability-based measurement expresses development as age percentiles for a reference population. It is useful for individual milestones but does not support multiple items or a latent variable interpretation.

Score-based measurement quantifies development by summing the number of passes. Different instruments make different selections of milestones, so the scores taken are unique to the tool. Thus comparing the measurement obtained by different devices is difficult. Skipping or adding items require corrections.

Unit-based measurement defines a unit by a theoretical model. When the data fit the model, we are able to construct instruments that produce values in a standard metric.

Chapter 3

The D-score

Section 1 provided historical background on the nature of child development. Section 2 discussed three general quantification approaches. This section explains how to apply the unit-based approach to arrive at the D-score scale. The text illustrates the process with real data.

- Dutch Development Instrument (DDI) (3.1)
- Milestone passing by age and by D-score (3.2, 3.3)
- How do age and D-score relate? (3.4)
- Role of the measurement model (3.5)
- Item and person response functions (3.6)
- Engelhard invariance criteria (3.7)
- Why the Rasch model? (3.8)

3.1 The Dutch Development Instrument (DDI)

3.1.1 Setting

The Dutch Youth Health Care (YHC) routinely monitors the development of almost all children living in The Netherlands. During the first four years, there are 13 scheduled visits. During these visits, the YHC professionals evaluate the growth and development of the child.

The *Dutch Development Instrument* (DDI; in Dutch: *Van Wiechenschema*) is the standard instrument used to measure development during the ages 0-4 years. The DDI consists of 75 milestones. The instrument assesses three developmental domains:

1. Fine motor, adaptation, personality and social behaviour;

2. Communication;
3. Gross motor.

The milestones form two sets, one for children aged 0-15 months, and another for children aged 15-54 months. The YHC professionals administer an age-appropriate subset of milestones at each of the scheduled visits, thus building a *longitudinal developmental profile* for each child.

3.1.2 Description of SMOCC study

The Social Medical Survey of Children Attending Child Health Clinics (SMOCC) study is a nationally representative cohort of 2,151 children born in The Netherlands during the years 1988–1989 (Herngreen et al. 1994). The study monitored child development using observations made on the DDI during nine visits covering the first 24 months of life. The SMOCC study collected information during the first two years on 57 (out of 75) milestones.

The *standard* set in the DDI consists of relatively easy milestones that 90 per cent of the children can pass at the scheduled age. This set is designed to have maximal sensitivity for picking up delays in development. A distinctive feature of the SMOCC study was the inclusion of more difficult milestones beyond the standard set. The *additional* set originates from the next time point. The success rate on these milestones is about 50 per cent.

3.1.3 Codebook of DDI 0-30 months

```
Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
R Markdown document.
```

Table 3.1: Codebook of DDI as used in the SMOCC study

Item	Debut	Domain	Label
ddicmm029	1m	Communication	Reacts when spoken to
ddifmd001	1m	Fine motor	Eyes fixate
ddigmd052	1m	Gross motor	Moves arms equally well
ddigmd053	1m	Gross motor	Moves legs equally well
ddigmd056	1m	Gross motor	Lifts chin off table for a moment

Item	Debut	Domain	Label
ddicmm030	2m	Communication	Smiles in response (M; can ask parents)
ddifmd002	2m	Fine motor	Follows with eyes and head 30d < 0 > 30d
ddicmm031	3m	Communication	vocalizes in response
ddifmd003	3m	Fine motor	Hands open occasionally
ddifmm004	3m	Fine motor	Watches own hands
ddigmd054	3m	Gross motor	Stays suspended when lifted under the armpits
ddigmd057	3m	Gross motor	Lifts head to 45 degrees on prone position
ddicmd116	6m	Communication	Turn head to sound
ddifmd005	6m	Fine motor	Plays with hands in midline
ddigmd006	6m	Gross motor	Grasps object within reach
ddigmd055	6m	Gross motor	No head lag if pulled to sitting
ddigmd058	6m	Gross motor	Looks around to side with angle face-table 90
ddigmd059	6m	Gross motor	Flexes or stomps legs while being swung
ddicmm033	9m	Communication	Says dada, baba, gaga
ddifmd007	9m	Fine motor	Passes cube from hand to hand
ddifmd008	9m	Fine motor	Holds cube, grasps another one with other hand
ddifmm009	9m	Fine motor	Plays with both feet
ddigmm060	9m	Gross motor	Rolls over back to front
ddigmd061	9m	Gross motor	Balances head well while sitting
ddigmd062	9m	Gross motor	Sits on buttocks while legs stretched
ddicmm034	12m	Communication	Babbles while playing
ddicmm036	12m	Communication	Waves 'bye-bye' (M; can ask parents)
ddifmd010	12m	Fine motor	Picks up pellet between thumb and index finger
ddigmd063	12m	Gross motor	Sits in stable position without support
ddigmm064	12m	Gross motor	Crawls forward, abdomen on the floor
ddigmm065	12m	Gross motor	Pulls up to standing position
ddicmm037	15m	Communication	Uses two words with comprehension
ddicmd136	15m	Communication	Reacts to verbal request (M; can ask parents)
ddifmd011	15m	Fine motor	Puts cube in and out of a box

Item	Debut	Domain	Label
ddifmm012	15m	Fine motor	Plays 'give and take' (M; can ask parents)
ddigmm066	15m	Gross motor	Crawls, abdomen off the floor (M; can ask parents)
ddigmm067	15m	Gross motor	Walks while holding onto play-pen or furniture
ddicmm039	18m	Communication	Says three 'words'
ddicmd141	18m	Communication	Identifies two named objects
ddifmd013	18m	Fine motor	Tower of 2 cubes
ddifmm014	18m	Fine motor	Explores environment energetically (M; can ask parents)
ddigmd068	18m	Gross motor	Walks alone
ddigmd069	18m	Gross motor	Throws ball without falling
ddicmm041	24m	Communication	Says sentences with 2 words
ddicmd148	24m	Communication	Understands 'play' orders
ddifmd015	24m	Fine motor	Builds tower of 3 cubes
ddifmm016	24m	Fine motor	Imitates everyday activities (M; can ask parents)
ddigmd070	24m	Gross motor	Squats or bends to pick things up
ddigmd146	24m	Gross motor	Drinks from cup (M; can ask parents)
ddigmd168	24m	Gross motor	Walks well
ddicmm043	30m	Communication	Refers to self using 'me' or 'I' (M; can ask parents)
ddicmd044	30m	Communication	Points at 5 pictures in the book
ddifmd017	30m	Fine motor	Tower of 6 cubes
ddifmd018	30m	Fine motor	Places round block in board
ddifmm019	30m	Fine motor	Takes off shoes and socks (M; can ask parents)
ddifmd154	30m	Fine motor	Eats with spoon without help (M; can ask parents)
ddigmd071	30m	Gross motor	Kicks ball

Table 3.1 shows the 57 milestones from the DDI for ages 0 - 30 months as administered in the SMOCC study. Items are sorted according to *debut*, the age at which the item appears in the DDI. The response to each milestone is either a PASS (1) or a FAIL (0). Children who did not pass a milestone at the debut age were re-measured on that milestone during the next visit. The process continued until the child passed the milestone.

3.2 Probability of passing a milestone given age

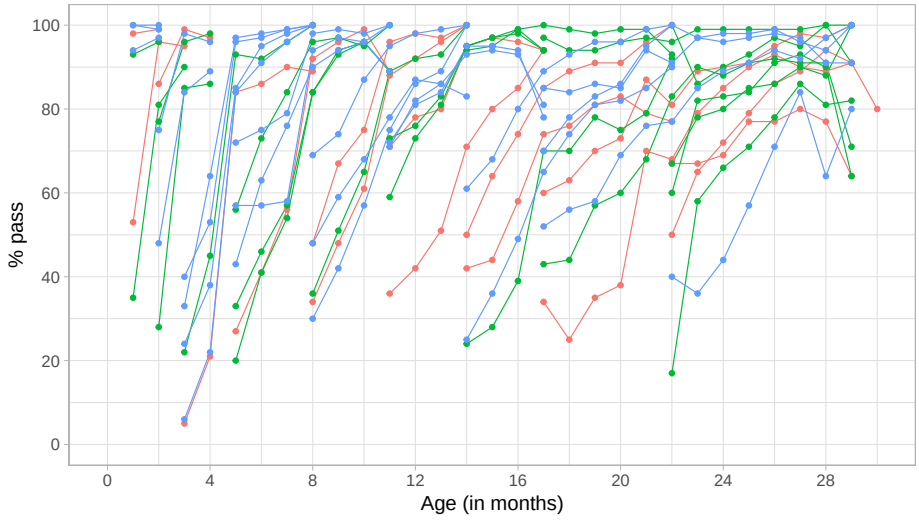

Figure 3.1: Empirical percentage of passing each milestone in the DDI against age (Source: SMOCC data, $n = 2151$, 9 occasions).

Figure 3.1 summarizes the response obtained on each milestone as a curve against age. The percentage of pass scores increases with age for all milestones. Note that curves on the left have steeper slopes than those on the right, thus indicating that development is faster for younger children.

The domain determines the coloured (blue: gross motor, green: fine motor, red: communication). In general, domains are well mixed across age, though around some ages, e.g., at four months, multiple milestones from the same domain appear.

3.3 Probability of passing a milestone given D-score

Figure 3.2 is similar to Figure 3.1, but with the horizontal axis replaced by the D-score. The D-score summarizes development into one number. See 4.3 for a detailed explanation on how to calculate the D-score. The vertical axis with per cent pass is unchanged.

The percentage of successes increases with D-score for all milestones. In contrast to Figure 3.1 all curves have a similar slope, a desirable property needed for an interval scale with a constant unit of measurement (c.f. Section 2.4).

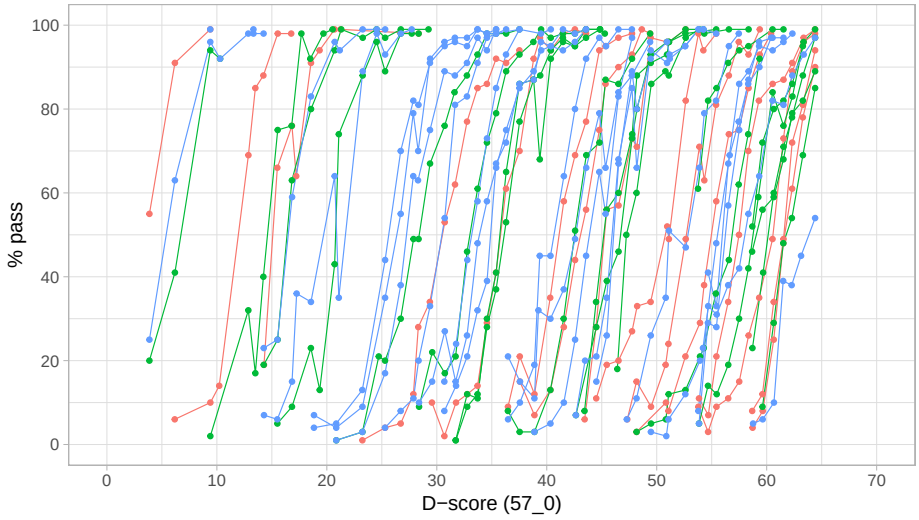

Figure 3.2: Empirical percentage of passing each milestone in the DDI against the D-score (Source: SMOCC data, 2151 children, 9 occasions).

How can the relation between per cent pass and age be so different from the relation between per cent pass and the D-score? The next section explains the reason.

3.4 Relation between age and the D-score

Figure 3.3 shows that the relation between D-score and age is nonlinear. Development in the first year is more rapid than in the second year. During the first year, infants gain about $40D$, whereas in the second year they gain about $20D$. A similar change in growth rate occurs in length (first year: 23 cm, second year: 12 cm, for Dutch children).

Figure 3.4 shows the mutual relations between age, percentage of milestone passing and the D-score. There are three main orientations.

- In the default orientation (age on the horizontal axis, D-score on the vertical axis), we see a curvilinear relation between the age and item difficulty.
- Rotate the graph (age on the horizontal axis, passing percentage on the vertical axis). Observe that this is the same pattern as in Figure 3.1 (with *unequal slopes*). Curves are coloured by domain.
- Rotate the graph (D-score on the horizontal axis, passing percentage on the vertical axis). Observe that this pattern is the same as in Figure 3.2 (with *equal slopes*).

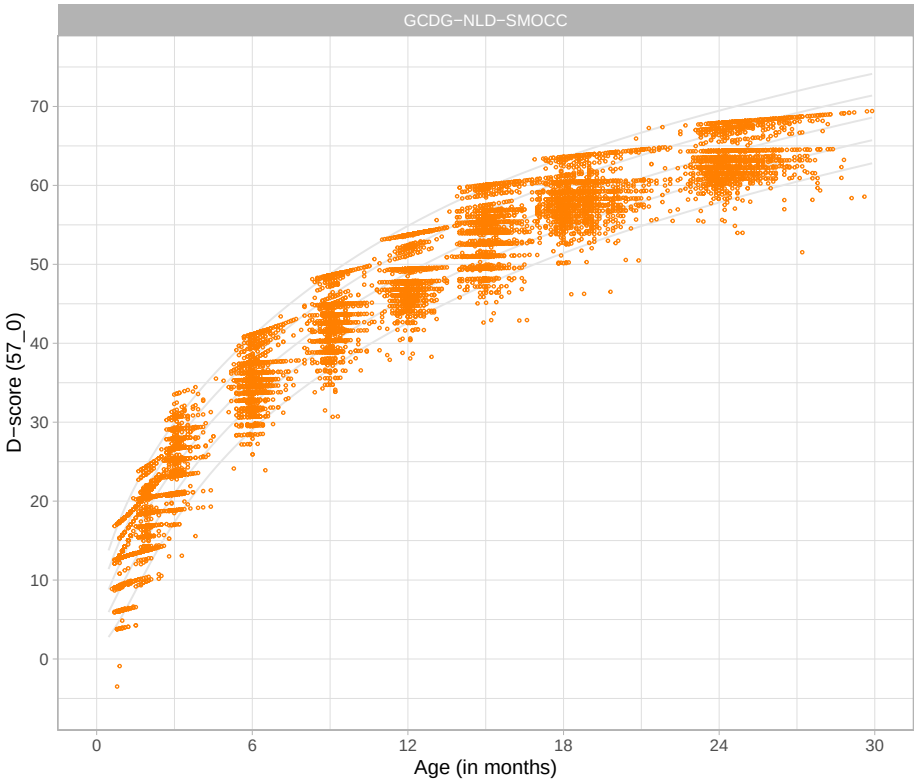

Figure 3.3: Relation between child D-score and child age in a cohort of Dutch children (Source: SMOCC data, $n = 2151$, 9 occasions).

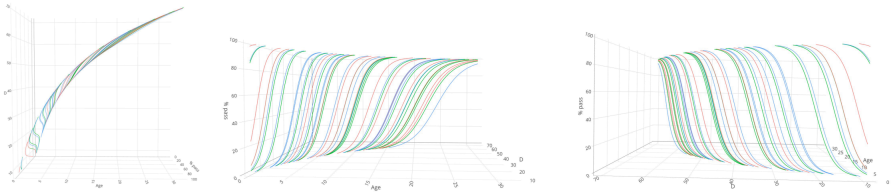

Figure 3.4: 3D-line graph illustrating how the patterns in Figures 3.1 and 3.2 induce the curvature in the relation between D-score and age. The printed version shows three orientations of the relation between age, percent pass and D-score. The online version holds an interactive 3D graph that the reader can actively manipulate the orientation of the graph by click-hold-drag mouse operations.

All patterns can co-exist because of the curvature in the relation between D-score and age. The curvature is never explicitly modelled or defined, but a consequence of the equal-slopes assumption in the relation between the D-score and the passing percentage of a milestone.

3.5 Measurement model for the D-score

3.5.1 What are measurement models?

From section 2.5 we quote:

> The measurement model specifies the relations between the data and the latent variable.

The term *Item Response Theory* (IRT) refers to the scientific theory of measurement models. Good introductory works include Wright and Masters (1982), Embretsen and Reise (2000) and Engelhard Jr. (2013).

IRT models enable quantification of the locations of both *items (milestones) and* persons* on the latent variable. We reserve the term *item* for generic properties, and *milestone* for child development. In general, items are part of the measurement instrument, persons are the objects to be measured.

An IRT model has three major structural components:

- Specification of the underlying *latent variable(s)*. In this work, we restrict ourselves to models with just one latent variable. Multi-dimensional IRT models do have their uses, but they are complicated to fit and not widely used;
- For a given item, a specification of the *probability of success* given a value on the latent variables. This specification can take many forms. Section 3.6 focuses on this in more detail;
- Specification how probability models for the different items should be combined. In this work, we will restrict to models that assume *local independence* of the probabilities. In that case, the probability of passing two items is equal to the product of success probabilities.

3.5.2 Adapt the model? Or adapt the data?

The measurement model induces a predictable pattern in the observed items. We can test this pattern against the observed data. When there is misfit between the expected and observed data, we can follow two strategies:

- Make the measurement model more general;

- Discard items (and sometimes persons) to make the model fit.

These are very different strategies that have led to heated debates among psychometricians. See Engelhard Jr. (2013) for an overview.

In this work, we opt for the - rigorous - Rasch model (Rasch (1960)) and will adapt the data to reduce discrepancies between model and data. Arguments for this choice are given later, in Section 3.8.

3.6 Item response functions

Most measurement models describe the probability of passing an item as a function of the *difference* between the person's ability and the item's difficulty. A person with low ability will almost inevitably fail a heavy item, whereas a highly able person will almost surely pass an easy item.

Let us now introduce a few symbols. We adopt the notation used in Wright and Masters (1982). We use β_n (ability) to refer to the true (but unknown) developmental score of child n. Symbol δ_i (difficulty) is the true (but unknown) difficulty of an item i, and π_{ni} is the probability that child n passes item i. See Appendix A for a complete list.

The difference between the ability of child n and difficulty of item i is

$$\beta_n - \delta_i$$

In the special case that $\beta_n = \delta_i$, the person will have a probability of 0.5 of passing the item.

3.6.1 Logistic model

A widely used method is to express differences on the latent scale in terms of *logistic units* (or *logits*) (Berkson 1944). The reason preferring the logistic over the linear unit is that its output returns a probability value that maps to discrete events. In our case, we can describe the probability of passing an item (milestone) as a function of the difference between β_n and δ_i expressed in logits.

Figure 3.5 shows how the percentage of children that pass the item varies in terms of the ability-difficulty gap $\beta_n - \delta_i$. The gap can vary either by β_n or δ_i so that we may use the graph in two ways:

- To find the probability of passing items with various difficulties for a child with ability β_n. If $\delta_i = \beta_n$ then $\pi_{ni} = 0.5$. If $\delta_i < \beta_n$ then $\pi_{ni} > 0.5$, and if $\delta_i > \beta_n$ then $\pi_{ni} < 0.5$. In words: If the difficulty of the item is

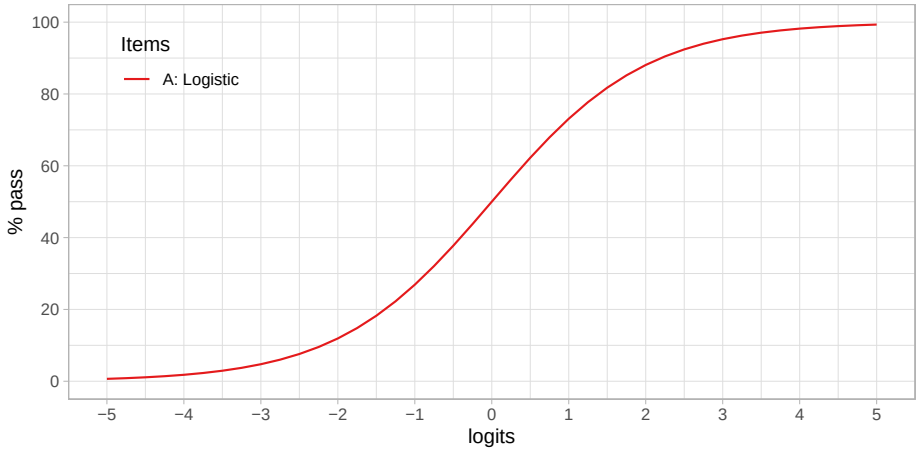

Figure 3.5: Standard logistic curve. Percentage of children passing an item for a given ability-difficulty gap $\beta_n - \delta_i$.

equal to the child's ability, then the child has a 50/50 chance to pass. The child will have a higher than 50/50 chance of passing for items with lower difficulty and have a lower than 50/50 chance of passing for items with difficulties that exceed the child's ability.

- To find the probability of passing a given item δ_i for children that vary in ability. If $\beta_n < \delta_i$ then $\pi_{ni} < 0.5$, and if $\beta_n > \delta_i$ then $\pi_{ni} > 0.5$. In words: Children with abilities lower than the item's difficulty will have lower than 50/50 chance of passing, whereas children with abilities that exceed the item's difficulty will have a higher than 50/50 chance of passing.

Formula (3.1) defines the standard logistic curve:

$$\pi_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \tag{3.1}$$

One way to interpret the formula is as follows. The logarithm of the odds that a person with ability β_n passes an item of difficulty δ_i is equal to the difference $\beta_n - \delta_i$ (Wright and Masters 1982). For example, suppose that the probability that person n passes milestone i is $\pi_{ni} = 0.5$. In that case, the odds of passing is equal to $0.5/(1 - 0.5) = 1$, so $\log(1) = 0$ and thus $\beta_n = \delta_i$. If $\beta_n - \delta_i = \log(2) = 0.693$ person n is *two* times more likely to pass than to fail. Likewise, if the difference is $\beta_n - \delta_i = \log(3) = 1.1$, then person n is *three* more likely to pass. And so on.

3.6.2 Types of item response functions

The standard logistic function is by no means the only option to map the relationship between the latent variable and the probability of passing an item. The logistic function is the dominant choice in IRT, but it is instructive to study some other mappings. The *item response function* maps success probability against ability.

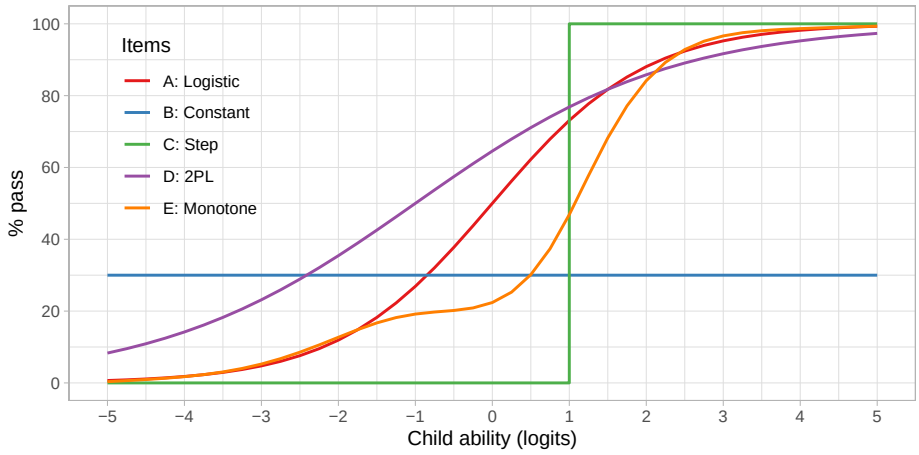

Figure 3.6: Item response functions for five hypothetical items, each demonstrating a positive relation between ability and probability to pass.

Figure 3.6 illustrates several other possibilities. Let us consider five hypothetical items, A-E. Note that the horizontal axis now refers to the ability, instead of the ability-item gap in 3.5.

- A: Item A is the logistic function discussed in Section 3.6.
- B: For item B, the probability of passing is constant at 30 per cent. This 30 per cent is not related to ability. Item B does not measure ability, only adds to the noise, and is of low quality.
- C: Item C is a step function centred at an ability level of 1, so *all* children with an ability below 1 logit fail and *all* children with ability above 1 logit pass. Item C is the ideal item for discriminating children with abilities above and below 1. The item is not sensitive to differences at other ability levels, and often not so realistic in practice.
- D: Like A, item D is a smoothly increasing logistic function, but it has an extra parameter that allows it to vary its slope (or discrimination). The extra parameter can make the curve steeper (more discriminatory) than the red curve, in the limit approaching a step curve. It can also become shallower (less discriminatory) than the red curve (as plotted here), in the limit approaching a constant curve (item B). Thus, item D generalizes items A, B or C.

- E: Item E is even more general in the sense that it need not be logistic, but a general monotonically increasing function. As plotted, the item is insensitive to abilities between -1 and 0 logits, and more sensitive to abilities between 0 to 2 logits.

These are just some examples of how the relationship between the child's ability and passing probability could look. In practice, the curves need not start at 0 per cent or end at 100 per cent. They could also be U-shaped, or have other non-monotonic forms. See Coombs (1964) for a thorough overview of such models. In practice, most models are restricted to shapes A-D.

3.6.3 Person response functions

We can reverse the roles of persons and items. The *person response function* tells us how likely it is that a single person can pass an item, or more commonly, a set of items.

Let us continue with items A, C and D from Figure 3.6, and calculate the response function for three children, respectively with abilities $\beta_1 = -2$, $\beta_2 = 0$ and $\beta_3 = 2$.

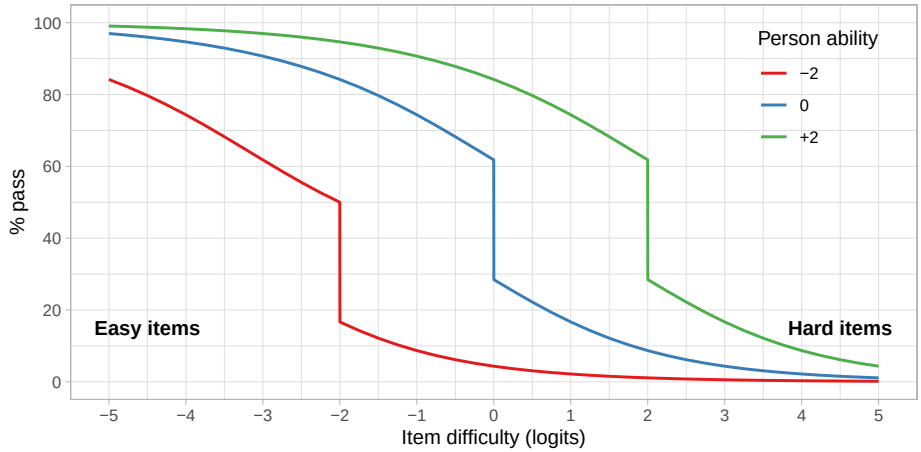

Figure 3.7: Person response functions for three children with abilities -2, 0 and +2, using a small test of items A, C and D.

Figure 3.7 presents the person response functions from three persons with abilities of -2, 0 and +2 logits. We calculate the functions as the average of response probabilities on items A, C and D. Thus, on average, we expect that child 1 logit will pass an easy item of difficulty -3 in about 60 per cent of the time, whereas for an intermediate item of difficulty of -1 the passing probability would be 10 per cent. For child 3, with higher ability, these probabilities are quite different:

97% and 90%. The substantial drop in the middle of the curve is due to the step function of item A.

3.7 Engelhard criteria for invariant measurement

In this work, we strive to achieve *invariant measurement*, a strict form of measurements that is subject to the following requirements (Engelhard Jr. 2013, 14):

1. *Item-invariant measurement of persons*: The measurement of persons must be independent of the particular items used for the measuring.
2. *Non-crossing person response functions*: A more able person must always have a better chance of success on an item that a less able person.
3. *Person-invariant calibration of test items*: The calibration of the items must be independent of the particular persons used for calibration.
4. *Non-crossing item response functions*: Any person must have a better chance of success on an easy item than on a more difficult item.
5. *Unidimensionality*: Items and persons take on values on a *single* latent variable. Under this assumption, the relations between the items are fully explainable by the scores on the latent scale. In practice, the requirement implies that items should measure the same construct. (Hattie 1985)

Three families of IRT models support invariant measurement:

1. Scalogram model (Guttman 1950)
2. Rasch model (Rasch 1960; Andrich 1978; Wright and Masters 1982)
3. Mokken scaling model (Mokken 1971; Molenaar 1997)

The Guttman and Mokken models yield an ordinal latent scale, while the Rasch model yields an interval scale (with a constant unit).

3.8 Why take the Rasch model?

- *Invariant measurement*: The Rasch model meets the five Engelhard criteria (c.f. Section 3.7).
- *Interval scale*: When it fits, the Rasch model provides an interval scale, the de-facto requirement for any numerical comparisons (c.f. Section 2.4.1).
- *Parsimonious*: The Rasch model has one parameter for each item and one parameter for each person. The Rash model one of the most parsimonious IRT models, and can easily be applied to thousands of items and millions of persons.

- *Specific objectivity*: Person and item parameters are mathematically separate entities in the Rasch model. In practice, this means that the estimated difference in ability between two persons does not depend on the difficulty of the test. Also, the estimated differences in difficulties between two items do not depend on the abilities in the calibration sample. The property is especially important in the analysis of combined data, where abilities can vary widely between sources. See Rasch (1977) for derivations and examples.
- *Unified model*: The Rasch model unifies distinct traditions in measurement theory. One may derive the Rasch model from
  - Thorndike's 1904 criteria
  - Guttman scalogram model
  - Ratio-scale counts
  - Raw scores as sufficient statistics
  - Thurstone's scaling requirements
  - Campbell concatenation
  - Rasch's specific objectivity
- *Fits child development data*: Last but not least, as we will see in Section **??**, the Rasch model provides an excellent fit to child development milestones.

Chapter 4

Computation

This section explains the basic computations needed for fitting and evaluating the Rasch model. We distinguish the following steps:

- Identify nature of the problem (4.1)
- Estimation of item parameters (4.2)
- Anchoring (4.2.2)
- Estimation of the D-score (4.3)
- Estimation of age-conditional references (4.4)

Readers not interested in these details may continue to model evaluation in Section **??**.

4.1 Identify nature of the problem

The SMOCC dataset, introduced in Section 3.1.2, contains scores on the DDI of Dutch children aged 0-2 years made during nine visits.

Table 4.1: SMOCC DDI milestones, first three children, 0-2 years.

29	30	31	33	34	36	37	39	41	43	44	16	36	41	48	01	02	03	04	05	07	08	09	10	11	12	13	14	15	16	17	18	19	54	06	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	46	68
1	0														1	0																					1	1			1																	
	1															1																																												
		1									0						1	1	0																0			1	1		1	1	1															
			0								1								1	1	0	1													1				1			1	1	0	1	1												
			1																		1	0	1																					1	1	1	1												
				1	1	1						1											1																							1	0	1													
					1							1												1	1																								0	0								
						1																				0	1																								0	0				
							1																																																			
								1						1															1																														0	1
1	1														1	1																				1	1			1																	
	1															1																																										
		1									0						1	1	0																0			1	0		1	0	0													
			0								1								1	0	0	1													1				1			1	0	1	0	1												
			1																		1	1	1																					1	1	1	0		1									
				1	1																			1	1																							1	1	1	1	1					
								1		1				1															1	1		1		1																					1		1	1
							1																				0	1																														
1	0														1	0																				1	1			1																	
	1															0																																										
		1									0						1	1	0																0			1	0		1	1	0													
			1								1								1	0	0	1													0				1			1	1	1	1	1												
			1																		1	1	1																					1	1	1	0	0											
				1	1																			1																							1	1	0									
						0						0													1	1																									1	1						
							1							1													0	1		1																									1	0	1			1
								1	1	1					1														1	1		1	1	1																						1		1	1

Table 4.1 contains data of three children, measured on nine visits between ages 0 - 2 years. The DDI scores take values 0 (FAIL) and 1 (PASS). In order to save horizontal space, we truncated the column headers to the last two digits of the item names.

Since the selection of milestones depends on age, the dataset contains a large number of empty cells. Naive use of sum scores as a proxy to ability is therefore problematic. An empty cell is not a FAIL, so it is incorrect to impute those cells by zeroes.

Note that some rows contain only 1's, e.g., in row 2. Many computer programs for Rasch analysis routinely remove such *perfect scores* before fitting. However, unless the number of perfect scores is very small, this is not recommended because doing so can severely affect the ability distribution.

In order to effectively handle the missing data and to preserve all persons in the analysis we separate estimation of item difficulties (c.f. section 4.2) and person abilities (c.f. section 4.3).

4.2 Item parameter estimation

4.2.1 Pairwise estimation of item difficulties

There are many methods for estimating the difficulty parameters of the Rasch estimation. See Linacre (2004) for an overview.

We will use the pairwise estimation method. This method writes the probability that child n passes item i but not item j given that the child passed one of them as $\exp(\delta_i)/(\exp(\delta_i) + \exp(\delta_j))$. The method optimizes the pseudo-likelihood of all item pairs over the difficulty estimates by a simple iterative procedure.

Zwinderman (1995) has shown that this procedure provides consistent estimates with similar efficiency computationally more-intensive conditional and marginal maximum likelihood methods.

The beauty of the method is that it is independent of the ability distribution, so there is no need to remove perfect scores. We use the function `rasch.pairwise.itemcluster()` as implemented in the `sirt` package (Robitzsch 2016).

Figure 4.1 summarizes the estimated item difficulty parameters. Although the model makes no distinction between domains, the results have been ordered to ease spotting of the natural progression of the milestones per domain. The figure also suggests that not all domain have equal representation across the scale. For example, there are no communication milestones around the logit of -10.

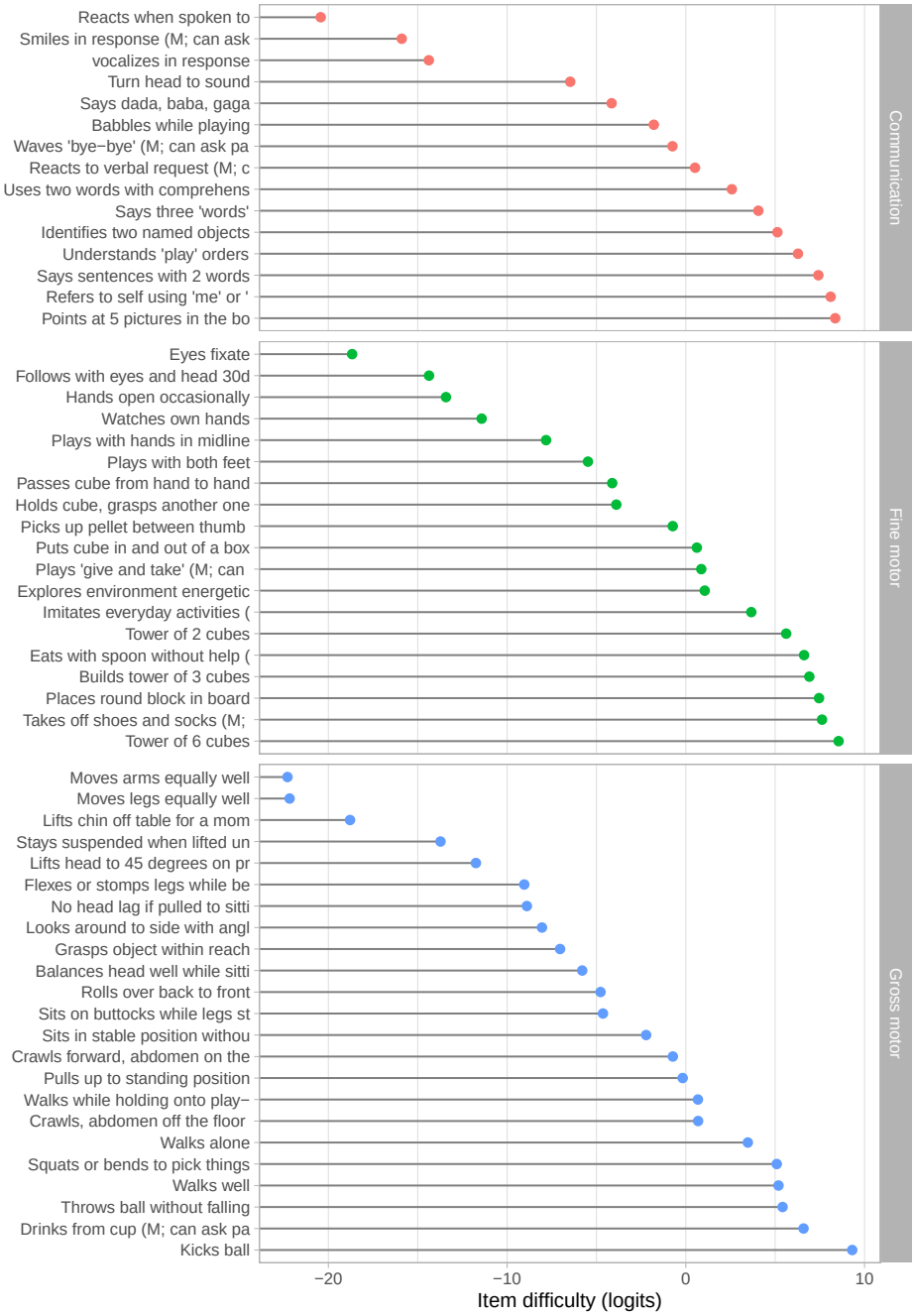

Figure 4.1: Estimated item difficulty parameters (d_i) for 57 milestones of the DDI (0 - 2 years).

Table 4.2: Anchoring values used to identify the D-score scale

Item	Label	Value
`ddigmd057`	Lifts head to 45 degrees on prone position	20
`ddigmd063`	Sits in stable position without support	40

4.2.2 Anchoring

The Rasch model identifies the item difficulties up to a linear transformation. By default, the software produces estimates in the logit scale (c.f. Figure 4.1). The logit scale is inconvenient for two reasons:

- The logit scale has negative values. Negative values do not have a sensible interpretation in child development, and are likely to introduce errors in practice;
- Both the zero in the logit scale, as well as its variance, depend on the sample used to calibrate the item difficulties.

Rescaling preserves the properties of the Rasch model. To make the scale independent of the specified sample, we transform the scale so that two items will always have the same value on the transformed scale. The choice of the two anchor items is essentially arbitrary, but they should correspond to milestones that are easy to measure with small error. In the sequel, we use the two milestones to anchor the D-score scale:

With the choice of Table 4.2, D-score values are approximately $0D$ around birth. At the age of 1 year, the score will around $50D$, so during the first year of life, one D unit corresponds to approximately a one-week interval. Figure 4.2 shows the difficulty estimates in the D-score scale.

4.3 Estimation of the D-score

The second part of the estimation process is to estimate a D-score. The D-score quantifies the development of a child at a given age. Whereas the instrument developer is responsible for the estimation of item parameters, D-score estimation is more of a task for the user. To calculate the D-score, we need the following ingredients:

- Child's PASS/FAIL scores on the milestones administered;

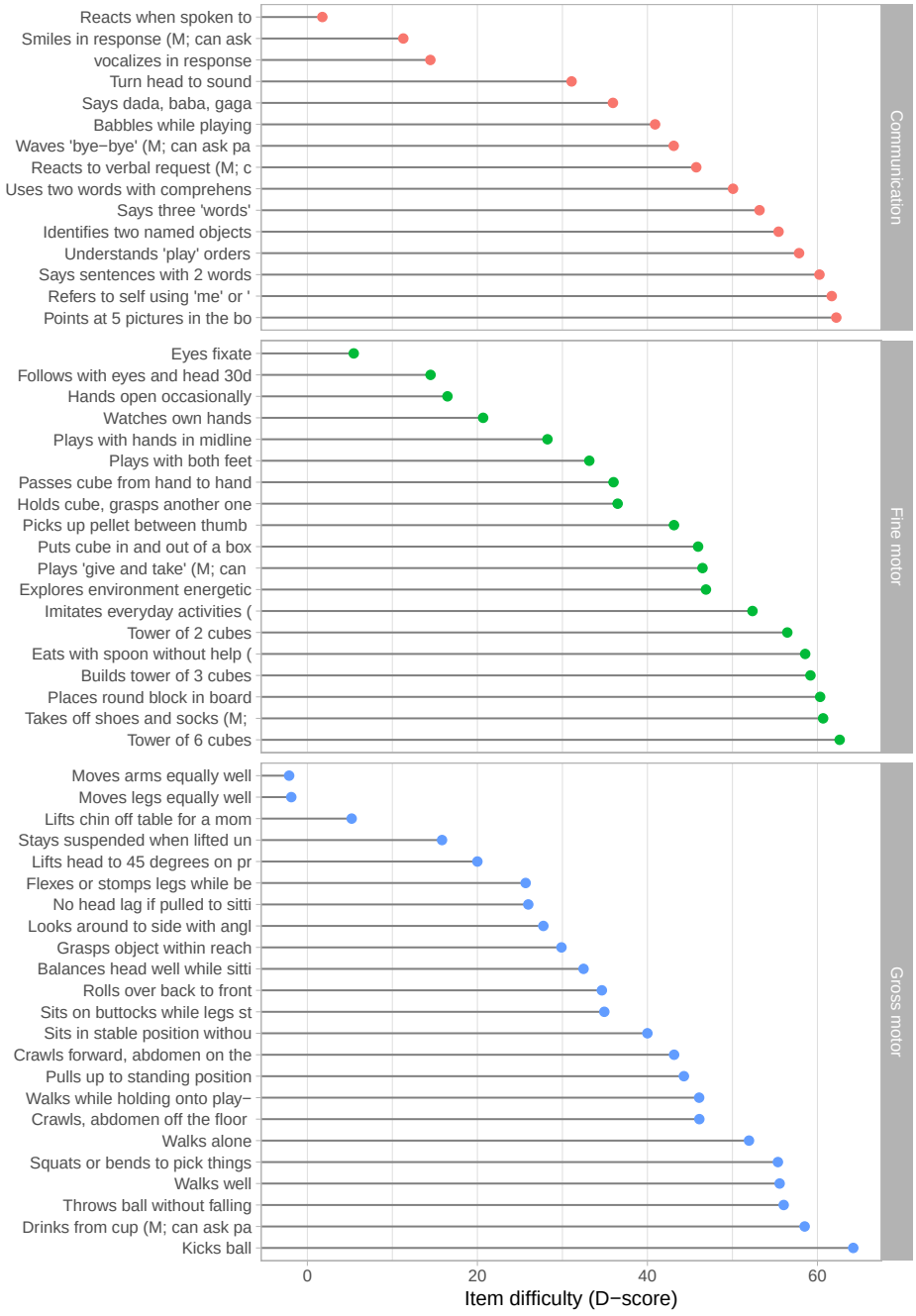

Figure 4.2: Estimated item difficulty parameters (d_i) for 57 milestones of the DDI (0 - 2 years). Milestones `ddigmd057` and `ddigmd063` are anchored at values of $20D$ and $40D$, respectively.

- The difficulty estimates of each milestone administered;
- A prior distribution, an estimate of the D-score distribution before seeing any PASS/FAIL score.

Using these inputs, we may use Bayes theorem to calculate the position of the person on the latent variable.

4.3.1 Role of the starting prior

The first two inputs to the D-score will be self-evident. The third component, the prior distribution, is needed to be able to deal with perfect responses. The prior distribution summarizes our knowledge about the D-score before we see any of the child's PASS/FAIL scores. In general, we like the prior to be non-informative, so that the observed responses and item difficulties entirely determine the value of the D-score. In practice, we cannot use truly non-informative prior because that would leave the D-score for perfect responses (i.e., all PASS or all FAIL) undefined. The choice of the prior is essentially arbitrary, but we can make it in such a way that its impact on the value D-score is negligible, especially for tests where we have more than, say, four items.

Since we know that the D-score depends on age, a logical choice for the prior is to make it dependent on age. In particular, we will define the prior as a normal distribution equal to the expected mean in Figure 3.3 at the child's age, and with a standard deviation that considerably higher than in Figure 3.3. Numerical example: the mean D-score at the age of 15 months is equal to $53.6D$. The standard deviation in Figure 3.3 varies between $2.6D$ and $3.0D$, with an average of $2.9D$. After some experimentation, we found that using a value of $5.0D$ for the prior yields a good compromise between non-informativeness and robustness of D-score estimates for perfect patterns. The resulting starting prior for a child aged 15 months is thus $N(53.6, 5)$.

The reader now probably wonders about a chicken-and-egg problem: To calculate the D-score, we need a prior, and to determine the prior we need the D-score. So how did we calculate the D-scores in Figure 3.3? The answer is that we first took at rougher prior, and calculated two temporary models in succession using the D-scores obtained after solution 1 to inform the prior before solution 2, and so on. It turned out that D-scores in Figure 3.3 hardly changed after two steps, and so there we stopped.

4.3.2 Starting prior: Numerical example

Figure 4.3 illustrates starting distributions (priors) chosen according to the principles set above for the ages of 1, 15 and 24 months. As expected, the assumed ability of an infant aged one month is much lower than that of a child aged 15 months, which in turn is lower than the ability of a toddler aged 24 months.

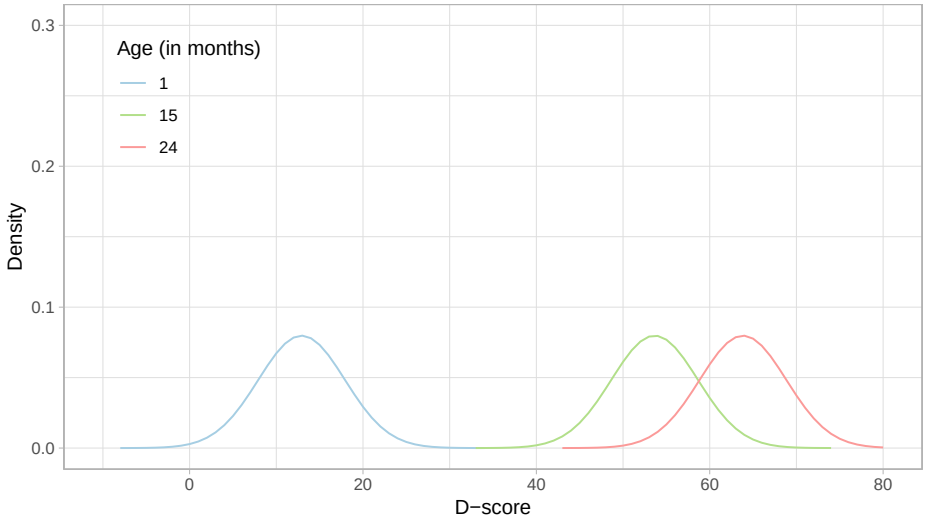

Figure 4.3: Age-dependent starting priors for the D-score at the ages of 1, 15 and 24 months.

The green distribution for 15 months corresponds to the normal distribution $N(53.6, 5)$.

Another choice that we need to make is the grid of points on which we calculate the prior and posterior distributions. Figure 4.3 uses a grid from -10D to +80D, with a step size of 1D. These are fixed *quadrature points*, and there are 91 of them. While these quadrature points are sufficient to estimate D-score for ages up to 2.5 years, it is wise to extend the range for older children with higher D-scores.

4.3.3 EAP algorithm

The algorithm for estimating the D-score is known as the Expected a posteriori (EAP) method, first described by Bock and Mislevy (1982). Calculation of the D-score proceeds item by item. Suppose we have some vague and preliminary idea about the distribution of D, the starting prior (c.f. section 4.3.1), based on age. The procedure uses Bayes rule to update this prior knowledge with data from the first item (using the child's FAIL/PASS score and the estimated item difficulty) to calculate the posterior. The next step uses this posterior as prior before processing the next item, and so on. The procedure stops when the item pool is exhausted. The order in which items enter does not matter for the result. The D-score is equal to the mean of the posterior calculated after the last question.

Table 4.3: Scores of David and Rob on five milestones from the DDI

Item	Label	Delta	David	Rob
ddifmd011	Puts cube in and out of a box	46.0	1	1
ddifmm012	Plays 'give and take' (M; can ask parents)	46.5	1	0
ddicmm037	Uses two words with comprehension	50.1	1	1
ddigmm066	Crawls, abdomen off the floor (M; can ask parents)	46.1	1	1
ddigmm067	Walks while holding onto play-pen or furniture	46.1		0

4.3.4 EAP algorithm: Numerical example

Suppose we measure two boys aged 15 months, David and Rob, by the DDI. David passes the first four milestones but does not complete the test. Rob completes the test but fails on two out of five items.

Table 4.3 shows the difficulty of each milestone (in the column labelled "Delta"), and the responses of David and Rob for the standard five DDI milestones for the age of 15 months.

The mean D-score for Dutch children aged 15 months is $53.6D$, so the milestones are easy to pass at this age, with the most difficult is `ddicmm037`. David passed all milestones but has no score on the last. Rob fails on `ddifmm012` and `ddigmm067`. How do we calculate the D-score for David and Rob?

Figure 4.4 shows how the prior transforms into the posterior after we successively feed the measurements into the calculation. There are five milestones, so the calculation comprises five steps:

1. Both David and Rob pass `ddifmd011`. The prior (light green) is the same as in Figure 4.3. After a PASS, the posterior will be located more to the right, and will often be more peaked. Both happen here, but the change is small. The reason is that a PASS on this milestone is not very informative. For a child with a true D-score of $53D$, the probability of passing `ddifmd011` is equal to 0.966. If passing is so common, there is not much information in the measurement.
2. David passes `ddifmm012`, but Rob does not. Observe that the prior is identical to the posterior of `ddifmd011`. For David, the posterior is only slightly different from the prior, for the same reason as above. For Rob, we find a considerable change to the left, both for location (from $54.3D$ to $47.1D$) and peakedness. This one FAIL lowers Rob's score by $7.2D$.
3. Milestone `ddicmm037` is more difficult than the previous two milestones, so a pass on `ddicmm037` does have a definite effect on the posterior for both David and Rob.

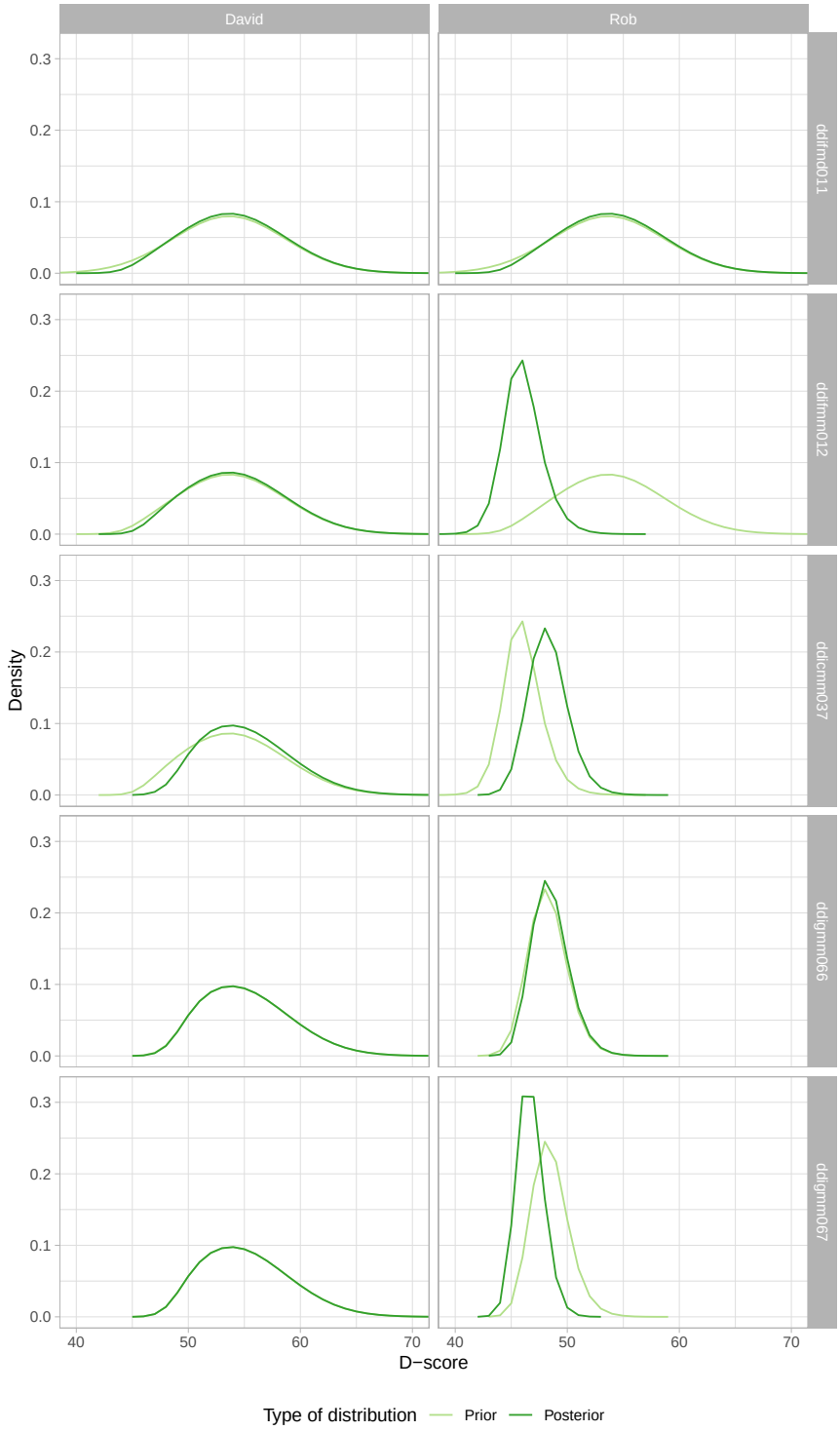

Figure 4.4: D-score distribution for David and Rob before (prior) and after (posterior) a milestone is taken into account.

4. David's PASS on `ddigmm066` does not bring any additional information, so his prior and posterior are virtually indistinguishable. For Rob, we find a slight shift to the right.
5. There is no measurement for David on `ddigmm067`, so the prior and posterior are equivalent. For Rob, we observe a FAIL, which shifts his posterior to the left.

We calculate the D-score as the mean of the posterior. David's D-score is equal to $55.7D$. Note that the measurement error, as estimated from the variance of the posterior, is relatively large. Rob's D-score is equal to $47.7D$, with a much smaller measurement error. This result is consistent with the design principles of the DDI, which is meant to detect children with developmental delay.

The example illustrates that the quality of the D-score depends on two factors, the match between the true (but unknown) D-score of the child and the difficulty of the milestone.

4.3.5 Technical observations on D-score estimation

- Administration of a too easy set of milestones introduces a *ceiling* with children that pass all milestones, but whose true D-score could extend well beyond the maximum. Depending on the goal of the measurement, this may or may not be a problem.
- The specification of the prior and posterior distributions requires a set of quadrature points. The quadrature points are taken here as the static and evenly-spaced set of integers between -10 and +80. Using other quadrature points may affect the estimate, especially if the range of the quadrature points does not cover the entire D-score range.
- The actual calculations are here done item by item. A more efficient method is to handle all responses at once. The result will be the same.

4.4 Age-conditional references

4.4.1 Motivation

The last step involves estimation an age-conditional reference distribution for the D-score. This distribution can be used to construct growth charts that portray the normal variation in development. Also, the references can be used to calculate age-standardized D-scores, called DAZ, that emphasize the location of the measurement in comparison to age peers.

Estimation of reference centiles is reasonably standard. Here we follow van Buuren (2014) to fit age-conditional references of the D-score for boys and girls combined by the LMS method. The LMS method by Cole and Green (1992)

assumes that the outcome has a normal distribution after a Box-Cox transformation. The reference distribution has three parameters, which model respectively the location (M), the spread (S), and the skewness (L) of the distribution. Each of the three parameters can vary smoothly with age.

4.4.2 Estimation of the reference distribution

The parameters are estimated using the BCCG distribution of `gamlss 5.1-3` (Stasinopoulos and Rigby 2007) using cubic splines smoothers. The final solution used a log-transformed age scale and fitted the model with smoothing parameters $\mathrm{df}(M) = 2$, $\mathrm{df}(S) = 2$ and $\mathrm{df}(L) = 1$.

Figure 3.3 plots the D-scores together with five grey lines, corresponding to the centiles -2SD (P2), -1SD (P16), 0SD (P50), +1SD (P84) and +2SD (P98). The area between the -2SD and +2SD lines delineates the D-score expected if development is healthy. 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.

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness).

age	M	S	L
0.0383	8.81	0.3126	1.3917
0.0575	10.59	0.2801	1.4418
0.0767	12.27	0.2526	1.4891
0.0958	13.87	0.2291	1.5331
0.1150	15.39	0.2089	1.5722
0.1342	16.83	0.1916	1.6049
0.1533	18.20	0.1767	1.6304
0.1725	19.50	0.1640	1.6487
0.1916	20.75	0.1531	1.6607
0.2108	21.94	0.1436	1.6676
0.2300	23.07	0.1354	1.6706
0.2491	24.16	0.1283	1.6711
0.2683	25.21	0.1220	1.6698
0.2875	26.21	0.1165	1.6673

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness). *(continued)*

age	M	S	L
0.3066	27.17	0.1117	1.6636
0.3258	28.10	0.1074	1.6589
0.3450	28.99	0.1035	1.6533
0.3641	29.86	0.1001	1.6471
0.3833	30.70	0.0970	1.6403
0.4025	31.50	0.0942	1.6330
0.4216	32.29	0.0917	1.6255
0.4408	33.05	0.0894	1.6178
0.4600	33.79	0.0873	1.6100
0.4791	34.51	0.0854	1.6022
0.4983	35.21	0.0837	1.5946
0.5175	35.89	0.0821	1.5870
0.5366	36.55	0.0807	1.5797
0.5558	37.20	0.0793	1.5725
0.5749	37.83	0.0781	1.5656
0.5941	38.44	0.0770	1.5588
0.6133	39.04	0.0759	1.5523
0.6324	39.63	0.0749	1.5460
0.6516	40.21	0.0740	1.5399
0.6708	40.77	0.0731	1.5340
0.6899	41.32	0.0723	1.5284
0.7091	41.86	0.0715	1.5230
0.7283	42.39	0.0707	1.5178
0.7474	42.91	0.0700	1.5128
0.7666	43.42	0.0693	1.5081

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness). *(continued)*

age	M	S	L
0.7858	43.92	0.0687	1.5036
0.8049	44.40	0.0681	1.4993
0.8241	44.88	0.0674	1.4952
0.8433	45.36	0.0669	1.4913
0.8624	45.82	0.0663	1.4876
0.8816	46.27	0.0657	1.4841
0.9008	46.72	0.0652	1.4809
0.9199	47.16	0.0647	1.4778
0.9391	47.59	0.0642	1.4749
0.9582	48.01	0.0637	1.4723
0.9774	48.43	0.0632	1.4698
0.9966	48.84	0.0627	1.4676
1.0157	49.24	0.0622	1.4655
1.0349	49.64	0.0618	1.4637
1.0541	50.03	0.0613	1.4620
1.0732	50.41	0.0608	1.4605
1.0924	50.79	0.0604	1.4592
1.1116	51.16	0.0600	1.4580
1.1307	51.53	0.0595	1.4570
1.1499	51.89	0.0591	1.4561
1.1691	52.24	0.0587	1.4553
1.1882	52.59	0.0583	1.4547
1.2074	52.94	0.0578	1.4542
1.2266	53.27	0.0574	1.4538
1.2457	53.61	0.0570	1.4535

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness). *(continued)*

age	M	S	L
1.2649	53.94	0.0566	1.4534
1.2841	54.26	0.0562	1.4533
1.3032	54.58	0.0559	1.4533
1.3224	54.89	0.0555	1.4533
1.3415	55.20	0.0551	1.4535
1.3607	55.50	0.0547	1.4537
1.3799	55.81	0.0544	1.4539
1.3990	56.10	0.0540	1.4542
1.4182	56.39	0.0536	1.4546
1.4374	56.68	0.0533	1.4551
1.4565	56.97	0.0530	1.4555
1.4757	57.25	0.0526	1.4561
1.4949	57.52	0.0523	1.4567
1.5140	57.80	0.0520	1.4573
1.5332	58.06	0.0517	1.4580
1.5524	58.33	0.0514	1.4587
1.5715	58.59	0.0510	1.4595
1.5907	58.85	0.0508	1.4603
1.6099	59.11	0.0505	1.4612
1.6290	59.36	0.0502	1.4620
1.6482	59.61	0.0499	1.4630
1.6674	59.86	0.0496	1.4639
1.6865	60.11	0.0494	1.4649
1.7057	60.35	0.0491	1.4660
1.7248	60.59	0.0488	1.4670

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness). *(continued)*

age	M	S	L
1.7440	60.82	0.0486	1.4681
1.7632	61.06	0.0483	1.4692
1.7823	61.29	0.0481	1.4704
1.8015	61.52	0.0478	1.4716
1.8207	61.75	0.0476	1.4728
1.8398	61.97	0.0474	1.4740
1.8590	62.20	0.0471	1.4752
1.8782	62.42	0.0469	1.4765
1.8973	62.64	0.0467	1.4778
1.9165	62.85	0.0465	1.4791
1.9357	63.07	0.0463	1.4805
1.9548	63.28	0.0461	1.4818
1.9740	63.49	0.0459	1.4832
1.9932	63.70	0.0457	1.4846
2.0123	63.91	0.0455	1.4861
2.0315	64.11	0.0453	1.4875
2.0507	64.32	0.0451	1.4890
2.0698	64.52	0.0449	1.4904
2.0890	64.72	0.0447	1.4919
2.1081	64.92	0.0445	1.4934
2.1273	65.11	0.0443	1.4949
2.1465	65.31	0.0441	1.4964
2.1656	65.50	0.0440	1.4979
2.1848	65.70	0.0438	1.4994
2.2040	65.89	0.0436	1.5009

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness). *(continued)*

age	M	S	L
2.2231	66.08	0.0434	1.5024
2.2423	66.26	0.0433	1.5039
2.2615	66.45	0.0431	1.5054
2.2806	66.64	0.0429	1.5069
2.2998	66.82	0.0428	1.5084
2.3190	67.00	0.0426	1.5098
2.3381	67.18	0.0425	1.5113
2.3573	67.36	0.0423	1.5127
2.3765	67.54	0.0421	1.5142
2.3956	67.72	0.0420	1.5156
2.4148	67.89	0.0418	1.5170
2.4339	68.07	0.0417	1.5185
2.4531	68.24	0.0415	1.5199
2.4723	68.41	0.0414	1.5213
2.4914	68.59	0.0412	1.5226
2.5106	68.75	0.0411	1.5240
2.5298	68.92	0.0410	1.5254
2.5489	69.09	0.0408	1.5267
2.5681	69.26	0.0407	1.5281
2.5873	69.42	0.0405	1.5294
2.6064	69.59	0.0404	1.5308
2.6256	69.75	0.0403	1.5321
2.6448	69.91	0.0401	1.5334
2.6639	70.07	0.0400	1.5347
2.6831	70.23	0.0399	1.5360

Table 4.4: Dutch reference values for the D-score: M-curve (median), S-curve (spread) and L-curve (skewness). *(continued)*

age	M	S	L
2.7023	70.39	0.0397	1.5373
2.7214	70.55	0.0396	1.5386
2.7406	70.71	0.0395	1.5398
2.7598	70.86	0.0394	1.5411
2.7789	71.02	0.0392	1.5423

Table 4.4 defines age-conditional references for Dutch children as the M-curve (median), S-curve (spread) and L-curve (skewness) by age. This table can be used to calculate centile lines and Z-scores.

The references are purely cross-sectional and do not account for the correlation structure between ages. For prediction purposes, it is useful to extend the modelling to include velocities and change scores.

4.4.3 Conversion of D to DAZ, and vice versa

Suppose that M_t, S_t and L_t are the parameter values at age t. Cole (1988) shows that the transformation

$$Z = \frac{(D_t/M_t)^{L_t} - 1}{L_t S_t}$$

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

$$Z = \frac{\ln(D_t/M_t)}{S_t}$$

We may derive any required centile curve from Table 4.4. First, choose Z_α as the Z-score that delineates 100α per cent of the distribution, for example, $Z_{0.05} = -1.64$. The D-score that defines the 100α centile is equal to

$$D_t(\alpha) = M_t(1 + L_t S_t Z_\alpha)^{1/L_t}$$

If L_t is close to zero, we use

$$D_t(\alpha) = M_t \exp(S_t Z_\alpha).$$

References

Andrich, D. 1978. "A Rating Formulation for Ordered Response Categories." *Psychometrika* 43: 561–73.

Berk, L. E. 2011. *Child Development. 9th Ed.* Boston, MA: Pearson.

Berkson, Joseph. 1944. "Application of the Logistic Function to Bio-Assay." *Journal of the American Statistical Association* 39 (227): 357–65. http://www.jstor.org/stable/2280041.

Bock, D. D., and R. J. Mislevy. 1982. "Adaptive EAP Estimation of Ability in a Microcomputer Environment." *Applied Psychological Measurement* 6 (4): 431–44.

Cameron, N., and B. Bogin. 2012. *Human Growth and Development.* London: Academic Press.

Cole, T. J. 1988. "Fitting Smoothed Centile Curves to Reference Data (with Discussion)." *Journal of the Royal Statistical Society, Series A* 151: 385–418.

Cole, T. J., and P. J. Green. 1992. "Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood." *Statistics in Medicine* 11 (10): 1305–19.

Coombs, C. H. 1964. *A Theory of Data.* New York: Wiley.

Embretsen, S. E., and S. P. Reise. 2000. *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Engelhard Jr., G. 2013. *Invariant Measurement.* New York: Routledge.

Erikson, E. H. 1963. *Childhood and Society. 2d Ed., Rev. And Enl.* New York, NJ: Norton.

Frankenburg, W. K., J. Dodds, P. Archer, H. Shapiro, and B. Bresnick. 1992. "The Denver II: A Major Revision and Restandardization of the Denver Developmental Screening Test." *Pediatrics* 89 (1): 91–97.

Gesell, A. 1943. *Infant and Child in the Culture of Today.* Los Angeles, CA: Read Book Ltd.

Guttman, L. 1950. "The Basis for Scalogram Snalysis." In *Measurement and Prediction, Vol. IV*, edited by S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen, 60–90. Princeton, NJ: Princeton University Press.

Hattie, J. 1985. "Methodology Review: Assessing Unidimensionality of Tests and Items." *Applied Psychological Measurement* 9 (2): 139–64.

Herngreen, W. P., S. van Buuren, J. C. van Wieringen, J. D. Reerink, S. P. Verloove-Vanhorick, and J. H. Ruys. 1994. "Growth in Length and Weight from Birth to 2 Years of a Representative Sample of Netherlands Children (born in 1988-89) Related to Socio-Economic Status and Other Background Characteristics." *Annals of Human Biology* 21 (5): 449–63.

Jacobusse, G., and S. van Buuren. 2007. "Computerized Adaptive Testing for Measuring Development of Young Children." *Statistics in Medicine* 26 (13): 2629–38. https://stefvanbuuren.name/publication/2007-01-01_jacobusse2007/.

Jacobusse, G., S. van Buuren, and P. H. Verkerk. 2006. "An Interval Scale for Development of Children Aged 0-2 Years." *Statistics in Medicine* 25 (13): 2272–83. https://stefvanbuuren.name/publication/2006-01-01_jacobusse2006/.

Kohlberg, L. 1984. *The Psychology of Moral Development: The Nature and Validity of Moral Stages: Vol. 2.* San Francisco: Harpen & Row.

Liebert, R. M., R. W. Poulos, and G. D. Strauss. 1974. *Developmental Psychology.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Linacre, J. M. 2004. "Rasch Model Estimation: Further Topics." *Journal of Applied Measurement* 5 (1): 95–110.

Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis: With Applications in Political Research.* Berlin: Walter de Gruyter.

Molenaar, I. W. 1997. "Nonparametric Models for Polytomous Responses." In *Handbook of Modern Item Response Theory*, 369–80. Springer.

Piaget, J., and B. Inhelder. 1969. *The Psychology of the Child.* New York, NJ: Basic Books.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

———. 1977. "On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements." *The Danish Yearbook of Philosophy* 14: 58–93. https://www.rasch.org/memo18.htm.

Robitzsch, A. 2016. *Sirt: Supplementary Item Response Theory Models.* https://CRAN.R-project.org/package=sirt.

Salkind, N. J., ed. 2002. *Child Development.* Macmillan Library Reference.

Santrock, J. 2011. *Child Development: An Introduction. 13th Ed.* New York, NJ: McGraw-Hill Higher Education.

Shirley, M. M. 1931. *The First Two Years: A Study of Twenty-Five Babies.* Vol. I: Postural and Locomotor Development. Minneapolis: University of Minnesota Press.

———. 1933. *The First Two Years: A Study of Twenty-Five Babies.* Vol. II: Intellectual Development. Minneapolis: University of Minnesota Press.

Stasinopoulos, D., and R. Rigby. 2007. "Generalized Additive Models for Location Scale and Shape (GAMLSS) in r." *Journal of Statistical Software* 23 (7): 1–46. https://doi.org/10.18637/jss.v023.i07.

Stott, L. H. 1967. *Child Development: An Individual Longitudinal Approach.* New York, NJ: Holt, Rinehart; Winston, Inc.

van Buuren, S. 2014. "Growth Charts of Human Development." *Statistical Methods in Medical Research* 23 (4): 346–68. https://stefvanbuuren.name/publication/2014-01-01_vanbuuren2014gc/.

Wright, B. D., and G. N. Masters. 1982. *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.

Zwinderman, A. H. 1995. "Pairwise Estimation in the Rasch Models." *Applied Psychological Measurement* 19 (4): 369–75.