Child Development with the D-score: Tuning instruments to Unity

Authors: Iris Eekhout & Stef van Buuren

Contents

Tuning instruments to unity

Children learn to walk, speak, and think at an astonishing pace. The D-score captures this process as a one-number summary. The D-score book explains why we need the D-score, how we construct it, and how we calculate it. Application of the D-score enables comparisons in child development across populations, groups and individuals.

We are preparing four D-score book chapters under the following titles:

I. Turning milestones into measurement
II. Tuning instruments to unity
III. Tailoring tests to fit the occasion
IV. Taking off the hood

This chapter is written by Iris Eekhout and Stef van Buuren

Chapter I and Chapter II are currently available as a complete draft. We still need to write most of Chapter III and IV. The series addresses conceptual aspects of the D-score, discusses practical issues, and introduces a dedicated set of `R` packages.

The *Health Birth Growth and Development knowledge integration (ki)* program of the *Bill & Melinda Gates Foundation* kindly supports the work.

If you have any suggestions or comments, please let us know.

About our work

The first 1000 days of human life cover the period between conception and the second birthday. Proper development during this period contributes to future health, happiness, and productivity, so it is essential to track the child's progress during infancy and early childhood. But did you know that more than 150 instruments exist that quantify child development? And are you aware that many of these tools produce not just one, but many scores? Such an overwhelming choice may seem a luxury until you realise that we cannot compare their ratings. Of course, we could settle on just one instrument . . . , but that's never going to happen.

Our work on the D-score explores an alternative strategy—modern data science methods aid in connecting instruments through shared milestones. We present a unified framework that places children and milestones from different tools onto the same scale. As a result, we can measure child development by just one number, the D-score. Separating the scale from the instrument is a revolutionary concept. Application of the D-score enables comparisons in child development across populations, groups and individuals, even when we measure by different tools.

The new "unit for child development" has exciting implications. We may:

- Track child development over time, as in growth charts;
- Construct age-related references for healthy development;
- Adjust the D-score for age;
- Select an instrument that is precise enough for the setting at hand;
- Compare developmental trajectories between children;
- Compare child development between countries;
- Derive concise tools by picking only well-targeted milestones;
- Study the impact of interventions on child development;
- Predict future health from the current D-score.

Our ongoing work addresses conceptual aspects of the D-score, discusses practical issues, and introduces a dedicated set of R packages.

We aim for three audiences:

1. Professionals in child development who wish to familiarise themselves with a new approach to measure child development in early childhood. Separating the tools from the scales allows the professional to select the means most suited for a particular setting. These chapters give professionals the conceptual background of the D-score.

2. Policymakers in international settings who need to weigh the effect of interventions on child development. The existence of different instruments severely hampers their ability to obtain insight into the results of these interventions. The ability to place measurements onto the same scale allows for a more accurate understanding of policy effects, thus supporting the setting of priority levels.

3. Data scientists who can transform a vector of milestone data into a one-number summary with an unambiguous unit. The techniques have a solid psychometric backing, and also apply to other types of problems. These chapters explain this conversion process in detail, thereby opening up the way for the application of precise analytic techniques in many different settings.

Additionally, parents are always eager to follow every step of their child. While we do not target this work to parents, our methodology may spark the interest of authors, app writers, and instrument creators that do address the interests and needs of parents. Hence, the publication of these chapters may have additional societal impact.

About the Authors

Stef van Buuren is a professor of Statistical Analysis of Incomplete Data at the University of Utrecht and statistician at the Netherlands Organisation for Applied Scientific Research TNO in Leiden. His interests include the analysis of incomplete data and child growth and development (h-index 61). Van Buuren is the inventor of the MICE algorithm for multiple imputation of missing data (>85.000 downloads per month) and has written the accessible monograph Flexible Imputation of Missing Data. Second Edition, CRC/Chapman & Hall. He designed the growth charts for the Dutch child health care system and invented the D-score, a new method for expressing child development on a quantitative scale. He consults for the World Health Organization and the Bill & Melinda Gates Foundation. More background at https://stefvanbuuren.name, and software at https://github.com/stefvanbuuren.

Iris Eekhout holds a double masters in clinical psychology and methodology and statistics of psychology (Leiden University). She obtained her PhD at the Department of Epidemiology and Biostatistics of the VU University medical centre in Amsterdam. Her dissertation work resulted in novel ways of dealing with missing data in questionnaire items and total scores. Currently, Iris teaches a course on missing data analysis in the epidemiology master's program at VU University medical centre. At TNO, Iris works on a variety of projects as a methodologist and statistical analyst related to child health, e.g., measuring child development (D-score) and adaptive screenings for psycho-social problems (psycat). More background at https://www.iriseekhout.com, and software at https://github.com/iriseekhout.

Chapter 1

Introduction

This introductory section

- briefly summarises our previous work on the D-score (1.1)
- introduces the main topic of the chapter (1.2)
- highlights the relevance of work (1.3)
- explains why we have written this chapter (1.4)
- delineates the intended audience (1.5)

1.1 Previous work on the D-score

Chapter I - Turning milestones into measurement - highlights the concepts and tools needed to obtain a quantitative score from a set of developmental milestones.

In practice, we typically want to make the following types of comparisons:

- Compare development within the same child over time;
- Compare the development of two children of the same age;
- Compare the development of two children of different ages;
- Compare the development of groups of children of different ages.

To do this well, we need an *interval scale with a fixed unit of development*. We argued that the simple Rasch model is a very suitable candidate to provide us with such a unit. The Rasch model is simple, fast, and we found that it fits child developmental data very well (Jacobusse, van Buuren, and Verkerk 2006)(van Buuren 2014). The Rasch model has a long history, but -unfortunately- it is almost unknown outside the field of psychometrics. We highlighted the concepts of the model that are of direct relevance to child development. Using data collected by the Dutch Development Instrument, we demonstrated that the model and its estimates behave as intended for children in the open population, for prematurely born-children, and children living in a low- and middle-income country.

As our approach breaks with the traditional paradigm that emphasises different domains of child development, we expected a slow uphill battle for acceptance. We have now gained the interest from various prominent authors in the field, and from organisations who recognise the value of a one-number-summary for child development. In analogy to traditional growth charts, it is entirely possible to track children, or groups of children, on a developmental chart over time. Those and other applications of the technology may eventually win over some more souls.

1.2 What this volume is about

It is straightforward to apply the D-score methodology, as explained in Chapter I: Turning milestones into measurement, for measurements observed by one instrument. In practice, however, there is a complication. We often need to deal with multiple, partially overlapping tools. For example, our data may contain

- different versions of the same instrument (e.g., Bayley I, II and III);
- different language versions of the same tool;
- different tools administered to the same sample;
- different tools administered to different samples;
- and so on.

Since there are over 150 different instruments to measure child development (L. C. H. Fernald et al. 2017), the chances are high that our data also hold data observed by multiple tools.

It is not apparent how to obtain comparable scores from different instruments. Tools may have idiosyncratic instructions to calculate total scores, distinctive domain definitions, unique compositions of norm groups, different floors and ceilings, or combinations of these.

This chapter addresses the problem *how to define and calculate the D-score based on data coming from multiple sources, using various instruments administered at varying ages*. We explain techniques that systematically exploit the overlap between tools to create comparable scores. For example, many instruments have variations on milestones like *Can stack two blocks*, *Can stand* or *Says baba*. By carefully mapping out the similarities between instruments, we can construct a constrained measurement model informed by subject matter knowledge. As a result, we can map different instruments onto the same scale.

Many of the techniques are well known within psychometrics and educational research. This chapter translates the concepts to the field of child development.

1.3 Relevance of the work

We all like our children to grow and prosper. The *first 1000 days* refers to the time needed for a child to grow from conception to its second birthday. During this period, the architecture of the developing brain is very open to the influence of relationships and experiences. It is a time of rapid change that lays the groundwork for later health and happiness.

Professionals and parents consider it necessary to monitor children's development. While we can track the child's physical growth by growth charts to identify children with signs of potential delay, there are no charts for monitoring child development. To create such charts, we need to have a unit of development, similar to units like centimetres or kilograms.

The D-score is a way to define a unit of child development. With the D-score, we see that progress is much faster during infancy, and that different children develop at different rates. The D-score also allows us to define a "normal" range that we can use to filter out those who are following a more pathological course. There is good evidence that early identification and early intervention improve the outcomes of children (Britto et al. 2017). Early intervention is crucial for children with developmental disabilities because barriers to healthy development early in life impede progress at each subsequent stage.

Monitoring child development provides caregivers and parents with reliable information about the child and an opportunity to intervene at an early age. Understanding the developmental health of populations of children allows organisations and policymakers to make informed decisions about programmes that support children's greatest needs (Bellman, Byrne, and Sege 2013).

1.4 Why this chapter?

We believe that *there can be one scale* for measuring child development and that this scale is useful for many applications. We also believe that *there cannot be one instrument* for measuring child development that is suitable for all situations. In general, the tool needs tailoring to the setting.

We see that practitioners often view instruments and scales as exchangeable. In daily practice, the practitioner would pick a particular tool to measure a specific faculty, which then effectively produces a "scale score." Each tool produces its own score, which then feeds into the diagnostic and monitoring process.

We have always found it difficult to explain that scales and instruments are different things. For us, a scale is a continuous concept, like "distance," "temperature" or "child development," and the instrument is the way to assign values to the particular object being measured. For measuring distance, we use devices like rods, tapes, sonar, radar, geo-location, or red-shift detection, and we can express the results as the location under the underlying scale (e.g., number of meters). It would undoubtedly be an advance if we could establish a *unit of child development*, and express the measurement as the number of units. If we succeed, we can compare child development scores, that are measured through different devices. This chapter explores the theory and practice for making that happen.

1.5 Intended audience

We aim for three broad audiences:

- Professionals in the field of child growth and development;
- Policymakers in international settings;
- Statisticians, methodologists, and data scientists.

Professionals in child development are constantly faced with the problem that different instruments for measuring child development yield incomparable scores. This chapter introduces and illustrates sound psychometric techniques *for extracting comparable scores from existing instruments.* We hope that our approach will ease communication between professionals.

Policymakers in international settings are looking for simple, versatile, and cheap instruments to gain insight into the effectiveness of interventions. The ability to measure child development by a single number *enhances priority setting and leads to a more accurate understanding of policy effects.*

The text may appeal to statisticians and data scientists for *the simplicity of the concepts, for the (somewhat unusual) application of statistical models to discard data, for the ease of interpretation of the result, and for the availability of software.*

Chapter 2

Data

This chapter explains the methodology for obtaining a comparable developmental score (D-score) from different instruments. This section introduces the data that will illustrate our approach. The data originates from a study by the Global Child Development Group (GCDG), that brought together longitudinal measurement on child development data from 16 cohorts worldwide.

- Overview of cohorts and instrument (2.1)
- Cohort descriptions (2.2)
- Instruments (2.3)

2.1 Overview of cohorts and instruments

The Global Child Development Group (GCDG) collected longitudinal data from 16 cohorts. The objective of the study was to develop a population-based measure to monitor early child development across ages and countries. The requirements for inclusion were

1. direct assessment of child development;
2. availability of individual milestone scores;
3. spanning ages between 0-5 years;
4. availability of follow-up measures, at ages 5-10 years.

The effort resulted in a database containing individual data from over 16,000 children from 11 countries. The world map below (Figure: 2.1) colors the countries included in the study. Section 2.2 briefly describes each cohort. Section 2.3 reviews the measurement instruments.

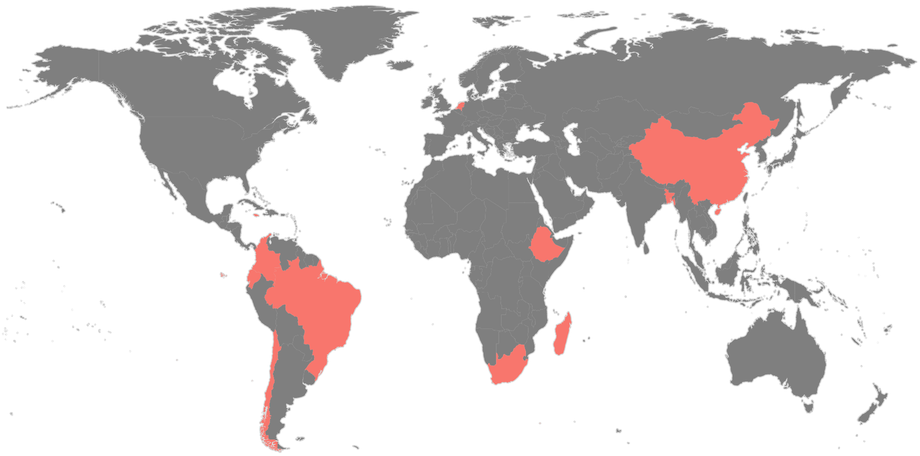

Figure 2.1: Coverage of countries included in the study.

The GCDG data comprises of birth cohorts, impact evaluation studies and instrument evaluation studies. Table 2.1 displays a brief overview of the instruments used in each sub-study.

Table 2.1: Overview of instruments administered in the cohorts.

Cohort	`by`	`den`	`gri`	`bat`	`vin`	`ddi`	`bar`	`tep`	`aqi`	`sbi`
Bangladesh	x									
Brazil 1		x								
Brazil 2				x						
Chile 1	x									
Chile 2				x				x		
China	x									
Colombia 1	x									
Colombia 2	x	x		x					x	
Ecuador							x			
Ethiopia	x									
Jamaica 1			x							
Jamaica 2			x							
Madagascar										x
Netherlands1						x				
Netherlands2						x				
South Africa	x		x		x					

2.2 Cohort descriptions

The cohorts have different designs, age ranges and assessment instruments. Figure 2.2 displays the age range of developmental assessments per cohort, coloured according to the instruments.

A brief description of each cohort follows:

The **Bangladesh** study (GCDG-BGD-7MO) was an impact evaluation study including 1862 children around the age of 18 months. The Bayley Scale for Infant and Toddler Development-II (`by2`) was administered and long-term follow-up data were available for the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) at 5 years (Tofail et al. 2008).

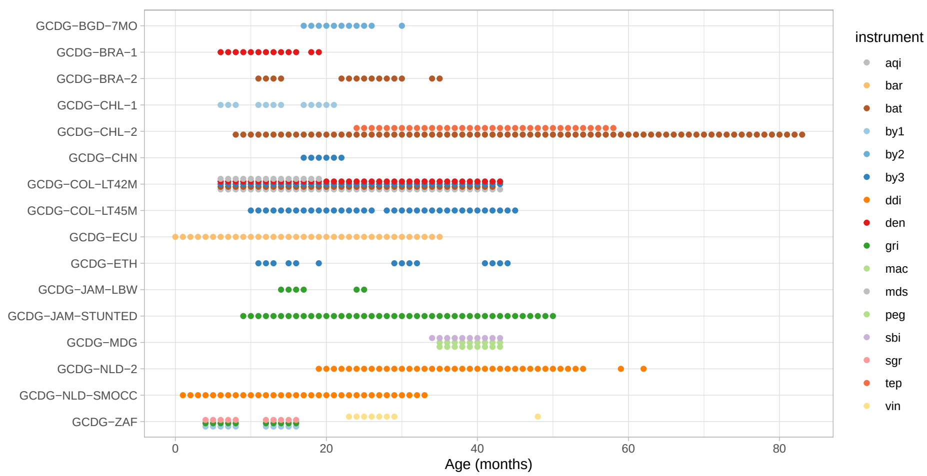

Figure 2.2: Age range and assessment instrument of included data for each GCDG cohort

The **Brazil 2** study (GCDG-BRA-2) was a birth-cohort with measurements of 3907 children at 12 months and 3869 children at 24 months. Both occasions collected data on the Battelle Development Inventory (`bat`) (Moura et al. 2010).

The **Chile 1** study (GCDG-CHL-1) was an impact evaluation study of 128 children assessed at 6 months, 1732 children at 12 months and 279 at 18 months. The `by1` was administered at each of the three waves. Long-term follow-up data were available for the WPPSI at 5-6 years (Lozoff et al. 2003).

The **Chile 2** study (GCDG-CHL-2) consists of a birth-cohort of 4869 children. The investigators measured child development by the Battelle Developmental Inventory (`bat`) at 7-23 months. A total of 9201 children aged 24-58 responded to the Test de Desarrollo Psicomotor (`tep`) at 24-58 months. For the latter group, follow-up data were available for the Peabody Picture Vocabulary Test (PPVT) at 5-6 years (Conteras and González 2015).

The **China** study (GCDG-CHN) was an impact evaluation study that contained 990 children assessed with the by3 at 18 months (Lozoff et al. 2016).

The **Colombia 1** study (GCDG-COL-LT45M) was an impact evaluation study that comprised two waves. Wave 1 contained 704 children at 12-24 months and wave 2 631 children at 24-41 months. The `by3` was administered at each wave. Long-term follow-up data were available for PPVT at 4-6 years (Attanasio et al. 2014).

The **Colombia 2** study (GCDG-COL-LT42M) was an instrument validation study where all 1311 children aged 6-42 months were measured the `by3`. Also, there are data for a subgroup of 658 children on `den`, the Ages and Stages Questionnaire (`aqi`), and the `bat` screener. Long-term follow-up data were available for the Fifth Wechsler Intelligence Scale for Children (WISC-V) and the PPVT (Rubio-Codina et al. 2016).

An impact evaluation study The **Jamaica 1** study (GCDG-JAM-LBW) was an impact evaluation study that collected data on the Griffiths Mental Development Scales (`gri`) for 225 children aged 15 months (first wave), and 218 children of aged 24 months (second wave). Long-term follow-up data were available for WPPSI and PPVT at 6 years (Walker et al. 2004).

The **Jamaica 2** study (GCDG-JAM-STUNTED) was an impact evaluation study with data on the `gri` for 159 children at 9-24 months, 21-36 months, and at 33-48 months. Long-term follow-up data were available for `sbi`, Raven's Coloured Progressive Matrices (Ravens), and PPVT at 7-8 years and the WAIS at 17-18 years (Grantham-McGregor et al. 1991).

The **Madagascar** study (GCDG-MDG) was an impact evaluation study that used the `sbi` for 205 children aged 34-42 months. Long-term follow-up data were available for `sbi` and PPVT at 7-11 years (Lia C. H. Fernald et al. 2011).

The **Netherlands 1** study (GCDG-NLD-SMOCC) was an instrument validation study with a total of 9 waves. At each wave the Dutch Developmental instrument (`ddi`) (In the Netherlands known as Van Wiechenschema) was administered. The first wave included 1985 children at 1 month, wave 2 1807 children at 2 months, wave 3 1963 children at 3 months, wave 4 1919 children at 6 months, wave 5 1881 children at 9 months, wave 6 1802 children at 12 months, wave 7 1776 children at 15 months, wave 8 1787 children at 18 months, and wave 9 1815 children at 24 months (Herngreen et al. 1992).

The **Netherlands 2** study (GCDG-NLD-2) was an instrument validation study with a total of five waves. This study resembles GCDG-NLD-SMOCC but for older children. Wave 1 included 1016 children at 24 months, wave 2 995 children at 30 months, wave 3 1592 children at 36 months, wave 4 1592 children at 42 months, and wave 5 1024 children at 48 months (Doove 2010).

The **South Africa** study (GCDG-ZAF) was a birth cohort with four waves. The first wave included 485 children and second wave 275 children, who were assessed at 6 and 12 months, respectively, with the `by1` and the `gri`. The third wave included 1802 children and the fourth wave 1614 children, assessed at 24 and 48 months, respectively, with the Vineland Social Maturity Scale (`vin`) (Richter et al. 2007).

2.3 Instruments

The **Bayley Scales for Infant and Toddler Development (`by1`,`by2`, `by3`)** aim to assess infants and toddlers, aged 1-42 months. The current version is the `by3`, but some GCDG cohorts used earlier versions (i.e. `by1` and `by2`) (Bayley 1969)(Bayley 1993)(Bayley 2006). The 326 items of the `by3` measure three domains: Cognitive items, Motor items (with fine and gross motor items), and Language items (with expressive and receptive items). The `by2` contains 277 items and has two additional subscales: Social-Emotional and Adaptive Behavior. by1 contains 229 items.

The **Denver Developmental Screening Test (`den`)** is aimed to identify developmental problems in children up to age six. The 125 dichotomous test items are distributed over the age range from birth to six years. The Denver covers four domains: personal-social, fine motor and adaptive, language, and gross motor. The test items are all directly observed by an examiner and are not dependent on parent report (Frankenburg et al. 1992) (Frankenburg et al. 1990).

The **Griffiths Mental Development Scales (`gri`)** measure the rate of development in infants and young children in six developmental areas: locomotor, personal-social, hearing and language, eye and hand coordination, performance and practical reasoning (Griffiths 1967).

The **Battelle Developmental Inventory (`bat`)** measures key developmental skills in children from birth to 7 years, 11 months. The instrument contains 450 items distributed over five domains: adaptive, personal-social, communication, motor, and cognitive (Newborg 2005).

The **Vineland Social Maturity Scale (`vin`)** is a test to assess social competence. The instrument contains eight subscales that measure communication skills, general self-help ability, locomotion skills, occupation skills, self-direction, self-help eating, self-help dressing and socialisation skills (Doll 1953).

The **Dutch Developmental Instrument (`ddi`)** measures early child development during the ages 0-4 years. The instrument consists of 75 milestones spread over three domains: fine motor, adaptive, personal and social behaviour; communication; and gross motor (Schlesinger-Was 1981).

The **Barrera Moncada (`bar`)** is a Spanish instrument that measures the growth and psychological development of children (Barrera Moncada 1981).

The **Test de Desarrollo Psicomotor (`tep`)** is an instrument to evaluate toddlers aged 2 to 5 years on their development. The items come from three sub-tests: 16 items assess coordination; 24 items measure language skills and 12 items tap into motor skills (Haeussler and Marchant 1999).

The **Ages and Stages Questionnaire (`aqi`)** measures developmental progress in children aged 2 mo – 5.5 yrs. The instrument distinguishes development in five areas: personal-social, gross motor, fine motor, problem solving, and communication. The caregiver completes 30 items per age intervals and (Squires and Bricker 2009).

The **Stanford Binet Intelligence Scales (`sbi`)** is a cognitive ability and intelligence test to diagnose developmental deficiencies in young children. The items divide into five subtests: fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory (Roid 2003)(Hagen and Stattler 1986).

Chapter 3

Comparability

This section describes challenges and methodologies to harmonize child development measurements obtained by different instruments:

- Are instruments connected? (3.1)
- Bridging instruments by mapping items (3.2)
- Overview of promising item mappings (3.3)

3.1 Are instruments connected?

The ultimate goal is to compare child development across populations and cultures. A complication is that measurements are made by different instruments. To do deal with this issue, we harmonize the data included in the GCDG cohorts. In particular, we process the milestone responses such that the following requirements hold:

- Every milestone in an instrument has a unique name and a descriptive label;
- Every milestone occupies one column in the dataset;
- Item scores are (re)coded as: 1 = PASS; 0 = FAIL;
- Items not administered or not answered are a missing value;
- Every row in the dataset corresponds to a unique cohort-child-age combination.

Cohorts and milestones need to be *connected*. There are several ways to connect cohorts:

- Two cohorts are directly connected if they use the same instrument;

Table 3.1: Linkage pattern indicating combinations of cohorts and instruments.

	aqi	bar	bat	by1	by2	by3	ddi	den	gri	mac	peg	sbi	sgr	tep	vi
Bangladesh					X										
Brazil 1								X							
Brazil 2			X												
Chile 1				X											
Chile 2			X											X	
China						X									
Colombia 1						X									
Colombia 2	X		X			X		X							
Ecuador		X													
Ethiopia						X									
Jamaica 1									X						
Jamaica 2									X						
Madagascar										X	X	X			
Netherlands1							X								
Netherlands2							X								
South Africa					X				X				X		X

- Two cohorts are indirectly connected if both connect to a third cohort that connects them.

Likewise, instruments can be connected:

- Two instruments are directly connected if the same cohort measures both;
- Two instruments are indirectly connected if both connect to a third instrument that connects them.

An `X` in Table 3.1 identifies which cohorts use which instruments. The linkage table shows that studies from China, Colombia, and Ethiopia are directly connected (by `by3`). Brazil 1 indirectly connects to these studies through `den`. Some cohorts (e.g., Chile 1 and Ecuador) do not link to any other study. Likewise, we might say that `aqi`, `bat`, `by3`, and `den` are directly connected. Note that no indirect connections exist to this instrument group.

Table 3.1 is a somewhat simplified version of the linkage pattern. As we saw in section 2.2, there are substantial age differences between the cohorts. The linked instrument linkage table shows the counts of the number of registered scores per age group. What appears in Table 3.1 as one test may comprise of two disjoint subsets, and hence some cohorts may not be connected after all.

Table 3.2: Example of similar items from different instruments.

Item	Label
`by1mdd136`	sentence of 2 words
`by2mdd114`	Uses a two-word utterance
`ddicmm041`	Says sentences with 2 words
`denlgd019`	Combine Words
`grihsd217`	Uses word combinations
`vinxxc016`	use a short sentence

Connectedness is a necessary - though not sufficient - requirement for parameter identification. If two cohorts are not connected, we cannot distinguish between the following two alternative explanations:

- Any differences between studies can be attributed to the ability of the children;
- Any differences between studies can be attributed to the difficulties of the instruments.

The data do not contain the necessary information to discriminate between these two explanations. Since many cohorts in Table 3.1 are unconnected, it seems that we are stuck.

The next section suggests a way out of the dilemma.

3.2 Bridging instruments by mapping items

Many instruments for measuring child development have appeared since the works of Shirley (1933) and Gesell (1943). It is no surprise that their contents show substantial overlap. All tools assess events like starting to see, hear, smile, fetch, crawl, walk, speak, and think. We will exploit this overlap to bridge different instruments. For example, Table 3.2 displays the labels of milestones from six instruments. All items probe the ability of the child to formulate "sentences" of two words.

The idea is to check whether these milestones measure development in the same way. If this is found to be true, then we may formally restrict the difficulty levels of these milestones to be identical. This restriction provides a formal bridge between the instruments. We repeat the process for all groups of similar-looking items.

A first step in the bridging process is to group items from different instruments by similarity. As the `by3` is relatively long and is the most often used instrument, it provides a convenient starting point. Subject matter experts experienced in child development mapped items from other tools to `by3` items. These experts evaluated the similarity of wordings and descriptions in reference manuals. Also, they mapped same-skill items across other instruments into groups if these did not map onto `by3` items.

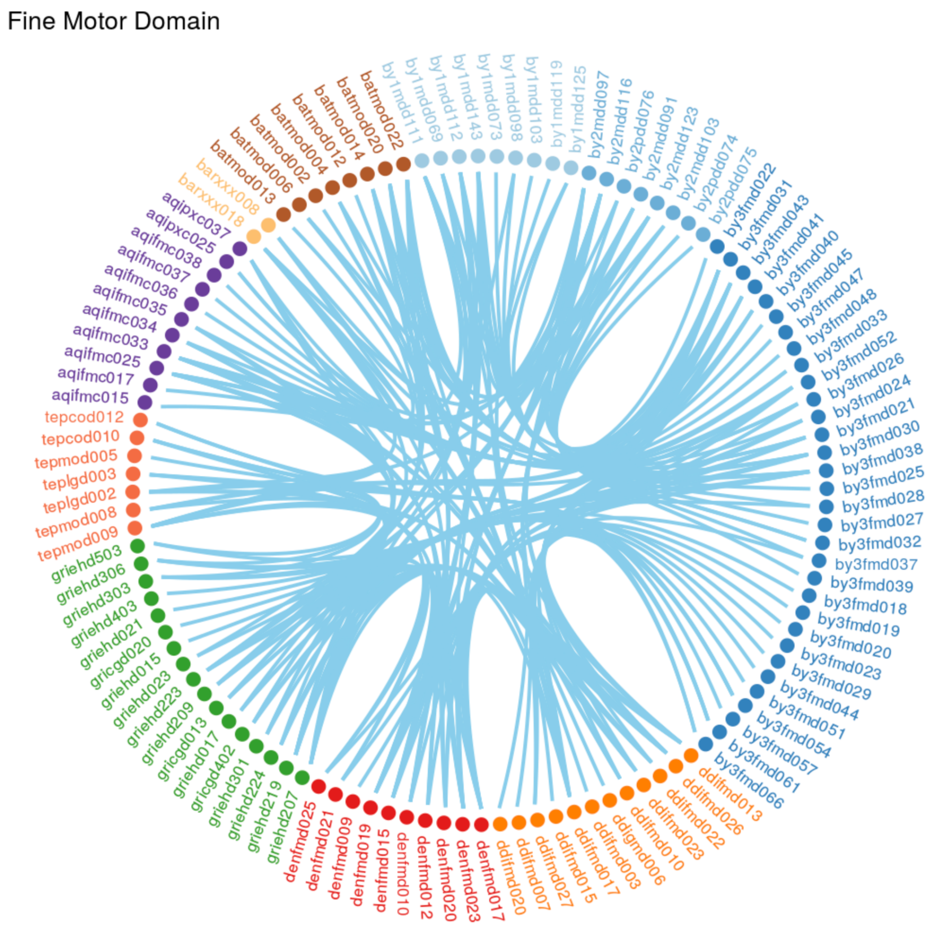

Figure 3.1: Connections between the instruments via mapped item groups by domain (https://tnochildhealthstatistics.shinyapps.io/GCDG_mapping/).

Figure 3.1 connects similar items and hence visualises connections between instruments. Items are displayed in the wheel, coloured by instrument. We organised item mappings into five domains: fine motor (FM), gross motor (GM), cognitive (COG), receptive (REC), and expressive (EXP). The `Prev` and `Next` buttons allow us to visit other domains.

3.3 Age profile of item mappings

Another way to explore the similarity of milestones from different instruments is to plot the probability of passing by age. Figure 3.2 shows two examples. The first graph presents the age curves of a group of four cognitive items for assessing the ability to put a cube or block in a cup or box. The milestones are administered in different studies and seem to work similarly. The second plot shows a similar graph for items that assess the ability to build a tower of six cubes or blocks. These milestones have similar age patterns as well.

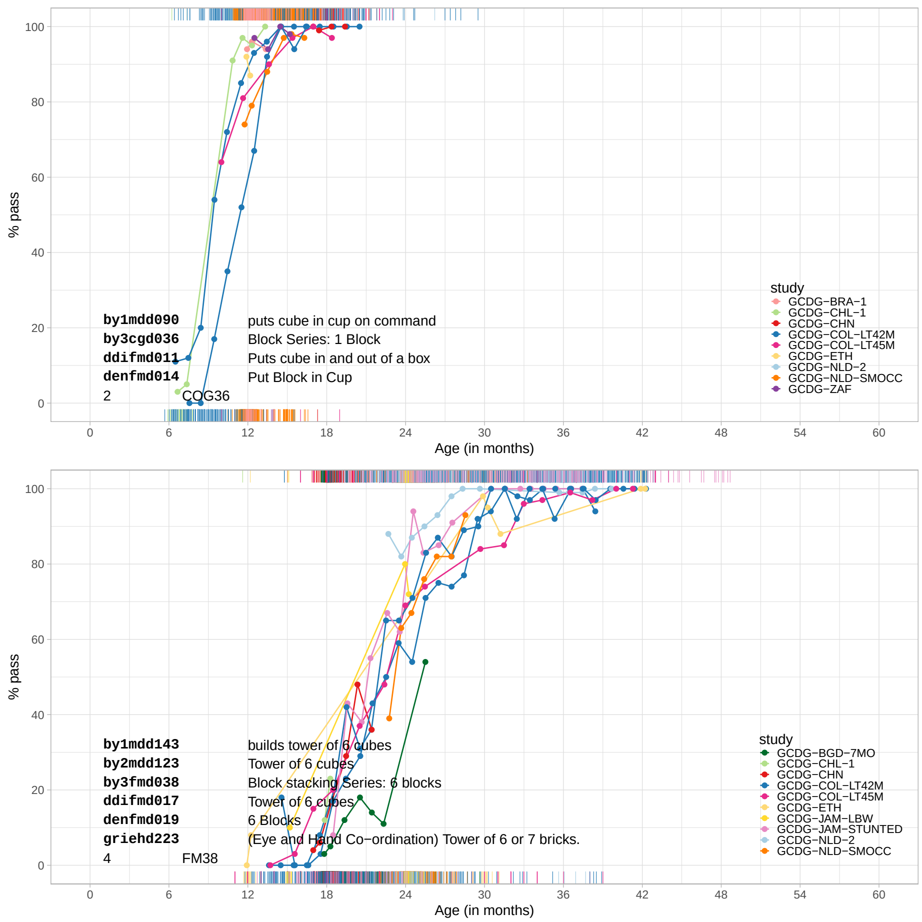

Figure 3.2: The probability of passing by age in potential bridging items.

Figure 3.3 presents two examples of weak item mappings. Notable timing differences exist for the “babbles” and “bangs” milestones, which suggests that we should not take these as bridges.

While these plots are suggestive, their interpretation is surprisingly complicated.

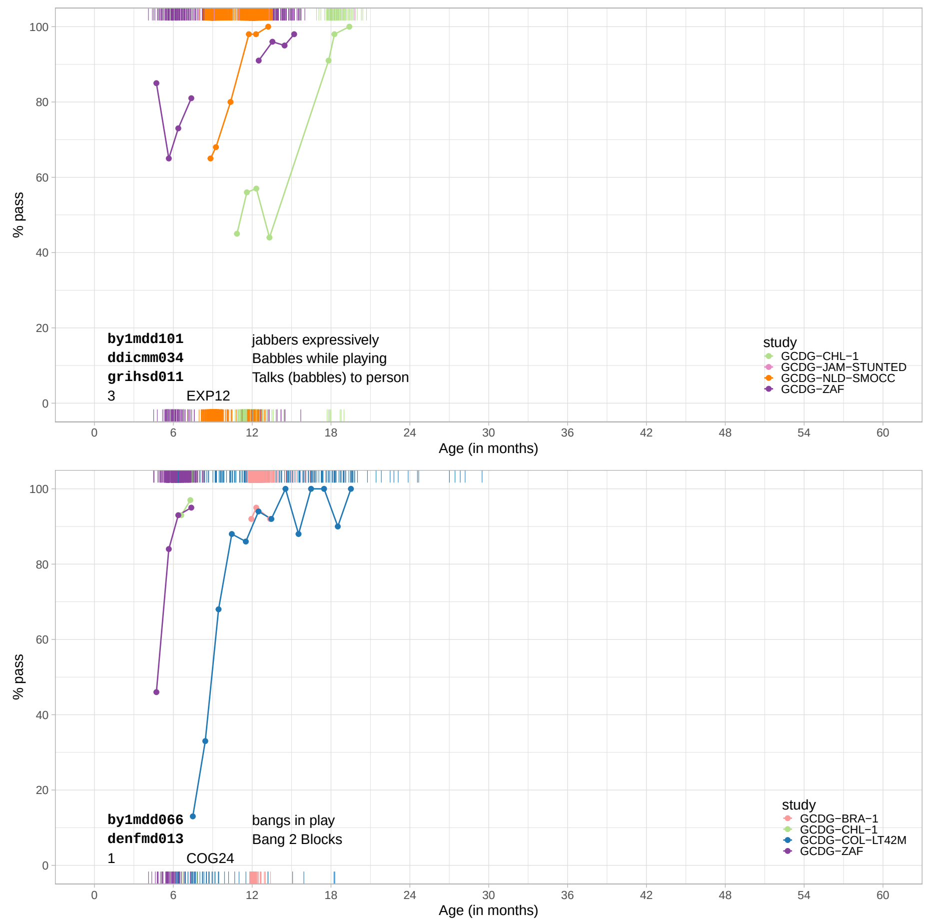

Figure 3.3: Probability to pass items for age in poor bridges.

We may find that age profiles of two milestones *A and B* administered in samples 1 and 2 respectively *are identical* if

- A and B are equally difficult and samples 1 and 2 have the same maturation level;
- A is more difficult than B and sample 1 is more advanced than 2.

Similarly, we may find that the age profile for *A is earlier than B* if

- A is easier than B and if samples 1 and 2 have the same level of maturation;
- A and B are equally difficult and if sample 1 is more advanced than sample 2.

Note that the age curves confound difficulty and ability, and hence cannot be used to evaluate the quality of the item map.

What we need to do is separate difficulty and ability. For this, we need a formal statistical model. The next section introduces the concepts required in such a model.

Chapter 4

Equate groups

This section introduces the concepts and tools needed to link assessments made by different instruments administered across multiple cohorts. Our methodology introduces the idea of an equate group. Systematic application of equate groups provides a robust yet flexible methodology to link different instruments. Once the links are in place, we may combine the data to enable meta-analyses and related methods.

- What is an equate group? (4.1)
- Concurrent calibration (4.2)
- Strategy to form and test equate groups (4.3)
- Statistical framework (4.4)
- Common latent scale (4.5)
- Quantifying equate fit (4.6)
- Differential Item Functioning (4.7)

4.1 What is an equate group?

An *equate group* is a set of two or more milestones that measure the same thing in (perhaps slightly) different ways. Table 3.2 contains an example of an equate group, containing items that measure the ability to form two-word sentences. Also, Figures 3.2 and 3.3 show examples of equate groups.

Equate groups vary in quality. We can use high-quality equate groups to link instruments by restricting the difficulty of all milestones in the equate group to be identical. Equate groups thus provide a method for bridging different tools.

Figure 4.1 displays items from three different instruments with overlapping sets of milestones. The shared items make up equate groups, as presented by the arrows between them. In the example, all three instruments share one milestone

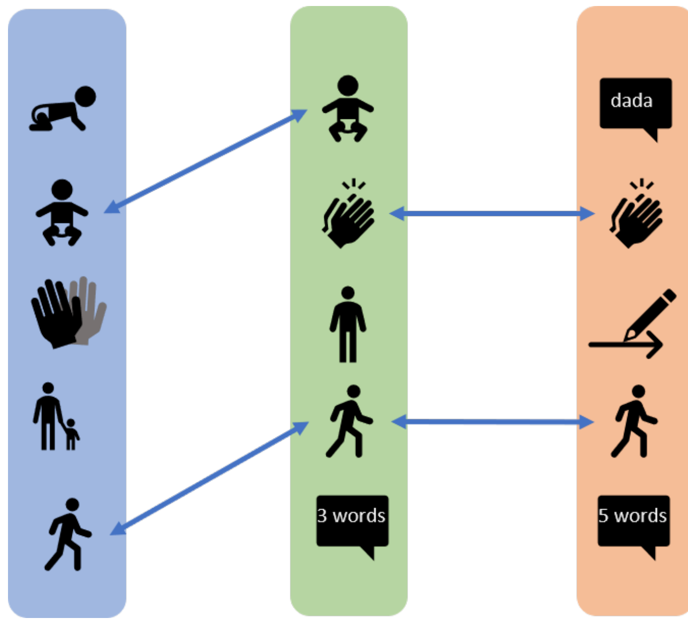

Figure 4.1: Example of three instruments that are bridged by common items in equate groups.

("walk alone"). The "sitting" and "clap hand" items appear in two tools. So in total, there are three equate groups.

4.2 Concurrent calibration

Patterns as in Figure 4.1 occur if we have multiple forms of the same instrument. Although in theory, there might be sequence effects, the usual working assumption is that we may ignore them. Equate groups with truly shared items that work in the same way across samples are of high quality. We may collect the responses on identical items into the same column of the data matrix. As a consequence, usual estimation methods will automatically produce one difficulty estimate for that column (i.e. common item).

The procedure described above is known as *concurrent calibration*. See Kim and Cohen (1998) for more background. The method simultaneously estimates the item parameters for all instruments. Concurrent calibration is an attractive option for various reasons:

- It yields a common latent scale across all instruments;
- It is efficient because it calibrates all items in a single run;
- It produces more stable estimates for common items in small samples.

However, concurrent calibration depends on a strict distinction between items that are indeed the same across instruments and items that differ.

In practice, strict black-white distinctions may not be possible. Items that measure the same skill may have been adapted to suit the format of the instrument (e.g. number of response options, question formulation, and so on). Also, investigators may have altered the item to suit the local language and cultural context. Such changes may or may not affect the measurement properties. The challenge is to find out whether items measure the underlying construct in the same way.

In practice, we may need to perform concurrent calibration to multiple - perhaps slightly dissimilar - milestones. When confronted with similar - but not identical - items, our strategy is first to form provisional equate groups. We then explore, test and rearrange these equate groups, in the hope of finding enough high-quality equate groups that will bridge instruments.

4.3 Strategy to form and test equate groups

An equate group is a collection of items. Content matter experts may form equate groups by evaluating the contents of items and organising them into groups with similar meaning. The modelling phase takes this set of equate groups (which may be hundreds) as input. Based on the analytic result, we may activate or modify equate groups. It is useful to distinguish between *active* and *passive* equate groups. What do we mean by these terms?

- *Active equate group*: The analysis treats all items within an active equate group as one super-item. The items obtain the same difficulty estimate and are assumed to yield equivalent measurements. As the items in an active equate group may originate from different instruments, such a group acts as a bridge between instruments.
- *Passive equate group*: Any non-active equate groups are called passive. The model does not restrict the difficulty estimates, i.e., the milestones within a passive equate group will have separate difficulty estimates.

Since active equate groups bridge different instruments, they have an essential role in the analysis. In general, we will set the status of an equate group to active *only* if we believe that the milestones in that group measure the underlying construct in the same way. Note that this does not necessarily imply that all items need to be identical. In Table 3.2, for example, small differences exist in item formulation. We may nevertheless believe that these are irrelevant and ignore these in practice. Reversely, there is no guarantee that the same milestone will measure child development in the same way in different samples. For example, a milestone like “climb stairs” could be more difficult (and more dangerous) for children who have never seen a staircase.

Figure 4.2: One year old child climbs stairs.

The data analysis informs decisions to activate equate groups. The following steps implement our strategy for forming and enabling equate groups:

- Content matter experts compare milestones from different instruments and sort similar milestones into equate groups. It may be convenient to select one instrument as a starting point, and map items from others to that (see section 3.2);
- Visualise age profiles of mapped items (see section 3.3). Verify the plausibility of potential matches through similar age profiles. Break up mappings for which age profiles appear implausible. This step requires both statistical and subject matter expertise;
- Fit the model to the data using a subset of equate groups as active. Review the quality of the solution and optimise the quality of the links between tools by editing the equate group structure. The technical details of this model are explained in section 4.4. Refit the model until (1) active equate groups link all cohorts and instruments, (2) active equate groups are distributed over the full-scale range (rather than being centred at one point);
- Assess the quality of equate groups by the infit and outfit (see section 4.6).
- Test performance of the equate groups across subgroups or cohorts by methods designed to detect differential item functioning (see section 4.7).

The application of equate groups is needed to connect different instruments to a universal scale. The technique is especially helpful in the situation where abilities differ across cohorts.

If the cohort abilities are relatively uniform (for example as a result of experimental design) and if the risk of misspecification of the equate groups is high, a good alternative is to rely on the equality of ability distribution. In our application, this was not an option due to the substantial age variation between cohorts.

4.4 Parameter estimation with equate groups

The Rasch model is the preferred measurement model for child development data. Chapter I, Section 4 provides an introduction of the Rasch model geared towards the D-score.

The Rasch model expresses the probability of passing an item as a logistic function of the difference between the person ability β_n and the item difficulty δ_i. Table 4.1 explains the symbols used in equation (4.1). Formula (4.1) defines the model as

$$\pi_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \tag{4.1}$$

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). See the logistic model in Chapter 1, Section 4.6.1 for more detail.

In model (4.1) every milestone i has one parameter δ_i. We extend the Rasch model by restricting the δ_i of all items within the same equate group to the same value. We thereby effectively say that these items are interchangeable measures of child development.

Estimation of the parameter for the equate group is straightforward. Wright and Masters (1982) present a simple method for aligning two test forms with common items. There are three steps:

- Estimate the separate δ_i's per item;
- Combine these estimates into δ_q by calculating their weighted average;
- Overwrite each δ_i by δ_q.

Suppose that Q is the collection of items in equate group q, and that w_i is the number of respondents for item i. The parameter estimate δ_q for the equate group is

$$\delta_q = \frac{\sum_{i \in Q} \delta_i w_i}{\sum_{i \in Q} w_i} \tag{4.2}$$

Table 4.1: Overview the symbols used in equations (4.1) and (4.2).

Symbol	Term	Description
β_n	Ability	True (but unknown) developmental score of child n
δ_i	Difficulty	True (but unknown) difficulty of item i
δ_q	Difficulty	The combined difficulty of the items in equate group q
π_{ni}	Probability	Probability that child n passes item i
l		The number of items in the equate group
w_i		The number of respondents with an observed score on item i

4.5 Common latent scale

The end goal for using the equate group method to model development items is to measure development on one common latent scale, the D-score. That way, the measure (i.e. D-score) can be obtained, irrespective of which instrument is used in which population.

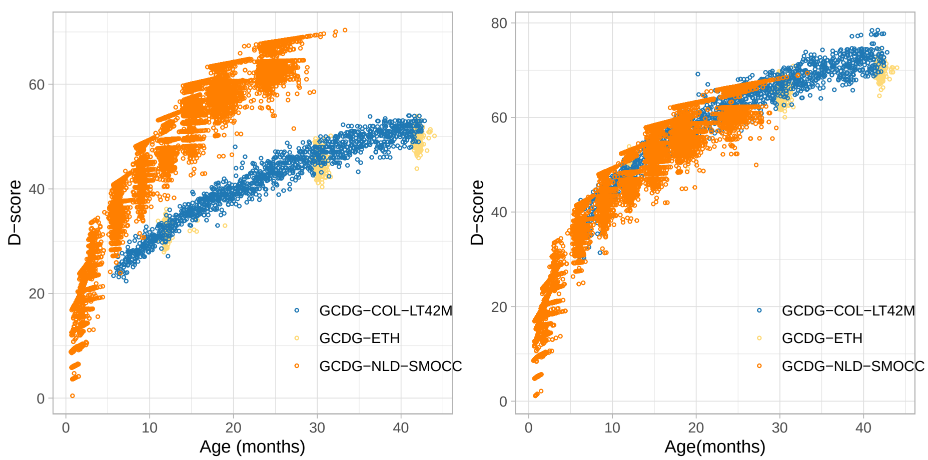

Figure 4.3: Example of three cohorts with and without equate group linking.

Figure 4.3 displays the D-score estimates by age in three cohorts from the GCDG study: Netherlands 1 (GCDG-NLS-SMOCC), Ethiopia (GCDG-ETH) and Colombia 2 (GCDG-COL-LT42M) for two different analyses. As described in section 2.2, the Netherlands 1 study administered the `ddi`; Ethiopia measured children by the `by3`; and Colombia collected data on `by3`, `den`, `aqi` and `bdi`. Accordingly, there is an overlap in items between Ethiopia and Colombia via the `by3`, but the Netherlands 1 cohort is not linked.

We created the plot on the left-handed side without active equate groups. The large overlap between Ethiopian and Columbian children occurs because the scales for these studies are linked naturally via shared items from `by3`. Since the `ddi` instrument is not connected, the Dutch cohort follows a different track. While we can compare D-scores between Ethiopia and Colombia, it is nonsensical to compare Dutch to either Ethiopia or Colombia. The right-handed side plot is based on an analysis that used active equate groups to link the cohorts. Since the analysis connected the scales for all three cohorts, we can now compare D-scores obtained between all three cohorts.

This example demonstrates that active equate groups form the key for converting ability estimates for children from different cohorts using different instruments onto the same scale.

4.6 Quantifying equate fit

It is essential to activate only those equate groups for which the assumption of equivalent measurement holds. We have already seen the *item fit* and *person fit* diagnostics of the Rasch model. This section describes a similar measure for the quality of an active equate group.

4.6.1 Equate fit

Section 6 of Chapter I defines the observed response of person n on item i as x_{ni}. The accompanying standardized residual z_{ni} is the difference between x_{ni} and the expected response P_{ni}, divided by the expected binomial standard deviation,

$$z_{ni} = \frac{x_{ni} - P_{ni}}{\sqrt{W_{ni}}},$$

with variances $W_{ni} = P_{ni}(1 - P_{ni})$.

Equate infit is an extension of item infit that takes an aggregate over all items i in active equate group q, i.e.,

$$\text{Equate infit} = \frac{\sum_{i \in q} \sum_n^N (x_{ni} - P_{ni})^2}{\sum_{i \in q} \sum_n^N W_{ni}}.$$

Likewise, we calculate *Equate outfit* of group q as

$$\text{Equate outfit} = \frac{\sum_{i \in q} \sum_n^{N_i} z_{ni}^2}{\sum_{i \in q} N_i},$$

where N_i is the total number of responses observed on item i. The interpretation of these diagnostics is the same as for item infit and item outfit.

Note that these definitions implicitly assume that the expected response P_{ni} is calculated under a model in which all items in equate group q have the same difficulty. This is not true for passive equate groups. Of course, no one can stop us from calculating the above equate fit statistics for passive groups, but such estimates would ignore the between-item variation in difficulties, and hence gives a too optimistic estimate of quality. The bottom line is: *The interpretation of the equate fit statistics should be restricted to active equate groups only.*

4.6.2 Examples of well fitting equate groups

The evaluation of *equate fit* involves comparing the observed probabilities of endorsing the items in the equate group to the estimated probability of endorsing the items in the equate group. For an equate group there is an empirical curve for each item in the equate group and one shared estimated curve. The empirical curves should all be close together, and close to the estimated curve for a good equate fit.

Figure 4.4 shows a diagnostic plot for equate groups `REC6` (Turns head to sound of bell) and `GM42` (Walks alone). The items within `REC6` have slightly different formats in the Bayley I (`by1`), Dutch Development Instrument (`ddi`), and the Denver (`den`). The empirical curves in the upper figure show good overlap, but note that hardly any negative responses were recorded for four of the five studies, so the shared estimate depends primarily on the Dutch sample. Items from equate group `GM42` appear in six instruments: `bar`, `by1`, `by2`, `by3`, `ddi`, and `gri`. Also, here the empirical data are close together, and even a little steeper than the fitted dashed line, which indicates a good equate fit. The infit and outfit indices, shown in the upper left corners, confirm the good fit (fit < 1).

4.6.3 Examples of equate groups with poor equate fit

Poor fitting equate groups are best treated as passive equate groups, so that items in those groups are not restricted to the same difficulty. Empirical item curves with different locations and slopes indicate a poor fit. Additionally, the equate fit indices will indicate a poor fit (fit > 1).

Figure 4.5 shows examples for groups `COG24` (Bangs in play / Bangs 2 blocks) and `EXP12` (Babbles). In both cases there is substantial variation in location between the empirical curves. For `COG24` we find that the fitted curve is closer to the `den` item, which suggests that the equate difficulty is mostly based on the `den` item. Items from equate group `EXP12` have a different format in instruments `by1`, `ddi`, and `gri`. The empirical curves, with different colours for each instrument, are not close to each other, nor close to the fitted curve. Note that all infit and

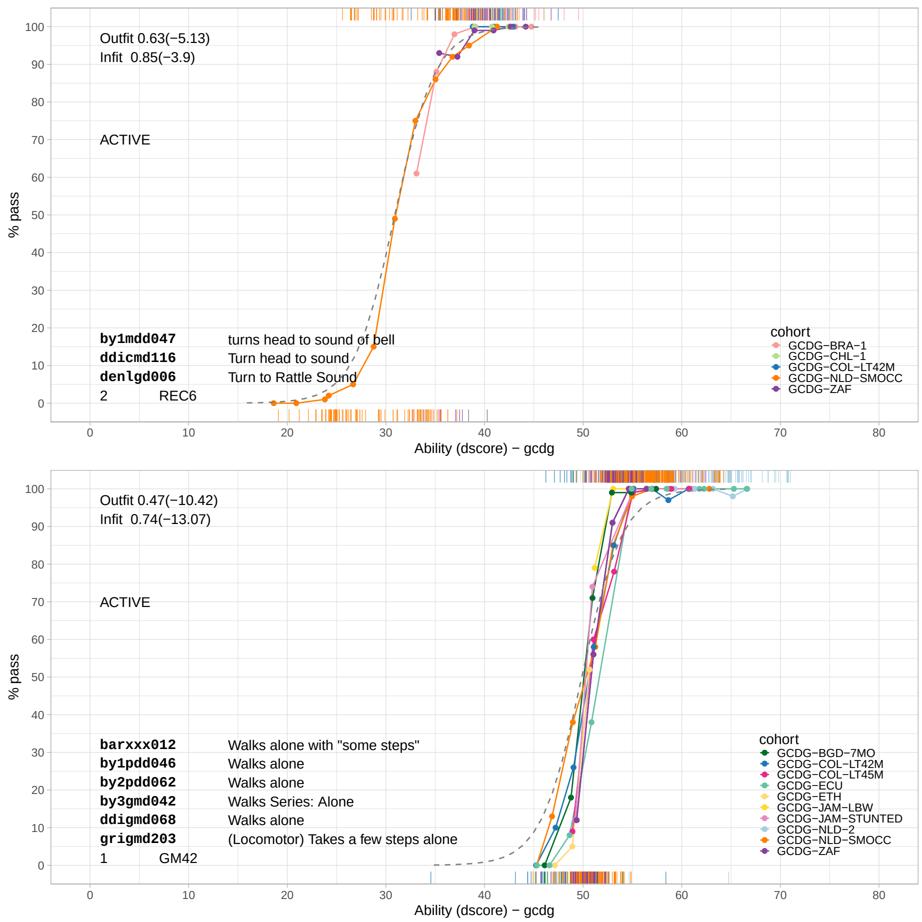

Figure 4.4: Two equate groups that present a good equate fit.

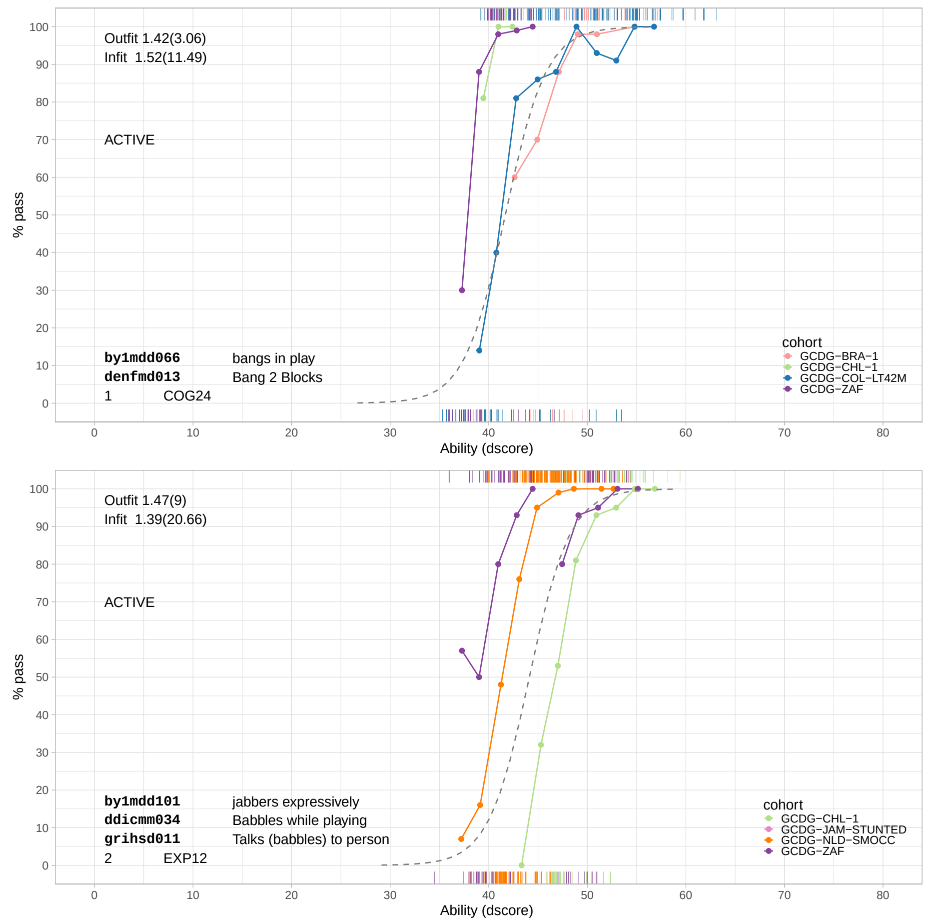

Figure 4.5: Two equate groups that present a poor equate fit.

outfit statistics are fairly high, indicating poor fit. Both equates are candidates for deactivation in a next modelling step.

4.7 Differential item functioning

Items within an active equate group should work in the same way across the different cohorts, i.e., they have no differential item functioning (DIF). The assumption of no DIF is critical for active equate groups. If violated, restricting the difficulty parameters as equal across cohorts may introduce unwanted bias in comparisons between cohorts. This section illustrates the role of DIF in equate groups.

4.7.1 Good equate groups without DIF

Chapter I discusses the role of DIF in the evaluation of the fit of items to the Rasch model. This section illustrates similar issues in the context of equate groups.

Figure 4.6 shows the empirical curves of two equate groups, `FM31` (two cubes) and `EXP26` (two-word sentence). All curves are close to each other, so there is no differential item functioning here.

4.7.2 Poor equate groups with DIF for study

Figure 4.7 plots the empirical curves for equate groups `GM44` (throws ball) and `EXP23` (5 or more words). The substantial variation between these curves is a sign of differential item functioning. For example, *Throws ball* is easier for children in the South-Africa cohort (purple curve; GCDG-ZAF) and more difficult for children in Colombia (blue curve; GCDG-COL-LT42M). In other words, the probability of passing the item given the D-score (i.e. item difficulty) differs between the cohorts. Likewise, there is differential item functioning for *Says more than 5 words*. This milestone is easier for children in Jamaica (yellow and pink curves; GCDG-JAM-LBW and GCDG-JAM-STUNTED) than for children from Ecuador (green; GCDG-ECU).

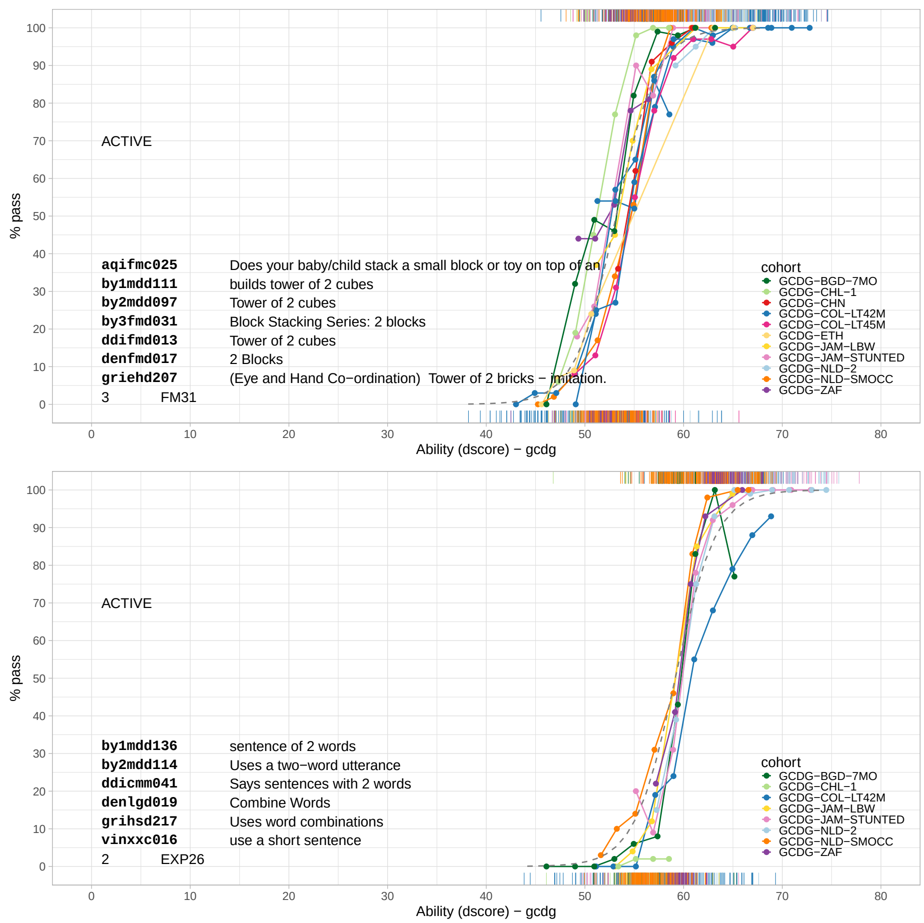

Figure 4.6: Two equate groups that present no differential item functioning between cohorts.

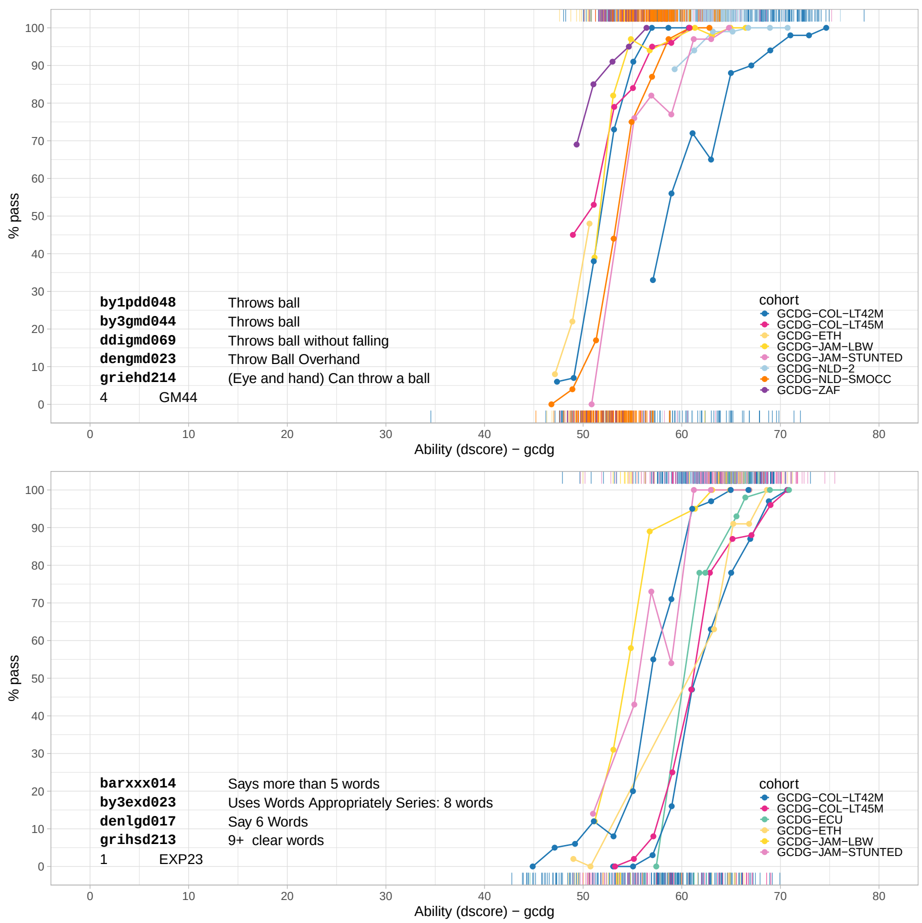

Figure 4.7: Two equate groups that present differential item functioning between cohorts.

Chapter 5

Modelling equates

This section deals with the nitty-gritty of the modelling strategy used for the GCDG data introduced in Section 2. This section

- provides a high-level description of the GCDG data (5.1)
- discusses various modelling strategies (5.2)
- shows the impact of equate groups on the model in extreme cases (5.3)
- demonstrates visualisation of age profiles to select promising equate groups (5.4)
- introduces a helpful visualisation of the quality of the equate group (5.5)
- highlights infit and outfit for removing misfitting milestones (5.6)
- discusses instrument fit and equate group editing (5.7)
- introduces a grading system for equate groups (5.8)
- provides pointers to the final model (5.9)

5.1 GCDG data: design and description

5.1.1 Data combination

Section 2.1 provides an overview of the data collected by Global Child Development Group. The group collected item level measurements obtained on 12 instruments for measuring child development across 16 cohorts.

We coded every item as 0 (FAIL), 1 (PASS) or missing. For some instrument we did some additional recoding to restrict to these two response categories. The Battelle Developmental Inventory scores items as 0 (FAIL), 1, or 2, depending on the level of skill demonstrated or time taken to complete the task. We joined categories 1 and 2 for these items. The ASQ items were originally scored as 0 (not yet), 5 (sometimes) and 10 (succeeds). We recoded both 5 and 10 to 1.

We concatenated the datasets from the GCDG cohorts cohort. The resulting data matrix has 71403 rows (child-visit combinations) and 1572 columns (items) collected from 36345 unique children. We removed 233 items that had fewer than 10 observations in a category. The remaining 1339 items were candidates for analysis. The total number of observed scores was equal to about 2.8 million pass/fail responses. While this is a large number of measurements, about 97 percent of the entries in the matrix are missing.

5.1.2 Equate group formation

A group of 13 subject-matter experts from the Global Child Development Group cross-walked the available instruments for similar milestones. This group

- developed an item coding schema;
- matched similarly appearing items stemming from different instruments;
- formed an opinion about the quality of each match;
- noted peculiarities of the matches;
- reported the results as a series of detailed Excel spreadsheets.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1			Expressive																
2		Bayley 3		Griffiths		Bayley 2		Dutch		WM Memory of phrases (Madagascar)		Bayley 1		Denver					
3	Item	Label	Description	Item	Description	Item	Description	Item	Description	Item	Description	Item	Description	Item	label	administration	scoring	age (mo)	notes
4	b3e01	Undifferentiated throaty sounds					new												
5	b3e02	Social smile				b2m14	Smiles when examiner speaks												
6	b3e03	Vocalizes mood				b2m31	Vocalizes attitude												
7	b3e04	Undifferentiated nasal sounds	Child produces nasal vocalizations.	gh1	vocalisation other than crying		new							DL2	Vocalizes	Also pass if they child passes more advanced language questions.	Child makes small throaty sound or short sounds like "uh" or "eh"?	1-4	Not exactly the same
8	b3e05	Social vocalization or laughing	Child vocalizes or laughs in response to speaker's attention.			b2m21	vocalizes when examiner speaks	n31	vocalizes in response					DL4	Laughs	This does not have to be a deep belly laugh like an adult, but may be a soft giggle.	Child sometimes laugh out loud?	2-6	
9	b3e06	2 Vowel sounds	Child vocalizes at least two different, distinct vowel sounds.			b2m22	Vocalizes vowel sounds							DL3	"Ooo" / "Aah"		Child makes vowel sounds like "ooo" or "aah"?	1-5	
10	b3e07	Gets attention	Child tries to get attention from you or others.				new												
11	b3e08	2 Consonant sounds	Child produces at least two different, distinct consonant sounds.	gh5	(Hearing and Speech) Makes 2 + different speech sounds		new							DL8	Single Syllables		Child sometimes says 1 syllable like "ba", "da", "ga", "ma"?	4-10	
12	b3e09	Uses gestures	Child uses at least one gesture to make wants known.			b2m68	Uses gestures					b1m116	uses gestures to make wants known	DP8	Indicate Wants	During the test, note if the child lets someone know if he wants something, without cry.	Child indicates their wants other than by crying (e.g. pointing, reaching, making sounds, reaching up with arms)?	7-15	
13	b3e10	Consonant-Vowel Combination Series: 1 Combination	Child imitates at least one repetitive consonant-vowel combination.	gh14	2 syllable babble	b2m71	Consonant vowel combination 1 combination	n33	Says dada, baba, gaga			b1m85	says "da-da" or equivalent	DL10	Abba / Mama Nonspecific	10. Abba / Mama Nonspecific	10. Abba / Mama Nonspecific	6-13	
14	b3e11	Participates in play routine	Child actively participates in at least one play routine.			b2m64	Participates in play routine							DP7	7. Clapping game "pat a cake"	Without touching the child, play "Pat a cake" and clap your hands together. You can ask the guardian to try.	Child plays clapping games such as "pat a cake" with you?	7-15	
15	b3e12	Jabbers expressively	Child produces at least one vocalization that contains inflections and is expressive.	gh11	Talks (babbles) to person	b2m76	Jabbers expressively	n34	Babbles while playing			b1m101	jabbers expressively						
16	b3e13	Consonant-Vowel Combination Series: 4 Combination	Child imitates at least four different repetitive consonant-vowel combinations.	gh16	(Hearing and Speech) Babbled phrases: 4 + syllables	b2m78	Consonant-Vowel Combination Series: 4 Combination							DL11	Combine Syllables		Child repeats the same syllable 3 or more times, eg. "Dadada" "Gagaga"?	6-15	
17	b3e14	Uses one-word approximations	Child produces at least 1 one-word approximation.	gh17	(Hearing and Speech) Says Mama or Dada, etc. (one word clear)		new							DL14	Say 1 Word	Interviewer write the words here:_________________ (Also fill in Q15, Q16, Q17)	How many words does the child say and what are the words? (Pass if say 1 or more word)	10-28	
18	b3e15	Directs attention of other	Child points to or shows you at least one object.				new												
19	b3e16	Imitates word	Child imitates at least one word, even if imitation consists of vowels only.							w1-w6	Ask child to repeat back very simple words	b1m106	imitates words	DL9	Imitate Speech Sounds	Repeat a sound several times (eg, a click of the tongue or the sound of a kiss).	Child sometimes repeats a sound that it hears imitate sounds of an adult)?	4-12	Not exactly the same

Figure 5.1: A snapshot of information generated by subject-matter experts.

The group evaluated around 1500 milestones. After several days, this highly-skilled, intensive labour resulted in a series of spreadsheets. Figure 5.1 shows an example. These sheets formed the basis of an initial list of 184 equate groups, each consisting of at least two items.

5.2 Modelling strategies

The analytic challenge is twofold:

- to find a subset of items that form a scale;
- to find a subset of equate groups with items similar enough to bridge instruments.

Note that both subsets are related, i.e., changing one affects the other. Thus, we cannot first identify items and then equate groups, or first identify equate groups followed by the items. Rather we need to find the two subsets in an iterative fashion, primarily by hand. This section describes some of the modelling issues the analyst needs to confront.

In general, we look for a final model that

- preserves the items that best fit the Rasch model;
- uses active equate groups with items that behave the same across many cohorts and instruments;
- displays reasonable age-conditional distributions of the D-scores;
- has difficulty estimates that are similar to previous estimates.

The modelling strategy is a delicate balancing act to achieve all of the above objectives. Particular actions that we could take to improve a given model are:

- remove bad items;
- inactivate bad equate groups;
- break up bad equate groups;
- move items from one equate group to another;
- create new equate groups;
- remove entire instruments;
- remove persons;
- remove studies.

In order to steer our actions, we look at the following diagnostics (in order of importance):

- quality of equate groups (both visually and through infit);
- plausibility of the distribution of the D-score by age per study;
- correspondence of difficulty estimates from published (single study) Dutch data and the new model;
- infit of the items remaining in the model.

Various routes are possible and may result in different final models. The strategy adopted here is to thicken active equate groups by covering as many studies as possible, in the hope of minimizing the number of active equates needed.

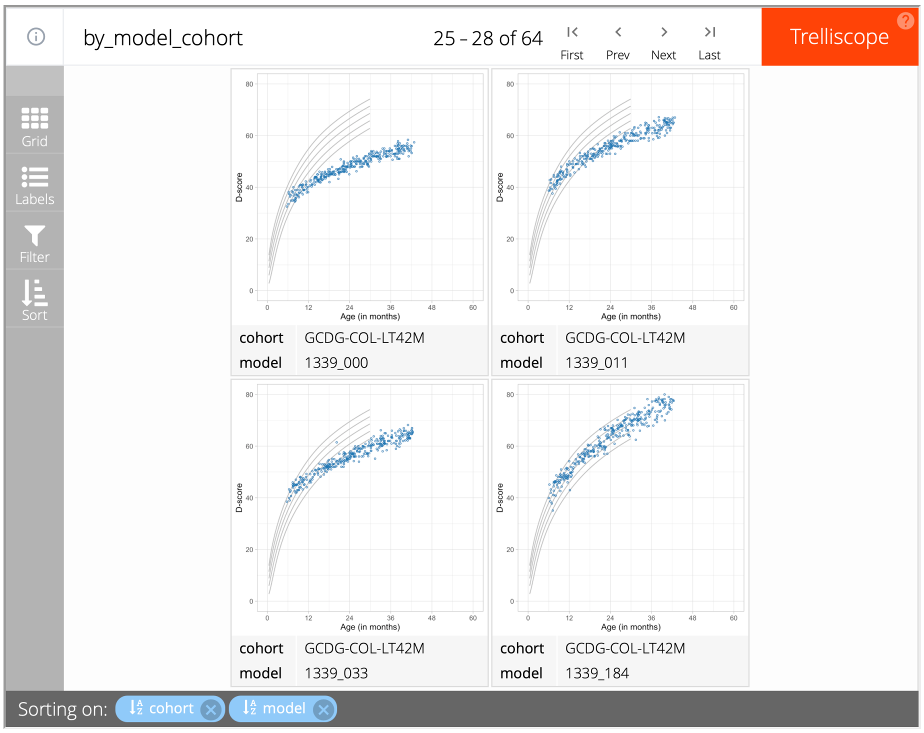

Figure 5.2: D-score by age of four models with all 1339 items using 0, 11, 33 and 184 active equate groups. The number of equate groups has a substantial effect on the D-score distribution. Use the arrows to see other cohorts (https://d-score.org/dbook-apps/models1339/).

5.3 Impact of number of active equate groups

Figure 5.2 is a display of the D-score by age for all 16 cohorts under four models. As a rough reference to compare, the grey curves in the back represent the Dutch model as calculated from the SMOCC study. In order to speed up the calculations, the figure shows a random subsample of 25% of all points. Manipulate the plot controls to switch cohorts.

All models contain 1339 items, but differ in the number of active equate groups. The most salient features per model are:

- `1339_0`: No equate groups, so different instruments in different cohorts are fitted independently;
- `1339_11`: Connects all cohorts through one or more equated items using 11 equate groups in total;
- `1339_33`: There are 33 equate groups that bridge cohort and instruments;
- `1339_184`: Maximally connects instruments and cohort by all equate groups.

Comparison of the D-score distribution by age across these models yields various insights:

- The location of cohorts on the vertical scale depends on the number of active equate groups. For example, for Madagascar (MDG) the points are located around 52 when no equate groups are activated, whereas if all are activated it is about 68.

- The age trend depends on the number of active equate groups. For example, for Colombia (COL) or Ethiopia (ETH), the model without equate groups has a shallow age trend, whereas it is steep for the `1339_184` model.

- The vertical spread depends on the number of equate groups. For example, the spread in the Chile-2 (CHL-2) cohort substantially increases with the number of active equates.

- Model `1339_0` for the Dutch NLD-SMOCC cohort is equivalent to the model fitted to the SMOCC study alone. Introducing equate groups compresses the range of scores, especially at the higher end.

We have now seen that the number of active equate groups has a large effect on the model. The next sections look into the equate groups in more detail.

5.4 Age profiles of similar milestones

Figure 5.3 displays the percentage of children that pass milestones at various ages. Subject matter experts clustered similar items stemming from different

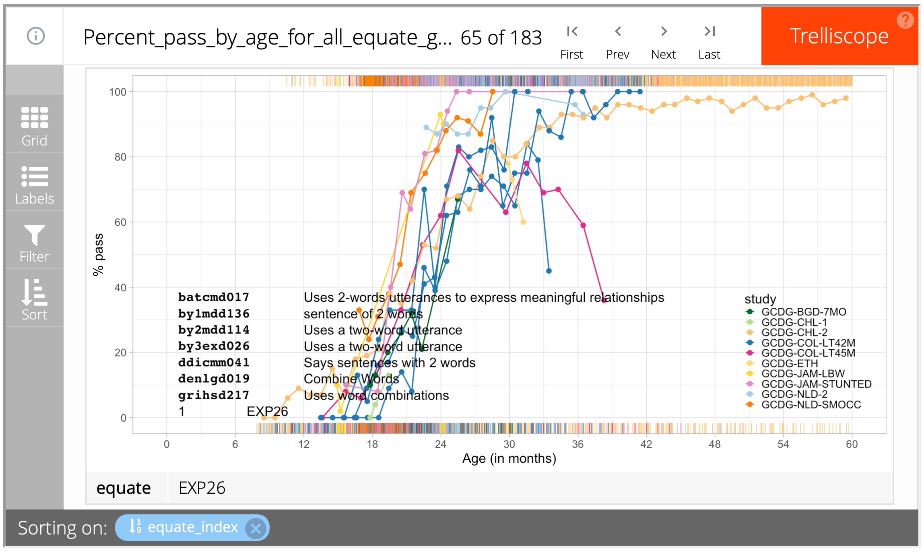

Figure 5.3: Percentage of children that pass similar milestones at a given age (https://d-score.org/dbook-apps/p-a-equate-1339/).

instruments into equate groups. There are 184 equate groups that contain two or more milestones.

Most age profiles show a rising pattern, as expected, though some (e.g. `FM17` or `EXP11`) have one item showing a negative relation with age. Equate `EXP26` combines `two-word sentences` items from seven instruments into one plot. The item difficulties expressed as age-equivalents (c.f. Section 3.1.2, Chapter I) for these cohorts vary between 20-25 months. By comparison, equate group `EXP18` (`says two words`) shows more heterogeneity across cohorts, and is therefore, less likely to be useful for equating. Equate group `FM31` (`stack two blocks`) is another example of a promising example. By comparison, `FM38` (`stack 6-8 blocks`) shows additional heterogeneity. As a last example, consider `GM42` (`walks alone`), which has a similar age profile across cohorts, whereas `GM44` (`throws ball`) or GM49 (`walk down stairs`) are more heterogeneous.

We could follow different strategies in selecting which equate groups to activate. One strategy would be to include as many equate groups as possible (e.g. all 184 equates) so as to build as many bridges as possible between different instruments. A more selective strategy would be to activate a subset of promising equates and leave others inactive. The following section compares four different approaches.

Figure 5.4: Percentage of children that pass similar milestones given their D-score as calculated under four models (1339 items, and 0, 11, 33 and 184 equate groups, respectively (https://d-score.org/dbook-apps/p-d-equate-1339/).

5.5 Quality of equate groups

Figure 5.4 shows how the passing percentage depends on the child's D-score as calculated under four models. All models include the same 1339 milestones, but differ in the number of active equates. The grey curve corresponds to the estimate made under the assumption that milestones are equally difficult. Good milestones for bridging instruments will have a tight bundle of curves. For example, equate `EXP26` has tight bundles especially in models `1339_11` and `1339_33`. By comparison, the curves of the two extreme models vary considerably: the model without any bridges (`1339_0`) or the model with all bridges (`1339_184`) are thus less than ideal. The shallow grey curve of model `1339_184` indicates a poorer overall fit.

Outfit and infit statistics measure the residual deviation of the items to the grey curve. High values (e.g. above 1.4) are undesirable and indicate lack of fit to the model. For example, the fit statistics for `EXP26` in model `1339_184` (1.70 and 1.25) indicate a mediocre fit, whereas `EXP26` in models `1339_33` and `1339_11` fits well. Sometimes the individual item curves are steeper than the grey curve. This indicates that these milestones are more discriminative than the combined item. Model `1339_0` lacks a grey curve and has no fit statistics for equate groups, because in that model, the combined item is not activated.

The probability curves provide a quick visual method for spotting promising and problematic equate groups. Examples of promising equate groups include

`COG36`, `FM31`, `GM26` and `GM42`. A little more weak are `FM26` (has more variability), `FM52` (looks promising, but has a problem with the item `grigcd402` from the `GCDG_JAM_STUNTED` cohort), and `GM35` (does not align cohort `GCDG-ZAF`). In such cases, one may wish to move an item out of an equate group, combine equate groups, or inactivate troublesome links.

Until now we only looked at models that include all 1339 items. In practice, we may improve upon the model by selecting the subset of milestones that fit the Rasch model. The next section looks in this modelling step in more detail.

5.6 Milestone selection

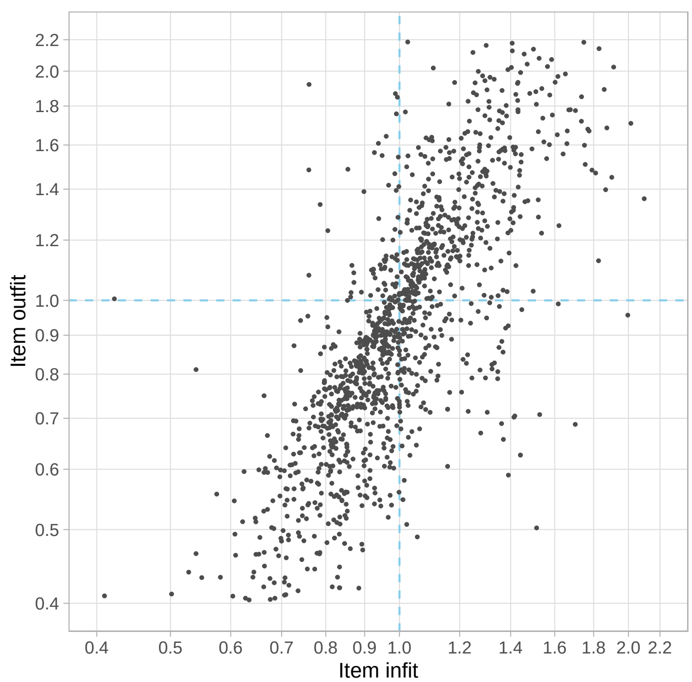

Figure 5.5: Infit and outfit of 1339 items in model `1339_11`. About 8 percent of the points falls outside the plot.

Item infit and outfit are convenient statistics for selecting the milestones that fit the model. Figure 5.5 displays the infit and outfit statistics of model `1339_11`. The correlation between infit and outfit is high ($r = 0.84$). The expected value of the infit and outfit statistics for a perfect fit is 1.0. The centre of infit and outfit in Figure 5.5 is approximately 1.0, so on average one could say the items fit the model. Note however that fit values above and below the values of 1.0 are qualitatively different. Item with fit statistics exceeding 1.0 fit the model less well than expected (**underfit**), whereas items with fit statistics lower than 1.0 fit the model better than expected (**overfit**). See Chapter 1, Section 6.1 for more details.

Some practitioners remove both underfitting and overfitting items. However, we like to preserve overfitting items and be more strict in removing items that underfit. The idea is that preservation of the best fitting items may increase scale length, and hence reliability and measurement precision. Figure 5.5 draws two cut-off lines at 1.0. Taking items with infit < 1.0 and outfit < 1.0 will select **631 out of 1339** items for further modelling.

A practical problem of item removal is that it also affects equate group composition. By default, a removed item will also be removed from the equate group, so item removal may reduce the size of an equate group below two items. For passive equates this is no problem, since passive equates do no affect the estimates. However, removal of an underfitting item from an active equate group will break the bridge between the instrument it pertains to and the rest of the item set. Potentially this can result in substantial effects on the D-score distribution of the cohort, as demonstrated in Figure 5.2. As a solution, we force any items that are members of active equate groups to remain in the analysis. If that leads to substantially worse equate fit in the next model, we must search for alternative equate groups that bridge the same instruments and that are less sensitive to misfit.

5.7 Other modelling actions

5.7.1 Instrument fit

Some instruments fit better than others. Figure 5.6 shows the box plots of outfit per instrument. Instruments `bar`, `by1`, `ddi` and `vin` generally fit well, whereas discrepancies between model and data are larger for `bat`, `by2` and `sbi`. Through additional modelling, we found that it was extremely difficult to get enough high-quality bridge items that could link `bat` (Battelle Development Inventory) to the other instruments. We also found that models without the Battelle were able to better discriminate children in the upper range of the D-score scale. We therefore opted to remove `bat` from the model, even though this meant that one cohort (`GCDG-BRA-2`) had to be dropped from the analysis.

It is not clear why `bat` does not fit. Perhaps the scoring system of the Battelle in three categories invokes scoring behaviour that is different from the PASS/FAIL scoring used by most other instruments, even though this appears to be less of a troublesome aspect in `aqi`, which also uses three response categories.

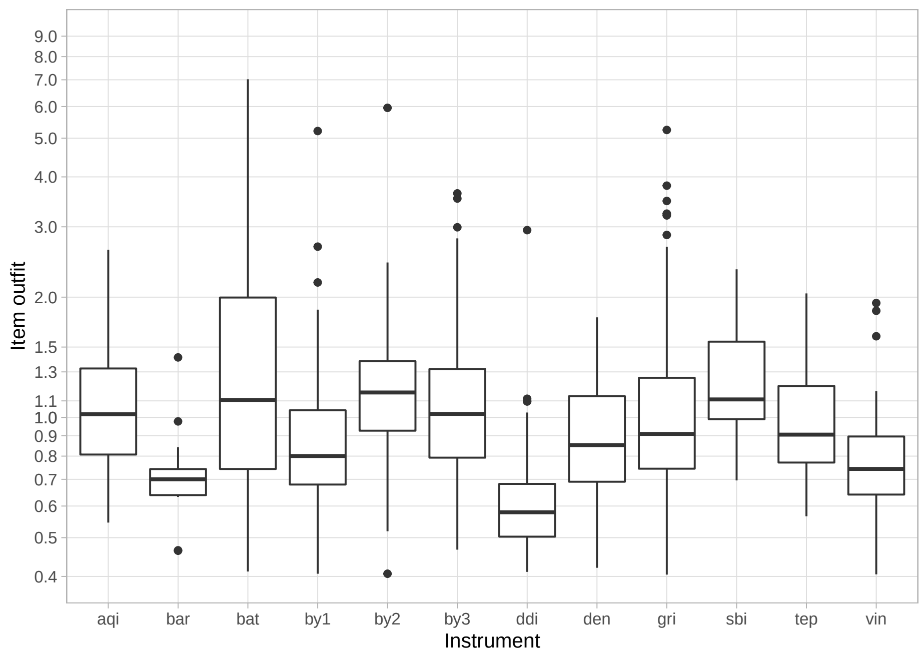

Figure 5.6: Box plot of the distribution of item outfit per instrument in model `1339_11`.

5.7.2 Splitting, combining and selecting equate groups

Most of the modelling effort went into finding a set of high-quality equate groups that link the instruments. For example, we tried to bridge the South-African study placing `vinxxc016` (uses a short sentence) into `EXP26` (two-word sentences) and `EXP36` (sentences of 3 or more words), but neither option led to a reasonable model. On the surface, milestone `by3gmd060` (balances on right foot, 2 seconds) appears to fit within `GM60` (balances on foot), but the analysis showed large discrepancies with the other items in the groups, so it had to be taken out.

Subject-matter experts identified 38 items that were thought to be cross-culturally incompatible. Table 5.1 provides an overview. Many of such milestones involve a specific language concept (such as a pronoun), refer to stairs (less common in rural settings), help in house or clothing behaviour. These items have different meanings in different contexts, so they were not used to bridge instruments.

5.8 Item information

Item information is a psychometric measure that quantifies the sensitivity of the item to changes in the person's ability. An item is most sensitive around the D-score value where the PASS probability equals the FAIL probability, which corresponds to the item difficulty (δ_i). One unit change around δ_i has a large effect on the probability of endorsing, while one unit change far away from δ_i has negligible impact. Suppose person A had passing probability 0.7 for some item. The information delivered by that item for person A is the product $0.7 \times (1.0 - 0.7) = 0.21$. Suppose person B has a D-score that coincides with the difficulty level of the item. In that case, the information for B equals $0.5 \times (1 - 0.5) = 0.25$, the maximum. Likewise, for a person C with high ability, the information could be $0.98 * 0.02 = 0.02$, so that item carries almost no information for person C.

The information is inversely related to the error of measurement. More information amounts to less measurement error. For each response in the data, we can compute the amount of information it contributed to the model D-score. By summing the information over persons, we obtain a measure of certainty about the difficulty estimate of the item. This sum of information incorporates both the number of administrations and the quality of the match between person abilities and item difficulty.

Figure 5.7 displays the summed information for each item, divided into four grades: A(best) to D (worst). The information grade measures the stability of the difficulty estimate. Most items receive grades higher than C. In total, 30 milestones have grade D. Adding these items to future studies may yield important additional information.

Table 5.1: Milestones not used for equating because of limited cross-cultural validity

Item	Label
aqislc023	When you dress your baby does she lift her foot for her shoe, sock, or pant leg?
aqislc041	Using these exact words, ask your child, "Are you a girl or a boy?" Does your child answw
by1mdd050	Washes and dries hands
by1pdd053	Bowel and bladder control
by1pdd054	manipulates table edge actively
by2pdd069	Walsk up stairs with help
by3cgd043	Walks down stairs with help
by3cgd052	Walks down stairs with help
by3gmd047	Clear Box: Front
by3gmd049	Clear Box: Sides
by3gmd057	Uses pronouns
by3gmd058	Walks Up Stairs Series: Both feet on each step, with support.
by3red030	Walks Down Stairs Series: Both feet on each step, with support
by3exd030	Walks Up Stairs Series: Both feet on each step, alone.
barxxx016	Walks Down Stairs Series: Both feet on each step, alone
barxxx020	Understands pronouns (him, me, my, you, your)
dengmd020	Eats with spoon without help (M; can ask parents)
densld012	Takes off shoes and socks (M; can ask parents)
densld013	Can dress (one piece) (M; can ask parents)
grigmd219	Walk Up Stairs
grigmd222	Drink from a cup
mdsgmd002	help in house
mdsgmd003	(Locomotor) Walks up and down stairs.
mdsgmd004	(Locomotor) Goes alone on the stairs (any method)
mdsgmd005	Hands-and-knees crawling
mdsgmd006	Standing with assistance
ddifmm019	Walking with assistance
ddifmd154	Standing alone
vinxxc002	Walking alone
vinxxc003	chew solid foods
vinxxc009	take off socks / shoes
vinxxc012	get on with other children
vinxxc014	know what's edible
vinxxc022	walk upstairs
vinxxc028	avoid simple danger - knife / hot
vinxxc031	help around the house / clear table
vinxxc040	Play or do things with other children of same age eg sing song
ddifmm025	Help with little things around the house eg pick up things

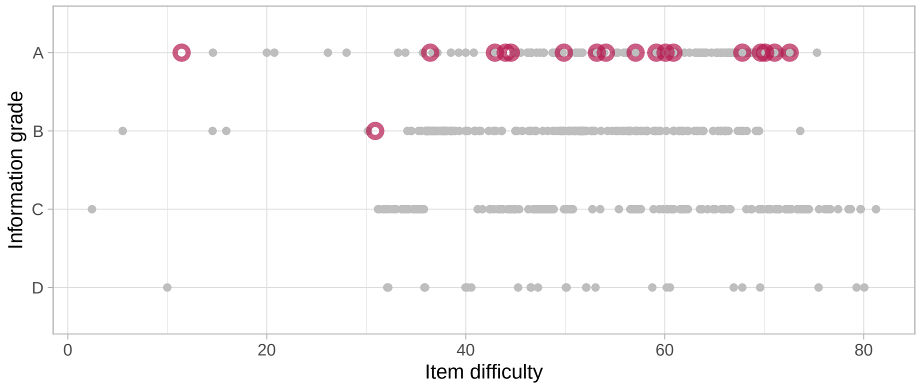

Figure 5.7: Item information grade by item difficulty for the final model

The red circles indicate active equate groups. Most have grade A, so we have a lot of information about the items that form the active equate groups. Table 5.2 displays more detailed information for the active equate groups. The sample sizes are reasonably large. Many information statistics are well is above 100; the criterion for Grade A. The interpretation of this criterion is as follows. Suppose that we obtain a sample of 400 persons who are all perfectly calibrated to the item of interest. In that case, the information for that item will be equal to 100.

5.9 Final model

Unfortunately, there is no single index of model fit that we can optimise. Modelling is more like a balancing act among multiple competing objectives, such as

- preserving as many items as possible that fit the model;
- finding high-quality active equate groups that span many cohorts and instruments;
- picking active equate groups for which we have enough information;
- providing reasonable age-conditional distributions of the D-score;
- representing various developmental domains in a fair way;
- preserving well-fitting historical models as new data become available;
- maintaining a reasonable calculation time.

This section showed various modelling techniques and ways to assess the validity of the model. In real life, we fitted a total number of 140 models on the data and made many choices that weigh the above objectives. The final model for the GCDG data consists of 565 items (originating from 14 instruments) that fit the Rasch model and that connect through 18 equate groups. Due to the

Table 5.2: Equate group information in the final model.

Equate group	Difficulty	Sample Size	Information	Grade
EXP2	11.4	3608	162.3	A
REC6	30.9	5428	95.4	B
GM25	36.4	6380	470.6	A
FM26	42.9	4155	296.8	A
GM35	44.0	5522	356.0	A
COG36	44.5	7912	230.0	A
GM42	49.9	5953	327.7	A
FM31	53.2	10991	731.7	A
COG55	54.1	5647	420.3	A
FM72	57.1	5430	253.6	A
EXP26	59.1	9119	578.8	A
SA1	60.1	3363	172.1	A
FM38	60.9	10236	491.7	A
FM52	67.8	13487	1159.9	A
FM43	69.7	15765	1563.9	A
GM60	70.1	9519	1070.6	A
REC40	71.0	10393	1182.9	A
FM61	72.6	10612	945.9	A

sparseness of data at the very young ages, the quality of the model is best for ages between 4-36 months.

Model `565_18` formed the basis of the publication by Weber et al. (2019). Additional detail on model `565_18` is available through the `dmodel` shiny app.

Chapter 6

Comparing ability

Once we identified a satisfactory D-score model, we may calculate the D-score for children from different cohorts and compare their values. This section highlights various techniques and issues for comparing D-score distributions between studies. We will address the following topics:

- Comparing child development across studies (6.1)
- Precision of the D-score (6.2)
- Domain coverage (6.3)

6.1 Comparing child development across studies

Figure 6.1 shows the scatterplot of the D-score by age separately for each cohort. Remember from section 2.1 that each study selected its own set of instruments to collect the data. The scatterplots demonstrate a significant advance made possible by the D-score: We can plot the developmental scores of children from **different** cohorts, with **different** ages, using **different** instruments, on the **same** vertical axis.

The five blue lines guide the eye. These lines indicate the locations of the -2SD, -1SD, 0SD, +1SD and +2SD quantiles at each age in the combined data. Section 5.4, in Chapter I motivates the idea and provides some technical details. We'll come back to these lines in section 7.2.

By and large, the data in every study follow the blue lines. Perhaps the most obvious exception is the `GCDG-JAM-STUNTED` cohort, where older children somewhat exceed the D-score range. It is unknown whether this is real, or due to a sub-optimal calibration of the instrument.

Figure 6.2 plots the same data with D-score transformed into age standardized scores (DAZ). Replacing the D-score by the DAZ emphasises the differences

Figure 6.1: D-score distributions by study (https://d-score.org/dbook-apps/gcdgdscores/).

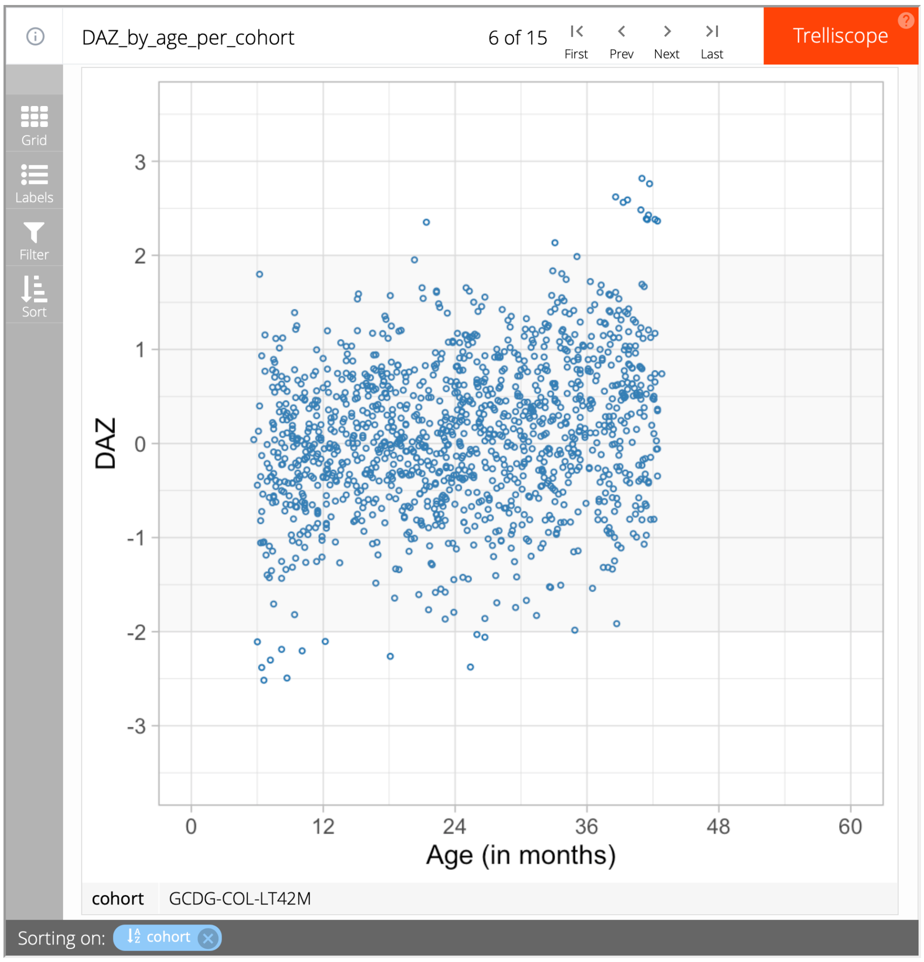

Figure 6.2: DAZ distributions by study (https://d-score.org/dbook-apps/gcdgdaz/).

both within and between studies. The majority of observations lies between the -2 SD and +2 SD lines in all cohorts. Using DAZ makes is easier to spot deviating trends, e.g., for the Jamaican or Ethiopian data.

6.2 Precision of the D-score

The EAP algorithm estimates the D-score from a set of PASS/FAIL scores. The standard deviation of the posterior distribution (or *sem*: standard error of measurement) quantifies the imprecision of the D-score estimate. The *sem* is inversely related to the number of items. Thus, when we administer more milestones, the *sem* of the D-score drops.

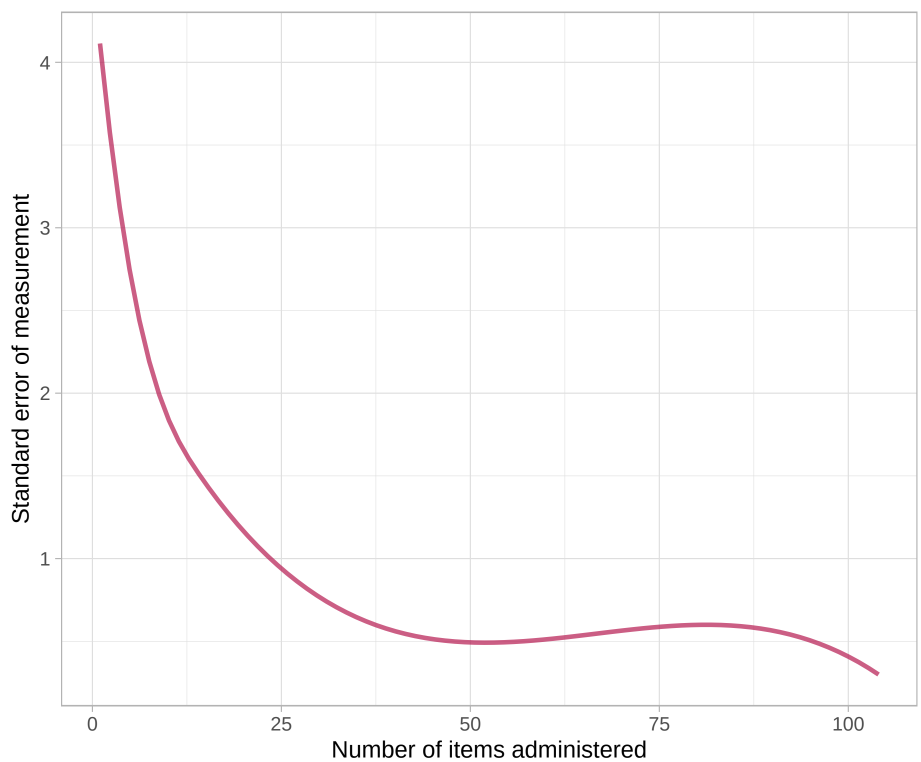

Figure 6.3: Standard error of measurement (*sem*) as a function of the number of items.

Figure 6.3 shows that the *sem* drops off rapidly when the number of items is low and stabilises after about 35 items. Apart from test length, the precision of the D-score also depends on item information (c.f. section 5.8). Administering items that are too easy, or too difficult, does not improve precision. The figure suggests that - in practice - a single D-score cannot be more precise than 0.5 D-score units.

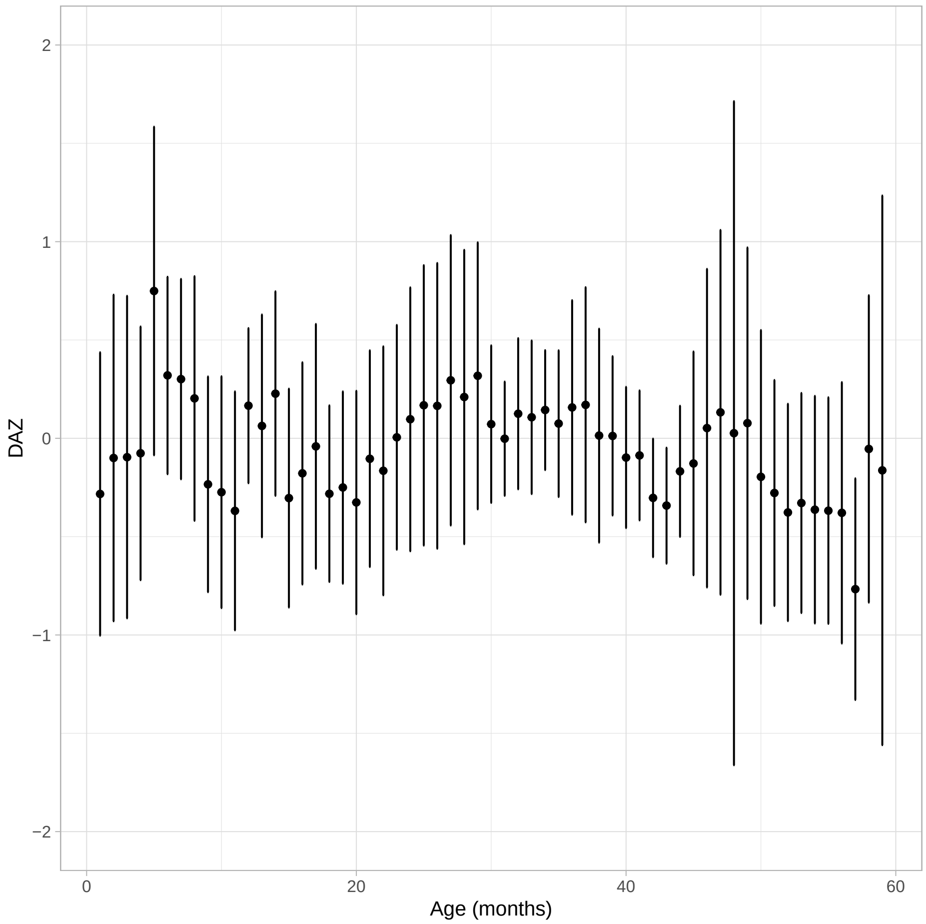

Figure 6.4: Mean DAZ ± *sem* as a function of age.

One may wonder whether the *sem* depends on age. Figure 6.4 suggests that this is not the case. The average DAZ is close to zero everywhere, as expected. The interval DAZ ± *sem* will cover the true, but unknown, DAZ in about 68% of the cases. While the interval varies somewhat across ages, there is no systematic age trend.

Does precision vary with studies? The answer is yes. Figure 6.5 plots the same information as before but now broken down according to cohort. Individual data points are added to give a feel for the design. The Colombia cohort `GCDG-COL-LT45M` administered the Bayley-III, where each child answered on average 45 items, so the *sem* is small. In contrast, the Dutch cohort GCDG-NLD-SMOCC collected data on a screener consisting of about ten relatively easy

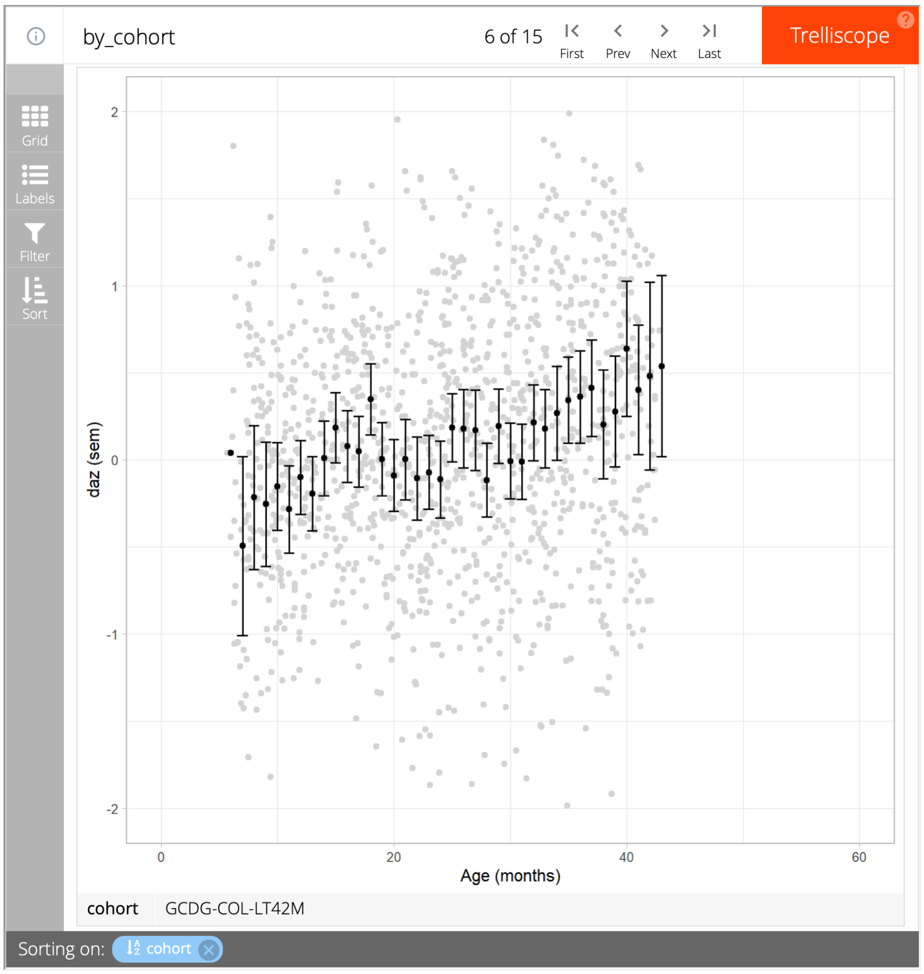

Figure 6.5: The standard error of measurement (*sem*) around the age-standardized D-scores (DAZ) per cohort (https://d-score.org/dbook-apps/gcdgsem).

milestones, so the *sem* is relatively large. As a result, the Colombian D-scores are much more precise than the Dutch.

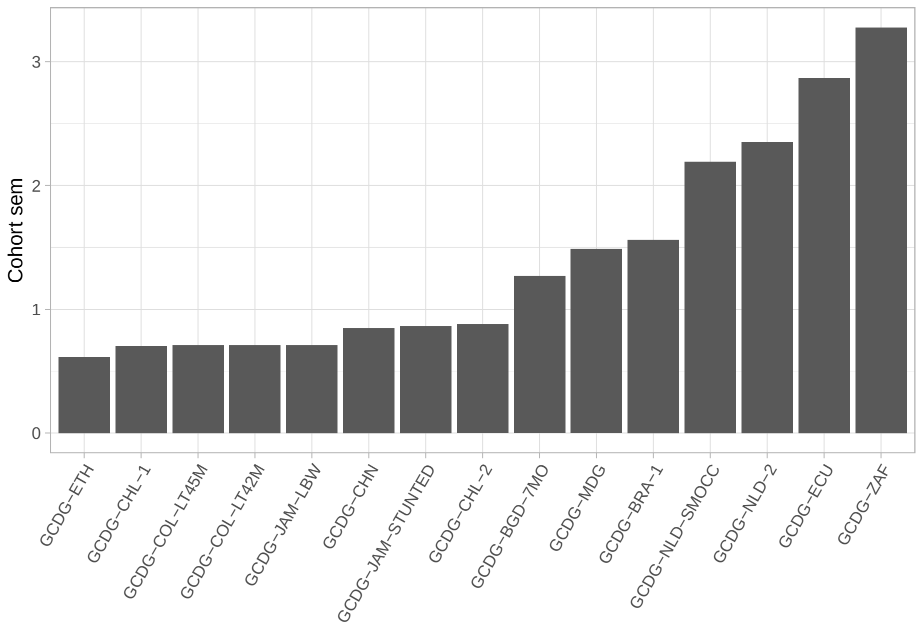

Figure 6.6: Cohort Standard Error of Measurement (*sem*).

The ordering of studies depends on test length and item information. Table 6.1 shows the median number of items per child (test length) and the probability to pass the item. The Ethiopian cohort `GCDG-ETH` administered 39 milestones with a median probability of 0.66. In contrast, the South Africa study `GCDG-ZAF` measures 12 items which were all very easy for the sample at hand (median probability of 1.0). One may thus well explain the extremes by test length and item information.

In general, the design of the study has a significant impact on the precision of the measurement. Our ongoing work addresses the question how one may construct a measurement instrument that will be optimally precise given the goals of the research.

6.3 Domain coverage

The D-score is a one-number summary of early child development. Traditional instruments distinguish domains (like motor, communication, language and cognitive development) and some provide ways to calculate a total score. The

Table 6.1: Test length and probability to pass the items per cohort

Cohort	Test length	Pass probability
GCDG-ETH	39	0.66
GCDG-CHL-1	32	0.67
GCDG-COL-LT45M	45	0.64
GCDG-COL-LT42M	61	0.62
GCDG-JAM-LBW	43	0.55
GCDG-CHN	27	0.50
GCDG-JAM-STUNTED	38	0.65
GCDG-CHL-2	33	0.48
GCDG-BGD-7MO	14	0.38
GCDG-MDG	8	0.35
GCDG-BRA-1	18	0.89
GCDG-NLD-SMOCC	10	0.80
GCDG-NLD-2	11	1.00
GCDG-ECU	3	0.67
GCDG-ZAF	12	1.00

D-score, on the other hand, is based on the notion that child development is a unidimensional latent construct and hence does not provide domain scores. And thus, the question is how the D-score represents domains.

This section explores the following two questions:

- Can we break down the D-score by domain contribution, and if so, can we evaluate whether the D-score fairly represents all domains?
- Can we calculate domain-specific D-scores?

6.3.1 Domain coverage of the scale

For many items in the D-score model, we had expert information available as to which domain the item belongs. For each item, we calculated the proportion of times the experts assigned it to one of five domains: Fine Motor, Gross Motor, Expressive, Receptive, Cognitive. We then calculated the distribution of domain by age.

Figure 6.7 shows the domain composition of the D-score across different levels of ability. Note that we miss domain information for a few items. The share of gross-motor is large in early development (e.g., between 15 and 30 months), and gradually tapers off at higher levels. Reversely, the percentage of cognition and language is relatively small before 30 months but rapidly rises as the child

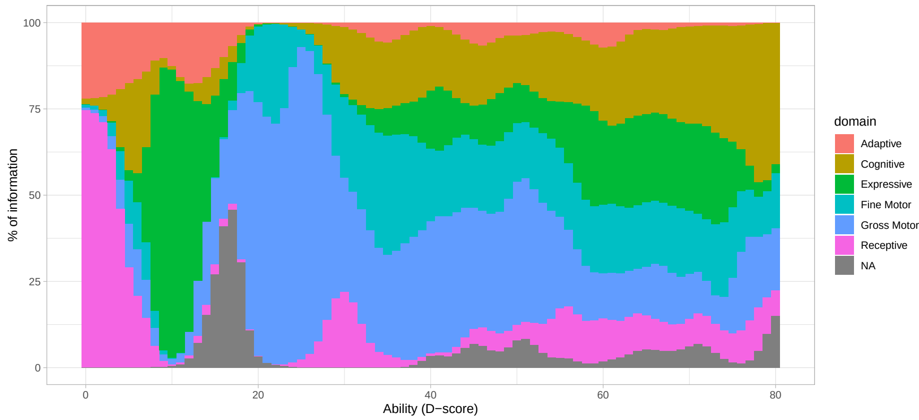

Figure 6.7: Domain coverage of the D-score scale.

matures. These transitions in domain composition look both reasonable and valid.

6.3.2 Domain-specific D-scores

Suppose we select a domain of interest and calculate the D-score only from items that substantially load onto that domain. We then get a domain-specific D-score. Items that relate to multiple domains contribute to multiple domain-specific D-scores.

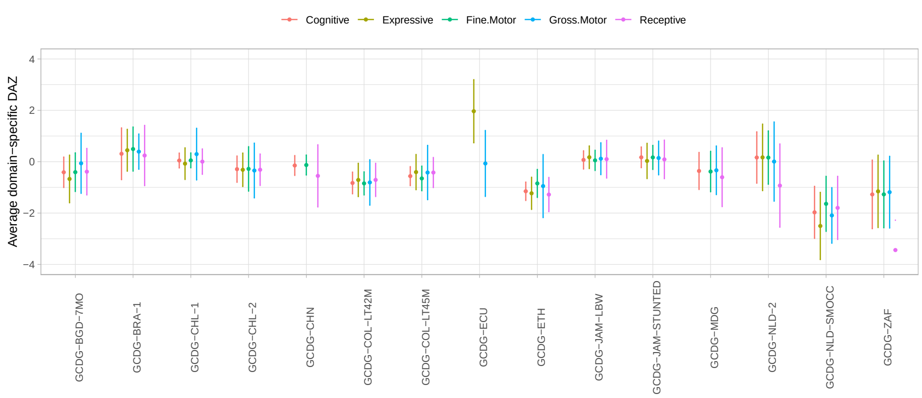

Figure 6.8: Average domain-specific DAZ ± *sem* by cohort.

Figure 6.8 displays the standardized domain-specific D-score (i.e. DAZ) per cohort. The DAZ strips out irrelevant age variation, and thus enhances comparability between cohorts. The error bars around the scores depict the *sem*

Table 6.2: Pearson correlation of the DAZ and five domain-specific DAZ scores

	DAZ	Fine motor	Gross Motor	Cognitive	Receptive	Expressive
DAZ	1.00	0.69	0.57	0.84	0.70	0.69
Fine motor	0.69	1.00	0.40	0.74	0.50	0.39
Gross Motor	0.57	0.40	1.00	0.43	0.34	0.30
Cognitive	0.84	0.74	0.43	1.00	0.76	0.59
Receptive	0.70	0.50	0.34	0.76	1.00	0.63
Expressive	0.69	0.39	0.30	0.59	0.63	1.00

interval. We observe some variation in domain-specific DAZ scores within cohorts. Still, these differences are relatively small and well within the margins of error. This analysis suggests that the D-score is an excellent overall summary of the domain-specific D-scores.

The D-score methodology assumes that child development is a unidimensional scale. As a consequence, the correlations between different domain-specific D-scores are extremely high ($r > 0.95$). It is more interesting to study the correlation between the DAZ equivalent of the domain-specific scores.

Table 6.2 lists the Pearson correlation matrix of the DAZ and the five domain-specific DAZ scores. All correlations between the DAZ and the domain-specific scores are high, thus confirming the generic character of the D-score and DAZ. We find high inter-domain correlations for the cognitive-receptive, cognitive-fine motor and expressive-receptive pairs. The gross motor domain appears as somewhat distinct from the four other domains. Its position may be genuine, but could also be related to the smaller number of responses on gross motor milestones in the GCDG data.

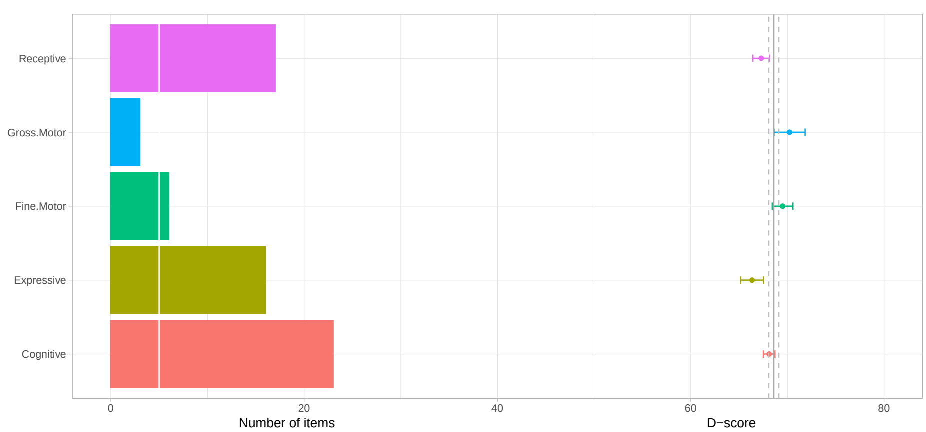

Figure 6.9: Domain-specific D-scores for a 3 year old boy.

Figure 6.9 displays individual scores for a 3 year old boy. The filled bars indicate the number of available items per domain. The vertical white line that crosses the horizontal axis at value 5 indicates a threshold for a minimum number of items needed for a D-score. Note that the number of items for Gross Motor in this example is meagre (only three items). The grey vertical line indicates the value of the overall D-score ($68.55D$). The nearby dashed lines are located at one *sem* ($0.53D$) distance. The coloured points are the domain-specific D-scores with the *sem* around in error bars. The plot visualises that the boys' scores on language domains (i.e. Expressive and Receptive) are low as compared to the motor and cognitive domains. A systematic discrepancy between various domain-specific scores might be an early warning sign for developmental delay.

Chapter 7

Application I: Tracking a Sustainable Development Goal

The Sustainable Development Goals (SDG) formulated by the United Nations (UN) set targets to promote prosperity while protecting the planet. One or more indicators quantify the progress towards each target.

This section explores the use of the D-score to monitor the progress of the indicator for healthy child development, SDG 4.2.1. We propose a method to define on-track development and show how the application of this method pans out for the GCDG data. More in detail, the section deals with the following topics:

- Estimating SDG 4.2.1 indicator from existing data (7.1)
- Defining *developmentally on track* (7.2)
- Country-level estimations (7.3)
- Relation to other estimates (7.4)

7.1 Estimating SDG 4.2.1 indicator from existing data

The UN Sustainable Development Goals form a universal call to action to end poverty, protect the planet and improve the lives and prospects of everyone, everywhere. All UN Member States adopted the 17 Goals in 2015. The SDG 4 target to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. SDG 4.2 reads as:

> By 2030, ensure that all girls and boys have access to quality early childhood development, care and preprimary education so that they are ready for primary education.

To measure progress, the UN defined indicator 4.2.1 as follows:

> Proportion of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being, by sex.

On July 22, 2020, the indicator was changed into

> Proportion of children aged 24-59 months who are developmentally on track in health, learning and psychosocial well-being, by sex.

The exclusion of children 0-24 months is at variance with the importance of healthy growth and development during the first 1000 days of life. Indeed, the UN restricted the age range for practical concerns. Loizillon et al. (2017) report:

> The initial recommendation was for the ECDI to measure child development from birth–5 years, but the range was restricted to 3–5 years due to time and resource constraints and limited availability of comparable measurement tools for children under age 3.

The careful scientific approach underlying the D-score fills the gap for children aged 0-24 months. Also, the D-score methodology enables extensions to ages beyond 24 months, permits back-calculation of D-scores from existing data, and acts as a linking pin to compare child development from birth onwards.

The cohorts included in the GCDG study represent a wide range of countries and instruments (see Section 2.1). Combining existing data from such a wide range of countries to create the D-score, is undoubtedly challenging, but doable. Although, in all fairness, we note that obtaining accurate comparisons between world-wide populations requires additional representative (existing) data beyond what is available here.

7.2 Defining *developmentally on track*

In 2006, the World Health Organisation (WHO) published the WHO Child Growth Standards. These standards specify "how children should grow" and form the basis for widely used anthropometric indicators such as stunting and wasting. We advocate a similar approach for child development. More in particular, the following steps:

1. Measure child development on an interval scale;
2. Estimate the age-conditional reference distribution for normal child development;
3. Define the indicator *developmentally on track* as the proportion above a chosen cut-off.

Step 1 is solved by the D-score. Step 2 borrows from well-tested statistical methodology for constructing growth standards (Borghi et al. 2006). Step 3 can be done in different ways, but a applying a simple cut-off fits easily with regular practice in reporting international comparisons.

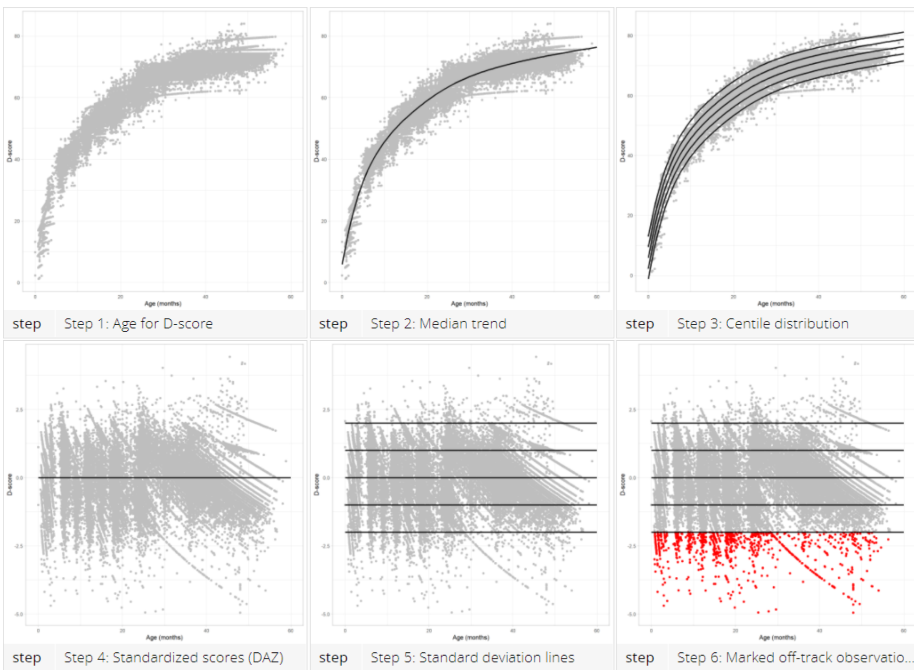

Figure 7.1: Illustration of the method to define on-track development (https://d-score.org/dbook-apps/gcdgreferences/).

Figure 7.1 demonstrates steps 2 and 3 in more detail. Click 'Next' to advance a series of six steps:

1. Plot the D-score by age;
2. Model the relation between age and D-score by an LMS model. In practice, this amounts to smoothing three curves representing the median, coefficient of variation and the skewness.
3. Present the centile lines for the model;
4. Plot the age-standardized scores for development (DAZ);

Table 7.1: Percentage of on-track children per country

Country	Percentage on-track
BGD	94.9
BRA	99.5
CHL	98.3
CHN	99.9
COL	98.9
ECU	93.9
ETH	99.4
JAM	99.6
MDG	96.6
NLD	96.8
ZAF	97.4

5. Draw standard deviation lines to indicate the location at ± 1 and ± 2 standard deviation from the mean;
6. Count observations above the -2 SD line as on-track. Count observation below the -2 SD lines as off-track (red dots).

Note: These SD lines build upon on a convenience sample. The GCDG cohorts are not representative samples, and the countries are not representative of the world. While we should not over-interpret these references, they play a central role in a stepwise, principled approach to define "developmentally on track."

7.3 Country-level estimations

Using the definition from the previous section, we can calculate the percentage of children that are developmentally on track. Table 7.1 summarises this statistic by country. At a cut-off value of -2 SD, we expect that about 97.7% of the children will be on track. The actual country estimates fall into the range 93.9 - 99.9 and are thus near the theoretical value. This close correspondence shows that the definition and estimation procedure work as expected.

Bear in mind that the measurements leading up to these estimates come from different instruments. It is gratifying to see how well we can do with historical data, thanks to the robust underlying measurement model. Of course, comparability only gets better if all countries would use the same instrument. However, using the same tool everywhere is not a requirement.

7.4 Off-track development and stunted growth

Weber et al. (2019) thoroughly discuss concurrent, discriminant and predictive validity of the D-score using the GCDG data. In this section, we concentrate on the relation between the D-score and stunting, a popular measure of impaired height growth in children due to nutrition problems. The WHO defines stunted growth as a height-for-age Z-score below the -2 SD line of the WHO Child Growth Standards (HAZ < -2.0).

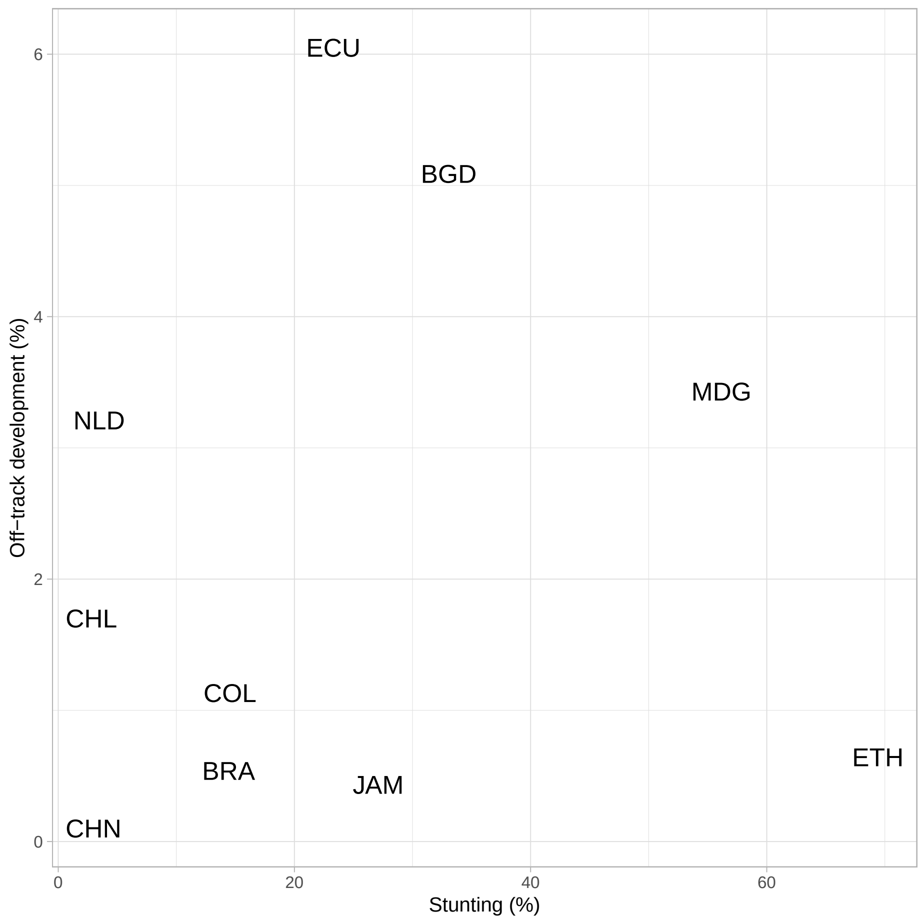

Figure 7.2: Off-track development (%) versus stunting (%) per country

Figure 7.2 plots the percentage off-track and percentage stunting per country. This plot reveals two exciting features:

- *The variation in stunting is much larger than the variation off-track development.* One might speculate that height is more dependent on the environment than off-track development, and hence more variable.

- *Stunted growth and off-track development are unrelated.* Ranking countries by stunting or by off-track development yields substantially different orders. This finding provides clear counter-evidence to the argument that stunted growth is as a proxy for delayed development. It may even be the case the child development and physical growth are different maturation processes that develop largely independently.

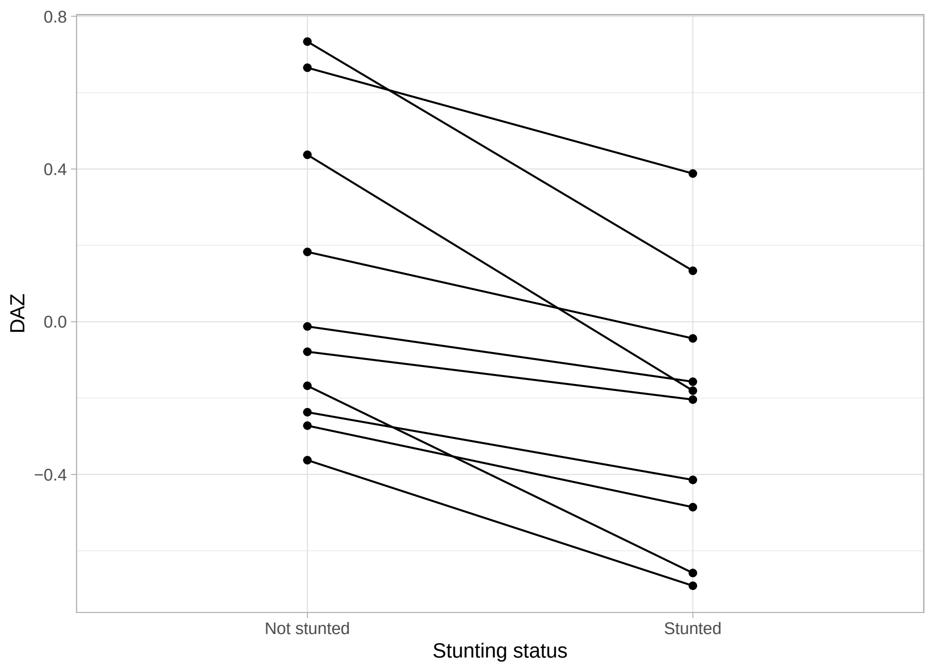

Figure 7.3: Difference in mean DAZ per country between stunted and not stunted children.

However, this is not the whole story. Figure 7.3 reveals a consistent difference in DAZ between stunted and non-stunted children of about 0.2 - 0.3 SD. There could be factors at the child level that affect both development and height growth. For example, low-income families may lack the resources for adequate nutrition, which may impact both child development and physical growth.

The exact nature of the relation between stunting and development is still obscure. The D-score provides a means to study the intriguing interplay between both measures in more detail.

Chapter 8

Application II: Who is on-track?

Section 7 described a method to define and estimate off-track development. The current section highlights strategies to find factors that discriminate between children that are on-track and off-track. We order explanatory factors relative to their importance and discuss opportunities for interventions.

- What determines who is developmentally on-track (8.1)
- Factors that impact child development (8.2)

8.1 What determines who is developmentally on-track?

There are multiple ways to define on-track development. Here we will use the method outlined in section 7.2. Ideally, we would like to fit the age-conditional reference distribution on a sample of children with normal, healthy development. As noted before, we calculated the references used in section 7.2 from a convenience sample. They may not be representative of healthy development.

Assuming we place the cut-off value at -2 SD, we may subdivide the observed D-scores into off-track and on-track. Figure 8.1 colours the regions of the D-score for children considered on-track (green) and off-track (red). The regions indicate the expected locations of D-scores in practice. Although one could find D-score outside the coloured areas, such should be very rare. The occurrence of such cases may indicate an error in the calculation of the D-score, most likely caused by setting an incorrect age variable.

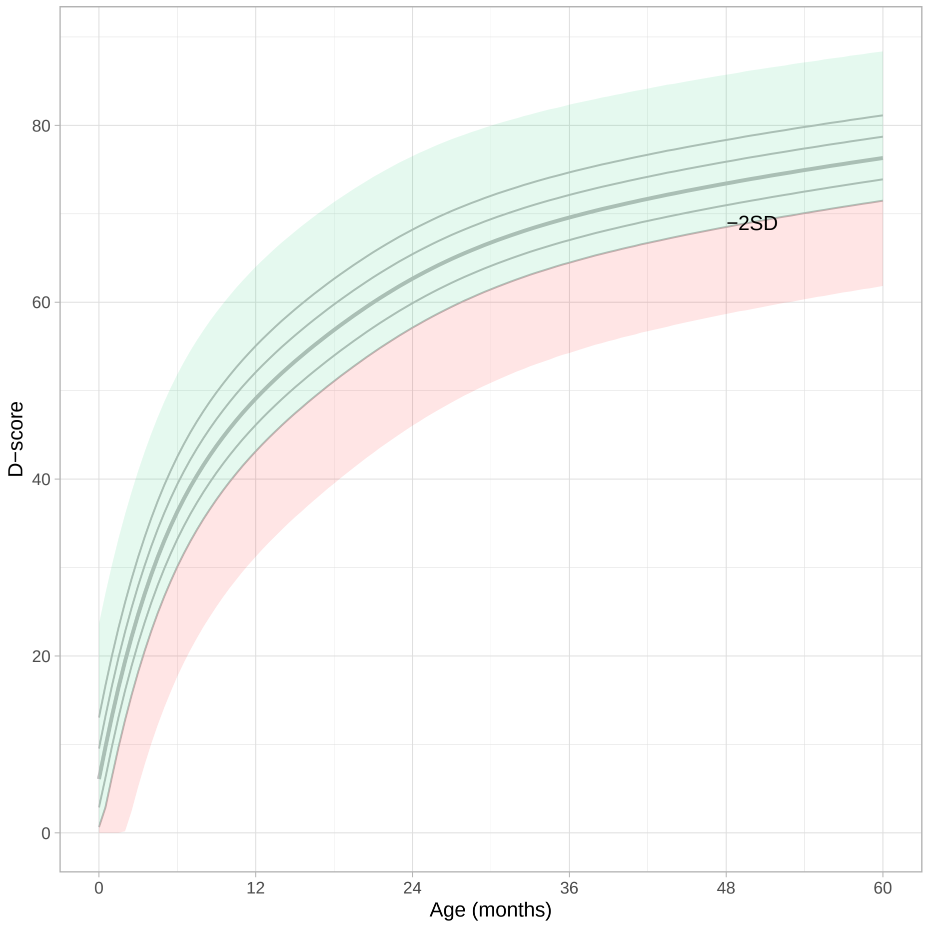

Figure 8.1: D-score observatations that are on-track according the current references.

Table 8.1: Comparisons between on-track and off-track development

		On-track		Off-track	
		n	%	n	%
sex	female	21136	0.977	489	0.023
	male	20805	0.972	595	0.028
birth weight	<2500gr	3388	0.948	185	0.052
	>2500gr	36375	0.978	821	0.022
maternal education	no education	1907	0.967	66	0.033
	any primary	11764	0.967	398	0.033
	any secondary	21576	0.977	503	0.023
	higher secondary	6263	0.984	101	0.016
residence	rural	1251	0.989	14	0.011
	semi-urban	2236	0.990	23	0.010
	urban	18740	0.971	566	0.029
	metropolitan	11122	0.979	234	0.021

* Exludes children with missing DAZ or missing factor

Preventing observations in the red region requires us to form an idea about the factors that determine the off-track probability. The next section looks into this topic.

8.2 Factors that impact child development

We already know many of the factors that influence early child development. A higher level of education in the family promotes development. Infectious diseases like malaria slow down growth. Access to adequate nutrition, clean water and a stimulating, prosperous and safe environment is favourable for healthy development. And so on. Unfortunately, we do not have data on most factors, so we need to limit ourselves to a few background characteristics.

Table 8.1 compares the frequency distributions of various factors for children on-track versus off-track. There are only tiny differences between boys and girls. Children with low birth weight ($<$ 2500 gr) are more at risk for off-track development. This estimate does not correct for gestational age. We discussed techniques for such corrections elsewhere.

The influence of maternal education on off-track development follows the expected trend. Interestingly, it seems that a rural environment could prevent off-track development. We note that original measures of maternal education and residence were harmonised across studies. It would, therefore, also be interesting to study the impact per cohort using the actual factor coding.

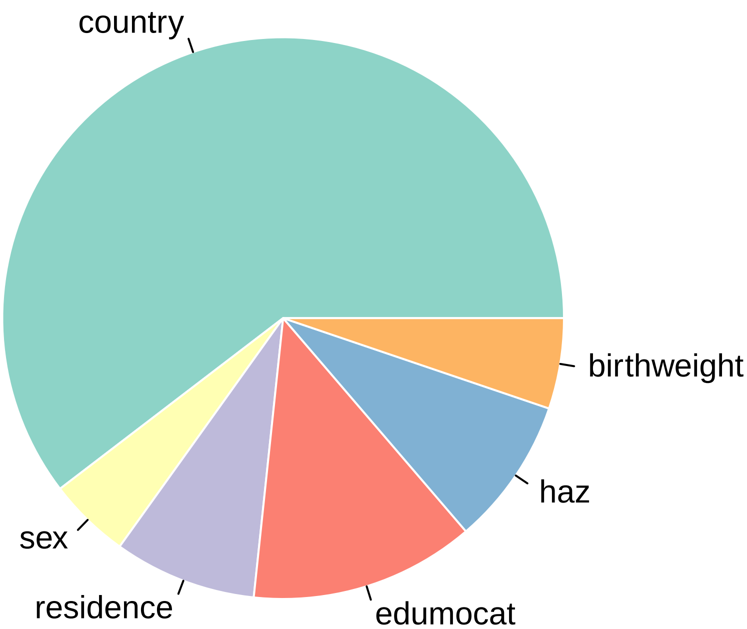

Figure 8.2: Relative importance of the explanatory factors in this study

We predicted DAZ by linear regressions with predictors country, sex, birth weight, maternal education, height for age and residential area. The percentage of explained variance was 11 percent. Figure 8.2 depicts the relative contributions of the individual factors to the prediction. Country differences explain over half the variances, followed by maternal education. Contributions of height-for-age (HAZ), low birth weight and residence are about equal in magnitude.

These analyses only scratch the surface. It is nowadays common to analyse the impact of interventions on height and HAZ by multivariate techniques and machine learning methods. The D-score and DAZ are drop-in replacements that allow similar procedures to study which factors contribute to healthy child development worldwide.

Chapter 9

Discussion

This closing section briefly summarises the key lessons from this section. The section covers:

- D-score from multiple instruments (9.1)
- Variability within and between cohorts (9.2)
- D-score for international comparisons (9.3)
- Better measurement (9.4)

9.1 D-score from multiple instruments

We developed the initial D-score methodology for just one instrument. In practice, however, we need to deal with data collected on multiple, partially overlapping tools. This chapter addressed the problem *how to define and calculate the D-score based on data coming from various sources, using multiple instruments administered at varying ages.*

We had longitudinal data available from 16 cohorts, collected with 15 tools to measure child development at various ages. Our analytic strategy to define a D-score from these data consists of the following steps:

1. Make an inventory of instruments and cohorts;
2. Combine all measurements into one dataset;
3. Find out which shared instruments connect cohorts;
4. Place similar items from different instruments into equate groups;
5. Find the best set of *active* equate groups;
6. Estimate item difficulty using a restricted Rasch model that requires the estimates of all items within an active equate group to be identical;
7. Weed out items that do not fit the model.

We need to perform steps 5, 6 and 7 in an iterative fashion. Depending on the result, we may also need to redefine, combine or break up equate groups (step 4).

These techniques are well-known within psychometrics and educational research. Our approach builds upon a well-grounded and robust theory of psychological measurement. We, therefore, expect that repeating our method on other data will lead to very similar results.

A novel aspect in our methodology is the systematic formation of candidate equate groups by subject-matter experts based on similarity in concept and content. Our subsequent testing and tailoring of each equate group given the data provide empirical evidence of its quality for connecting instruments. While anchoring tests by itself is not novel, we are not aware of any work aimed at identifying the best set of active equate groups on this scale.

9.2 Variability within and between cohorts

The final model retains 565 items and employs 18 equate groups. Given the difficulty estimates from that model, we can estimate the D-score and DAZ for each measurement.

Figure 6.1 reveals that all cohorts show a rapidly rising age trend in the D-score, which matches the earlier finding that child development is faster in younger children.

Figure 6.2 shows large overlaps in the DAZ distributions between cohorts. This finding suggests that the level of child development is similar in different regions of the world. Some studies display more variability in DAZ than others, which is likely to be related to differences in measurement error, as the number of milestones differs widely.

Observe that we used all cohorts for modelling, which may have made them appear more similar than they are. It would be good if we could verify the apparent similarities in level and variability of child development in different regions by other data that were not part of the modelling.

9.3 D-score for international comparisons

The D-score is a universal scale of early child development. The D-score does not depend on a particular instrument. Instead, we can calculate a D-score as long as appropriate difficulty estimates are available for the tool at hand. This feature makes the D-score methodology flexible and helpful for international comparisons.

Of course, the ideal situation for international comparisons would be that all countries collect child development data in the same way. In practice, this ideal may be difficult to achieve. Also, we cannot change past data. In these less-than-ideal worlds, the D-score presents a convenient, conscientious and timely alternative.

As an example, we outlined a generic strategy on how to advance on SDG 4.2.1. We use the D-score to operationalise the concept *developmentally on track*. We calculated age-conditional references of the D-score, analogous to the WHO Multicentre Growth Reference Study. We may then define cut-off values. Children above the cut-off then count as developmentally on track.

While we highlighted the principles, much work still needs to be done. First, there are over 150 instruments for child development, and our current key covers only a fraction of these. We are actively expanding the key using additional data, so as time passes the coverage of tools will go up. Second, we calculated the references on a mix of studies, some of which include special populations. Thus, we cannot interpret the current reference values as portraying normal development. We hope that the inclusion of healthy population data will improve the usefulness of the references as a standard for child development.

9.4 Better measurement

The D-score metric is a generic measure of child development. It summarises child development by *one number*. We found that D-score fairly represents development domains over the entire scale. Due to its generic nature, the D-score is less suitable for measuring a specific domain. It may then be better to use a specialised tool that accesses motor, cognitive or communication faculties. For example, think of sub-scales from the Bayley, ASQ, Griffiths, and so on. Note that also in those cases, one still has the option of calculating a D-score.

The opposite scenario may also be of interest. Suppose we want to measure generic development AND identify any areas of slow growth. Extending the measurement by adding more items from domains with a higher failure rate will then increase precision in areas of suspected delay.

Since we based the D-score on a statistical model, we may create instruments customised to the exact needs of the study. Population-based studies may require a short measure consisting of a handful of items per child, and aggregate scores over many children to achieve precision. Intervention studies aim for a precise estimate for the intervention effect. If group sizes are small, we may administer a more extended test to achieve the same precision and vice versa. At the other end of the spectrum, for clinical purposes, we want a precise estimate for one particular person, so here we will administer a relatively long test. The good news is: As long as we pick items from the statistical model, the D-score in those three cases are all values on the same scale.

Our ongoing work targets tailoring instruments to a study design and discusses all of these options. And more.

Appendix A

Abbreviations

Section	Abbreviation	Description
2.2	GCDG-BGD-7MO	The Bangladesh study of the GCDG (Tofail et al. 2008)
2.2	GCDG-BRA-1	The Brazil 1 study of the GCDG (Victora et al. 2006)
2.2	GCDG-BRA-2	The Brazil 2 study of the GCDG (Moura et al. 2010)
2.2	GCDG-CHL-1	The Chile 1 study of the GCDG (Lozoff et al. 2003)
2.2	GCDG-CHL-2	The Chile 2 study of the GCDG (Conteras and González 2015)
2.2	GCDG-CHN	The China study of the GCDG (Lozoff et al. 2016)
2.2	GCDG-COL-LT45M	The Colombia 1 study of the GCDG (Attanasio et al. 2014)
2.2	GCDG-COL-LT42M	The Colombia 2 study of the GCDG (Rubio-Codina et al. 2016)
2.2	GCDG-ECU	The Ecuador study of the GCDG (Paxson and Schady 2010)
2.2	GCDG-ETH	The Ethiopia study of the GCDG (Hanlon et al. 2009)
2.2	GCDG-JAM-LBW	The Jamaica 1 study of the GCDG (Walker et al. 2004)
2.2	GCDG-JAM-STUNTED	The Jamaica 2 study of the GCDG (Grantham-McGregor et al. 1991)
2.2	GCDG-MDG	The Madagascar study of the GCDG (Lia C. H. Fernald et al. 2011)
2.2	GCDG-NLD-SMOCC	The Netherlands 1 study of the GCDG (Herngreen et al. 1992)

Section	Abbreviation	Description
2.2	GCDG-NLD-2	The Netherlands 2 study of the GCDG (Doove 2010)
2.2	GCDG-ZAF	The South Africa study of the GCDG (Richter et al. 2007)
2.3	by1	Bayley Scale for Infant and Todler Development version 1 (Bayley 1969)
2.3	by2	Bayley Scale for Infant and Todler Development version 2 (Bayley 1993)
2.3	by3	Bayley Scale for Infant and Todler Development version 3 (Bayley 2006)
2.3	den	Denver Developmental Screening Test (Frankenburg et al. 1992)
2.3	gri	Griffiths Mental Development Scales (Griffiths 1967)
2.3	bat	Battelle Developmental Inventory (Newborg 2005)
2.3	vin	Vineland Social Maturity Scale (Doll 1953)
2.3	ddi	Dutch Developmental Instrument (Schlesinger-Was 1981)
2.3	bar	Barrera Moncada (Barrera Moncada 1981)
2.3	tep	Test de Desarrollo Psicomotor (Haeussler and Marchant 1999)
2.3	aqi	Ages and Stages Questionnaire (Squires and Bricker 2009)
2.3	sbi	Stanford Binet Intelligence Scales (Roid 2003)

Appendix B

Notation

The notation in this chapter follows Wright and Masters (1982).

Section	Symbol	Term	Description
4.4	β_n	Ability	True (but unknown) developmental score of child n
4.4	δ_i	Difficulty	True (but unknown) difficulty of item i
4.4	δ_q	Difficulty	The combined difficulty of the items in equate group q
4.4	π_{ni}	Probability	True (but unknown) probability that child n passes item i
4.4	l	Count	The number of items in the equate group
4.4	w_i	Count	The number of respondents with an observed score on item i
4.6	P_{ni}	Probability	Estimated probability that child n passes item i
4.6	x_{ni}	Data	Observed response of child n on item i, 0 or 1
4.6	W_{ni}	Variance	Variance of x_{ni}
4.6	z_{ni}	Residual	Standardized residual between x_{ni} and P_{ni}
4.6	N_i	Count	Number of responses on item i
5.6	r	Correlation	Correlation coefficient
6	D	Score	Developmental score of a child: D-score
6.2	sem	Error	Standard Error of Measurement: precision of the D-score

Appendix C

Acknowledgements

This book was supported by the Bill & Melinda Gates Foundation. The contents are the sole responsibility of the authors and may not necessarily represent the official views of the Bill & Melinda Gates Foundation or other agencies that may have supported the primary data studies used in the present chapter. The authors wish to recognize the principal investigators and their study team members for their generous contribution of the data that were used to illustrate this booklet and the members of the Global Child Development Group who directly or indirectly contributed to the contents of this chapter: Maureen M. Black, Stef van Buuren, Gary L. Darmstadt, M. Caridad Araujo, Susan M. Changm, Bernice M. Doove, Iris Eekhout, Lia C.H. Fernald, Emanuela Galasso, Sally Grantham-McGregor, Pamela Jervis, Jena D. Hamadani, Charlotte Hanlon, Simone M. Karam, Betsy Lozoff, Orazio Attanasio, Girmay Medhin, Ana M. B. Menezes and the 1993 Pelotas cohort team, Helen Pitchik, Lisy Ratsifandrihamanana, Sarah Reynolds, Linda Richter, Marta Rubio-Codina, Norbert Schady, Susan P. Walker, Ann M. Weber.

References

Attanasio, Orazio P, Camila Fernández, Emla O A Fitzsimons, Sally M Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2014. "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial." *BMJ (Clinical Research Ed.)* 349 (sep29 5): g5785. https://doi.org/10.1136/bmj.g5785.

Barrera Moncada, G. 1981. "Crecimiento y Desarrollo Psicológico Del Niño Venezolano."

Bayley, N. 1969. "Bayley Scales of Infant Development."

———. 1993. "The Bayley Scales of Infant Development-II."

———. 2006. "Bayley Scales of Infant and Toddler Development–Third Edition: Technical Manual."

Bellman, M., O. Byrne, and R. Sege. 2013. "Developmental Assessment of Children." *BMJ* 346 (e8687).

Borghi, E, M de Onis, C Garza, J Van den Broeck, E A Frongillo, L Grummer-Strawn, S Van Buuren, et al. 2006. "Construction of the World Health Organization child growth standards: selection of methods for attained growth curves." *Statistics in Medicine* 25 (2): 247–65. https://doi.org/10.1002/sim.2227.

Britto, P. R., S. J. Lye, K. Proulx, A. K. Yousafzai, S. G. Matthews, T. Vaivada, R. Perez-Escamilla, N. Rao, P. Ip, and L. C. H. Fernald. 2017. "Nurturing Care: Promoting Early Childhood Development." *The Lancet* 389 (10064): 91–102.

Conteras, D., and S. González. 2015. "Determinants of early child development in Chile: Health, cognitive and demographic factors." *International Journal of Education and Development* 40: 217–30.

Doll, E. A. 1953. "The Measurement of Social Competence: A Manual for the Vineland Social Maturity Scale."

Doove, B. M. 2010. "Ontwikkeling kinderen in Maastricht en Heuvelland (MOM), Evaluatie integraal kindvolgsysteem voor signalering in de Jeugdgezondheidszorg: MOMknowsbest."

Fernald, L. C. H., E. Prado, P. Kariger, and A. Raikes. 2017. "A Toolkit for Measuring Early Childhood Development in Low and Middle-Income Countries."

Fernald, Lia C H, Ann Weber, Emanuela Galasso, and Lisy Ratsifandrihamanana. 2011. "Socioeconomic gradients and child development in a very low income population: evidence from Madagascar." *Developmental Science* 14 (4): 832–47. https://doi.org/10.1111/j.1467-7687.2010.01032.x.

Frankenburg, W. K., J. Dodds, P. Archer, H. Shapiro, and Bresnick B. 1992. "The Denver II: A Major Revision and Restandardization of the Denver Developmental Screening Test." *Pediatrics* 89: 91–98.

Frankenburg, W. K., J. Dodds, P. Archer, H. Shapiro, and B. Bresnick. 1990. "The DENVER II Technical Manual."

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

Grantham-McGregor, S M, C A Powell, S P Walker, and J H Himes. 1991. "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study." *Lancet (London, England)* 338 (8758): 1–5. https://doi.org/10.1016/0140-6736(91)90001-6.

Griffiths, R. 1967. "The Abilities of Babies: A Study in Mental Measurement."

Haeussler, I. M., and T. Marchant. 1999. "Tepsi: Test de Desarrollo Psicomotor 2-5 años."

Hagen, E., and J. Stattler. 1986. "Stanford–Binet Intelligence Scales, Fourth Edition."

Hanlon, Charlotte, Girmay Medhin, Atalay Alem, Fikru Tesfaye, Zufan Lakew, Bogale Worku, Michael Dewey, et al. 2009. "Impact of antenatal common mental disorders upon perinatal outcomes in Ethiopia: the P-MaMiE population-based cohort study." *Tropical Medicine & International Health* 14 (2): 156–66. https://doi.org/10.1111/j.1365-3156.2008.02198.x.

Herngreen, W. P., J. D. Reerink, B. M. van Noord-Zaadstra, S. P. Verloove-Vanhorick, and J. H. Ruys. 1992. "The SMOCC-Study: Design of a Representative Cohort of Live-Born Infants in the Netherlands." *European Journal of Public Health* 2: 117–22.

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/.

Kim, Seock-Ho, and Allan S Cohen. 1998. "A Comparison of Linking and Concurrent Calibration Under Item Response Theory." *Applied Psychological Measurement* 22 (2): 131–43.

Loizillon, A, N Petrowski, P Britto, and C Cappa. 2017. "Development of the Early Childhood Development Index in MICS Surveys. MICS Methodological Papers, No. 6. Data and Analytics Section, Division of Data." *Research and Policy. New York: UNICEF.*

Lozoff, Betsy, Isidora De Andraca, Marcela Castillo, Julia B Smith, Tomas Walter, and Paulina Pino. 2003. "Behavioral and developmental effects of preventing iron-deficiency anemia in healthy full-term infants." *Pediatrics* 112 (4): 846–54. http://www.ncbi.nlm.nih.gov/pubmed/14523176.

Lozoff, Betsy, Yaping Jiang, Xing Li, Min Zhou, Blair Richards, Guobin Xu, Katy M Clark, et al. 2016. "Low-Dose Iron Supplementation in Infancy Modestly Increases Infant Iron Status at 9 Mo without Decreasing Growth or Increasing Illness in a Randomized Clinical Trial in Rural China." *The Journal of Nutrition* 146 (3): 612–21. https://doi.org/10.3945/jn.115.223917.

Moura, Danilo R, Jaderson C Costa, Iná S Santos, Aluísio J D Barros, Alicia Matijasevich, Ricardo Halpern, Samuel Dumith, Simone Karam, and Fernando C Barros. 2010. "Natural history of suspected developmental delay between 12 and 24 months of age in the 2004 Pelotas birth cohort." *Journal of Paediatrics and Child Health* 46 (6): 329–36. https://doi.org/10.1111/j.1440-1754.2010.01717.x.

Newborg, J. 2005. "Battelle Developmental Inventory-2nd Edition."

Paxson, Christina, and Norbert Schady. 2010. "Does money matter? The effects of cash transfers on child development in rural Ecuador." *Economic Development and Cultural Change* 59 (1): 187–229. http://www.ncbi.nlm.nih.gov/pubmed/20821896.

Richter, Linda, Shane Norris, John Pettifor, Derek Yach, and Noel Cameron. 2007. "Cohort Profile: Mandela's children: the 1990 Birth to Twenty study in South Africa." *International Journal of Epidemiology* 36 (3): 504–11. https://doi.org/10.1093/ije/dym016.

Roid, G. H. 2003. "Stanford–Binet Intelligence Scales, Fifth Edition."

Rubio-Codina, Marta, M Caridad Araujo, Orazio Attanasio, Pablo Muñoz, and Sally Grantham-McGregor. 2016. "Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies." Edited by David O. Carpenter. *PloS One* 11 (8): e0160962. https://doi.org/10.1371/journal.pone.0160962.

Schlesinger-Was, E. A. 1981. "Ontwikkelingsonderzoek van Zuigelingen En Kleuters Op Het Consultatiebureau."

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

Press.

Squires, J., and D. Bricker. 2009. “Ages & Stages Questionnaires, Third Edition (ASQ- 3). A Parent-Completed Child-Monitoring System.”

Tofail, Fahmida, Lars Åke Persson, Shams El Arifeen, Jena D Hamadani, Ferdina Mehrin, Deborah Ridout, Eva-Charlotte Ekström, Syed N Huda, and Sally M Grantham-McGregor. 2008. “Effects of prenatal food and micronutrient supplementation on infant development: a randomized trial from the Maternal and Infant Nutrition Interventions, Matlab (MINIMat) study.” *The American Journal of Clinical Nutrition* 87 (3): 704–11. https://doi.org/10.1093/ajcn/87.3.704.

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/.

Victora, Cesar Gomes, Cora Luiza Pavin Araújo, Ana Maria Batista Menezes, Pedro Curi Hallal, Maria de Fátima Vieira, Marilda Borges Neutzling, Helen Gonçalves, et al. 2006. “Methodological aspects of the 1993 Pelotas (Brazil) Birth Cohort Study.” *Revista de Saude Publica* 40 (1): 39–46. https://doi.org/10.1590/s0034-89102006000100008.

Walker, Susan P, Susan M Chang, Christine A Powell, and Sally M Grantham-McGregor. 2004. “Psychosocial intervention improves the development of term low-birth-weight infants.” *The Journal of Nutrition* 134 (6): 1417–23. https://doi.org/10.1093/jn/134.6.1417.

Weber, Ann M, Marta Rubio-Codina, Susan P Walker, Stef van Buuren, Iris Eekhout, Sally M Grantham-McGregor, Maria Caridad Araujo, et al. 2019. “The D-score: A Metric for Interpreting the Early Development of Infants and Toddlers Across Global Settings.” *BMJ Global Health* 4 (6). https://doi.org/10.1136/bmjgh-2019-001724.

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