ELSEVIER

Infant Behavior & Development 25 (2002) 340–374

Infant Behavior & Development

Focus on variability: New tools to study intra-individual variability in developmental data

P. van Geert*, M. van Dijk

Department of Developmental Psychology, The Heijmans Institute, University of Groningen, Grote Kruisstraat 2/1, Groningen TS 9712, The Netherlands

Received 1 February 2001; received in revised form 28 January 2002; accepted 15 February 2002

Abstract

In accordance with dynamic systems theory, we assume that variability is an important developmental phenomenon. However, the standard methodological toolkit of the developmental psychologist offers few instruments for the study of variability. In this article we will present several new methods that are especially useful for visualizing and describing intra-individual variability in individual time-serial data of repeated observations. In order to illustrate these methods, we apply them to data of early language development. After reviewing the common techniques and measures, we present new methods that show variability in developmental time-series data: the moving min–max graph, and the progmax–regmin graph. In addition, we demonstrate a technique that is able to detect sudden increases of variability: the critical frequency method. Also, we propose a technique that is based on a central assumption of the measurement-error-hypothesis: namely the symmetric distribution of error. Finally, as traditional statistical techniques have little to offer in testing variability hypotheses, we examine the possibilities that are provided by random sampling techniques. Our aim with the present discussion of variability and the demonstration of some simple yet illustrative techniques is to help researchers focus on rich additional sources of information that will lead to more interesting hypotheses and more powerful testing procedures, adapted to the unique nature of developmental data.

Keywords: Intra-individual variability; Dynamic systems theory; Early language acquisition; Development; MLU; Spatial prepositions; Statistical techniques; Random sampling techniques

* Corresponding author. Tel.: +31-50-3636397; fax: +31-50-3636304.
E-mail address: p.l.c.van.geert@ppsw.rug.nl (P. van Geert).

0163-6383/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S0163-6383(02)00140-6

1. Introduction

1.1. Overview

In the field of developmental psychology, intra-individual variability is often neglected as a meaningful phenomenon. In our perspective, which has been inspired by dynamic systems theory, variability is viewed as a potential driving force of development and a potential indicator of ongoing processes. It should therefore be treated as an important source of information.

In this introduction, we provide an overview of the theoretical issues and discuss the traditional vs. the more current approach to variability. We go on with a short overview of studies from the different domains of developmental psychology that have taken up on the study of variability. In spite of the potential importance of intra-individual variability, there are only few tools for presenting and studying variability in the context of developmental data sets. Our aim is to introduce and discuss a number of relatively simple techniques and approaches for specifying intra-individual developmental variability. In the major part of this article, we will introduce these techniques by applying them to a dataset from language development.

1.2. Theoretical issues: traditional vs. current views on variability

Developmental psychology has a long tradition of focusing on the regular, gradual aspects of change. Until today, the majority of developmental studies show smoothed developmental trajectories of the variable under investigation. Although it is seldom explicitly mentioned, the almost automatic retreat towards a smoothed trajectory testifies of a certain suspicion towards the meaningfulness of the actual data and a belief that the average captures the underlying true level better. In recent years, several authors have warned against the untimely use of statistical compression techniques and have strongly recommended a more descriptive, exploratory approach, with an emphasis on smart ways of graphically presenting the data (Loftus, 1996; Tukey, 1977). It is also striking that the majority of the developmental graphs found in the literature do not explicitly graph the ranges within which the data fluctuate.

Intra-individual variability can be defined as differences in the level of a developmental variable within individuals and between repeated measurements. In this article we will use the term “variability” to indicate these differences (in achievement or behavior) between measurement occasions. We also use the term “fluctuations” for the differences between consecutive points in a variable trajectory. The term “stability” is used to indicate the counterpart of (or the lack of) variability.

In recent years, an increasing number of researchers acknowledge the possible meaningfulness of intra-individual variability and show an increasing interest in these irregular aspects of change. The notion that people function at different levels of development at the same time and the belief that this variability can be an essential factor in promoting development, have become increasingly prominent in recent developmental literature. Examples from early development are the studies of De Weerth, van Geert, and Hoitink (1999) who focused on variability in infant emotional behavior, Bertenthal (1999) who studied variability in inter-limb coordination and postural control in infants, and Ruhland and van Geert (1998) who focused on variability in early language development. Variability also features in the microgenetic approach (e.g.,

Kuhn, 1995), which tries to increase the chances of observing developmental change by providing a subject frequent opportunities over a period of weeks or months to engage in the cognitive strategies under investigation. This increased density of the use of strategies may lead to change, allowing the researcher close observation of the process. This is shown, among others, in the work of Fischer, Bullock, Rotenberg, and Raya (1993), Fischer and Granott (1995), Goldin-Meadow, Alibali, and Church (1993), Granott (1993), Lautrey (1993), Lautrey, Bonthoux, and Pacteau (1996), Lautrey and Cibois (1994), and Siegler (1994, 1996, 1997).

One of the reasons variability is receiving increasing attention lies in the introduction of a new theoretical viewpoint, namely dynamic systems theory (e.g., Thelen & Smith, 1993; van Geert, 1994), and in particular, catastrophe theory (Thom, 1975; van der Maas & Molenaar, 1992). These theories share the importance they attribute to variability. Both take a radical departure from the measurement-error-hypothesis, which systematically considers variability (in the form of fluctuating developmental levels) as the result of measurement error. This error-hypothesis is closely related to true score theory (Cronbach, 1960; Lord & Novick, 1968; Nunnally, 1970) and is deeply rooted in psychology. The error-hypothesis is based on the assumption that every psychological measurement is subjected to random measurement error, which is expressed in the variability of repeatedly acquired scores. Since these random errors are, by definition, independent of the true value measured, they are symmetrically distributed around the true level. Thus, by averaging over these (supposedly random) fluctuations, the true underlying level can be approached. Again, dynamic systems theory radically rejects this automatic retreat to the error-hypothesis and claims that variability bears important information about the nature of the developmental process.

Thelen and Smith (1993) were among the first to apply the dynamic systems approach to developmental psychology. They build on the idea of development as a self-organizing system. Change is defined as the transition from one stable state to another. Highly attractive states are *dynamically stable* and exhibit fluctuations around the mean state that reflect the noisiness of their components. Thelen and Smith state that in self-organization, the system is attracted to one preferred configuration out of many possible states, but behavioral variability is an essential precursor (p. 56). Dynamic systems theory has specific predictions for the behavior close to a developmental transition. During a transition, variability is large and "*the system is free to explore new and more adaptive associations and configurations*" (p. 145). The dynamic approach turns variability within (and also between) individuals into an essential element in the developmental process. Variability is considered to be the harbinger of change. Variability is also the essential ground for exploration and selection. Thelen and Smith encourage researchers to investigate the variability in their data. They state: "*If errors of design or execution are not at fault, think dynamical and use the variability as data. Does the variability change over time?*" (Thelen & Smith, 1993, p. 342). Note that Thelen and Smith do not discard the existence of measurement error. Measurement error exists in the form of errors of design or execution.

Self-organization is also central in catastrophe theory (Thom, 1975), which can be considered as a specific branch of dynamic systems theory (van Geert, Savelsbergh, & van der Maas, 1997). According to catastrophe theory, self-organizational processes can be classified into a limited number of characteristic patterns of discontinuous change, depending on the number of fundamental variables that determine the change. As such, catastrophe theory offers concrete models and criteria for discontinuities in developmental processes. One of the types

of discontinuous change that has been applied is the *cusp model*. Catastrophe theory provides eight so-called catastrophe flags to test the presence of a cusp model (Gilmore, 1981; see also van der Maas & Molenaar, 1992, for an application to cognitive development). One of these catastrophe flags is "anomalous variance," which indicates that variability is expected to be greater in the vicinity of a phase transition, in the same sense as indicated by Thelen and Smith. However, catastrophe theory takes this reasoning one step further by taking variability as one of the criteria that indicates a discontinuous transition.

1.3. Recent findings in the field of infant motor and emotional development

1.3.1. Motor development

Several research domains have taken up on the ideas of dynamic systems theory and catastrophe theory. Initially, the field of motor development was most forceful in this pursuit. It is important to note that these studies are of different kinds: some are one-dimensional and quantitative, while others are multi-dimensional and qualitative. This distinction (between one-dimensional quantitative and multi-dimensional qualitative) has it implications for the methodology. We will elaborate on this issue later on.

The motor development domain provided many empirical studies on variability. We will only name a few for illustrative purposes. For instance Thelen (1985) documented the process of transition in the supine kicks of infants followed longitudinally from 2 weeks to 10 months (see also Thelen & Smith, 1993). One of the motor behaviors studied was the coordination between kicks. During the first few months, kicks were predominately alternating. However, this period is followed by a period with great variability. This instability led to new forms of coordination between legs, for instance simultaneous kicking of both legs. It appears that the infants must free themselves from the stable patterns of the newborn period before they can assemble new behavioral modes. It is clear that Thelen places great importance on behavioral variability as a precursor of a new behavioral repertoire. Furthermore, Wimmers (1996) studied transitions in the development from reaching without grasping to reaching with grasping. This occurs in most infants between 16 and 24 weeks of age. He used Gilmore's catastrophe flags to detect a phase transition in the development of grasping. One of the findings was the detection of the flag "anomalous variance," which indicates that the changes in question were accompanied by loss of stability (Wimmers, Savelsbergh, Beek, & Hopkins, 1998a, 1998b).

Bertenthal (1999) discusses the meaning of variability in the development of crawling patterns in infants. He states "[. . .] *this variability is not merely a correlate of change but instead a contributor to the change itself*" (p. 105) (also see Bertenthal & Clifton, 1998; Newell & Corcos, 1993). Bertenthal goes on stating that variability offers flexibility, which drives development following "Darwinian" principles. Principles of variation and selection cause successful behaviors to be stored and repeated more frequently than the less successful. Bertenthal believes that studying the change of variation patterns offers important insights into how children change with age.

1.3.2. Emotional development

The study of De Weerth et al. (1999) focused on variability in infant emotional behavior. After reviewing the literature on this subject, they concluded that there are indications that

infants display variable behavior both within and between observations. This is for instance the case in the field of visual behavior (Canfield, Wilken, Schmerl, & Smith, 1995), sleeping and waking patterns (Dittrichova, Tautermannova, & Vondracek, 1992), visual recognition behavior (Wachs, Morrow, & Slabach, 1990), and infantile emotions and temperament (e.g., Crockerberg & Smith, 1982; St. James Roberts & Wolke, 1984). However, the idea that variability could be an intrinsic characteristic of a normal developing system has seldom been recognized, much less explored. De Weerth et al. (1999) found considerable intra-individual variability in four different types of emotional behaviors in infants (crying, fretting/fussing, body-contact and smiling) that seemed to decline in the first year of life. Instead of attributing this variability to measurement error, they point at a possible adaptive strategy. They claim that variability in emotional behavior ensures the infant of continued maternal attention: "[. . .] *mother and infant try out new ways of communicating with each other, and also change them over time*. [. . .] [*They*] *tune into each other and influence each other with their moods attitudes and developing skills, etc*." (p. 11). As the infant grows older, he or she has access to more sophisticated means of communication, to insure himself or herself of maternal attention, and variability may decline.

1.4. Recent findings in the field of language development

In the domain of language development, the importance of variability has largely been neglected. However, there are solid indications that language development is characterized by large fluctuations. For instance, the classical study of Minifie, Darley, and Sherman (1963) found a lack of test–retest reliability in seven language measures, including the average utterance length. Furthermore, Chabon, Kent-Udolf, and Egolf (1982) found large temporal variability of MLU (Mean Length of Utterance as defined by Brown, 1973) in children from age 3;6 and older. Neither study elaborates on the meaning of this variability although the use of the term "reliability" in both studies suggests a tendency toward true score theory. The work of Ruhland and van Geert (1998), on the other hand, is inspired by dynamic systems theory and catastrophe theory. In their study, the language development of six children was followed from the first-word stage up to the differentiation stage. Focus of this study was the development of function words. The frequency with which function words occur in child language constitutes an important indicator of syntactic development, according to Ruhland and van Geert (1998). Although the shape of the developmental curves turned out to have great inter-individual differences, all children showed remarkable fluctuations. The peaks and wells immediately catch the eye. Inspired by dynamic systems theory and catastrophe theory, Ruhland and van Geert take variability as developmentally meaningful, by studying it in the context of the catastrophe flag "anomalous variance." Therefore, intra-session variability was investigated by dividing the observation sessions in two equal parts, and consequently comparing the first part with the second part. In two of the six subjects, these differences turned out to coincide with a sudden jump; the other four showed more moderate effects. In sum, the language development of all children showed considerable intra-individual variability. Recently, the study of Fenson et al. (2000) addresses variability (both between and within individuals) in the MacArthur Communicative Development Inventories (CDIs). The CDIs were criticized by Feldman et al. (2000) as having too little stability and insufficient ability to predict a possible language delay.

This critique would limit the utility of this parent report instrument for the study of language development. Fenson, however, offers the possibility that the finding is real, and that individual differences in language ability are quite unstable in this age range. He states: "*Language skills may simply not be sufficiently developed at age 1 to make accurate assessments*" (p. 325) and "[. . .] *the CDI is simply reflecting the non-linear character of development*" (p. 326).

In summary, although the literature on language development has so far largely neglected the issue of variability, there are solid indications that variability is prominently present and may bear theoretical and empirical importance.

1.5. Purpose of the article

Starting from the idea that variability is an undeservedly neglected and meaningful phenomenon, this article aims at presenting several new techniques for describing variability in developmental data. These techniques will be illustrated by applying them to data from the field of early language development. What these techniques have in common is that they focus on variability in individual trajectories. It is important to note we are convinced that this individual level is the starting point for analyzing patterns of variability. After the presentation of the techniques in an application to individual language trajectories, we will also indicate how they can be applied to cross-sectional data.

2. A case study of variability child language

2.1. Time-serial data of language development

Language development provides a good starting point for illustrating the techniques that will be introduced in the remainder of this paper. As we stated before, there are strong indications that language development is characterized by large instabilities. From a dynamic systems view, language development is especially relevant because of its dynamic interplay between the various linguistic elements and the non-linguistic domain. Moreover, the domain of language development shows several practical advantages for the study of variability. First of all, because the study of variability requires a relatively large collection of measurement points per individual, the measurement procedure itself must be as non-obtrusive as possible. The collection of spontaneous speech samples easily meets this demand. Secondly, language development provides quantitative data, which can be easily plotted and used for calculation. Thirdly, language is known to develop relatively quickly and shows a rapid increase in its complexity. Thus, meaningful data sets can be collected in relatively short periods of time (about 1–2 years, on average).

In this article, we will show the results of two developmental variables: MLU and spatial prepositions. The reason we present these two variables in particular is that they show very different developmental patterns, and thus illustrate different aspects of variability. While the data of MLU show a regular, continuous trend, the preposition data show a more irregular pathway.

2.2. Description of the study

2.2.1. Subjects

One subject (Heleen, a girl) was followed from age 1;6 to age 2;6. In the beginning of the study, Heleen was in the one-word stage. At the end of the observation period her language showed various characteristics of the differentiation stage (see for characteristics of the Dutch differentiation stage, Frijn & de Haan, 1994). Heleen was the first-born (and, during the observation period, only) child of middle class parents. The family lives in a suburban neighborhood in an average-size city in the North of The Netherlands. Heleen was raised in a monolingual Dutch environment. The family does not speak any apparent dialect. Heleen's general cognitive development was tested with the Bayley Developmental Scales 2/30 (Van der Meulen & Smrkovsky, 1983) a few months before her second birthday. She scored within the normal range (OI = 100).

2.2.2. The measurements

The study is based upon videotaped observations of spontaneous speech in a naturalistic environment (the child's home). The child and parent were free to follow their normal daily routine. There were a few practical restrictions given to the parents' activities (such as not watching the television, and not having extensive phone conversations). In addition to the child, one of the parents and the observer were present during the observations. Observations took 60 min each. The camera was positioned in a corner of the living room, overviewing much of the living room space. There was a warming-up time of 5 min. In practice, the child hardly noticed the camera and did not behave differently with or without the camera. All child language and all child-directed adult language were transcribed according to Childes conventions (MacWhinney, 1991).

2.2.3. The design

The measurement design was scheduled in such a way that variability could be optimally studied. The first level of variability is developmental variability, which in this case takes place over a timespan of a year or more. Questions that can be addressed at this level are for instance: what is the general shape of the development process, is it continuous or discontinuous, is it rapid or is it slow? Secondly, the measurement design was set up to also include short-term variability. At this level, we ask ourselves how capricious the developmental variables are within relatively small intervals. How large are the fluctuations in, for instance, a week? We will call this time scale day-to-day variability. At a still smaller time scale, we can study within-session variability, for instance by comparing the first half-hour of an observation with the second. For the record, we would like to point out once more that the present article is aimed at describing and illustrating a number of techniques for representing variability and does not intend to answer the previous questions.

It is important to incorporate all these time scales in our analyses, because it is highly likely that variability may be different on each time scale. For instance, a developmental variable may be slowly oscillating while gradually growing, while another variable may increase discontinuously with sharp day-to-day fluctuations. It is also conceivable that the kinds of differences in variability patterns, as described above, take place within the same variable, for different periods

in time. A variable may initially show slow oscillations, but when approaching a developmental transition grow to be very unpredictable and capricious. Although these possibilities are speculative, the point is that we wanted the design to be able to display these kinds of differences.

In order to capture long-term change, the longitudinal study covered the period of a whole year. The general format of the design is based on the common 2-weekly measurement design from the Childes-database samples (e.g., the Groningen Dutch Corpus). This measurement frequency is considered adequate to study developmental changes. In order to study day-to-day variability, we alternated the 2-weekly observations with six *intensive observation periods*. Each intensive observation period consists of six measurements in two consecutive weeks (three measurements in each week). The intensive periods were equally divided over the total observation period of a year. In total, we collected 55 samples. There was only one missing value (August 5, a 2-weekly observation).

Each observation lasted about 60 min, which is relatively long. A commonly used observation length in language acquisition studies is the unit of 100 utterances. Brown (1973), for instance, suggested calculating Mean Length of Utterance on the first 100 utterances. The period of 60 min consists of at least 200 utterances each, which means that there is enough room to study intra-observation variability, the smallest unit of variability.

2.3. Variables under study

2.3.1. Mean Length of Utterance

Mean Length of Utterance in words (MLU-w) was calculated by dividing the number of words in the total sample by the total number of utterances in the samples. Uninterpretable utterances, direct imitations, and yes/no-answers, songs and imitation games were excluded. MLU-w is not the same as Brown's original MLU in morphemes (MLU-m), but research has shown that the two measures are highly correlated in normally developing children (e.g., Arlman-Rupp, van Niekerk de Haan, & van de Sandt-Koenderman, 1976; Hickey, 1991; Thordardottir & Weismer, 1998). Because MLU-w is much simpler (both theoretically and in practice) MLU-w is considered the preferred measure of the two (Thordardottir & Weismer, 1998). In a pilot study, we also found a strong similarity between MLU-m and MLU-w, MLU-m being only a bit higher than MLU-w. Results in terms of variability did not differ significantly (van Dijk & van Geert, 1999).

2.3.2. Spatial prepositions

All prepositions that belong to the set of spatial prepositions were selected, even if the context was not spatial. This was done with LEGro (Language-analysis Excel add-in Groningen, van Geert, 2000), an Excel-macro that can select utterances with different kinds of criteria. We selected utterances with spatial prepositions. These selections were used for quantitative and qualitative analysis. First, we counted the total frequency of prepositions that were uttered in a particular spatial context. This means that if the context showed that the child referred to an object in a spatial relationship to another object, that preposition was included. We also included contexts that referred to spatial actions that had just occurred, or that still had to happen. So for instance, if a child said "*in chair*" in the context of "*I want to sit in my chair/mommy please put me in my chair?*" the preposition was also included, the same for "*ball under*"

in the context of "*the ball is under the table*." We counted all different spatial contexts. For instance if a child said "in chair" and "doll in bed," these were counted as two different spatial contexts. However, because we were interested in the child's ability to label spatial situations, we excluded repetitions. For instance, if a child repeatedly said "*in chair, in chair*," while the mother did not respond, this was counted as only one spatial preposition-in-context.

It might be argued that the variability we will eventually observe is based on the actual use of spatial prepositions and that this use is highly dependent on (linguistic and non-linguistic) context. Some contexts might be better suited to evoke spatial prepositions than others. Variability in the data therefore not only reflects the possible instability of the developing syntactical system, but also many situational factors. We agree that not all variability is a direct reflection of development, more precisely, of the processes of stabilization and destabilization that development might, among others, consist of. On the other hand, we must consider that there exists a mutual interaction between the developing infant and the spatial context. First of all, the behavioral repertoire of children of this age is filled with spatial activities. They climb on things, build with blocks, drive with cars around other objects, put dolls in beds and in chairs, etc. All these activities are in principle suited to evoke spatial prepositions and the child is an active agent in the selection and constitution of activities and topics. Thus, the variability is the result of this mutual relation between the developing child and the context. The child is not only dependent on and influenced by the linguistic and non-linguistic context, but also selects it and contributes to it. Therefore, we must not conceive of the development of prepositions as the development from one stable state (of no preposition use) to another stable state, namely that of having acquired prepositions and using them at a constant level of production. Instead, we must consider the fact that the end-state of development is not a "stable" category, but a category that is "dynamically stable" in the sense that the produced prepositions still show a considerably variable range dependent on situational factors. This range (of variability) of fully acquired preposition use should, however, be smaller than the range of preposition use in young children who are still acquiring this linguistic category. We have indications that this is indeed the case: children appear to be considerably more variable in their use of spatial prepositions than adults.[1] This indicates that at least part of the variability we wish to describe in this article is developmentally determined.

3. How can we describe variability in longitudinal data?

3.1. Focus on variability in developmental data

The question of whether intra-individual variability exists is not a subject of discussion in developmental psychology. In fact, variability is a well-known "problem" for many researchers. It is no coincidence, therefore, that there exists a broad spectrum of techniques to eliminate fluctuations in longitudinal data. We have already referred to them under the term *smoothing techniques*. There are, however, far less common techniques that allow us to specify and visualize variability in time-serial data, with a limited number of measurements.

The reason smoothing techniques are so well developed and commonly accepted, is that variability is often considered to represent error. This opinion is so widely spread that it is not

surprising that many developmental psychologists are only interested in revealing the "general developmental pathway," whatever that means. When analyzing a general trend, fluctuations are often considered to be inconvenient noise. In addition, developmental psychologists tend to use a rather limited set of idealized trend models (basically linear, quadratic and exponential growth models). By so doing, they reduce the information from the data even further. However, even if they acknowledge the possibility of *meaningful* variability, researchers will still be interested in describing a general developmental trend. They will still ask questions such as whether a developmental process is continuous or discontinuous, and whether there are developmental transitions. What techniques can be used to analyze both the general trend and take variability into account? In this article we will present several of these techniques that are essentially descriptive in nature and can be used with many kinds of individual time-serial developmental data.

3.2. Qualitative and quantitative variability

As we mentioned before, there is an important difference between (one-dimensional) quantitative and (multi-dimensional) qualitative variability. In the case of *quantitative* variability, each measurement consists of a level on a single dimension. This can be a frequency count, for instance the number of function words in an observation, but it can also be a number that expresses the level of some kind of psychometric variable, such as IQ. Variability in quantitative data shows itself in a fluctuating level of the variable at issue. This sort of data is typically obtained in the field of early infant emotions (for instance the percentage of crying time in De Weerth et al., 1999, see Fig. 1A), and language development (for instance the Mean Length of Utterance in Chabon et al., 1982). Fluctuating levels of this nature can easily be plotted in a line graph, such as Fig. 1A.

In the case of *qualitative* variability however, each measurement consists of a set of behaviors, which have a specific occurrence each. For example, a child uses one strategy A in 10% of all occasions, a second strategy B another 10%, while he or she predominantly uses strategy C (80% of all occasions; see also Siegler, 1996, 1997 for a similar model, applied to cognitive strategies). For instance the first measurement consists of the strategies A, B and C, the second of strategies A, D and E, and a third of B, E, F, and G. The most important

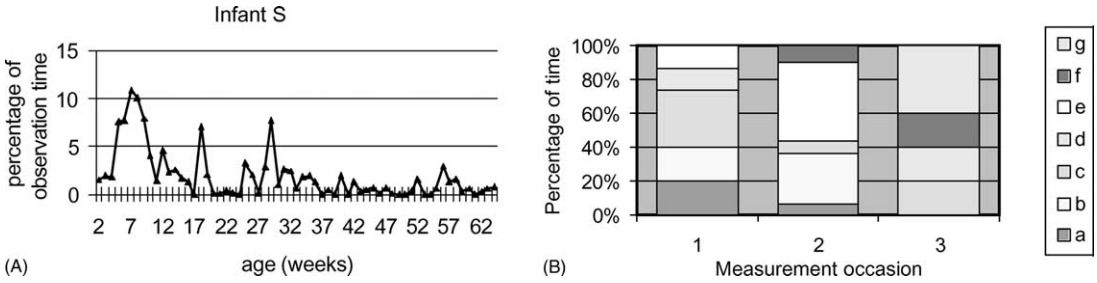


Fig. 1. (A) Example of one-dimensional quantitative variability: crying duration in percentage of observed time of one infant (infant S) (source: De Weerth, 1998). (B) Example of multi-dimensional variability: infant response patterns (labels a, b, c, d, e, and g given for present illustration) during Fw-translations of one infant (infant I) during postural response sitting tasks (source: Hadders-Algra et al., 1996).

difference is simply that this qualitative variability concerns additional dimensions. Not only are there differences in the level of the measured variables, but also completely new variables can be introduced. Variables can disappear, often to reappear later on. Note that this sort of variability is often studied in the field of motor development (for instance the different types of coordination between kicks in the development of supine kicking in Thelen, 1985 and the use of different muscle groups in the development of postural control, Hadders-Algra, Brogren, & Forssberg, 1996, see Fig. 1B). These differences in measurements can easily be expressed in a stacked bar graph, such as Fig. 1B.

Most of the techniques we are going to present in the remainder of this article, are particularly suited for one-dimensional quantitative data. However, it is important to know that it is possible to transform qualitative data into a quantitative format. At this point in time, there is no measure that expresses all dimensions of qualitative variability into one single number. However, there are some options for further analysis. For instance, one can study each dimension separately, by taking the occurrence of the variable per measurement occasion. Additionally, the total number of behavior classes (for instance strategies) can be counted, or the total number of new classes.

3.3. Raw data and smoothing techniques

3.3.1. Mean Length of Utterance

As our first example, we will focus on the growth of Heleen's Mean Length of Utterance. The simplest, and commonly used, way to present the data on MLU, or any type of developmental data for that matter, is by putting the data in a simple *XY*-diagram, the *X*-axis showing the date of measurement, the *Y*-axis showing MLU (Fig. 2). Some measurements are closer together (intensive periods) than other measurements (2-weekly measurement rate).

The graph shows two striking facts. The first is the existence of a general trend, MLU growing from a little over 1 (one-word stage) to almost 3 (differentiation stage). Second, visual inspection clearly shows large fluctuations between measurement days. Especially the sixth

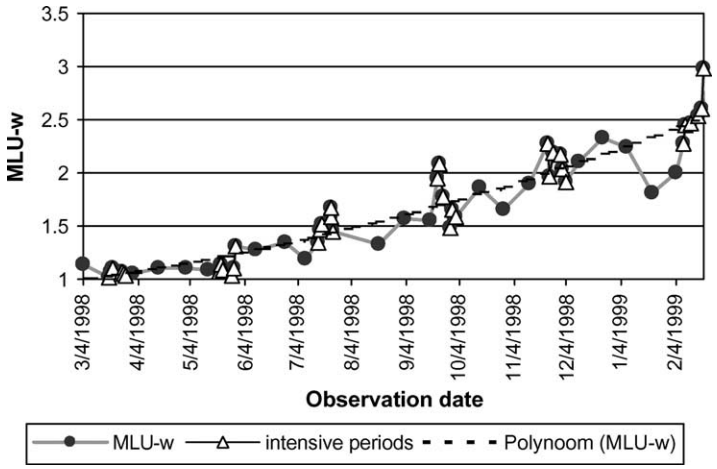


Fig. 2. Raw data of Heleen's MLU, including a linear trendline.

intensive period seems to show dramatic differences within only a couple of days. Keeping Brown's initial MLU-stages in mind, Heleen seems to fit in three different MLU-stages within the timeframe of no more than 10 days. This can at least be called remarkable. It also confirms the results of another study on variability in early language development (van Dijk, De Goede, Ruhland, & van Geert, 2000), where both subjects' MLU levels fell into three distinct MLU-stages within a period of several weeks.

In Fig. 2 we also used a commonly known technique to show the developmental trend in the data: we plotted a trendline (using a polynomial of the second degree). As can be seen in the graph, this trendline completely smoothes (as it is supposed to do) all the fluctuations in the curve, which are especially large at the end of the curve. This already shows that using a smoothing technique will indeed lead to loss of information—information that may be valuable. Obviously in the case of MLU we are not dealing with something minor: the variability we eliminate by using a smoothing technique is indeed considerable.

Researchers who use these smoothing techniques, probably consider the variability as relatively uninteresting in itself, for instance because it is seen as a form of error fluctuation. The assumption behind this approach is that of an underlying, true level that can be approached by averaging over the fluctuations. The most common technique for doing this is by using moving averages. The fluctuating levels over a pre-specified time window, e.g., the period of 1 month, are used to specify the fluctuating level's *central tendency*. This central tendency is supposed to contain more reliable and meaningful information than each of the separate observations, respectively. Another technique consists of polynomial regression models. They make an estimation of a trendline based on a function of time. The trendline that shows the smallest average (squared) distance is considered the best representation of the developmental trend present. Thus, by averaging over, for instance, six observations in an intensive period, we try to estimate a central MLU level that we think characteristic of the period at issue. Regression models over time have yet another function, in addition to representing a supposed true central score. This other function is that they are very well suited for representing a direction, i.e., a motion vector. This can be seen as the simplest possible general trend of a range of score levels over time, in a way similar to a meteorologist's representation of the direction of the wind by a single arrow.

We noted earlier that developmental psychologists tend to confine themselves to an unnecessarily small set of smoothed trends. The statistical literature contains far more sophisticated smoothing models, which follow the actual rise and fall of the data as faithfully as one wishes. Examples are spline models, but also local polynomial regression models or loess smoothers that follow any non-linear trend in the data (Simonoff, 1996). The point is, however, that the smoothing model we opt for implicitly expresses our view on what we consider essential or important in the data and what information can be safely disregarded. The first kind of data to usually fall victim to our smoothing activities are the data about variability.

Heleen's fluctuations between an MLU of 1.6 and 3 can be summarized by presenting an average, say 2.3. The question is: to what extent do we reliably characterize Heleen's language development at the period at issue, by specifying that the average is 2.3? We do not claim that this average score does not bear any information in itself. We believe, however, that the particular range of scores can be highly informative of a child's level of language development. We will get back to this question further on, where we will show how the score range can be used to analyze the developmental data at issue.

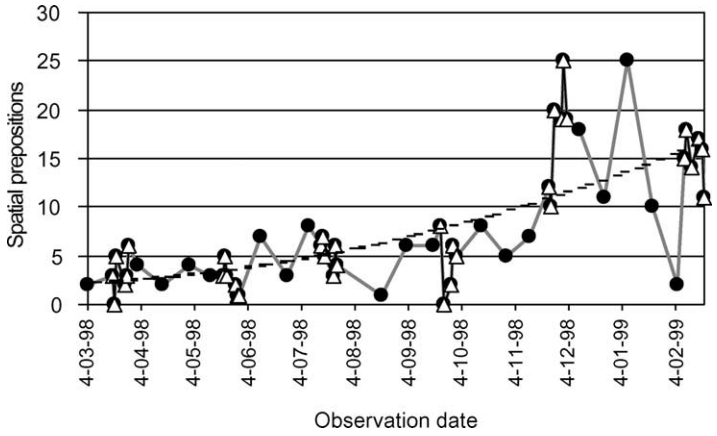


Fig. 3. Raw data of Heleen's spatial prepositions, including a linear trendline.

3.3.2. Spatial prepositions

Let us now turn to the data of Heleen's spatial prepositions. These data show a very different picture than the MLU data. The data points in Fig. 3, a simple *XY*-diagram, represent the total frequency of the prepositions per session, counted as the number of distinct spatial situations.

Based on the mere visual inspection of the data, it seems that the data can be cut into two clearly distinct periods. The first part of the data (up to observation number 38, November 12) shows a relatively stable, low frequency occurrence of prepositions. However, in the second part of the trajectory, we see a steep increase of prepositions. In this second part of the graph, large fluctuations immediately catch the eye. For instance note observation number 49 (February 4), where the frequency of prepositions is suddenly very low, and preceding and following measurement points show much higher numbers of prepositions.

When fitting a trendline to the data, using a second-degree polynomial of time, we obtain a continuous curve, with a moderate steepness. This also demonstrates the fact that the use of a standard smoothing technique indeed gives a completely different summary of the data than visual inspection of the actual data. While visual inspection points in the direction of two distinctive stages, the smoothing technique covers this fact up completely. Note that if we had used a less common smoother, namely a loess or weighted local regression smoother (Simonoff, 1996), the resulting smoothed curve would have displayed the stepwise increase in the data considerably more faithfully, thus supporting the suggestion of two distinct stages. The question of whether we indeed deal with two clearly distinctive parts of the developmental trajectory is in our opinion very important. Later on in this article, we will ask ourselves the question whether the two-stage pattern we see can also be justified by further statistical analysis.

3.3.3. Variability as a developing range

As we showed in the preceding part, standard smoothing techniques eliminate much information from the actual data. The more current view that embraces variability asks for different techniques that visualize the essence of a developmental trajectory. In this view, the fact that Heleen's MLU shows fluctuations between three different MLU-stages is seen as highly

informative of her present level of language acquisition. It is clear that she is capable of producing utterances that are of considerably higher relative complexity than the utterances she produced a few months earlier. However, although she is capable of doing so, it is not yet her habitual level of language production: there are days at which her level does not significantly exceed that of a few months earlier. In addition, during this period Heleen's MLU levels occupy almost any intermediary position between these two extremes. Although this observation seems trivial, it is not so in light of the fact that children sometimes oscillate between two different developmental states with no apparent states in between. For instance, in the study of van der Maas and Molenaar (1992) (where children were administered Piagetian conservation tasks) large variability in strategy use was found around the transition from the non-conserver stage to the conserver stage. At this point in time conservation and non-conservation strategies exist at the same time. Goldin-Meadow et al. (1993) found a similar situation in which two stages are simultaneously present in the system. They found that the shift from one (mathematical) strategy to a second, more advanced one, is characterized by the simultaneous presence of different strategies in different expressive modes (e.g., verbal and gestural). A final example of the simultaneity of distinct levels of problem solving is Siegler's model of overlapping waves in strategy use. Strategies at distinct levels of complexity are used in an alternating way. Learning and development amount to a change in the relative frequencies with which strategies are used, eventually resulting in the disappearance of less mature strategies (Rittle-Johnson & Siegler, 1999; Siegler, 1996, 1997).

In Heleen's MLU data, the phenomenon of interest is the broad range of Heleen's MLU levels, with the highest levels about twice as high as the lowest ones. One of the questions we could ask ourselves is whether or not this range is a developmental phenomenon in its own right. For instance, does the relative size of the range (let us say, the width in comparison to the central or average position) remain more or less stable across development? If this is so, the relative range itself is not specifically informative from a developmental viewpoint. However, if the range itself becomes either narrower or broader depending on various kinds of developmental phenomena or stages, the study of its properties becomes a worthwhile endeavor. We have seen that techniques for averaging or otherwise smoothing fluctuating data are readily available (moving averages, polynomials with time as a dependent variable, etc.). But what techniques do we have if we want to specify information that pertains to fluctuation and variability, as suggested by this new approach?

3.4. Showing variability in a graph

3.4.1. Moving minimums, maximums and averages

An elegant alternative technique, in which we can study the developmental trend, but that nevertheless also displays variability around a general trend is what we have called the *moving min–max graph*. This technique shows the data using the bandwidth of observed scores. Instead of displaying measurement points as simple dots, the moving min–max graph presents a score range for each measurement occasion. Instead of a single line graph, the data are presented in a bandwidth of scores. This method uses a moving window, a timeframe that moves up one position (measurement occasion) each time (the size of the window, e.g., five consecutive data points, 1 month, etc. is called its period). Each window partly overlaps the preceding

windows, using all the same measurement occasions minus the first and plus the next. For instance, for every set of seven consecutive measurements we calculate the maximum and the minimum values. This is done by way of a predetermined moving window, such that we obtain the following series:

$$\max(t1 \ldots t7), \max(t2 \ldots t8), \max(t3 \ldots t9), \text{ etc.}$$

$$\min(t1 \ldots t7), \min(t2 \ldots t8), \min(t3 \ldots t9), \text{ etc.}$$

Technically these values are very easy to plot. Any commercially available spreadsheet program offers functions such as max and min that can easily be computed over moving data windows. Once the moving minimums and maximums are plotted, one can visually inspect whether they too show considerable fluctuations over time. The question one should ask is whether these fluctuations are developmentally meaningful or not. The fluctuations should again be contrasted with the eventual long-term changes in the minimums and maximums.

In addition to plotting maximums and minimums, one can also plot some form of central score. One possibility is to plot the median, the value that has a similar number of values above and below it. Another possibility, which combines these techniques with smoothing approaches, is to compute a moving average corresponding with the moving minimums and maximums data. Fig. 4A shows the moving minimums, maximums and averages of Heleen's MLUs, with a moving window of period 5.

The moving max–min method can be used to specify a value, for instance a child's test score, with respect to upper and lower boundaries of a time window chosen in advance. As a reasonable rule-of-thumb, one could take windows of a size of about one-tenth of the entire data set, but in principle no less than five data points. In the present study an irregular measurement design was used, with 2-weekly measurements alternated with intensive periods. This design results in windows that are very different from each other in terms of time. Five consecutive measurement points can cover a period of only 10 days (during the intensive measurement periods) but also 10 weeks (between the intensive periods). Therefore it is better not to use an absolute number of measurement points, but to choose a moving window on the basis of time. For instance, Fig. 4B uses a moving window of 18 days. Because of the differing number of days between measurement points (recall the irregular measurement design) the number of measurement points per window can vary.

Looking at Fig. 4B, it is possible to compare the width of the band with the general developmental trend. In the MLU case (Fig. 4) there is no obvious widening or narrowing in the range, but instead we see a general increase in bandwidth, with several mild oscillations. However, it is not clear that these oscillations are meaningful because they seem to coincide with the intensive periods. The fact that the intensive periods have more measurement points in the moving time window, might very well explain the mild fluctuations in the bandwidth. During the intensive periods, more observations are carried out and this increases the probability of hitting upon an "extreme" value, which is conserved across the length of the moving window.

It is obvious from the figure that MLU shows a general increase in its level in addition to a generally increasing bandwidth. When interpreting this observation, it is important to take notice of the increasing mean in the timeframe. It is a well-known fact that variability is related to the general mean. For instance, a data series with a mean of 100, is expected to have a larger

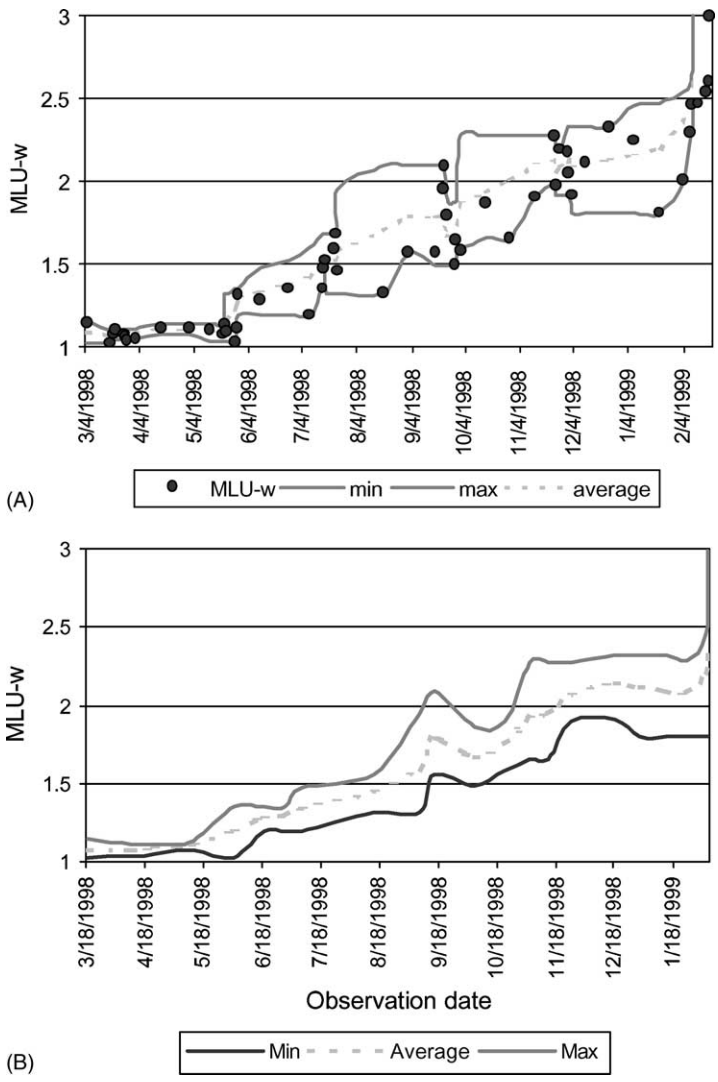


Fig. 4. (A) Application of the moving min–max methods to Heleen's MLU-w (5 positions). (B) Moving min–max graph of Heleen's MLU-w (timeframe 18 days, last window 15 days).

range (for instance expressed in terms of standard deviation (SD)) than a data series with a mean of 10. Therefore it is to be expected that variability in the data increases solely on the basis of the increasing mean MLU.

Earlier, we stated that the range might be an important developmental phenomenon. When the range itself becomes narrower or broader depending on various kinds of developmental phenomena or stages, the range is indeed a developmental phenomenon in its own right. In the case of MLU we do not see any obvious widening or narrowing, we predominantly see a general widening with the growth of MLU. However we do not know enough about the relation between the increasing trend and the increasing bandwidth. Is variability accruing more quickly

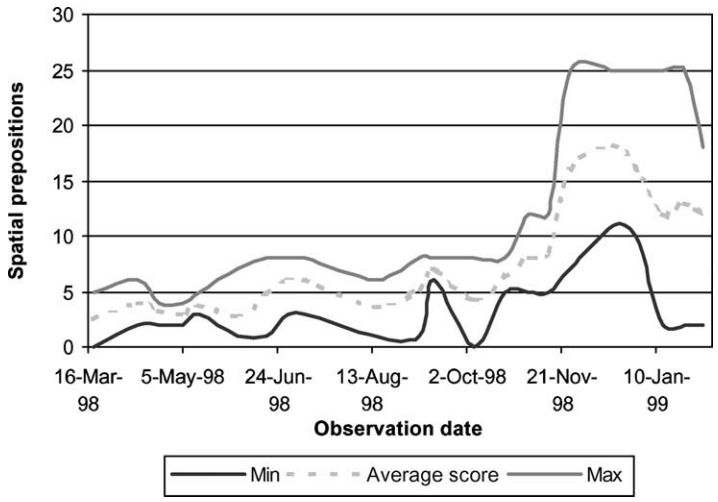


Fig. 5. Moving min–max graph of Heleen's spatial prepositions (time window of 18 days, last window 15 days).

than the MLU trend, or is the slope of the MLU trend actually steeper than that of variability? If the latter were the case, variability would in fact decline if we were to correct for this growing mean. The graphs we showed so far, however, make one wonder how MLU behaves later on in development. Does the bandwidth decrease again at some point in time? If variability is indeed a developmental phenomenon, we do expect exactly that to happen. In the light of our remarks about the dynamically stable state of adult prepositions, it should be noted that we do not expect a disappearing bandwidth. Although we expect that adults show a significantly smaller bandwidth than developing children, their output is also variable to some degree.

Fig. 5 shows the spatial preposition data in a moving min–max graph. Here, we see a completely different picture from what we saw with MLU. While MLU showed a general increasing bandwidth, with slow oscillations, prepositions show a moderate bandwidth in the beginning of the trajectory until observation 38. After this point, we first see a slight general increase, followed by a great widening of the range. This graph also suggests that something "different" occurs after observation number 38, which is probably a developmental transition in spatial prepositions.

3.4.2. Altitude lines

The moving min–max graph provides a general overview of the moving range along the trajectory. As intended, this technique is highly sensitive to so-called extreme values. For a more in-depth study of the distribution of the values in the range, one might consider the following extension of the use of moving minimums and maximums. This method incorporates also intermediate positions in a so-called "altitude line graph." In that case, we do not only plot the minimum and maximum values in the moving window, but also the second highest, third highest value, etc. We then connect the corresponding data points by a line, comparable with altitude lines on a geographical relief map.

For instance, consider a piece of Heleen's dataset of spatial prepositions (observations 22–27) with the values 7, 5, 3, 6, 4, 1. For the first window, the maximum value is 7, the second

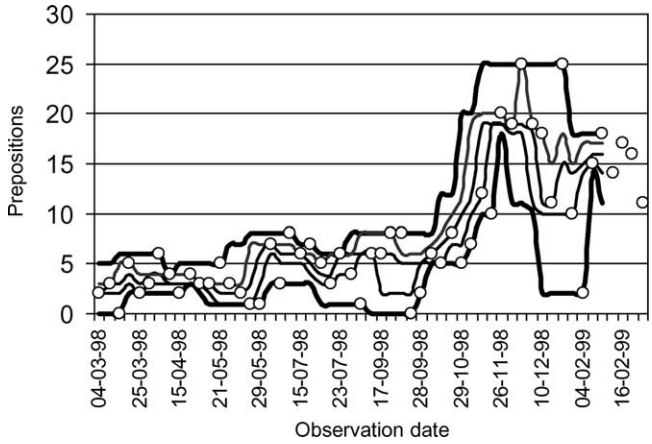


Fig. 6. Application of the altitude lines method to Heleen's spatial prepositions (window of 5 positions).

highest value is 6, the third highest value is 5, the fourth highest value is 4 and the fifth (or the minimum) is 3. For the second moving window, we get the values of, respectively 6, 5, 4, 3, 1. We can easily draw a graph linking all these (first, respectively second, third, etc.) positions with each other.

To give an illustration of these techniques we plotted Fig. 6 to show the altitude lines of the spatial preposition data of Heleen. The dots in this graph represent the actual data points, the outer lines are the moving minimums and maximums, and the intermediate lines are the intermediary positions.

In principle, these altitude lines can be interpreted in the same way as geographical lines representing the altitude and steepness of a physical relief. Our altitude lines, however, refer to the properties of the distribution of time-serial values. For instance, if they are concentrated either on the top or on the lower part of the range, they refer to a skewed distribution. Note that we can replace the actual data points by a polynomial approximation (e.g., in a linear regression model) or, more preferably, by a more flexible smoother such as a loess smoother. The polynomial or loess replacement results in a smoother and more easily interpretable representation of the longitudinal distribution of the data. Its disadvantage is that it conceals eventual sudden changes that may be indicative of discontinuities.

If the moving windows contain a sufficient number of data points, one may also plot the moving percentile scores, e.g., of the 90th, 10th and 50th percentiles. All these techniques serve to specify changes in the upper and lower boundaries of the scores and thus show the size of the range, in addition to the distribution of the actual data points over the range.

3.4.3. The progmax–regmin method for specifying a range

Assuming that a developmental process in general amounts to an overall increase in some phenomenon or variable of interest, we should take high values or levels that occur in the beginning of the process as particularly informative, simply because they are less expected to occur at an earlier stage than later on. Thus, if a high level occurs at some early point in time, high levels that occur later but that are not as high as the early one should not replace it in

our estimation of the variable's range or bandwidth. By the same token, low levels that occur at a later stage or point in time are particularly informative because they are less expected than higher values. They should not be concealed by low levels that occur at an earlier time and that are higher than the later one, if any such levels occur, of course. A simple way to implement this idea with longitudinal data points goes as follows. We specify a window with a period of 5, for instance, from the first data point on and compute the maximum value for that window. We then increase the window with one data point at a time, keeping its starting point (which is also the first point of our data series) constant and compute the maximum value of the extending window period. Similarly, we define a window, again with period 5, for instance, starting from the last point of our series and moving backwards. We compute the minimum value of that window and extend the window by one data point at the time, keeping the last point constant, which is also the last point of our data series. By doing so, we define the data series' progressive maximum and regressive minimum. The resulting line that circumscribes the collection of data points is closely related to the mathematical concept of an envelope or a so-called convex hull (see Fig. 7A and B). Note that this representation of the developmental range can again easily be achieved by means of any standard spreadsheet program.

Fig. 7A and B shows that both max–min methods in fact convey different kinds of information about or present a different perspective on a longitudinal data series. The prog(ressive)reg(ressive) method shows the range specified across the whole time period (up to the time point of interest, e.g., at time t_i it shows the range from the first point t_0 to that of time t_i, at time t_j it shows the range from t_0 to t_j, and so forth), whereas the ordinary max–min method shows ranges defined over considerably smaller time windows. The difference between the methods allows us to look for changes in the range's bandwidth, for instance in the form of temporary regressions, that show themselves as gaps between the ranges specified by the progreg and the ordinary windows method, respectively.

Concluding, in the previous sections we showed several techniques that specify and depict the range in which the developmental scores occur. These techniques can be used to obtain a first impression of the general trend of the developmental curve and the way variability is related to this general trend. They also give us an indication of where we can look for meaningful changes in variability. Before proceeding to a discussion of standard measures of variability, we will first briefly discuss the application of the preceding techniques to cross-sectional data.

3.4.4. An application to cross-sectional data

The techniques introduced so far do not only apply to individual trajectories or time series, collected with a single child. Also cross-sectional data can be described in terms of variability and changing ranges. The major difference is that the variability does not apply to fluctuations within a child but to differences between children of similar and different ages. By way of example, we present data from an ongoing study on the development of children's Theory-of-Mind (Blijd-Hoogewys et al., forthcoming). Theory-of-Mind refers to the child's ability to understand thoughts, beliefs, emotions, desires and so forth of other people and the relations between these mental phenomena and people's behavior. Theory-of-Mind is measured by means of a test, the Theory-of-Mind Story Books, which contains six parallel versions (suited for longitudinal research) with each 77 dichotomous items. The present results are based on a cross-sectional study of 220 children equally divided over both sexes, ranging from

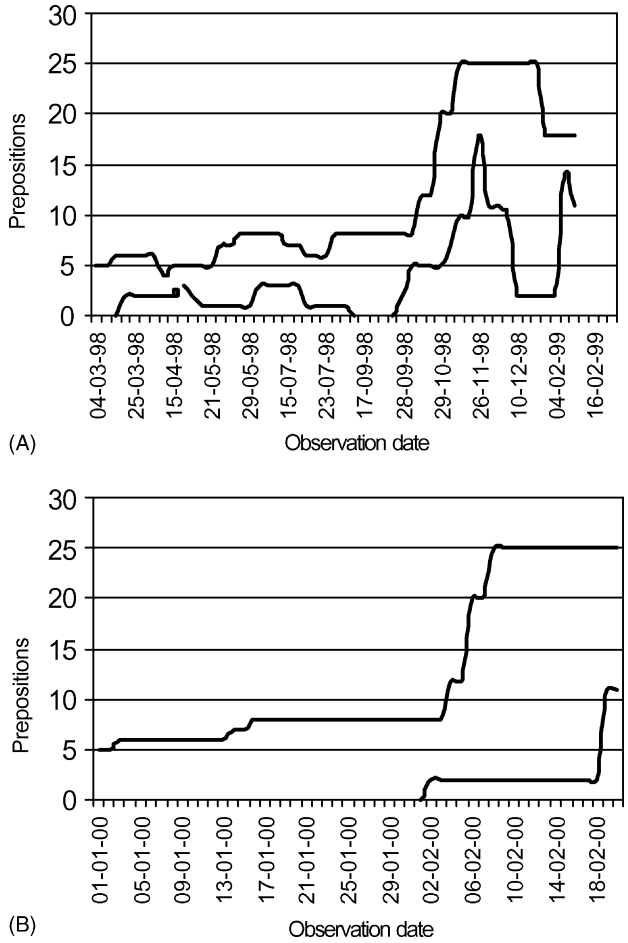
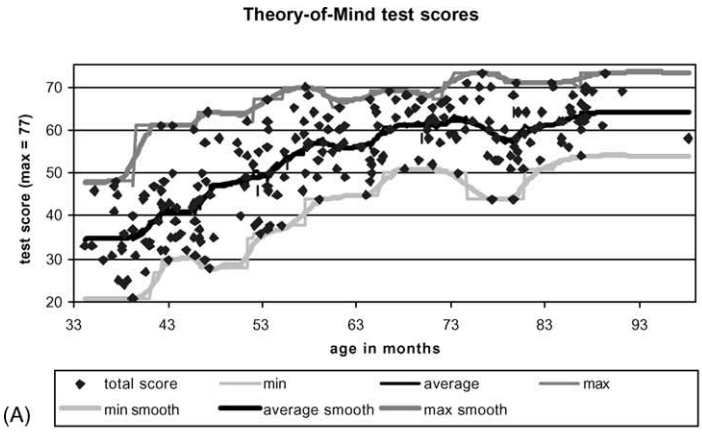


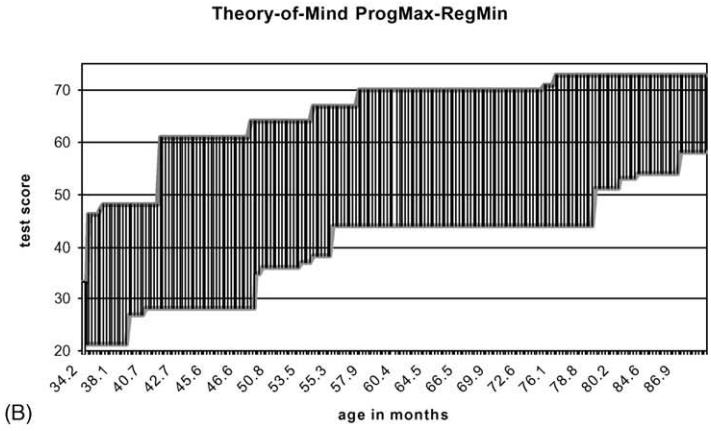
Fig. 7. (A, B) Application of the ordinary min–max method (window of 5 positions) (top), and the progmax–regmin method (bottom) to Heleen's spatial prepositions.

34 to 98 months. It is customary practice to present such data by first averaging the scores for age groups, the 3-year-olds, the 4-year-olds, and so forth, and then showing the increase of Theory-of-Mind understanding as the line through those averages. By so doing, interesting information about the relationship between inter-individual differences in children of approximately the same age and between children of different ages is lost. Instead, we argue for a representation of the data on the basis of the children's real ages. We can then apply the methods described earlier—the moving minmax, the progmax–regmin and the altitude lines—to the cross-sectional dataset. These methods show the quantitative change in Theory-of-Mind understanding in the form of a variable range (see Fig. 8 for various possibilities).

3.5. Standard measures of variability and their methodological problems

In addition to showing variability in graphs, we would also like to express variability in some sort of standard measure, because it can be used for the comparison of variability in





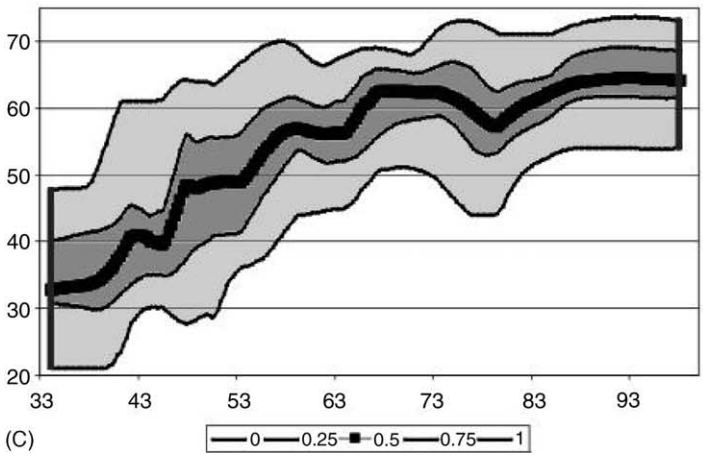


Fig. 8. (A–C) An application of the min–max, progmax–regmin and altitude line methods to cross-sectional data on the development of Theory-of-Mind. (C) Specifies altitude lines for the 0, 25th, 50th, 75th, and 100th percentile.

different samples. There are two common measures to describe variability in a sample: the SD and the coefficient of variation (CV). However, both have their own statistical problems when comparing variability in samples with different characteristics. The SD shows problems because of its sensitivity to the mean, and the CV shows problems with variables that have very low values. Data sets that begin with low values and show increasing means are called heteroscedastic (Kmenta, 1990). Heteroscedasticity, especially in the form of low initial values, increasing means and increasing variances, is likely to occur in developmental data.

Probably the best-known measure of variability is the SD. The SD is defined as the square root of the *variance*, which is in turn the average of the squared deviations from the mean. We can compute a SD for every meaningful data unit greater than one, and for every time unit. However, a problem arises if we want to compare SDs of different datasets. The reason for this is that the SD is very sensitive to the mean in a sample. We discussed this issue before when we showed the technique of moving minimums and maximums. The issue is that a higher mean is usually associated with a higher SD. Consequently, it is not possible to make a direct comparison between SDs of samples without taking the mean into account. In order to solve this issue, the CV is often used. The CV is defined as the standard deviation of a sample divided by its mean. We now have a measure that specifies the amount of SD in a standard unit of the mean, a measure that may indeed be helpful when comparing variability in different samples.

3.5.1. Mean Length of Utterance

One of the goals of our present study is to analyze variability in the developmental pathway. We ask ourselves if the amount of variability changes during the course of development. We want to investigate, for instance, if variability suddenly increases at some point in time. The study was also set up to analyze variability on different time scales. For the sake of illustration, we show here the results of the day-to-day measurements, and included only measurements from the intensive periods.

Table 1 shows the SDs, means and CVs of Heleen's MLU, for each of the six intensive periods. When we discussed the moving min–max graph, we asked ourselves whether variability increases with time for MLU. Looking at the pattern of SDs, we would indeed be inclined to conclude this. However, as can be seen in the table, the means also increase with time for almost every consecutive period. Consequently, the CVs show a subtler picture. While, on average, lower values occur in the first part of the set (periods 1–3) and higher values occur in the second part (periods 4–6), there exists no obvious simple increasing trend comparable to that of SDs and means. Given the small number of values (6) its seems hardly worthwhile trying to fit a regression line to see if the CVs show a statistically increasing trend. A simpler method is to check whether periods 1–3 are on average smaller than periods 4–6. This can be

Table 1
General measures of variability in time scale two: day-to-day variability in MLU-w

	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6
SD	0.03	0.10	0.12	0.23	0.14	0.24
M	1.06	1.13	1.51	1.76	2.09	2.55
CV	0.03	0.09	0.08	0.013	0.07	0.09

done with an exact permutation test, i.e., a test that computes all the different combinations of six elements into groups of three and compares the averages of those groups (Good, 1999). Since the number of groups that can be formed from six elements taken three by three is only 20, a *p*-value of 5% must correspond with the biggest difference out of the 20 possibilities. This difference occurs if the smallest values occur in periods 1, 2 and 3 and the larger values occur in periods 4, 5 and 6. It is easy to see that this is not the case (the second smallest CV, 0.07 occurs in period 5). Hence, the averages of periods 1, 2 and 3 and 4, 5 and 6 are not statistically significantly different at the 5% level. An exact permutation test reveals that the difference between the averages is the third smallest out of the 20 possibilities, which corresponds with a *p*-value of 15%. Further on in this article, we will get back to the possibilities random sampling techniques have to offer.

Concluding, in the case of MLU the SD as a general measure of variability would lead to an overestimation of the variability in the later part of the trajectory. Therefore, the SD is unsuitable for making comparisons between and within samples. So far, the CV seems to be the better candidate for such a goal.

3.5.2. Spatial prepositions

Now we turn to the data of *spatial prepositions*. Here we asked ourselves the question whether the growth curve shows a shift in growth pattern from measurement point 38 (November 12) onwards. Visual inspection of the *XY*-graph, and the moving min–max graph certainly points in that direction. In the early part of the trajectory, there seem to be small fluctuations, while the later part shows dramatic peaks and wells. The demarcation point falls on November 12 and lies between the fourth and fifth intensive period. As a consequence, we expect the later two intensive periods to be more variable than the first four.

When looking at Table 2 such an effect turns out to be absent. Even stronger, the effect seems to be reversed. The highest CV values are to be seen in the first, second and fourth intensive period, while periods 5 and 6 show moderate values. The *p*-value of the difference between the averages of periods 1, 2 and 3 and 4, 5 and 6, respectively, calculated with an exact permutation test is .70. This means that the difference is far from significant. It might seem surprising that the CVs do not confirm the pattern we thought we recognized with visual inspection. Is it true that especially the early intensive periods are the most variable ones? This hardly seems likely. In fact, the effect observed in the data is a good illustration of *heteroscedasticity*, a statistical problem that is very common if one deals with growth data. Heteroscedasticity is often associated with the presence of very low initial values. Such low values are unstable, because small absolute fluctuations are large in proportion to the values themselves. Consider for instance a variable where the values 1 and 2 succeed each other. These minimal fluctuations

Table 2
General measures of variability in time scale two: day-to-day variability in spatial prepositions

	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6
SD	2.14	1.52	1.47	3.25	5.54	2.48
M	3.16	2.5	5.17	4.83	17.5	15.17
CV	0.67	0.61	0.28	0.67	0.32	0.16

(there is no smaller measurement unit possible) lead to a very high CV, for the simple reason that this variable shows a fluctuation of a 100% (i.e., the proportion 1:1). The CV resulting for these data would be the same for a dataset in which the values 100 and 200 succeed each other in a similar fashion, which would amount to an oscillation which magnitude is rather unlikely. The reason heteroscedasticity occurs is because in the study of language acquisition (and many other developmental domains), there is by definition an absolute point zero. Also, the unit of measurement reaches its lower limit, for instance with spatial prepositions the minimal unit is one. Smaller units are simply not possible.

Recapitulating, we have seen that the two common measures to describe variability (the standard deviation and the coefficient of variation) have their own statistical problems when comparing variability in samples. While the standard deviation shows problems because of its sensitivity to the mean, the coefficient of variation shows problems with variables with very low values because of heteroscedasticity. These problems are especially serious when analyzing development. In order to observe development, the datasets should combine these two characteristics: they ideally start out with very low values and further show considerable growth. Therefore, both the standard deviation and the coefficient of variation are not suitable as general measures of to analyze variability patterns in developmental trajectories.

3.6. The critical moment method

There is another method to establish variability in data, or more precisely, to establish at what point in time variability significantly increases compared to a relatively stable period. This method is developed in the field of motor coordination (e.g., Verheul & Geuze, 1999) and can be applied to developmental research in general. The method is based on the following assumptions. First of all, a system is supposed to be relatively stable over some initial period of time. Second, this period must be followed by a period in which this system becomes "unstable," which results in large variability. The aim of this method is to establish if, and at what exact moment, the system loses its stability. We believe that the data of Heleen's early preposition use meet these assumptions in a satisfactory way. First of all, visual inspection of the developmental trajectory shows that there exists a period in which the system seems relatively stable. Secondly, at some point in time there is a period that shows larger fluctuations. This can very well be compared to a dynamic system that loses its stability. First, we ask ourselves if there is a point in time after which variability *critically* increases (comparable with the critical frequency which is used in the field of bimanual coordination, Kelso, Scholz, & Schöner, 1986; Verheul & Geuze, 1999). Secondly, we are interested in the pattern in which variability increases and eventually decreases again. Does the system gradually lose its stability or does this happen very suddenly? Does the system regain its stability at some point in time, and to what developmental incidence can this eventually be related?

In the application of this technique, we have to bear in mind again that we are dealing with a variable that shows a considerable increase, i.e., a growth trend. We do not want the general trend to influence the variability measure. In order to eliminate the influence of the general increase, we have to detrend the data, using a trendline. In order to obtain an optimal fit, we used a flexible regression model. With this model a moving linear regression equation was calculated on a moving window of 19 data points. Thus, slopes and intercepts were estimated

for each (moving) window of 19 data points (the result resembles a loess method smoothing, but the moving regression has the advantage that it can easily be implemented in a spreadsheet program, for instance). We proceeded by calculating the residuals of the original data for this regression model. The critical period was determined as follows. First, we calculated a moving standard deviation (using a moving window of five observation points) on these residuals. We took a timeframe that is relatively stable (in our case the first 21 observations), and calculated the 95% reliability interval. Secondly, we tried to establish at what moment the variability in the system increases. We defined this moment as the moment at which the moving SD exceeds the critical value (which is the upper limit of the reliability interval) for at least six consecutive moving SDs.

We applied this technique to the data of Heleen's spatial prepositions, and plotted both the preposition data and the moving SDs on the residuals of these data in Fig. 9. In the case of Heleen, the resulting critical value of the spatial prepositions was 3.61, based on 1.96 times the standard deviation of the first 21 measurement points. As can be seen in Fig. 9, this value of 3.16 is exceeded only once at measurement point 29 (September 16) and exceeded again at measurement point 37 (October 29). This value is exceeded for the seventh consecutive time only after measurement point 42 (December 1), which means that point of significant increase in variability is located at this date.

With this method, we do not have the problem of heteroscedasticity that we encountered when we used the SDs as a general index of variability. SDs are much less sensitive to these low values than CVs. We do however have the problem of a higher mean being associated with larger variability. The preposition data show mixed results. First, we see a slow, but somewhat irregular, increase in the moving CV. We cannot be sure that this is not caused by the increasing mean of the original data. However, we have some indication that the moving SDs decline at the end of the trajectory, while the mean of the original data remains high. Although the decline

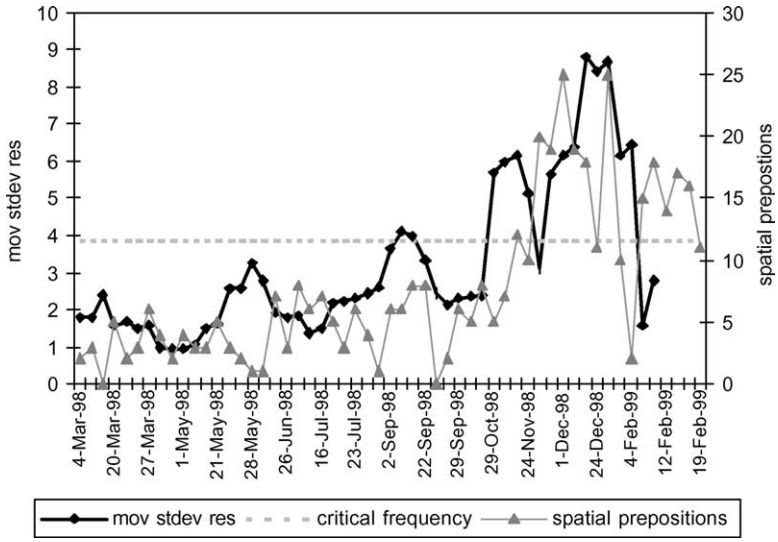


Fig. 9. Application of the critical moment method to Heleen's preposition data.

in our data is not very strong, we suggest that a decline of the moving standard deviation in general can be an important indication of a transition, since such an observation cannot be explained by the external causes such as measurement error. It seems that there has to be some other explanation for this decreasing variability, and this explanation can probably be found in the internal dynamics of the developmental process itself.

3.7. The distribution of fluctuations

The initial assumptions of the measurement-error-hypothesis suggest another technique to investigate the variability pattern. One of the assumptions is that error, or noise, is supposed to be symmetrically distributed around the central tendency, which represents the best estimation of the true underlying variable. A variability distribution that shows a different pattern does not agree with this assumption. For this reason, the distribution characteristic can give us additional information on the tenability of the error-hypothesis.

When studying the distribution characteristics, there is one property we are especially interested in, namely the skewness. We believe that the skewness can give us information about the degree of consolidation of the acquired developmental variable. When a child begins to discover that he or she can use prepositions to express spatial relations, we expect outliers on the positive side of the distribution. It is highly likely that the child uses this new linguistic category (in this case spatial prepositions) in outbursts, but most of the time this new category is not used and the child simply sticks to his or her usual repertoire (for instance by pointing to a location). This behavior would lead to a positively skewed distribution. In contrast, when the use of spatial prepositions is relatively well-consolidated, we expect the child to use these prepositional utterances most of the time, and only use less sophisticated ways of expressing spatial relations relatively rarely. This well-consolidated state in the acquisition of prepositions would result in a negative skewness.

In dynamic terms, a skewed distribution could possibly be an indication for the existence of bimodality (Alibali & Goldin-Meadow, 1994). Bimodality refers to a situation in which two equilibria exist in a developmental trajectory. In the case of language acquisition, these equilibria represent language rules (or linguistic categories). One equilibrium refers to the old rule and another equilibrium refers to the new rule. Bimodality can be an indication for a developmental transition between the old and the new rule or strategy. A skewed distribution might be indicative of a bimodal equilibrium with considerably different frequencies of occurrence.

What is the distribution of the variability of Heleen's spatial prepositions? In order to eliminate the influence of the general increase, it is important to detrend the data, using a trendline. In order to obtain an optimal fit, we used the flexible regression model we also applied with the critical moment method. We calculated the residuals for this regression line. Consequently, the distribution of these residuals was studied for skewness, using a "moving" skewness factor on a moving window of 13 data points (the choice of the period of the moving window is somewhat arbitrary; we have chosen a period of 13 because that contains enough data to make a reasonable estimation of the skewness possible, without covering too much of our data set and thus concealing changes in the skewness that might occur during the observed trajectory). Note that with this technique we combined the analysis of distribution characteristics with ideas we presented earlier discussing the moving window techniques (e.g.,

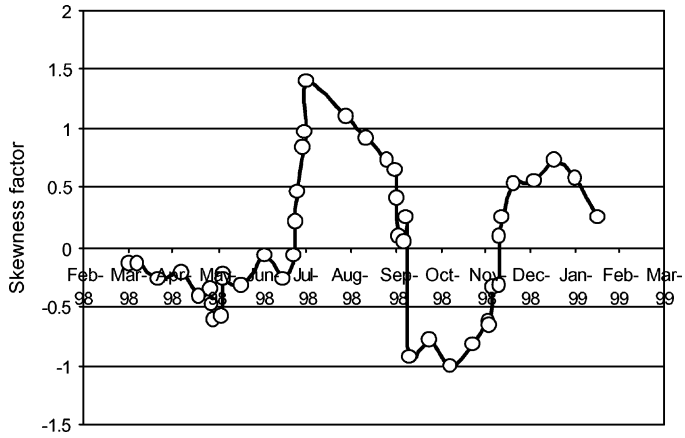


Fig. 10. Moving skewness graph of Heleen's spatial prepositions.

the moving min–max graph). Both the regression equation (slopes and intercepts) and the calculation of the skewness factor are based on so-called moving windows, which move up one position at a time. The result is a flexible symmetry analysis for each part of the developmental trajectory. The values of these moving skewness factors are plotted in a moving skewness graph (see Fig. 10).

Evidently, the degree of symmetry is not equal in each part of the developmental trajectory. Instead, an oscillating pattern is observable, with four distinctive parts. The data start out with a near symmetric (only slightly negatively skewed) distribution, which turns into a strong positively skewed distribution from June to September. Then, from September to November, a fairly strong negative skewness is detected, again followed by a part with only moderate positive skewness factor, after November. Note again that error is supposed to be symmetrically distributed around the central tendency, which is supposed to represent the best estimation of the true underlying variable. The oscillating skewness patterns that are found in Heleen's spatial prepositions do not support this symmetry-assumption.

A positive skewness, such as seen in the second part of the moving skewness graph, means that the right tail of the data distribution is longer. There are more outliers on the positive side of the graph. Positive skewness is likely to be associated with a developmental process where the degree of consolidation of a syntactic rule is still low. We expect that later on in the developmental process, this degree of consolidation will increase. This, in turn will lead to a situation in which the new rule will be used most of the time and negative outliers will occur in situations where the old rule is still used. This would lead to a negatively skewed distribution, as is observed in the third part of the moving skewness graph of Heleen's spatial prepositions. We expect that the use of the old rule will finally disappear, which will then lead to a symmetric distribution.

In a standard statistical package, such as SPSS, there is an option to test the normality of a distribution by means of a Kolmogorov–Smirnov test (for studies with $n = 50$ or more) or a Shapiro–Wilk test (for $n = 50$ or less). However in studies with a small number of cases per sample, the null-hypothesis of a normal distribution will seldom be rejected. Although it

goes well beyond the scope of this article, we would like to point out the possibilities that random sampling techniques have to offer. Bootstrapping techniques can be applied to test the significance of the differences in skewness values of, for instance, the four distinctive periods in Heleen's moving skewness graph. We will elaborate on the possibilities of random sampling techniques further on. Suffice it to say for now that we applied the bootstrap technique to our skewness data and found that the four skewness phases observed in the data correspond with two overlapping skewness distributions with different averages.

3.8. Pre-processing data for further analysis: the effect of different detrending models

It is important to note that the way the data are detrended in the application of the critical moment method and the moving skewness graph may be essential to the results of this analysis. This may especially be the case if the trendline chosen to detrend the data has a poor fit. For instance, if a simple linear regression is applied to data that show an obviously S-shaped curve, the outcome of the skewness factors can be greatly influenced. We must however, warn against detrending with a too complex model, which eliminates the variability we mean to analyze. The model should only be used to eliminate the effect of trend. As a general rule-of-thumb, we suggest to use the simplest model possible. If the trend looks linear, choose a linear model, if not, choose a simple flexible model, such as our flexible regression model or a loess smoother with a sufficiently long period. The choice of the detrending model is primarily a conceptual matter: one has to decide, on reasonably defendable grounds, what shall be conceived of as the main trend against which the variability will be plotted.

It might however be worthwhile to test the effect of the detrending model on the results. In the case of the critical frequency method for Heleen's prepositions, we applied various detrending models. The results showed that the effect of the model was very small. Although the critical frequency varied somewhat across models, the system lost stability at the same measurement point in all instances. This indicates that the effect found in the analysis is fairly robust. This might however not be the case for other datasets. We also tested the sensitivity of the skewness results by comparing various detrending models. The skewness pattern turns out to be independent on the exact method of detrending, as long as it results in a trend line that follows the general pattern of data sufficiently close. For instance, a qualitatively similar result was obtained when the data were detrended by means of a loess smoother. A completely different way of detrending, differencing, also yielded similar results (differencing means that the residuals are replaced by the difference between a measurement and the preceding measurement, Gottman, 1981).

3.9. Random sampling methods

Our study of developmental ranges and the variability within those ranges relies, by definition, on extreme values, namely the maxima and minima. If proper care is taken in the language data collection—and imitation games and songs have been removed from the data, for instance—the extremes are as reliable as the more central values. However, in the estimation of a—moving—range, extreme values have a considerably stronger effect on the estimation of the range's boundaries than on the estimation of a central value, such as an average. Take

for instance the following imaginary series of observed frequencies: 2, 3, 4, 10, 5, 6, 7. If we calculate the moving maximum for a window of period 4, the maximum is 10 throughout the series of seven measurements. If we compute the moving average with period 4, however, it amounts to 4.7, 5.5, 6.25, 7. This difference should not make us decide to remove the value of 10 from the series as a so-called outlier. An outlier it is, but it has been obtained in a reliable way and thus provides information about the observed child's abilities at that particular moment. In fact, we might have been lucky that we saw the child at a moment where it actually showed a glimpse of a rapidly increasing linguistic capacity. If we had come a day earlier or later, we might have found a frequency of 4 or 5.

3.9.1. Bootstrapping as a technique for estimating model probabilities in individual longitudinal data sets

Given this possibility, we wish to know to what extent our conclusions, for instance with regard to the average bandwidth of the range, are sensitive to sampling characteristics. In a standard design, we want to know to what extent sampling characteristics prevent us from estimating the correct value of an observed variable for the population from which the sample has been drawn. However, just like in any other comparable case, we have no other information about the "population" of observations than the sample itself. Furthermore, we cannot invoke assumptions about expected distributions, a normal distribution across the population, for instance. First, we usually have no idea of what the distribution of our longitudinally observed data should be. Second, the distribution characteristics (average, standard deviation, skewness, etc.) are not stationary over the observation period, because the variable at issue is rapidly developing. A method that allows us to nevertheless estimate the effect of sampling characteristics on our measurements is the so-called bootstrap method (Chernick, 1999; Efron & Tibshirani, 1993; Good, 1999). The method consists of randomly drawing a large number of subsamples (e.g., 1,000) from our original sample. For each subsample, we compute the test statistic of interest (for instance the average bandwidth of the range of frequencies of spatial prepositions before the "jump"). By comparing the value of the test statistic of the original sample with the distribution of the test statistics from the randomly drawn subsamples, we obtain a fairly good estimation of how subsamples relate to the original sample and thus of how the original sample relates to the population from which it is drawn.

In our design—2-weekly observations interspersed with intensive observation periods—we have a particularly good opportunity for testing the eventual effects of the standard 2-weeks approach on the estimation of the average bandwidth of the observed frequencies of spatial prepositions. For instance, we would like to know to what extent the bandwidth depends on the sampling frequency. Our current dataset with its combination of 2-weekly observations and intensive observation periods offers an interesting opportunity to try to answer that question. Our starting assumption is that each of the observations during the intensive periods could have been a potential 2-weekly observation, if all our observations had been scheduled according to the standard 2-weeks interval. We constructed 1,000 randomized observation samples, consisting of all our 2-weekly observations to which we added one observation from each intensive period, randomly drawn from each of those periods. For each randomly drawn sample we computed the average bandwidth of the time period before the jump in the use of spatial prepositions takes place (with observation 38 as the demarcation point) and the average

Table 3
Results of the subsampling procedure

	Average bandwidth	SD	Min	Max	Median
Subsamples					
Substage 1	6.69	0.62	5.48	8.69	6.62
Substage 2	20.05	0.5	19.67	21.17	19.89
Data					
Substage 1	8.03 ($p = 0.015$)				
Substage 2	21.17 ($p = 0.115$)				

bandwidth of the time after the jump. The bandwidths are based on windows that cover 58.4 days on average. Table 3 summarizes the results of the subsampling procedure.[2]

The table should be read as follows. For the bootstrapped samples, the average bandwidth is 6.69 and 20.05 for the first and second substage, respectively (standard deviations and additional statistics are given in the matrix). The average bandwidth of the original data is 8.03 and 21.17 for the first and second substage, respectively. In our 1,000 bootstrapped samples of the first substage, we found a value equal to or bigger than 8.03 in only 1.5% of the cases. However, the bootstrapped samples of the second substage were equal to or bigger than the empirical bandwidth in 11.5% of the cases. We can conclude, therefore, that the high sampling frequency (as defined by the intensive periods) has led to a significantly bigger bandwidth estimation than the standard sampling frequency of 2 weeks in the case of the first substage, but not in the case of the second substage. It should be noted, however, that also in the case of the second substage, the difference goes in the expected direction (bigger estimated bandwidth if sampling frequencies are higher).

The present subsampling procedure has been carried out for illustrative purposes only and differs from the standard bootstrap technique. A further elaboration of this issue, however, would far extend the scope of the present article.

In summary, bootstrap and resampling methods can be applied to longitudinal datasets of the kind described in the present article to help us understand the effect of sampling characteristics on our estimation of the ranges within which the observed variables vary.

3.9.2. Bootstrapping and generalization of models over children

So far, the techniques described applied to individual data trajectories. We have also shown that techniques for visualizing ranges and variability can be as easily applied to cross-sectional data. The problem with intensive individual research is that the number of cases that can be covered in one study is usually quite small (ranging from 1 to a few, e.g., four to five intensively studied children). The question is, how can we generalize our findings to the population (a basic question in social science) given we have only so few cases? Such generalization is possible if every single case is conceived of as a separate study. The generalization problem becomes one of meta-analysis: how can *p*-values or other test statistics obtained in independent studies, based on their own accidental samples, be combined into an overall *p*-value or test statistic? In the case of longitudinal studies, the sample is a time series of consecutive measurements of a single child. A simple but effective statistic suitable for meta-analysis is Fisher's combined

p-value (Glass, McGaw, & Smith, 1981; Snijders & Bosker, 1999). The technique of *p*-value combination can be approximated, if it is possible to test individual null-hypotheses by means of bootstrapping (or related permutation techniques). In the preceding paragraph, we have shown that a model of differences in variability over distinct substages can be tested by means of bootstrapping. For each bootstrapped sample of an individual child's data set, we calculate a bootstrap estimation of some statistic of interest (for instance, the aforementioned difference in bandwidths). If we do a bootstrap test for a small number of children, for instance five, we have five series of test statistics based on each child's null-hypothesis. We then randomly draw sets of five test statistics, one from each child, and repeat this a sufficient number of times (e.g., 1,000 times). We can easily compute how many times the average test statistic of the randomly drawn sets of five is equal to or exceeds the average of the five *observed* test statistics. This number is the combined *p*-value of our five separate studies, i.e., our five individual children (see for instance Good, 1999).

4. Summary and conclusion

The assumption that (intra-individual) variability in the data is basically an expression of measurement error, is a deeply rooted and often also tacit belief of many developmental psychologists and social science researchers in general. This belief is maintained and even amplified by the use of a standard toolbox of statistical techniques, each of which implicitly supports the error-hypothesis. In addition, developmental theories make little room for variability within individuals as a phenomenon of interest, either as an indicator of development or as cause or condition of change. Against this vicious circle of neglect, we have placed three alternatives. First, we have briefly pointed to an approach to development that conceives of variability as an important phenomenon, namely dynamic systems theory. Second, we have given an overview of studies that have shown that intra-individual variability is an interesting variable in its own right and that it occurs in various forms. Third, and most importantly, we have introduced a number of simple techniques for making variability visible, in order to help researchers explore this interesting source of information.

It is our firm belief that the starting point of developmental studies should be studies of individual trajectories, with as many repeated measurements as possible. Instead of conceptualizing a child's developmental level or developmental state as a hidden true value, concealed by the vagaries of error laden measurement, we invite researchers to look at a child's level as a range, a specific domain of variability, the properties of which change over the course of development. In order to help researchers achieve this goal, we introduced and discussed various techniques.

First, we discussed the most common techniques: *visual inspection of raw data* (vs. smoothing techniques) and two general variability measures: the *standard deviation* and the *coefficient of variation*. We reviewed the complications of these two measures when applying them to developmental (growth) data.

Second, we proposed a range of new techniques that were constructed specifically for the study of patterns of variability. First we presented several methods that show variability in a graph: the *moving min–max graph* (including the use of *altitude or percentile lines*), and the

progmax–regmin graph. The commonality between these graphs is the representation of the observed score within its score range in a specific time window, for each point in time. These methods are especially useful for obtaining a general impression of the variability pattern (e.g., is it generally increasing or decreasing; are there changes in the bandwidth?) that may be helpful in generating testable hypotheses.

Furthermore, we proposed a technique that is able to detect sudden increases of variability: *the critical frequency method*. With this technique, we calculated if and when fluctuations become "critically large," and the system loses its stability. We also proposed a technique that is based on a central assumption of measurement-error-hypothesis: namely the symmetric distribution of error. By investigating the *moving skewness of the distribution pattern* we can test the tenability of this hypothesis in the data. We argued that the direction of the skewness could give information about the degree of consolidation of a new ability (a positively skewed distribution suggests a low degree of consolidation, a negatively skewed distribution a high degree of consolidation). Finally, as traditional statistical techniques offer little in testing variability hypotheses, we suggested the potential benefits of employing *random sampling* techniques. We have given a example of how sampling techniques can be used to test hypotheses about the pattern of variability.

In line with authors such as Loftus (1996) and Tukey (1977), we believe that psychology in general and developmental psychology in particular will greatly benefit from a more exploratory approach to the data. The approach should be primarily aimed at making the interesting phenomena visible. In developmental psychology, variability is such an interesting phenomenon, although it has long been neglected. If researchers accustom themselves to begin their data analysis by inspecting the patterns of variability within and between individuals, developmental psychology will have a chance to overcome the largely static and in fact non-developmental image that has prevailed in the past decades.

Notes

1. A resampling procedure showed that, on average, variability of four infants (followed from age 1;6 to 2;6, among which Heleen) were larger than that of two adults samples. In all infant cases, samples of six sessions were selected randomly to calculate the CV. For these infants, we only used sessions after the first large increase in preposition use, because of heteroscedasticity (a statistical problem we will discuss later on). While the adults showed CVs of, respectively 0.242 and 0.144, the infants had average CVs of 0.534 (subject Heleen), 0.560 (subject Lisa), 0.459 (subject Jessica) and 0.513 (subject Berend). The resampling procedure based on 2,000 iterations, showed that the probability that these infants acquired the adult CV values of 0.242 and 0.144 and below, were, respectively $<.005$ and $<.005$ (subject Heleen), .068 and .01(subject Lisa), .04 and .01 (subject Jessica), .01 and $<.005$ (subject Berend). This means that the probability that the adult values come from a distribution similar to that of the infant is in seven out of eight cases below 5% and in 1 case 6.8%. These results indicate that it is highly likely that the adult use of prepositions-in-context shows a lower variability than that of the children.

2. Since the data set contains about twice as much observations as the subsamples, the average bandwidths of the data set were calculated on the basis of windows that were twice as big as those used for the subsamples. Since the average number of days covered by the windows in the subsamples was slightly smaller than that of the data set (with 7%) the average bandwidths of the subsamples were corrected by multiplying the values with 1.07.

References

Alibali, M., & Goldin-Meadow, S. (1994). Gesture–speech mismatch and mechanisms of learning: What the hands reveal about the child's state of mind. *Cognitive Psychology*, *25*(4), 468–523.

Arlman-Rupp, A. J. L., Van Niekerk de Haan, D., & Van de Sandt-Koenderman, M. (1976). Brown's early stages: Some evidence from Dutch. *Journal of Child Language*, *3*, 267–274.

Bertenthal, B. (1999). Variation and selection in the development of perception and action. In G. Savelsbergh, H. van der Maas, & P. van Geert (Eds.), *Non-linear developmental processes* (Vol. 175, pp. 105–121). Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.

Bertenthal, B., & Clifton, R. (1998). Perception and action. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology. Cognition perception and language* (Vol. 2, pp. 51–102). New York: Wiley.

Blijd-Hoogewys, E. M. A., Huyghen, A.-M. N., van Geert, P. L. C., Serra, M., Loth, F., & Minderaa, R. B. (forthcoming). Denken overdacht: De normering van het Theory-of-Mind Takenboek [Thinking about thoughts: Setting standard norms for the Theory-of-Mind Story Book]. Nederlands Tijdschrift voor de Psychologie.

Brown, R. (1973). *A first language: The early stages*. London, UK: Allen & Unwin.

Canfield, R. L., Wilken, J., Schmerl, L., & Smith, E. G. (1995). Age-related change and stability of individual differences in infant saccade reaction time. *Infant Behavior and Development*, *18*, 351–358.

Chabon, S., Kent-Udolf, L., & Egolf, D. (1982). The temporal reliability of Brown's Mean Length of Utterance (MLU-m) measure with post-stage V children. *Journal of Speech and Hearing Research*, *25*, 124–128.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley.

Crockerberg, S. B., & Smith, P. (1982). Antecedents of mother–infant interaction and infant irritability in the first 3 months of life. *Infant Behavior and Development*, *5*, 105–119.

Cronbach, L. J. (1960). *The essentials of psychological testing*. New York: Harper & Brothers.

De Weerth, C. (1998). *Emotion-related behaviors in infancy: A longitudinal study of patterns and variability*. Doctoral dissertation, University of Gorningen.

De Weerth, C., van Geert, P., & Hoitink, H. (1999). Intra-individual variability in infant behavior. *Developmental Psychology*, *35*(4), 1102–1112.

Dittrichova, J., Paul, K., Tautermannova, M., & Vondracek, J. (1992). Individual variability in infant's early behavior. *Studia Psychologica*, *34*, 199–210.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Feldman, H. M., Dolaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages 1 and 2 years. *Child Development*, *71*, 310–322.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., & Thal, D. (2000). Measuring variability in early child language: Don't shoot the messenger. *Child Development*, *71*(2), 323–328.

Fischer, K. W., Bullock, D., Rotenberg, E. J., & Raya, P. (1993). The dynamics of competence: How context contributes directly to skill. In R. H. Wozniak & K. W. Fisher (Eds.), *Development in context: Acting and thinking in specific environments* (pp. 93–117). Hillsdale, NJ: Erlbaum.

Fischer, K. W., & Granott, N. (1995). Beyond one-dimensional change: Multiple, concurrent, socially distributed processes in learning and development. *Human Development*, *38*(6), 302–314.

Frijn, J., & de Haan, G. J. (1994). *Het taallerend kind (The language-learning child)*. Dordrecht: ICG Publications.

Gilmore, R. (1981). *Catastrophe theory for scientists and engineers*. New York: Wiley.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. London: Sage Publications.
Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, *100*, 279–297.
Good, P. I. (1999). *Resampling methods: A practical guide to data analysis*. Boston: Birkhauser.
Gottman, J. M. (1981). *Time series analysis. A comprehensive introduction for social scientists*. Cambridge, UK: Cambridge University press.
Granott, N. (1993). Patterns of interaction in the co-construction of knowledge: Separate minds, joints efforts and weird creatures. In R. H. Wozniak & K. W. Fisher (Eds.), *Development in context: Acting and thinking in specific environments* (pp. 183–207). Hillsdale, NJ: Erlbaum.
Hadders-Algra, M., Brogren, E., & Forssberg, H. (1996). Ontogeny of postural adjustments during sitting in infancy: Variation, selection and modulation. *Journal of Physiology*, *493*(1), 273–288.
Hickey, T. (1991). Mean Length of Utterance and the acquisition of Irish. *Journal of Child Language*, *18*, 369–553.
Kelso, J. A. S., Scholz, J. P., & Schöner, G. (1986). Non-equilibrium phase transitions in cooridinated biological motion: Critical fluctuations. *Physics Letters A*, *118*(6), 279–284.
Kmenta, J. Heteroscedasticity. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Time series and statistics* (pp. 103–104). New York: Norton.
Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science*, *6*, 133–139.
Lautrey, J. (1993). Structure and variability: A plea for a pluralistic approach to cognitive development. In R. Case & W. Edelstein (Eds.), *The new structuralism in cognitive development: Theory and research on individual pathways* (pp. 101–114). Basel, Switzerland: Karger.
Lautrey, J., Bonthoux, F., & Pacteau, C. (1996). Le traitement holistique peut-il guider le traitement analytique dans la categorisation de visages? [Can holistic processing guide analytic processing for face categorization]. *Annee Psychologique*, *96*, 225–254.
Lautrey, J., & Cibois, P. (1994). Application of correspondence analysis to a longitudinal study of cognitive development. In D. Magnusson & L. R. Bergman (Eds.), *Problems and methods in longitudinal research: Stability and change* (pp. 190–211). Cambridge, UK: Cambridge University Press.
Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*(6), 161–171.
Lord F. M., & Novick, R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
MacWhinney, B. (1991). *The Childes project, tools for analyzing talk*. Hillsdale, NJ: Erlbaum.
Minifie, F., Darley, F., & Sherman, D. (1963). Temporal reliability of seven language measures. *Journal of Speech and Hearing Research*, *6*(2), 139–148.
Newell, K., & Corcos, D. M. (1993). *Variability and motor control*. Champaign, IL: Human Kinestics.
Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill
Rittle-Johnson, B., & Siegler, R. S. (1999). Learning to spell: Variability, choice, and change in children's strategy use. *Child Development*, *70*, 332–348.
Ruhland, R., & van Geert, P. (1998). Jumping into syntax: Transitions in the development of closed class words. *British Journal of Developmental Psychology*, *16*, 65–95.
Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science*, *3*, 1–5.
Siegler, R. S. (1996). *Emerging minds. The process of change in children's thinking*. New York: Oxford University Press.
Siegler, R. S. (1997). Concepts and methods for studying cognitive change. In E. Amsel & K. A. Renninger (Eds.), *Change and development: Issues of theory, method, and application* (pp. 77–97). Hillsdale, NJ: Erlbaum.
Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.
St. James Roberts, I., & Wolke, D. (1984). Comparison of mothers' with trained-observers' reports of neonatal behavioral style. *Infant Behavior and Development*, *7*, 299–310.
Thelen, E. (1985). Development origins of motor coordination: Leg movements in human infants. *Developmental Psychobiology*, *18*, 1–22.

Thelen, E., & Smith, L. B. (1993). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thom, R. (1975). *Structural stability and morphogenesis: An outline of a general theory of models*. Reading, MA: Benjamin.

Thordardottir, E., & Weismer, S. E. (1998). Mean Length of Utterance and other language sample measures in early Icelandic. *First Language*, *18*, 001–032.

Tukey, J. W. (1977). *Exploratory data analysis*. New York: Addison-Wesley.

van der Maas, H., & Molenaar, P. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*(3), 395–417.

Van der Meulen, B. F., & Smrkovsky, M. (1983). BOS 2-30. *Bayley ontwikkelingsschalen, handleiding* [Bayley developmental scales manual]. Lisse, The Netherlands: Swets & Zeitlinger.

van Dijk, M., De Goede, D., Ruhland, R., & van Geert, P. (2000). Kindertaal met bokkensprongen [Child language cuts capers]. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden*, *55*, 232–245.

van Dijk, M., & van Geert, P. (1999). *Short-term variability in child language*. Unpublished manuscript, The Netherlands: University of Groningen.

van Geert, P. (1994). *Dynamic systems of development: Change between complexity and chaos*. New York: Harvester wheatsheaf.

van Geert, P. (2000). *Language-analysis Excel add-inn Groningen (LEGro)* [Computer Software]. Groningen, The Netherlands: Author.

van Geert, P., Savelsbergh, G., & van der Maas, H. (1997). Transitions and non-linear dynamics in developmental psychology. In G. Savelsbergh, H. van der Maas, & P. van Geert (Eds.), *Non-linear developmental processes* (Vol. 175, pp. X–XX). Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.

Verheul, M., & Geuze, R. (1999). Constraints on the dynamics of bimanual finger tapping. In N. Gantchev & G. N. Gantchev (Eds.), *From basic motor control to function recovery* (pp. 358–362). Sofia, Bulgaria: Academic Publishing House "Prof. M. Drinov."

Wachs, T. D., Morrow, J., & Slabach, E. H. (1990). Intra-individual variability in infant visual recognition memory performance: Temperamental and environmental correlates. *Infant Behavior and Development*, *13*, 397–403.

Wimmers, R. H. (1996). *Grasping developmental change: Theory, methodology and data*. Doctoral dissertation, Free University of Amsterdam.

Wimmers, R. H., Savelsbergh, G.-J. P., Beek, P. J., & Hopkins, B. (1998a). Evidence for a phase transition in the early development of prehension. *Developmental Psychobiology*, *32*, 235–248.

Wimmers, R., Savelsbergh, G., Beek, P., & Hopkins, B. (1998b). A catastrophic change in the early development of prehension? In G. Savelsbergh, H. van der Maas, & P. van Geert (Eds.), *Non-linear developmental processes* (Vol. 175, pp. 125–136). Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.