UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Psicologia dello Sviluppo e dei Processi di Socializzazione

SCUOLA DI DOTTORATO DI RICERCA IN: Scienze Psicologiche
INDIRIZZO: Psicologia dello Sviluppo e dei Processi di Socializzazione
CICLO: XXIV

THE EFFECT OF STIMULUS VARIABILITY ON CHILDREN'S JUDGMENTS OF QUANTITY

Direttore della Scuola*:* Ch.mo Prof. Clara Casco
Coordinatore d'indirizzo: Ch.mo Prof. Maria Chiara Levorato
Supervisore: Ch.mo Prof. Franca Agnoli

Dottorando: Gianmarco Altoè

Dedicated to Martina and Betta

Gianmarco, 2012

TABLE OF CONTENTS

RIASSUNTO

Il concetto di variabilità, intesa come dispersione dei dati osservati, ha un ruolo centrale nelle scienze statistiche e nelle decisioni quantitative della vita quotidiana. Tuttavia, le ricerche in ambito educativo si sono focalizzate solo recentemente sullo studio del modo in cui si sviluppano le abilità di ragionamento riguardanti la variabilità (Garfield & Ben-Zvi, 2005). Da una prospettiva cognitivo-evolutiva, sono state condotte molte ricerche sul giudizio di quantità nei bambini a partire dai primi lavori di Piaget, ma pochi studi hanno indagato lo sviluppo del ragionamento sulla variabilità nei bambini in maniera sistematica.

Sulla base di queste lacune, il presente studio ha l'obiettivo di indagare lo sviluppo della capacità di formulare dei giudizi di quantità in presenza di variabilità.

Nel primo capitolo viene proposta una definizione di ragionamento statistico e si descrive l'importanza di questa abilità in situazioni di vita comuni, nonché nella ricerca empirica. Si introduce poi la prospettiva psicologica sul ragionamento statistico facendo riferimento ai lavori di Kahneman e Tversky. Successivamente ci focalizziamo sul ruolo della variabilità nel ragionamento statistico e definiamo in termini statistici il concetto di variabilità e le misure ad esso associate.

Nel secondo capitolo viene illustrato il ruolo cruciale della variabilità in diversi ambiti della statistica. Si descrivono le difficoltà più rilevanti incontrate dagli studenti nella comprensione della variabilità, e si conclude con una breve rassegna della letteratura sugli studi condotti in ambito educativo.

Poiché ragionare sulla variabilità implica sempre la presenza di quantità, nel terzo capitolo presentiamo una panoramica della letteratura psicologica riguardante lo sviluppo dell'abilità di formulare dei giudizi di quantità nei bambini, evidenziandone gli aspetti critici.

Nel quarto capitolo vengono descritti i risultati del primo esperimento. 241 bambini di 4, 5, 6, 8 e 12 anni e 82 studenti universitari hanno partecipato a un compito al computer, in cui veniva chiesto di confrontare due set contenenti 5 barre di cioccolata ciascuno. In un set, la media e la variabilità (intesa come deviazione standard) della lunghezza delle barre venivano mantenute costanti, mentre nel secondo set sono state manipolate. Ai partecipanti veniva chiesto di indicare quale set contenesse più cioccolata, o se la quantità fosse equivalente nei due set. Complessivamente i giudizi erano sorprendentemente difficili anche per gli adulti, che hanno fornito risposte non corrette al 29% degli stimoli. Più specificamente, dai risultati è emerso che 1) la performance nei giudizi di quantità aumenta significativamente con l'età, mostrando un aumento monotonico tra i 4 e i 12 anni. In particolare, i bambini di 8 anni hanno prestazioni significativamente migliori dei bambini più piccoli, e i dodicenni mostrano una performance migliore di quella di tutti gli altri bambini, ma non degli adulti; 2) nei dodicenni e negli adulti, la performance peggiora all'aumentare della variabilità dello stimolo, mentre un pattern diverso è emerso nei bambini di 4, 5 e 6 anni. In questi ultimi, le prestazioni migliori sono state rilevate in presenza di livelli intermedi di variabilità, dando luogo ad un effetto "a U rovesciata" piuttosto inaspettato. Nei bambini di 8 anni, l'effetto della variabilità era di tipo intermedio tra quello osservato

nei bambini più piccoli e negli adulti. Complessivamente, questi risultati suggeriscono la presenza di un cambiamento evolutivo nell'abilità di formulare giudizi di quantità in presenza di variabilità tra gli 8 e i 12 anni.

Uno dei risultati più salienti emersi dall'Esperimento 1 è l'effetto della variabilità sulla performance dei partecipanti in funzione dell'età. Nel capitolo 5 descriviamo un esperimento di controllo condotto per validare ulteriormente tale risultato, prendendo in considerazione dei possibili bias insiti nella procedura sperimentale (i bambini più piccoli apparentemente mostravano un bias nella risposta "uguale", da loro quasi mai utilizzata). L'esperimento è stato somministrato a 64 bambini (30 di 6 anni, 34 di 8 anni) eliminando tutti gli stimoli contenenti quantità uguali nonché l'opzione di risposta "uguale". I risultati di questo esperimento hanno confermato quanto già rilevato nell'esperimento principale.

Nel capitolo 6 riassumiamo brevemente e discutiamo i risultati della nostra ricerca. Per concludere, il giudizio di quantità in presenza di variabilità costituisce un compito rilevante e difficile. Riteniamo che la comprensione del modo in cui si sviluppa la capacità di formulare giudizi di quantità in presenza di variabilità può essere utile per implementare delle strategie d'insegnamento innovative e per prevenire la formazione di possibili bias di ragionamento negli adulti.

SUMMARY

The concept of variability (i.e., dispersion of observed data) has a central role in statistics and in quantitative decisions in everyday life. In the educational literature, however, scholars have only recently directed their attention to the study of how reasoning about variability develops (Garfield & Ben-Zvi, 2005). From a developmental cognitive perspective, much research has been conducted on children's quantity judgments since Piaget's seminal work, but few studies have systematically investigated the development of children's reasoning about variability.

To address these gaps, the present study investigated development of the ability to make quantity judgments in the presence of variability.

In the first Chapter, we define statistical reasoning and describe its importance in common life situations as well as in empirical research. We then introduce a psychological perspective on statistical reasoning based on the work of Kahneman and Tversky. We focus on the role of variability in statistical reasoning and define the concept of variability and its related measures in statistical terms.

In the second Chapter, we illustrate the crucial role of variability across different statistical domains. We describe the most relevant difficulties students encounter in understanding variability. Finally, a brief literature review of educational studies is provided.

Since reasoning about variability is always related to the presence of quantity, in the third Chapter we present a critical overview of the psychological literature concerning the development of children's ability to make quantity judgments.

In the fourth Chapter we describe the results of our first experiment. Two-hundred forty-one children aged 4, 5, 6, 8, and 12 years and 82 university students were assessed using a computerized task in which they were asked to compare two sets of five chocolate bars. The mean and variability of the chocolate bar lengths were held constant in one set and manipulated in the other set. Participants indicated which set contained more chocolate or that the amounts of chocolate were equal. Overall, the judgments were surprisingly difficult even for adults, who responded incorrectly to 29% of the stimuli. Quantity judgment performance significantly increased with age, with mean performance increasing monotonically between 4 and 12 years. In particular, 8-year-olds performed significantly better than their younger counterparts, and 12-year-olds performed significantly better compared to all the other children, but not compared to adults. The performance of 12-year-olds and adults decreased as stimulus variability increased, whereas a different pattern emerged for 4, 5, and 6-year-olds. In these age groups, performance was highest for intermediate levels of variability, resulting in a surprisingly inverted U-shaped effect of variability on performance. The effect of variability for 8-year-olds was intermediate between that observed among younger children and adults. Taken together, these results suggest that a developmental shift occurs between the ages of 8 and 12 in the ability to make quantity judgments in the presence of variability.

One of the most salient results emerging from Experiment 1 was the presence of an age-related effect of variability on participants' performance. In Chapter 5, we describe a control experiment conducted to further validate this finding by taking possible biases related to the specific experimental design into account (i.e., younger children apparently showed a response bias against selecting the equal response). In this experiment 64 children (30 six-year-olds, 34 eight-year-olds) performed the same task, but all stimuli with equal quantities and the equal response alternative were eliminated. Overall, the results of this study confirmed the findings of Experiment 1.

Chapter 6 briefly summarizes and discusses our findings. To conclude, judging quantity in the presence of variability is a relevant and difficult task. Understanding development of the ability to make quantity judgments in the presence of variability may suggest innovative teaching strategies and prevent possible reasoning biases in adults.

Chapter 1

The role of variability in statistical reasoning

"For Today's Graduate, Just One Word: Statistics"
The New York Times - August 6, 2009

In this chapter, we first define statistical reasoning and describe its importance in everyday life as well as in empirical research. We then introduce the psychological perspective on statistical reasoning based on the seminal work of Kahneman and Tversky. Next, we focus on the role of variability in statistical reasoning starting from a motivating example, and subsequently define the concept of variability and its related measures in statistical terms.

1.1 What is statistical reasoning?

Nowadays, quantitative information is everywhere and statistics are increasingly presented as a way to add credibility to advertisements, arguments, or advice. Being able to properly evaluate evidence (data) and claims based on data is an important skill that all students should learn as part of their educational programs. The study of statistics provides tools that both citizens and professional researchers in various fields need in order to react intelligently to quantitative information in the world around them (Ben-Zvi & Garfield, 2004). However, as first discovered by psychologists Kahneman and Tversky in the early 1970s, statistical thinking is remarkably difficult for people. Before analyzing the reasons for this difficulty (see section 1.1.1), we will provide a

definition of what statistical reasoning is.

In his illuminating article "Statistics among the liberal arts", David Moore (1988), a well-known American statistician, defined statistical reasoning as *a general, fundamental and independent mode of reasoning about data, variation and chance*. Statistical reasoning is a *general* method of reasoning because it can be applied wherever data, variation and chance appear. It is a *fundamental* method because data, variation and chance are omnipresent in modern life. Finally, from an epistemological perspective, statistics is an *independent* discipline with its own core ideas rather than, for example, a branch of mathematics (Moore, 1992). Indeed, as Hawkins (1996) suggested, "a mathematically educated person can be statistically illiterate" (pp.107-117).

Among the elements on which statistical reasoning is based, *data* are obviously the starting point. Collecting, checking, organizing, synthesizing, visualizing, and describing data is the first step of any given task involving statistical reasoning as well as of each statistical analysis. Historically, the first examples of statistics can be found in structured data collections such as the census in ancient Rome. Another major component is *variation*, that is, the variability that characterizes available data. This aspect always needs to be taken into account when interpreting data. As Wild and Pfannkuch (1999) put it: "Variation is the reason why people have had to develop sophisticated statistical methods to filter out any messages from the surrounding noise" (pp. 235–236). In brief, there is no statistical reasoning without variability. Finally, all methods of inferential statistics (i.e., obtaining information on a population based on a

sample drawn from that population) have to deal with probabilities and *chance*.

1.1.1 Statistical reasoning from a psychological perspective

From a psychological perspective, it can be argued that *inductive reasoning* is our most important and ubiquitous problem-solving activity. Concept formation, generalization from instances, and prediction are all examples of inductive reasoning (Nisbett, Krantz, Jepson, & Kunda, 1983). Since all these activities often involve tasks characterized by the presence of data, variation and chance, *statistical reasoning* has a crucial role in inductive reasoning. In other words, inductive reasoning must satisfy certain statistical principles to be correct. However, as noted above, Kahneman and Tversky demonstrated that statistical reasoning is remarkably difficult for people.

The authors' first findings were illustrated by the responses of professional psychologists to a questionnaire concerning research decisions (Tversky & Kahneman, 1971). In this work, they showed that even trained scientists regard a small sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. To describe this tendency, they ironically coined the term "Law of Small Numbers" as opposed to the statistical principle of the "Law of Large Numbers", which states that only very large samples are highly representative of the population from which they are drawn. The application of the Law of Small Numbers was later defined as *statistics heuristic*, i.e. a judgmental tool that is a rough intuitive equivalent of statistical principles and that people use in everyday inductive reasoning.

In a series of studies, Kahneman and Tversky found that people often solve inductive problems by means of a variety of intuitive heuristics, which include the *representativeness heuristic* (Kahneman & Tversky, 1972, 1973), the *availability heuristic* (Tversky & Kahneman, 1973), the *anchoring heuristic* (Tversky & Kahneman, 1974) and the *simulation heuristic* (Kahneman & Tversky, 1982). In tasks where these heuristics diverge from the correct statistical approach, individuals, from naive subjects to trained scientists, commit serious errors of inference.

The representativeness heuristic is the most studied and probably the most important heuristic. People often rely on it when making likelihood judgments, for example the likelihood that Object A belongs to Class B or the likelihood that Event A originates from Process B (Nisbett et al., 1983). Use of this heuristic involves basing such judgments on "*the degree to which A is representative of B, that is, by the degree to which A resembles B*" (Tversky & Kahneman, 1974, p. 1124). In one problem, for example, Kahneman and Tversky (1972) asked subjects whether days with 60% or more male births would be more common at a hospital with 15 births per day, or at a hospital with 45 births per day, or equally common at the two hospitals. Most subjects chose the latter alternative (56%), and the remainder divided about evenly between 15 (22%) and 45 (22%). The law of large numbers requires that, with a random variable such as sex of infant, deviant sample percentages should be less common as sample size increases. Hence, the correct answer would have been the smaller hospital. The representativeness heuristic, however, leads subjects to compare the similarities of the two sample proportions to the presumed population proportion (50%); because the two sample

proportions equally resemble the population proportion, they are deemed equally likely. The data indicate that, for this problem at least, most subjects (78%) used the representativeness heuristic and very few subjects (22%) referred to the Law of Large Numbers.

Other studies have confirmed and expanded the list of statistical failings documented by Kahneman and Tversky to other tasks. The failings seem particularly interesting in people's reasoning about social behavior. For example, Nisbett and Ross (1980) reported that people fail to apply necessary statistical principles to a very wide range of social judgments. Specifically, they demonstrated that subjects: 1) often make overconfident judgments about others, based on small and unreliable amounts of information; 2) are often insensitive to the possibility that their samples of information about people may be highly biased; 3) are often poor at judging covariation between events of different classes; and 4) are often little influenced by regression or base rate (i.e., a priori probabilistic information) considerations in their causal explanations for social events and their predictions of social outcomes.

Albeit influential, Kahneman and Tversky's theory has subsequently been criticized in some ways. First, it has been demonstrated that individuals' performance increases if task instructions are more explicitly stated or when more familiar stimuli are used (e.g., absolute frequencies vs. proportions) (Fiedler, 1988, Gigerenzer & Hoffrage, 1995). Second, other studies indicate that people also possess heuristics that are based on correct statistical concepts (Cosmides & Tooby, 1996).

Despite these criticisms, recent investigations have emphasized the importance

of Kahneman and Tversky's first intuitions. In particular, correct inductive reasoning may be improved by an effective statistical training to help people solve tasks requiring inductive abilities.

1.2 The central role of variability

In this section we illustrate the omnipresence and the importance of variability in statistical reasoning via an example concerning the evaluation of the difference between two treatments. We conclude by defining the standard deviation, i.e., the most widely used measure of variability, in statistical terms.

1.2.1 A motivating example

A researcher is interested in evaluating the difference between two new treatments, A and B, used to improve a cognitive skill (dependent variable). First, he/she randomly samples two groups of 30 subjects each (i.e., Group A and Group B), which can be considered homogeneous with regard to the dependent variable. At baseline, Group A is assigned the treatment A, and Group B the treatment B. After a pre-determined period of time, the researcher tests subjects on the dependent variable. Possible scores range from 0 to 10, where high scores indicate a high level of the investigated cognitive skill. Mean scores of the two groups (i.e., 6 for treatment A and 5 for treatment B) are reported in Figure 1.1.

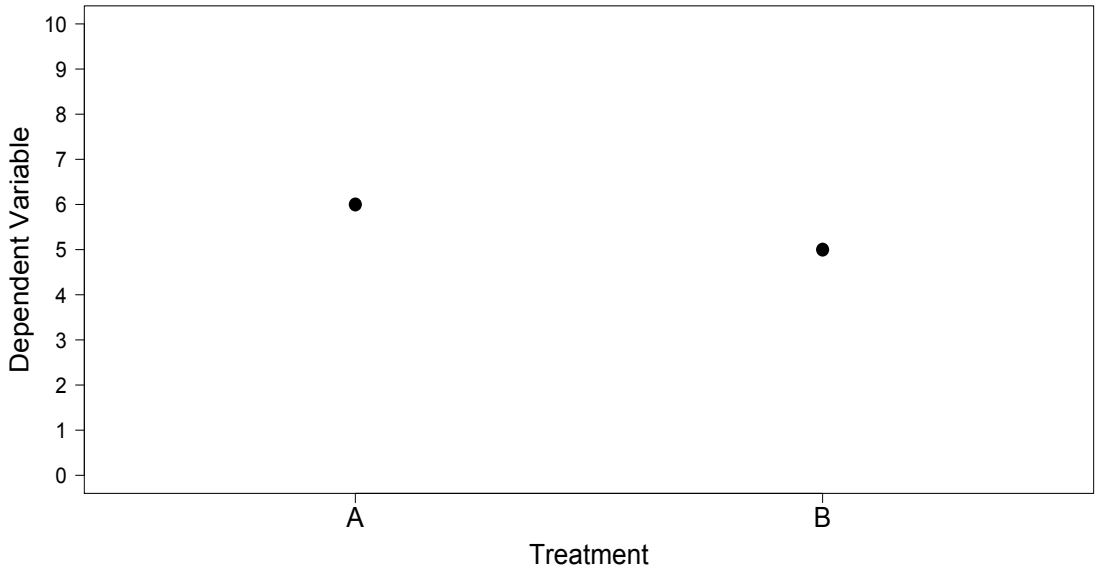

Figure 1.1. Mean values for two treatment groups A and B.

What information can be drawn from these results? Very little, if any. The researcher may note that the mean of treatment A is higher than the mean of treatment B, but to what extent? And what about the dispersion of scores in both treatments? Can these mean values be considered representative of the group scores? And finally, is there evidence supporting a difference between the two treatments?

In Figure 1.2 we present six scenarios that are all coherent with the situation shown above. In all scenarios, group means are fixed, whereas the variability of scores varies. As a measure of variability we use the sample standard deviation indicating the degree to which, on average, any given subject deviates from the group mean. As a measure of the difference between the two treatments we use a traditional measure of effect size introduced by Cohen (1988), that is, Cohen's *d*. This index is the ratio of the

difference between group means and the pooled standard deviation (i.e., the combined standard deviation of the two groups). Values above .3 indicate a small effect size (i.e., in our case a small evidence supporting the difference between treatments), above .5 a medium effect size, and .8 or greater a large effect size.

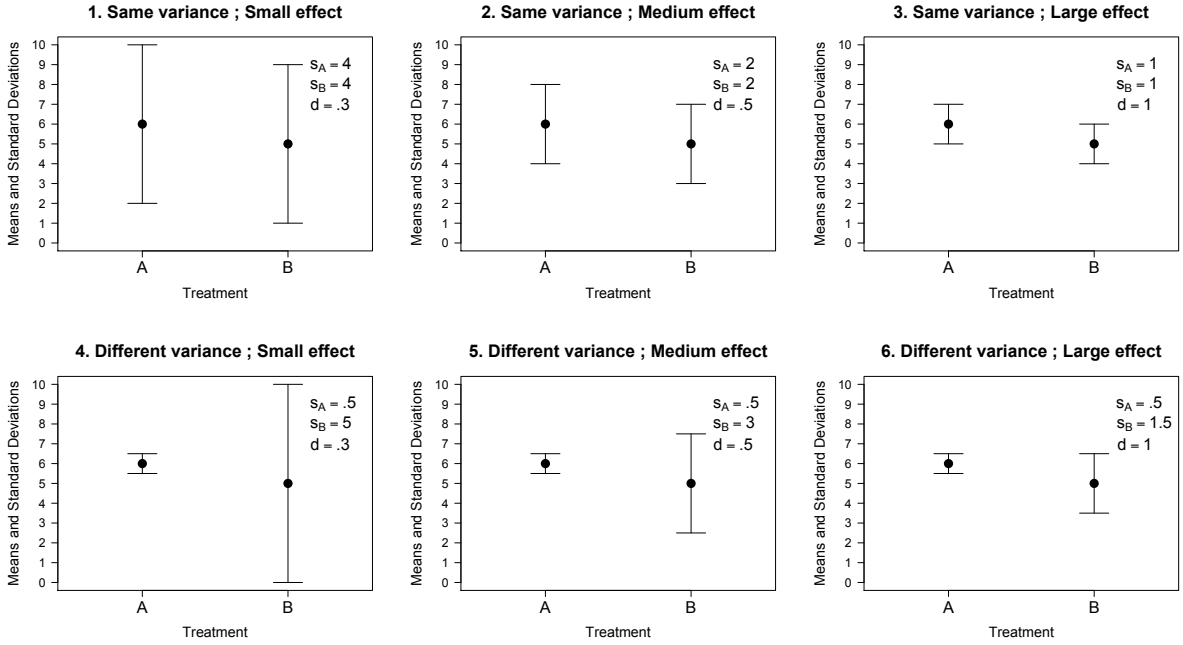

Figure 1.2. Different scenarios: fixed means, varying standard deviations, different effect sizes. Notes: S_A and S_B are standard deviations of group A and B, respectively. *d* is Cohen's *d*.

In the first three graphs presented in Figure 1.2 (first row), both groups have the same variability, and levels of variability decrease from left to right. In graphs 4-6 (second row), group means are characterized by different levels of variability, but the variability of group A is fixed, whereas the variability of group B decrease from left to

right. Looking at the columns in Figure 1.2 it is clear that, according to Cohen's guidelines, the difference between groups can be considered small in the first column, medium in the second column, and large in the third column. These six scenarios highlight the importance of taking variability into account when evaluating the size of the difference between groups.

We now move on to analyze each scenario:

Scenario 1) The two groups are characterized by the same marked variability. Each subject is on average 4 points away from his group mean. For example, in group B the standard deviation is 80% of the mean. Using a traditional metaphor, we can say that the signal we want to detect has a volume of 5, but this signal is surrounded by noise that has a volume of 4. In this case, it is obviously misleading to compare the group means. Indeed, in terms of size, the difference between the two groups is small (d = .3).

Scenario 2) The two groups have the same moderate variability. In group B, for example, the standard deviation is 40% of the mean. Hence, the mean may be considered an appropriate indicator of the group score. Indeed, in terms of size, the difference between the two groups can be considered medium (d = .5).

Scenario 3) The two groups are characterized by the same small variability. On average, subjects are one point away from their group mean; with regard to group B, the standard deviation represents 25% of the mean. The group mean can be considered highly representative of the group's score. In this case, the between-group difference is large (d = 1).

Scenario 4) The two groups have very different levels of variability. Rather than commenting on the group means, in this particular scenario it is more interesting and appropriate to consider group variabilities. Specifically, treatment A seems to have produced very homogeneous results. The standard deviation is .5, and thus each subject is on average 5 points away from the group mean. On the other hand, treatment B has yielded very heterogeneous results. Following the aforementioned metaphor, because the standard deviation equals the mean, both the signal we want to detect and the surrounding noise have the same volume. Overall, the variability among data is remarkable, resulting in a small effect size of the difference between treatments.

Scenario 5) The two groups have different levels of variability. Nonetheless, treatment B has yielded more heterogeneous scores than treatment A. Overall, the pooled variance is moderate and results in a medium effect size.

Scenario 6) The two groups are characterized by different, small levels of variability. The standard deviation of group A is one-third of the standard deviation of group B. However, because both standard deviations are small, both means may be deemed highly representative of the group scores. The pooled variability is low, resulting in a large effect size.

To summarize, in this example we showed how the difference between two group means can be dramatically influenced by group variability. The major implications are as follows:

1. To judge a mean, we need to know at least the variability of the underlying data.
2. When evaluating the difference between two means, it is necessary to consider the variability of the two groups.
3. Comparing two treatments involves not only comparing the respective means, but also interpreting their variability.
4. An increase in variability is related to a decrease in effect size.

1.2.2 Statistical definition of variability and of its measures

From a statistical perspective, when we deal with a quantitative[1] dataset, the first step is to compute indices that appropriately synthesize and describe the available information. Central tendency and variability are the two major kinds of indices. The purpose of the former is to describe, with one number, the center of the distribution of the dataset. The most popular among these indices is the arithmetic mean, often simply termed 'mean', defined as the sum of the values of data divided by the number of data. As we have seen in the previous example, this index is relevant but not sufficient to provide an accurate description of the data. In fact, one needs to know how data are spread with regard to the mean. To measure this dispersion, variability indices are used. Such

[1] Throughout this introduction, we will exclusively focus on quantitative variables without considering categorical variables.

indices are equal to zero when the data all have the same value, whereas they increase as the dispersion of values increases.

The most frequently used index of variability is the standard deviation, defined as the square root of the averaged square deviation from the mean:

$$\sigma = \sqrt{\frac{\sum_i^n (X_i - \overline{X})^2}{n}}$$

where $X_i - \bar{X}$ represents the deviation of each score from the group mean. As can be seen in the formula presented above, the value of the standard deviation increases as the single deviation from the mean increases. The standard deviation is indicated as σ if it is computed on the population, while we have to use an estimate if we need to compute it on a sample. In this case, the estimate is called 'sample standard deviation' and is indicated with *s*:

$$s = \sqrt{\frac{\sum_i^n (X_i - \overline{X})^2}{n-1}}$$

The advantage of using the standard deviation instead of the variance (i.e., the square of the standard deviation) when interpreting data is that it is expressed in the same unit of measurement as the phenomenon under study. In addition, this index can be viewed as a measure of how the mean is representative of the data. Specifically, for

low levels of standard deviation (indicating that data points tend to be very close to the mean) the mean can be considered a good summary index. In contrast, for high levels of standard deviation[2] (indicating that data points are spread out over a large range of values) the mean cannot be considered a good summary index of the data.

Given that data are always characterized by variability, *the concept of standard deviation is crucial in all statistical analyses, from the simplest to the more complex*.

[2] Overall, a standard deviation is considered high when it is more than half the value of its mean.

Chapter 2

Variability: An educational perspective

"Variability is … the essence of statistics as a discipline and it is not best understood by lecture. It must be experienced."

(Cobb, 1992)

In this chapter, we first illustrate the crucial role of variability across different statistical domains. Next, we describe the most relevant difficulties students encounter in understanding variability. Based on a recent study, we then discuss a list of components into which the concept of variability can be decomposed. Finally, a brief literature review of educational studies will be presented.

2.1 The omnipresence of variability in statistical analyses

One of the preeminent roles of all statistical analyses is to evaluate the variability of available data. Specifically, the aim of each statistical model is to decompose total observed variability into two additive parts: variability due to factors manipulated and/or measured by the researcher, i.e. *explained variability*; and variability not explained by factors under control, i.e. *unexplained or residual variability*. Residual variability can be generated from factors not controlled by the researcher, errors of measurement, and/or chance.

For example, the case of evaluating the difference between two treatments (see section 1.2.1) may be considered, from a statistical perspective, as the decomposition of

total observed variability of the dependent variables in two separate parts: variability due to treatments, and variability not due to treatments. As the ratio between these two quantities increases, support for the difference between treatments grows.

The quantification and interpretation of total variability as well as of explained and unexplained variability can be done in different statistical domains: *descriptive, inferential, and modeling*. Although these domains are related, they differ in their aims and in the statistical methods they use. In particular, descriptive statistics involve the use of indices and graphical representations to summarize and visualize data. Inferential statistics, via Probability and Distribution Theory, aims to obtain information on a population based upon a sample drawn from that population. Finally, statistical modeling uses mathematical models to investigate the relationships between a set of different variables. Without going into technical details, we will now provide a brief overview of the role of variability in each statistical domain.

The starting point of each statistical analysis is a descriptive analysis of data. To summarize the available information, both indices of central tendency (e.g., mean) and of variability (e.g., standard deviation) are used. These indices are equally important and must be interpreted together. Indeed, as Garfield and Ben-Zvi (2008) suggested, it is hard to imagine a situation in which one would summarize a dataset using only a measure of central tendency or a measure of spread.

In addition to indices, graphical representations play a crucial role. Before performing more complex analyses, it is paramount to visualize and explore the distribution of data at both univariate and multivariate levels (e.g., see Figures 2.1 and

2.2)[3].

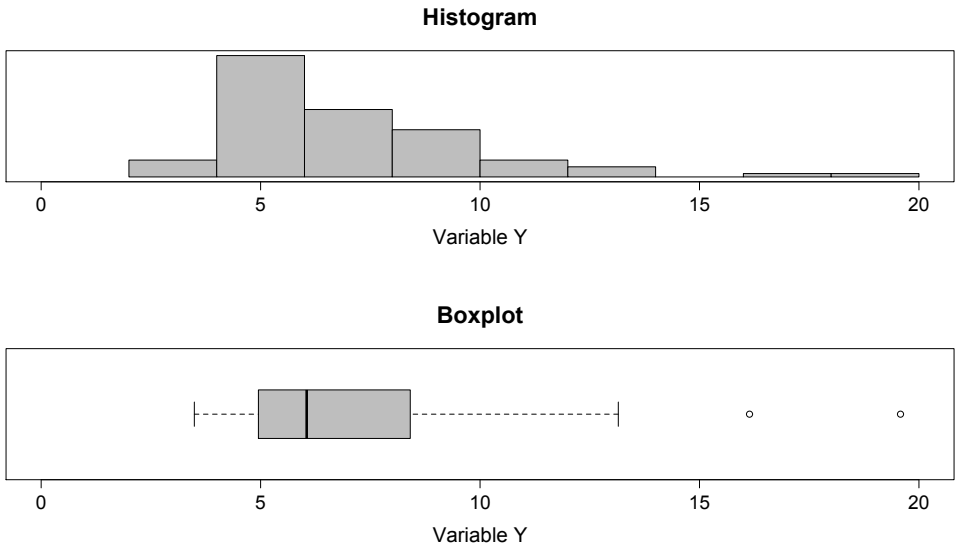

Figure 2.1. Examples of graphical representations at a univariate level.

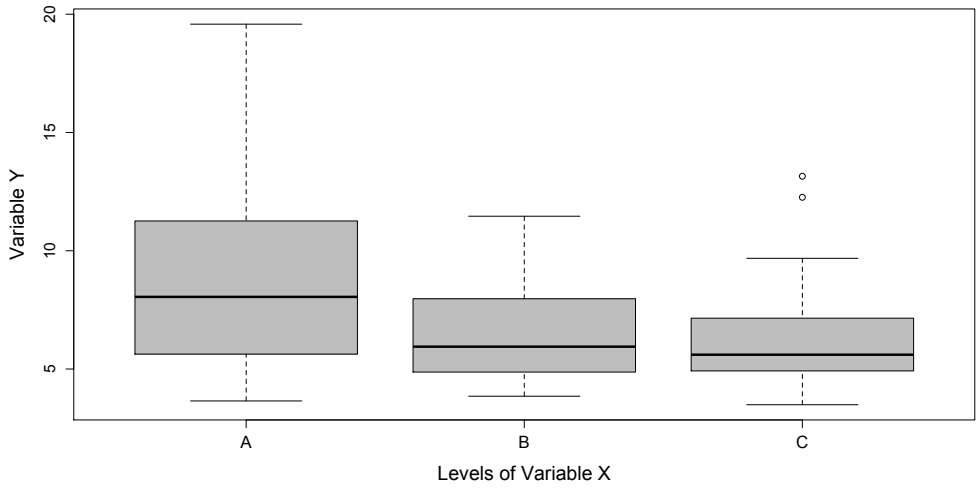

Figure 2.2. Example of graphical representation at a bivariate level.

[3] We present two of the most widely used graphical methods: 1) The histogram, which shows the frequencies of a quantitative variable, and 2) the boxplot, which is particularly useful to represent the spread of data. In this graph, the bottom and top of the box are the 25th and 75th percentile, respectively; the band inside the box is the median; the ends of the whiskers represent the smallest and the largest observations that are not anomalous values (i.e., outliers); and the little circles represent outliers. For details, see for example Chambers, Cleveland, Kleiner, and Tukey (1983).

Specifically, it is important to consider the observed range of data (i.e., observed minimum and maximum values), how and where the data tend to concentrate, the spread of data, and the presence of anomalous values (also called outliers; see Chambers et. al., 1983). With regard to variability, graphical representations are useful to become familiar with and understand the concept of variation as well as to investigate the sources and the impact of variance on observed data.

The classic statistical inferential approach consists of extracting information about unknown parameters of the population from the data of a sample drawn at random from this population. In most cases, the inferential approach is based upon the estimation of two quantities:

1) an estimate of the unknown parameter of interest, for example the sample mean if we need to make inferences about the population mean;

2) an estimate of the variability (i.e., *sample variability*) of the sample estimator used. This estimate is called *standard error*, and in the case of the sample mean[4], it is derived from the ratio between the sample standard deviation and the root mean square of the size of the sample.

Starting from these two estimates, and by means of Probability and Distribution Theory, it is possible to calculate an interval in which the true value of the unknown parameter of interest is included with a fixed level of probability (i. e., a confidence interval). The interpretation of the confidence interval gives information about both the true location of the unknown parameter and the variability of the estimator used.

[4] When the variance of the population is unknown.

With respect to statistical modeling, probably the most commonly used technique is *Analysis of Variance*. In this technique, the decomposition of total variance is used to make inferences about group means. The decomposition of variance into explained and residual variance has a preeminent role in most statistical models, such as *Linear Models* (e. g., regression models; Fox, 2008), *Generalized Linear Models* (e.g., logistic regression; Fox, 2008), and *Generalized Additive Models* (Hastie & Tibshirani, 1986). A relatively new kind of model of growing importance is termed *Mixed-Effects Models* (Pinheiro & Bates, 2000). These models are based on a deep investigation of the sources of variance. Specifically, they allow the researcher to simultaneously consider all factors that potentially contribute to variability of the dependent variable (Baayen, Davidson, & Bates, 2008). These factors comprise not only the traditional fixed-effects factors controlled by the experimenter (e.g., experimental conditions and gender of subjects), but also random effect factors with levels drawn at random from a population (e. g., subjects and stimuli). Indeed, total variability is decomposed into three additive parts: the variance explained by fixed factors, the variance explained by random factors, and the residual variance. This decomposition allows: 1) obtaining more consistent estimates of the fixed effects estimate; 2) controlling, quantifying, and interpreting the variability due to random effects; 3) investigating the possible interaction between fixed and random effects. It is interesting to note that, due to its efficacy in dealing with complex data, this statistical approach has become increasingly widespread in many areas of psychological science, such as educational psychology (e.g., Goldstein, Rasbash, Yang, Woodhouse, Pan, Nuttal, & Thomas, 1993), psycholinguistics (e.g.,

Baayen et al., 2008), neuropsychology (e.g., Marangolo, Bonifazi, Tomaiuolo, Craighero, Coccia, Altoè, Provinciali, & Cantagallo, 2010), and developmental psychology (e.g., Marceau, Ram, Houts, Grimm, & Susman, 2011).

2.2 Reasoning about variability

Although variability is omnipresent in statistical analyses in both curriculum design and statistics education research, human reasoning about variability has not been given enough attention (Reading & Shaughnessy, 2004). Indeed, despite the widespread belief in the importance of this concept, only recently have educational researchers devoted their attention to the study of how reasoning about variability develops (Garfield & Ben-Zvi, 2005).

In this section, we present the most relevant difficulties in understanding variability, and subsequently illustrate the core components of the concept of variability.

2.2.1 Students' difficulties in understanding variability

Current research demonstrates that it is extremely difficult for students to reason about variability, and educational researchers are just beginning to learn how reasoning about this concept develops (Garfield & Ben-Zvi, 2005).

Reasoning about variability has both informal and formal aspects, ranging from understanding that data vary (e.g., differences in data values) to calculating and interpreting formal measures of variability (e.g., variance and standard deviation). While students learn relatively easily how to compute formal measures of variability,

they rarely understand what these summary statistics represent, either numerically or graphically (Garfield & Ben-Zvi, 2008). Furthermore, students show difficulties in recognizing the importance of these formal measures and in connecting them with other summary statistics. For example, in an article by DelMas and Liu (2007) that first inspired our empirical research, the authors demonstrated that even trained college students have strong difficulties in relating the concept of variability to the concept of center, and in comparing the degree of variability across groups.

From an inferential perspective, the understanding of the relationships between variability and the concepts of distribution, estimation, and sampling is one of the most problematic issues in students' statistical reasoning (Cobb, McClain, & Gravemeijer, 2003).

Finally, what makes reasoning about variability even more complex is that, as shown in Section 2.1, variability may sometimes be desired and of interest, whereas sometimes it is considered error or noise (Gould, 2004; Konold & Pollatsek, 2002). Indeed, the tasks of recognizing, modeling, and interpreting explained and residual variability are difficult not only for students, but even for well-trained researchers, and can be considered the final stage of any advanced statistical course.

2.2.2 Core components of the concept of variability

In the past decade, several theoretical frameworks have been proposed to describe the levels of cognitive development of the concept of variability (Reading, 2004; Watson, Kelly, Callingham, & Shaughnessy, 2003). In particular, a recent contribution presented

at the International Conference On Teaching Statistics (ICOTS) by Reading and Reid (2010) provides an exhaustive synthesis of the components of the concept of variation. The authors identified nine major components shared across a number of theoretical models in the field (see Table 2.1).

Table 2.1

List of core components of the concept of variability

1 Developing intuitive ideas of variability
2 Describing and representing variability
3 Using variability to make comparison
4 Recognizing variability in special type of distribution
5 Identifying patterns of variability in fitting models
6 Using variability to predict random samples or outcomes
7 Considering variability as a part of statistical thinking
8 Recognizing sources of variation
9 Resolving expectations with observed variation

Note. Adapted from Reading and Reid (2010).

Although highly interrelated, the nine components have a hierarchical structure in terms of cognitive complexity, ranging from the development of intuitive ideas concerning variability as well as the description and representation of this concept (# 1, #2), to the resolution of expectations with observed variation (#9). The first two

components can be deemed *prerequisites* that are necessary for all subsequent components, while the last becomes possible only once the concept has been consolidated.

Overall, the nine components are crucial for a "deep" understanding of the concept of variability. In addition to being theoretically relevant, the list of components proposed by Reading and Reid (2010) has at least three practical implications. First, students may refer to this list for an organic overview of the preeminent components of the concept of variability. Second, teachers may take it into account to organize their statistical courses more efficiently. Third, educational researchers could make use of it to develop high-quality statistics curricula.

In accordance with Cobb's (1992) statement reported at the beginning of this chapter, we strongly believe that a deep understanding of most statistical concepts cannot occur without practice. In our opinion, students would benefit from early training in learning how to deal with real or ad hoc simulated data sets, thus experiencing the theoretical concepts they learn. With regard to the concept of variability and the components presented above, students should be involved in practical activities during lessons, such as investigating how single values affect variability, comparing variability across groups, and learning how sample variability changes according to sample size. All these activities can be relatively easily implemented using personal computers. Although recent studies have documented the efficacy of computer-based activities both individually and in group settings (for a

review, see Garfield & Ben-Zvi, 2007), they are still underemployed - at least in the Italian educational context.

2.3 Overview of educational studies concerning variability

As recently pointed out by Garfield and Ben-Zvi (2008), a variety of contexts have been used in statistics education to investigate students' reasoning about variability at different ages.

A first example is a study conducted by Lehrer and Schauble (2007), who assessed elementary school children's reasoning about variability in two contrasting contexts. In the measurement context, participants were asked to measure the heights of several objects. Results showed that children had the ability to create rudimentary statistical indices of central tendency and spread. In the natural or 'biological' context (i.e., growth of plants), however, subjects manifested difficulties in handling sources of natural variation and related statistics. Interestingly, the authors found that students' understanding of variability could be improved by specific activities involving explorations of sampling and comparing distributions.

Similar findings were reported by Bakker (2004) in a study of adolescents attending grades 7 and 8. By proposing a number of instructional activities to foster participants' reasoning about key statistical concepts including variability, the author found that activities in which students were encouraged to reason about sampling and shape of data were the most effective.

Overall, most studies have been conducted on undergraduate and graduate students. In these studies, the understanding of variability was examined in different contexts, including variability in data (Ben-Zvi, 2004a; Groth, 2005; Konold & Pollatsek, 2002; Petrosino, Lehrer, & Schauble, 2003), bivariate relationships (Cobb et. al, 2003; Hammerman & Rubin, 2003), comparing groups (Ben-Zvi, 2004b; Lehrer & Schauble, 2002; Makar & Confrey, 2005), probability and sampling (Chance, DelMas, & Garfield, 2004; Reading & Shaughnessy, 2004; Watson, 2004). On one hand, these studies have shown that students have major difficulties in reasoning appropriately about variability in many different contexts. On the other hand, they indicate that such reasoning can be significantly improved via the implementation of specific activities.

At least three critical issues can be identified in the literature reviewed above. First, many of the studies have focused on variability as embedded in more complex contexts and in connection with other statistical concepts, rather than on variability per se. Much effort has been devoted to the understanding of variability from an inferential perspective (e.g., sampling; see Reading & Shaughnessy, 2004), while the descriptive perspective has been relatively neglected despite being the key prerequisite for reasoning correctly about variation. Second, most of these studies concern students' performance and the development of effective training programs, whereas little is known about the cognitive mechanisms underlying statistical reasoning from a psychological perspective. An increased integration of the educational and the cognitive literature is necessary to improve our knowledge about the processes pertaining to the development of reasoning about variability. Finally there is a lack of studies focusing on

young children and on their first intuitions concerning variation. A developmental approach may shed light on the origins of the concept of variability as well as its course across different ages.

To examine how children and young adults reason about variability in a descriptive context using a cognitive-developmental approach, we conducted a series of experiments (see Chapters 4 and 5). Given the notable difficulties experienced by students at varying levels of schooling, we believe that the early identification of the cognitive processes involved in reasoning about variability may help teachers and other professionals plan more effective learning activities.

Chapter 3

Quantity judgments: A cognitive developmental perspective

Reasoning about variability is always related to the presence of quantity. Therefore, in this chapter we provide a critical overview of the psychological literature concerning the development of children's ability to make quantity judgments. This ability has been the focus of a large number of studies due to its centrality in human life as well as its important evolutionary function. Indeed, as noted by Opfer and Siegler (in press), quantitative abilities allow making rational judgments and decisions, and choosing among alternative strategies or behaviours in a rational way.

In the current review, we will specifically focus our attention on area judgments for two main reasons. First, this ability has played an essential role in the study of how quantitative concepts develop in children (e.g., Anderson & Cuneo, 1978; Cuneo, 1980; Gigerenzer & Richter, 1990; Wilkening, 1980). Second, educational researchers commonly use tasks involving graphical representations of areas (i.e., histograms; see for example DelMas & Liu, 2007) to investigate how people reason about variability. From this perspective, area judgment tasks may be viewed as a useful tool to explore the effect of variability on children's quantity judgments.

3.1 Piaget's seminal work on quantity judgments

Until the late 1970s, most developmental scholars agreed that children's quantitative concepts are predominantly unidimensional (Silverman & Paskewitz, 1988). This idea was largely based on Piaget's theory (1941/1952; Piaget & Inhelder, 1948/1967) and was demonstrated in a series of conservation experiments conducted by the Swiss scientist in the 1950s. In these experiments, children were presented with two stimuli that were quantitatively equal. Initially, the stimuli were presented so as to appear perceptually equal. Subsequently, one of them was modified perceptually but not quantitatively. After this transformation, young children judged the stimuli as no longer being equal quantitatively, noting, by way of justification, that the stimuli were different on one dimension (Silverman & Paskewitz, 1988).

One of the most well-known tasks used by Piaget involves estimating amounts of liquid. In this task, children are shown two standard glasses containing the same amount of liquid. The content of one glass is then poured into a taller, thinner glass, and children are asked whether both glasses contain the same amount of liquid, or if one glass has more liquid than the other. Piaget observed that children aged up to 7 years typically judge the glass with the higher liquid level to have more, thus focusing their attention only on this dimension of the stimulus. In contrast, older children are able to "conserve" in that they recognize that the amount of liquid remains constant albeit undergoing perceptual change.

According to Piaget, conservation is the realization that quantity does not change when nothing has been added or taken away from an object despite changes in

its form. This ability emerges between 7-8 years, when children reach the concrete operational stage of development. Before this stage, children have difficulties in focusing on more than one dimension of a stimulus because their thought is centered. Piaget proposed that centration is a consequence of the absence of mobility that characterizes early thought.

Later work highlighted several critical aspects of this view. One major concern was related to methodological issues, such as the instructions and materials used in Piagetian tasks. Indeed, several researchers demonstrated that children's performance significantly improved if the experimenter simplified the instructions (Miller, 1976) or utilized less ambiguous materials (Anderson & Cuneo, 1978). Overall, these studies showed that children younger than 7-8 years are able to solve tasks involving conservation, and that the ability to make quantity judgments is more complex and develops more gradually than was previously believed.

Among the scholars who challenged Piaget's findings, Anderson and Cuneo (1978) and Gigerenzer and Richter (1990) provided notable insights into how young children make quantity judgments. In the next sections we will present an overview of their work and discuss the implications for our study.

3.2 Relevant contributions to Piaget's work on quantity judgments

In a series of experiments, Anderson and Cuneo (1978) showed that 5-year-olds judged the areas of rectangles by a rule termed 'height + width', that is, children were able to consider two dimensions rather than one when asked to make quantity judgments.

Although the judgments of 5-year-olds were not correct (i.e, they did not follow the correct rule 'height X width' used by adults), the results showed that children were able to 'decenter' their thought and consider more than one salient aspect of the presented stimulus long before the age of 7-8 years, in contrast to what Piaget had suggested.

In their main experiment, Anderson and Cuneo presented a series of 9 rectangular cookies to 30 children (10 5-year-olds, 10 8-year-olds, and 10 11-year-olds) following a 3 X 3, height X width, factorial design[5]. For each factor, values were 7, 9 and 11 cm. Participants were told that a fictional child was very hungry, and were asked to evaluate how sad or happy the child would be with that much cookie to eat on a graphic rating scale (see Figure 3.1).

Figure 3.1. Graphic rating scale used by Anderson and Cuneo (1978).

In contrast with the classical Piagetian tasks, which involved a categorical response (i.e., less, more or the same), the authors used an ordinal response scale. Indeed, Anderson

[5] Each child completed two random replications of the factorial design, thus resulting in a total number of 18 trials.

and Cuneo (1978) argued that categorical choice tasks were unable to differentiate between continuous and all-or-none processes. For example, in the liquid conservation task, they suggested that a child's choice of the glass with the higher level of liquid does not necessarily mean that the height is the only dimension considered, but only that it has a greater effect than the diameter.

Analysis of responses was done within the theoretical framework of the Information Integration Theory (Anderson, 1981). With regard to this experiment, the aim of the theoretical approach was to determine how participants integrate information concerning the perceived height and width of each rectangle to formulate a quantity judgment. The authors suggested three possible alternatives:

1) *a one-dimensional rule*, i.e. area estimation is based only on one dimension, which could be the height or width of rectangles;

2) *an additive rule "height +width",* i.e. area estimation is the result of considering two dimensions (i.e., height and width), which are integrated following an additive model;

3) *a multiplicative rule "height X width",* i.e. area estimation is the result of the integration of height and width following a multiplicative model. This rule yields correct estimates and is typically used by adults in area judgments.

Figure 3.2 illustrates children's area judgments by height and width separately for each age group.

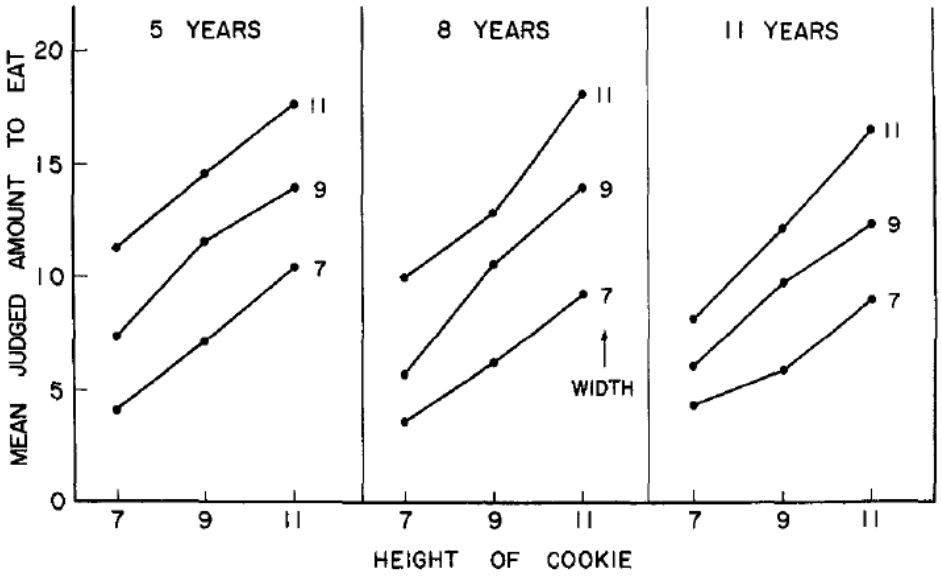

Figure 3.2. Mean judgment of amount to eat as a function of height and width of a rectangular cookie for the three age groups (Anderson & Cuneo, 1978, p. 344).

Visual inspection of children's performance indicates that:

- In 5-year-olds, both height and width have large and almost equal effects. In the words of Anderson and Cuneo, "No sign of centration is visible" (1978, p. 345). Furthermore, the three curves representing children's judgments by height are almost parallel, suggesting the presence of an additive rule of integration (i.e, height + weight).
- In 11-year-olds, the three curves representing children's judgments by height are not parallel, thus suggesting that 11-years-olds adopt the correct multiplicative rule.

- Eight-year-olds show a less regular pattern of response. This pattern appears to be somewhere in between that of 5- and 11-year-olds, supporting the idea that the developmental process involved in quantitative judgments is in a transitional phase.

Descriptive analyses were confirmed by inferential models. Specifically, for each age group the authors conducted a 3X3 repeated measures analysis of variance, with child's judgments as dependent variable, and height and width as within factors. For 5-year-olds, the main effects of height and width were statistically significant but their interaction was not, indicating the presence of an additive model. For 11-year-olds, a significant interaction between height and width emerged. According to Figure 3.2, this result suggests that older children use a multiplicative model. Finally, for 8-year-olds the interaction term did not reach statistical significance. Although in this age group the pattern of response is qualitatively similar to that of 11-year-olds, the additive model cannot be rejected. Hence, the authors conclude that 8-year old children are in a 'transitional state' in terms of stimulus integration rules.

In a series of collateral experiments, Anderson and Cuneo used the same methodological approach to investigate children's conservation of liquids. The authors found that 5-year-olds' judgments of amount of liquid in a glass obeyed a "height only" rule, that is, children were not able to integrate perceptually salient information from more than one stimulus dimension. Although this finding apparently supports Piaget's notion of centration, it does not account for the height + width rule found in children's area judgments. Hence, Anderson and Cuneo (1978) suggested that the height-only rule derives from children's familiarity with daily drinking. This hypothesis was confirmed

by several experiments showing that the height-only rule was specific to quantities inside of glass containers.

A few years later, Cuneo (1980) discovered that the height + width rule could be observed also among 3- and 4-year-olds when performing area quantity judgments. Based on the Information Integration Theory approach, other researchers further demonstrated that young children are able to integrate many kinds of information through the application of the additive rule (e.g., Miller, 1982; Wilkening, 1980). Taken together, these studies provide evidence for the hypothesis of a general-purpose integration rule in young children's quantity judgments.

A decade later, Anderson and Cuneo's findings were seriously criticized by Gigerenzer and Richter (1990). These authors proposed an alternative theoretical approach, called Perceptual Constancy (Brunswick, 1934), to study the problem of area quantity judgments. To summarize, this approach postulates that: 1) children perceive rectangular areas dependent on shape; 2) the degree of this dependency varies with age; 3) area perception in adults matches physical area, independent of shape. While Anderson and Cuneo assumed that young children are capable of formulating quantity judgments based on a decomposition of the area into two dimensions (i.e., height and width), Gigerenzer and Richter considered the whole perceived area as a basic theoretical concept.

The two theoretical frameworks were compared in 3 experiments involving 130 children aged between 4 and 9 years and 80 adults. Participants were asked to judge the area of a series of rectangles presented as chocolate bars in two different task

conditions. In the first condition, each chocolate bar was judged separately using a rating scale. In the second condition, two chocolate bars were showed simultaneously, and participants were asked which of the two bars contained more chocolate. The choice "equal" was allowed, but not explicitly cited in the instructions of experiment 1, whereas in experiments 2 and 3 it was clearly presented from the beginning. Responses were analyzed considering 5 alternative rules of judgment:

- *a longer-side rule*, which predicts that judgments of area are based on the longer dimension between the height and width of rectangles;
- *a width-only rule*, which predicts that judgments are based only on the width of rectangles;
- *a height-only rule*, which predicts that judgments are based only on the height of rectangles;
- *a height + width rule*, which predicts that judgments are based on the sum of height and width;
- *an area constancy rule*, which predicts that judgments are based on the physical area of rectangles.

In terms of data analysis and statistical testing, several improvements were introduced by the authors (for a detailed discussion, see Gigerenzer, Krauss & Vitouch, 2004). First, each theoretical rule was translated into a specific statistical hypothesis. Differently from Anderson and Cuneo, each hypothesis was further reformulated and considered as a null hypothesis, thus avoiding the problem of asymmetric testing (Gigerenzer & Richter, 1990). Furthermore, given the high variability in individual

patterns of children's responses, hypotheses were also tested at the individual level. The aim of this analysis was to control for potential bias deriving from the interpretation of performance only at the group level.

Results indicated that young children's use of the height + width rule was not supported. Indeed, all 130 children aged 4-9 showed judgments that significantly deviated from the predictions of the height + width rule. Among the hypothesized rules, the area constancy rule provided the best prediction in all age groups including 5- to 6 year-olds. In addition, area was judged as dependent on shape of rectangle by children. Finally, in accordance with Piaget's first findings, some of the younger children consistently centered.

In an additional experiment conducted on 20 children aged 4-6 years, the authors discovered a previously unknown effect: in contrast with both Anderson and Cuneo's approach and with Perceptual Constancy theory (Brunswick, 1934), children's judgment of the area of a single rectangle was influenced by the specific series of rectangles presented during the whole task. This context effect on quantity judgment was defined as "limited perceptual constancy".

Based on their findings, the authors proposed a three-step process model to account for the development of area quantity judgments:

1) *Centering*. The most frequent centering strategy is one in which children focus on the longer side of the rectangle. According to the authors, this does not mean that children's thought is one-dimensional, as most researchers erroneously inferred from Piaget's first findings. Indeed, children pay attention to both

dimensions of the rectangle, but tend to overestimate the effect of the longer dimension to formulate their judgments. This process was first recognized by Piaget himself and termed "law of relative centrations" (Piaget, 1969/1961; p. 8 -12).

2) *Limited Perceptual Constancy*. Children's area perception is influenced by shape, but this dependency is related both to the rectangle under judgment and to the series of rectangles presented during the experiment.
3) *Perceptual Constancy.* Area perception matches physical area, independent of shape.

From a developmental perspective, the authors' findings indicate that Centering and Limited Perceptual Constancy strongly overlap in young children aged between 4 and 9 years, whereas Perceptual Constancy is typically observable in adults.

To conclude, Gigerenzer and Richter's work contributed to the reappraisal of Piaget's initial findings concerning the law of relative centration in young children. However, their research also underlined how the ability to make quantity judgments is more complex and develops more gradually than was previously believed.

Chapter 4

Experiment 1. The Chocolate Study: Exploring the effect of stimulus variability on children's performance in a quantity judgment task

As shown in the previous chapters, the concept of variability has a central role in statistics and in quantitative decisions in everyday life. Understanding variability may improve inductive reasoning, which often involves tasks characterized by the presence of variation in observed data (Nisbett et al, 1983). In the educational literature, however, only recently have scholars directed their attention to the study of how reasoning about variability develops (Garfield & Ben-Zvi, 2005). Current educational research has mainly focused on high school and college students (e.g., DelMas & Liu, 2007), indicating that it is difficult for students and even for their teachers to reason about variability. From a developmental cognitive perspective, much research has been conducted on children's quantity judgments since Piaget's seminal work. However, few studies have systematically investigated the development of children's reasoning about variability.

To address these gaps, the present study aims to investigate the effect of variability in a quantity judgment task using 241 children aged between 4 and 12 years,

and 82 university students. In the next sections we will provide a detailed description of our study.

4.1 Introduction

This study investigates development of the ability to make quantity judgments in the presence of variability. Specifically, we evaluate how people compare two sets of vertical bars, similar in appearance to histograms, to determine which set represents a greater quantity.

To explore the influence of variability on the judgments of children from 4 to 12 years old and adults, we asked them to compare the quantities represented by two sets of five vertical bars, while keeping the mean and variance of bar heights constant in one set (the constant set) and manipulating the mean and variance of the bar heights in the other set (the comparison set). These two sets of bars appear similar to histograms, and the quantity represented by a set of bars is equivalent to the number of observations represented by a histogram. The bars are not histograms, however, because their horizontal positions are irrelevant. Indeed, the vertical bars were described as chocolate, and participants were asked to say which set contained the most chocolate.

Comparing the quantities represented by two sets of bars involves cognitive processes that develop over a longer span than the development of simple perceptual processes. Children should be able to perform this task easily if all the chocolate bars in each set were equally long; they would simply select the set containing longer bars. The task is more difficult when bar lengths are variable; some bars may be longer and some

shorter than the bars in the same position in the other set. Indeed, when the variability of bar lengths is very high in the comparison set, it may contain both bars longer and shorter than any in the other set, making the comparison between sets difficult. We predict, therefore, that the accuracy of quantity judgments will decrease as the variability of bar lengths in a comparison set increases. We also predict that quantity judgments will improve with age as children develop the capability to selectively attend only to the bars of unequal length and to combine the results of multiple potentially inconsistent comparisons.

4.2 Method

4.2.1 Participants

Participants included 241 children and 82 adults residing in Northern Italy. The children (38 four-year-olds, 33 five-year-olds, 39 six-years-olds, 67 eight-year-olds and 64 twelve-year-olds) were recruited at two kindergartens, four primary schools, and a middle school. Adults (M = 23.71 years, SD = 2.48) were undergraduate students majoring in psychology at the University of Padova. We obtained written informed consent from the school principals and both parents as well as verbal assent from all the children.

4.2.2 Materials and design

Materials were constructed for three training trials and 15 experimental trials. The stimulus for each trial consisted of two sets of five bars, as shown in Figure 4.1.

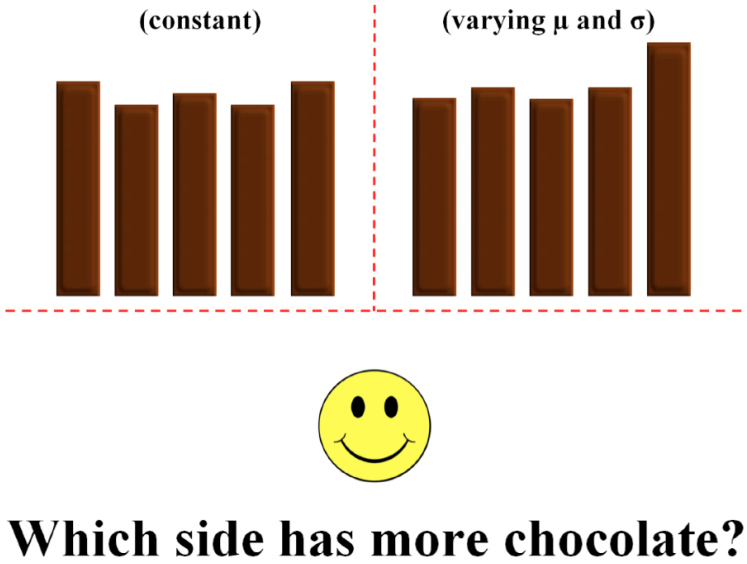

Figure 4.1. Example stimulus with the constant set on the left and comparison set on the right.

The bar widths were all 2 cm. In the constant set the mean bar length μ was 7.50 cm and the standard deviation σ was 0.36 cm. The 15 comparison sets were constructed by factorially combining three means and five standard deviations of bar lengths[6]. The

[6] In the constant set, we chose a mean of 7.5 cm because this value allowed us to create bars that were clearly recognizable on a computer screen of 15 inches. The standard deviation of .36 cm was selected to have a low impact of variability in the constant set (the standard deviation is 5% of the mean). In addition, this value allowed us to create both lower and higher levels of variability in the comparison set. For each of the 15 experimental conditions in the comparison set, there were an infinite number of possible combinations in terms of chocolate bar lengths. Hence, we fixed two constraints: 1) the lowest bar > 4.5 cm to be clearly detectable; 2) the highest bar < 10.5 cm to be contained in the screen. Based on these constraints, the bar lengths in the comparison set were fixed by hand according to the mean and standard deviation used in each experimental condition. It should be noted that the adequacy of the experimental apparatus was confirmed in a pilot study conducted on 10 4-and 5-year-old children before running the main experiment.

mean bar length in the comparison set was 7.86 cm (i.e., more chocolate than the constant set), 7.14 cm (i.e., less chocolate than the constant set), or 7.50 cm (i.e. the same amount of chocolate as the constant set). Figure 4.2 shows all 15 stimuli with the constant set on the left and the comparison set on the right. The quantity represented by the comparison set is larger in the first column, smaller in the second column, and equal to the constant set in the third column. The standard deviation of bar lengths in the comparison set was 0 (all equal height) in the first row, 0.36 cm (the same as the constant set) in the second row, 0.72 cm (twice the standard deviation of the constant set) in the third row, 1.08 cm (three times the standard deviation of the constant set) in the fourth row, or 1.55 cm (more than 4 times the standard deviation of the constant set) in the fifth row. The stimuli in the fifth row contain outlier bars, using Tukey's (1977) definition of an outlier as a point that falls more than 1.5 times the interquartile range above the third quartile or below the first quartile of data. Specifically, in stimulus 13 the comparison set contains a very short bar, in stimulus 14 a very high bar, and in stimulus 15 both a very short and a very long bar. The mean bar length of the comparison set was greater than μ in the first column, less than μ in the second column, and equal to μ in the third column. The standard deviation of the constant set was σ = 0.36 cm. The standard deviation of the comparison set increases from 0 in the first row to more than 4σ in the fifth row.

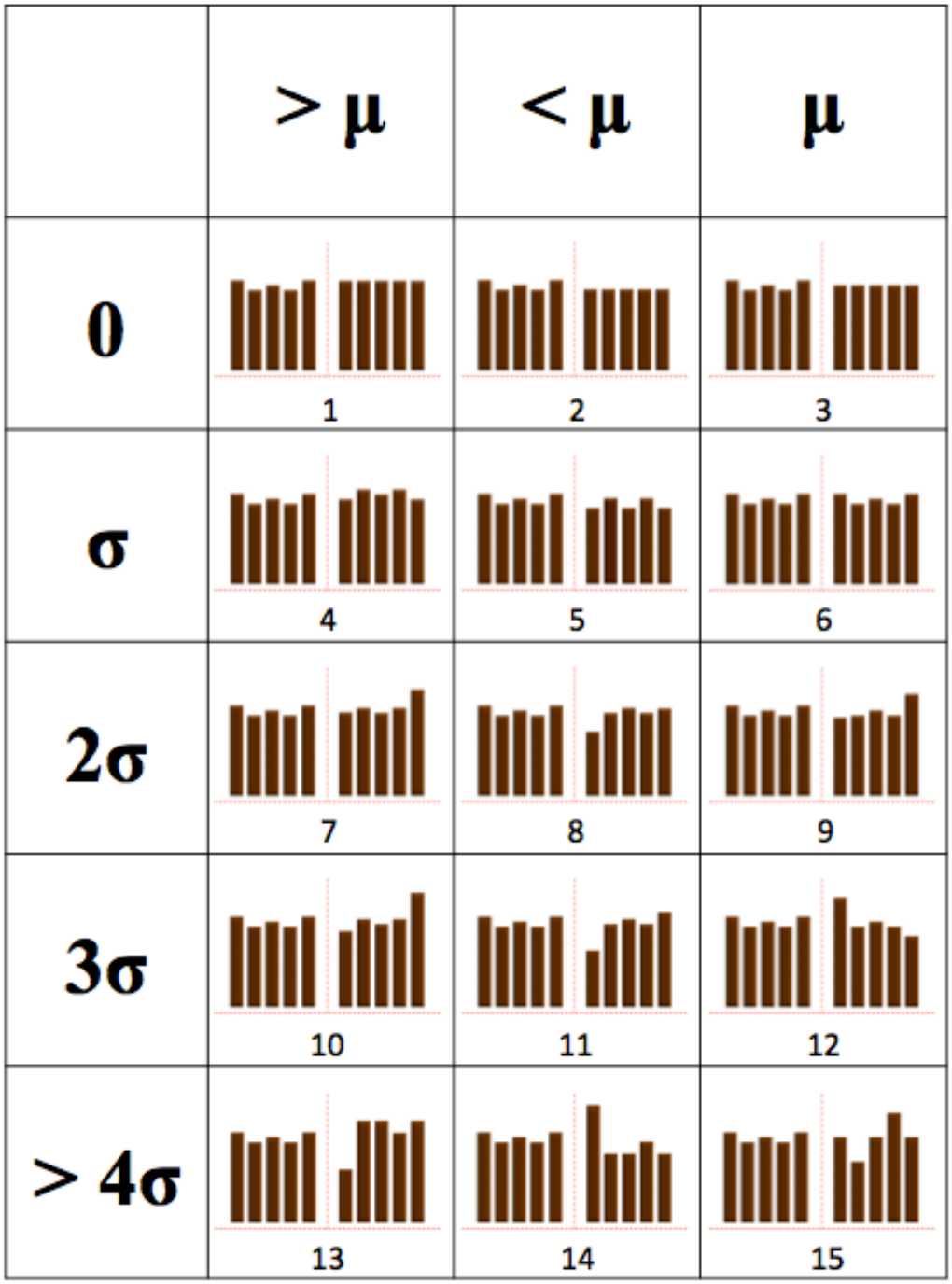

Figure 4.2. The 15 stimuli with the comparison set on the right. The mean bar length of the constant set was μ = 7.50 cm and the standard deviation was σ = 0.36.

4.2.3 Procedure

Children participated individually during school hours in a quiet room. The children were seated in front of a computer at a comfortable reading distance (approximately 60 cm). They were instructed that they would be comparing the quantity of two sets of chocolate bars in a series of graphical representations. Specifically, children were asked "Which side has more chocolate? This side, that side, or are they the same?" The children responded orally and by pointing, and their responses were audiotaped and recorded by the experimenter using paper-and-pencil. No time limit was imposed. To ensure that participants understood the task, they completed three training trials in which the correct responses were left, right, and equal before proceeding to the 15 test trials. The side of appearance of the constant set and the order of the 15 experimental trials were counterbalanced across participants.

The oldest children (8 and 12 years) were asked to perform two additional tasks after completing all 15 trials. In this task they again judged which set contained the most chocolate in Stimuli 1 and 13, presented in this order. After responding to Stimuli 1 and 13, both stimuli were presented simultaneously with Stimulus 1 above Stimulus 13, and children were asked which judgment was more difficult. The children responded by pointing at either Stimulus 1 or Stimulus 13. Then they judged which set contained more chocolate in Stimulus 3 and in Stimulus 15. After these two judgments, both stimuli were presented simultaneously and children were asked which of these two judgments was more difficult. Finally, participants were asked to judge again which set contained the most chocolate in Stimulus 14. After responding, the experimenter

exclaimed, “Well done! How did you do that? I don't know how to choose ... Can you explain how you decided?" Children’s responses were audio-recorded and there was no time constraint.

Adults participated in a classroom setting. Each adult was given a booklet containing three training trials and 15 experimental trials, one per page[7]. Four versions of the booklet were created combining two different presentation orders of the 15 stimuli and counterbalancing the side on which the constant set appeared. Adults were instructed to solve the three training and 15 experimental trials one at a time without turning back to prior pages. On each page adults judged which side contained the most chocolate and responded by checking one of three responses: left side, equal, or right side.

4.2.4 Statistical analyses

Because the data were repeated measurements of a categorical response, a logistic mixed-effects model was used (Baayen, 2008; Jaeger, 2008; Pinheiro & Bates, 2000). In this model the dependent variable was accuracy (correct or incorrect response). The fixed effects were age group (6 levels: 4-, 5-, 6-, 8-, 12 year-olds, and adults), the mean chocolate bar lengths in the comparison set (3 levels: more, less or the same as the constant set) and the standard deviation of the chocolate bar lengths in the comparison set (quantitative variable with 5 increasing values). To evaluate the potential quadratic

[7] We also performed a control experiment to rule out the possibility that adults’ performance differed as a function of means of administration of the task. To this end, 64 university students were individually assessed following the same procedure used with children. Analyses comparing adults’ performance in the two conditions (i.e., group assessment via paper-and-pencil vs. individual assessment via laptop) showed no significant difference.

effect of bar length variability on performance, the quadratic term of the standard deviation of the chocolate bars was also included. To evaluate if the effects of the manipulated experimental variables were constant across age groups, all two-way interactions including age were tested. Subjects were treated as random effect. To assess the significance of both fixed and random effects, we carried out a series of likelihood ratio tests (LRT) for nested models[8] based on the Chi-Square distribution (Pinheiro & Bates, 2000). As suggested by Wagenmakers (2007), we also considered the Bayesian Information Criterion (BIC). Since the two approaches always yielded the same results, only the results of the LRT are presented. Odds Ratios were reported as a measure of effect size. All analyses were performed using R software (R Core Development Team, 2010).

4.3 Results

This section presents the findings of Experiment 1. First, we report the results of the logistic mixed-effects model, in which we assessed the effects of age and the manipulated factors on participants' accuracy in the quantity judgment task. Second, we provide a qualitative analysis of 8- and 12-year-old children's judgments of difficulty in specific pairs of stimuli, as well as the reasoning strategies these children used when faced with a particular stimulus. Finally, we present a more in-depth analysis of stimuli features to assess the performance that would result from applying simple strategies and compared it with the observed performance of each age group.

[8]Each test was performed according to the principle of marginality (Type II tests), i.e., testing each term after all others, except ignoring the term's higher-order relatives (for details see Fox, 1997).

4.3.1 Effects of age and manipulated factors on participants' performance

Descriptive statistics of participants' responses by age group and stimulus are presented in Table 4.1.

Table 4.1

Percentage of left (L), equal (E), and right (R) responses for each stimulus and each age group. The standard deviation (SD) increases from 0 in the top row to more than 4σ in the bottom row. Correct responses are italicized

		Comparison Greater than Constant							Comparison Less than Constant							Comparison Equals Constant						
SD			Age							Age							Age					
		S#	4	5	6	8	12	A	S#	4	5	6	8	12	A	S#	4	5	6	8	12	A
	L	1	42	36	20	15	5	1	*2*	*50*	*58*	*74*	*87*	*80*	*83*	3	52	46	44	64	16	10
0	E		3	15	18	7	15	6		0	24	13	7	19	16		*3*	*18*	*15*	*9*	*68*	*86*
	R		*55*	*49*	*62*	*78*	*80*	*93*		50	18	13	6	1	1		45	36	41	27	16	4
	L	4	50	15	5	8	5	2	5	*68*	*61*	*69*	*73*	*89*	*88*	6	53	27	46	36	16	11
σ	E		5	21	31	28	25	20		8	18	18	19	9	12		*0*	*33*	*36*	*33*	*67*	*73*
	R		*45*	*64*	*64*	*64*	*70*	*78*		24	21	13	8	2	0		47	40	18	31	17	16
	L	7	21	6	5	12	2	1	8	*58*	*79*	*79*	*87*	*89*	*84*	9	34	12	23	15	8	10
2σ	E		0	15	8	3	12	17		10	15	13	3	9	16		*5*	*15*	*3*	*7*	*37*	*44*
	R		*79*	*79*	*87*	*85*	*86*	*82*		32	6	8	10	2	0		61	73	74	78	55	46
	L	10	21	9	21	13	5	9	11	*63*	*39*	*44*	*79*	*74*	*83*	12	29	15	13	15	12	12
3σ	E		0	15	10	3	14	18		0	15	18	5	13	15		*3*	*9*	*15*	*7*	*36*	*48*
	R		*79*	*76*	*69*	*84*	*81*	*73*		37	46	38	16	13	2		68	76	72	78	52	40
	L	13	34	12	10	22	16	12	14	*29*	*18*	*23*	*25*	*42*	*50*	15	39	20	21	30	22	26
>4σ	E		0	18	10	3	20	38		3	12	10	12	24	40		*0*	*15*	*8*	*10*	*48*	*56*
	R		*66*	*70*	*80*	*75*	*64*	*50*		68	70	67	63	34	10		61	64	72	60	30	18

Note. For each stimulus, the association between response and age was tested using Chi-Square. All Chi-Squares were significant at $p < .01$. The effect size for each test, measured as Cramer's Phi, ranged between .19 and .51 (mean = .32; median = .31).

Results of the logistic mixed-effects model are shown in Table 4.2.

Table 4.2

Results of the logistic mixed-effects model with accuracy as dependent variable

Fixed Effects	**B**	**SE**	***Odds Ratio***	$\chi^2(df)$
Age				214.09*** (5)
5 year-olds	.50	.71	1.64	
6 year-olds	1.45	.70	4.29*	
8 year-olds	1.51	.64	4.53*	
12 year-olds	2.35	.63	10.48***	
Adults	3.40	.64	30.01***	
Comparison Set Mean				713.54*** (2)
Less than the Constant Set	-.47	.21	.62*	
Same as the Constant Set	-4.55	.53	.01***	
Comparison Set Standard Deviation (Linear)	1.21	.38	3.36**	118.77*** (1)
Comparison Set Standard Deviation (Quadratic)	-.20	.06	.82**	35.06*** (1)
Age X Mean in the comparison set				191.54*** (8)
5 year-olds X Less than the Constant Set	-.24	.32	.79	
6 year-olds X Less than the Constant Set	-.20	.31	.82	
8 year-olds X Less than the Constant Set	.10	.28	1.11	
12 year-olds X Less than the Constant Set	.41	.28	1.50	
Adults X Less than the Constant Set	.62	.27	1.87*	
5 year-olds X Same as the Constant Set	2.25	.60	9.26***	
6 year-olds X Same as the Constant Set	1.77	.60	5.88**	
8 year-olds X Same as the Constant Set	1.31	.58	3.71*	
12 year-olds X Same as the Constant Set	3.38	.56	29.5***	
Adults X Same as the Constant Set	3.85	.56	47.0***	
Age X Comparison Set Standard Deviation (Linear)		.		29.234*** (5)
5 year-olds X Standard Deviation (Linear)	-.10	.53	.90	
6 year-olds X Standard Deviation (Linear)	-.57	.52	.57	
8 year-olds X Standard Deviation (Linear)	-.33	.48	.72	
12 year-olds X Standard Deviation (Linear)	-1.02	.47	.36*	
Adults X Standard Deviation (Linear)	-1.69	.47	.18***	
Age X Comparison Set Standard Deviation (Quadratic)				16.37** (5)
5 year-olds X Standard Deviation (Quadratic)	-.01	.09	.99	
6 year-olds X Standard Deviation (Quadratic)	.06	.09	1.06	
8 year-olds X Standard Deviation (Quadratic)	.01	.08	1.01	
12 year-olds X Standard Deviation (Quadratic)	.12	.08	1.13	
Adults X Standard Deviation (Quadratic)	.21	.07	1.23**	

Note. Baseline category for Age was "4 year-olds". Baseline category for Comparison Set Mean was "Greater than the Constant Set". For Comparison Set Standard Deviation, the degree of the estimated term (Linear or Quadratic) is reported in parentheses. Random effect was subject. Number of observations = 4845. Number of subjects = 323. *$p < .05$; **$p < .01$; *** $p < .001$.

We found a significant main effect of age on participants' performance ($\chi^2(4) = 214.09$, $p < .001$). As can be seen in Figure 4.3, the mean proportion of correct responses increased monotonically with age (.40 for 4-year-olds, .45 for 5-year-olds, .49 for 6-year-olds, .54 for 8-year-olds, .68 for 12-year-olds, and .71 for adults).

A planned comparison analysis was performed comparing each age group with the adjacent age group (i.e., 4-year-olds vs. 5 year-olds, 5-year-olds vs. 6 year-olds, and so on). Proportion correct was significantly greater for 12-year-olds than 8-year-olds ($p < .01$) and significantly greater for 8-year-olds than 6-year-olds ($p < .05$).

To assess the role of individual differences in participants' performance, we estimated the density of participants' accuracy for each age group (Sarkar, 2008) resulting in the distributions shown in Figure 4.4. Although mean performance significantly increased with age, there was large overlap of the estimated density curves across age groups, indicating the presence of marked individual variability. As Figure 4.4 shows, the distributions of correct responses were very similar for 5-, 6-, and 8-year-old children. The 4-year-old children performed strikingly worse than older participants on average, whereas 12-years old and adults performed substantially better than children of all ages.

We also found a significant main effect of mean amount of chocolate (χ^2 (2) = 713.54, $p < .001$), with participants performing better when the amount of chocolate in the constant and comparison sets was unequal. The proportion correct was .73 (SE = .01) when the comparison set was larger and .68 (SE = .01) when the comparison set was smaller. In contrast, the proportion correct was only .33 (SE =.01) when the

comparison and constant sets contained the same amount of chocolate. As shown in Figure 4.5, when the two sets contained equal amounts of chocolate, errors were strikingly greater for 4-, 5-, 6-, and 8-year-olds compared to 12-year-olds and adults, resulting in a significant interaction of mean chocolate bar length and age, χ^2 (2) = 191.54, $p < .001$. Judgments of every age group were much less accurate when the two sets represented equal quantities. The mean performance of the four youngest groups of children did not achieve 20% correct when the quantities were equal, and 4-year-old children achieved only 2% correct. The overall superior performance of 12-year-olds and adults was primarily due to their judgments of sets representing equal quantities.

As expected, the variabiliy of chocolate bar lengths affected performance (see Figure 4.6), but the relationship was complex. Both the linear (χ^2 (1) = 118.77, $p < .001$) and quadratic (χ^2 (1) = 35.06, $p < .001$) terms for the standard deviation were statistically significant. These effects were moderated, however, by age. As predicted, the performance of 12-year-olds and adults decreased as the standard deviation of chocolate bar lengths increased. The performance of 4-, 5-, 6- and 8-year-old children, however, was highest for intermediate levels of variability, resulting in a surprisingly inverted U-shaped effect of bar length variability on performance. These observations are supported by a significant interaction between both the linear term and age (χ^2 (5) = 29.23, $p < .001$) and the quadratic term and age (χ^2 (5) = 16.37, $p < .01$).

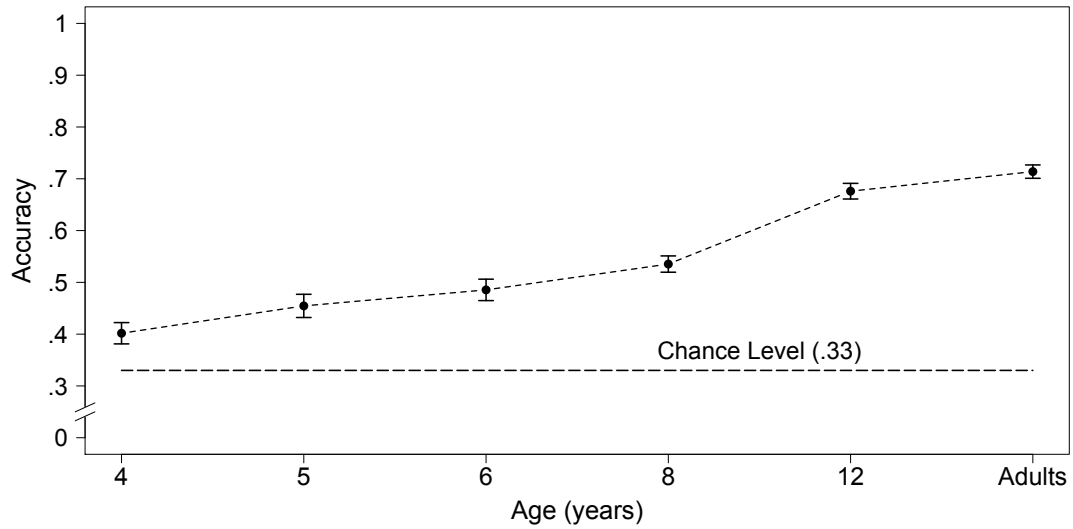

Figure 4.3. Mean proportion of correct responses by age for Experiment 1. Error bars represent standard errors of the mean ($N = 323$).

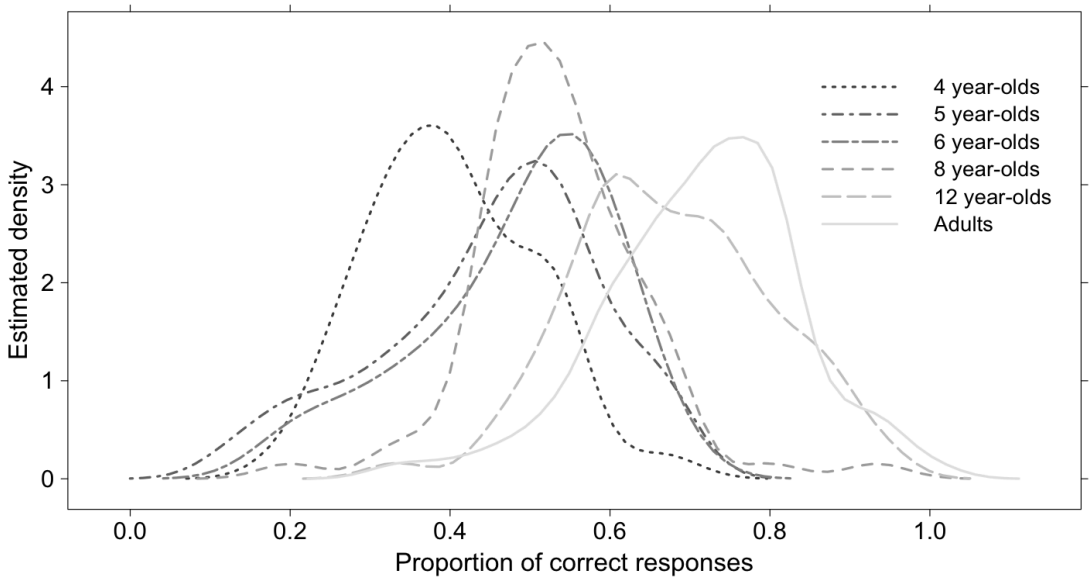

Figure 4.4. Estimated density of participants' accuracy by age ($N = 323$).

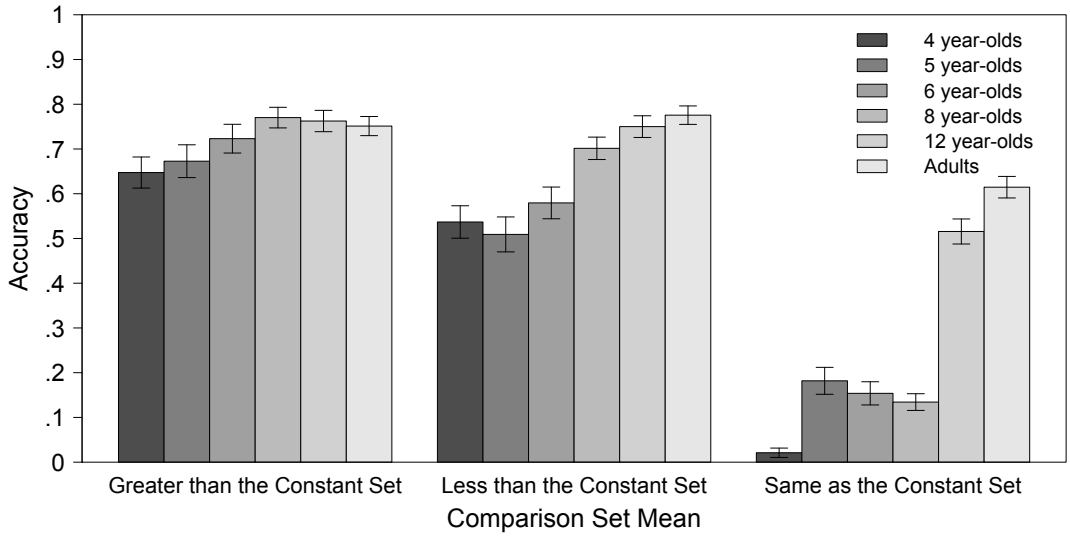

Figure 4.5. Mean proportion of correct responses by age and mean amount of chocolate in the comparison set for Experiment 1. Error bars represents standard errors of the mean ($N = 323$).

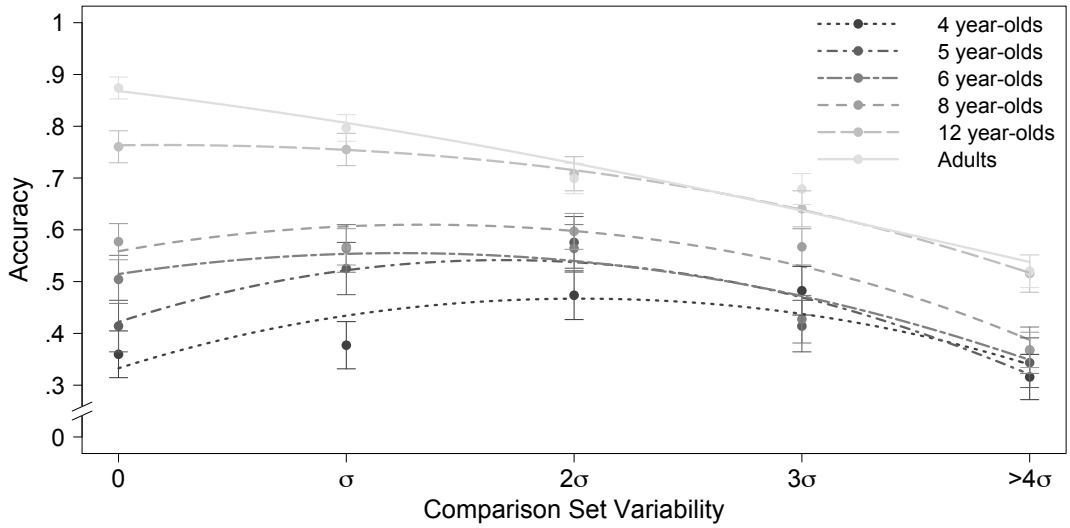

Figure 4.6. Mean proportion of correct responses by age and stimulus variability. Error bars represent standard errors of the mean. Lines represent estimated effects of the model ($N = 323$).

4.3.2 Difficulty judgments and solution strategies

The 8- and 12-year-old children were asked to judge the difficulty of comparing the quantities in two pairs of stimuli. First they compared Stimulus 1, in which all bars in the comparison set were of equal height, and Stimulus 13, in which the standard deviation of the comparison bar heights was more than 4 times as large as in the standard set. In both of these stimuli the mean bar lengths were greater in the comparison set. Then they compared Stimuli 3 and 15, in which the standard deviations of the comparison sets are the same as in Stimuli 1 and 13 but the mean bar lengths are the same in both the comparison and constant set. Most children judged Stimulus 13 to be more difficult than Stimulus 1 (63% of 8-year-old and 88% of 12-year-old children), and most children judged Stimulus 15 to be more difficult than Stimulus 3 (64% of 8-year-old and 86% of 12-year-old children). Although the perceived difficulty of the quantity judgment was greater for the stimuli with more variable bar lengths, the accuracy of 8-year-olds judgments was essentially unaffected by variability. The percentages of correct responses for the 8-year-old children were 78% for Stimulus 1, 75% for Stimulus 13, 9% for Stimulus 3, and 10% for Stimulus 15. The 12-year-old children, instead, achieved much greater accuracy overall and their judgments of difficulty were in greater alignment with their performance. The percentages of correct responses for the 12-year-old children were 80% for Stimulus 1, 64% for Stimulus 13, 69% for Stimulus 3, and 48% for Stimulus 15.

The 8- and 12-year-old children were also asked to explain how they decided which side had the most chocolate in Stimulus 14, which contained an exceptionally

long bar. Some 8-year-old (28%) and 12-year-old (5%) children simply said that one side appeared to have more chocolate and offered no more detailed explanation. Some 8-year-old (21%) but no 12-year-old children indicated that they chose the set containing the longest bar, thus centering their attention on one salient aspect of the stimulus. About half the 8-year-old children (51%) and most of the 12-year-old children (95%) explained that to formulate their judgment, they integrate the information of all bar lengths following various strategies characterized by different levels of complexity. Thus, these children appear to 'decenter' their attention by attending to multiple aspects of the presented stimulus in order to make the requested quantity judgment.

4.3.3 Stimulus feature analysis

Comparing the quantities represented by two sets of bars is difficult. The overall mean percentage correct was only 55%, and even the adults only achieved 71% correct. We know that children and adults can reliably compare the quantities represented by two rectangles of equal width and different heights. This experiment requires comparing two sets of five rectangles. Mathematically, this comparison could be accomplished by combining the quantities of all five rectangles in each set and then comparing these two results. Human judgments are more likely to rely on comparisons of features of the stimuli, such as bars of equal length, exceptionally long bars, and exceptionally short bars. Table 4.1 presents the responses of each age group to each of the 15 stimuli. We examined the stimuli to assess the performance that would result from applying simple strategies and compared it with the observed performance of each age group.

Stimuli 7, 9, 10, 12, 14, and 15 all have one bar that is substantially longer than the others. The majority of children in the four youngest age groups consistently selected the set containing this longer bar. Selecting the set with the longest bar yields the correct answer for Stimuli 7 and 10, and the percentage of children who selected this answer was 83% for Stimulus 7 and 77% for Stimulus 10. For the remaining four stimuli of this group the set containing the long bar is the incorrect answer, but the percentage of children (excluding the 12-year-olds) who selected it was 67% for Stimulus 14, 72% for Stimulus 9, 74% for Stimulus 12, and 64% for Stimulus 15. There were no systematic age differences in their responses to these stimuli. It appears that many children adopted a strategy of selecting the side with the long bar when there was a single exceptionally long bar. The 12-year-old children and adults, in contrast, were much less likely to base their judgments only on the longest bar. The percentages of 12-year-old children who incorrectly selected the longest bar for Stimuli 9, 12, 14, and 15 were 55%, 52%, 34%, and 30%, respectively. The percentages of adults who incorrectly selected the longest bar for these four stimuli were 46%, 40%, 10%, and 18%, respectively.

Of the remaining stimuli, 8, 11, and 13 all have one bar that is substantially shorter than the other bars. A possible strategy is to avoid the set with the shortest bar, and performance for Stimuli 8 and 11 suggests that this strategy developed relatively late. For these two stimuli, the short bar is in the comparison set and the correct response is the constant set. Averaging over these two stimuli, the percentages of correct responses for 4-, 5-, 6-, 8-, and 12-year-old children were 61%, 59%, 62%, 83%,

and 82%, respectively, and 84% of adults responded correctly. About 60% of children in the three youngest age groups responded correctly, and about 83% of 8-year-olds, 12-year-olds, and adults responded correctly, suggesting that many of the oldest children and adults adopted a similar strategy.

Stimulus 13 is interesting because the comparison set contains a short bar but also contains the most chocolate, so the strategy of avoiding the set with the shortest bar yields the wrong answer. The percentage of correct responses for 4-, 5-, 6-, 8-, and 12-year-old children were 66%, 70%, 80%, 75%, and 64%, respectively, but only 50% of adults responded correctly, resulting in a significant chi-square ($\chi^2(5) = 51.21$, $p < .001$, Cramer's Phi = .28). This is the only stimulus for which younger children were more accurate than older children and adults. Apparently many adults ruled out the comparison set because of its short bar.

The strongest evidence of age differences was found, surprisingly, in Stimuli 1, 2, and 3, in which all bars of the comparison set were equally long. Averaging across Stimuli 1 and 2, the percentage of correct responses for 4-, 5-, 6-, 8-, and 12-year-olds was 53%, 53%, 68%, 82%, and 80%, respectively. In striking contrast, for Stimulus 3, in which the two sets contain the same amount of chocolate, the mean accuracy for the four youngest groups was 3%, 18%, 15%, and 9%, respectively. Children in these four age groups were unable to recognize reliably that both sets of bars represented equal quantity, whereas 68% of 12-year-old children and 86% of adults responded correctly. Indeed, 4-year-old children almost never selected "equal" as a response; they responded

correctly to only 2.1% of the five equal stimulus sets. The oldest children, in contrast, responded correctly to 52% of the equal stimulus sets.

4.4 Discussion

The aim of this study was to investigate development of the ability to make quantity judgments in the presence of variability. Our findings showed that these judgments were surprisingly difficult even for adults, who responded incorrectly to 29% of the stimuli. Recognizing that two sets represented equal quantities proved especially difficult. The mean percentage correct when quantities were equal was only 12% for children and 61% for adults. There was a strong bias for responding that the sets represent different quantities.

As expected, we found that quantity judgment performance significantly increased with age, with mean performance increasing monotonically from 4 to 12 years. Despite this overall effect, there were also marked inter-individual differences indicating large overlap of performance between different age groups. Planned comparison analyses indicated that 8-year-olds performed significantly better than their younger counterparts; 12-year-olds performed significantly better compared to all the other children, but not compared to adults. In other words, the oldest children were more similar to adults in their performance than to the younger children, thus suggesting the presence of a developmental shift occurring between the ages of 8 and 12. This finding is supported by the qualitative analysis of children's difficulty judgments and use of strategies when faced with specific stimuli. Indeed, while most 12-year-olds

recognized that task difficulty increased with increasing levels of stimulus variability, only less than half of 8-year-olds did so. Furthermore, when asked to make a quantity judgment of a stimulus containing an exceptionally long bar, the majority of older children reported to consider multiple aspects; in contrast, more than half of 8-year-olds reported to use a strategy involving centration (i.e., focusing on the longest bar).

We also expected that participant performance would decrease as stimulus variability increased. However, this pattern held only among adults and 12-year-olds. In 4, 5, and 6-year-olds, children's performance was highest for intermediate levels of variability, thus resulting in a surprisingly inverted U-shaped effect of variability on performance. Finally, in 8-year-olds the effect of variability could be considered intermediate between that of younger children and adults. These findings suggest that increasing levels of stimulus variability are related to worse quantity judgments only in 12-year-olds and adults, whereas in younger children this relationship appears to be more complex.

The analysis of the performance for stimuli with similar features suggests that children, and adults to a lesser extent, adopt simple strategies for deciding which set represents the larger quantity. Two such strategies are to select a set that contains a bar longer than any others and to avoid a set that has a bar shorter than any others. These strategies constitute rules that children can follow when stimuli have these features, but they are not reliable. Three of the five stimuli with equal quantities contained a bar longer than any others, and most children and many adults selected the set containing

that bar, failing to recognize that the quantity represented by the long bar was offset by the shorter lengths of other bars.

These strategies are of no use, of course, when the bar lengths in the comparison set have little or no variability. In these conditions the bars in the comparison and constant set are of almost equal length, and quantity judgments depend on accurate perceptual judgments of these differences. The ability to judge quantity accurately for low-variability stimuli increased monotonically across all the ages from 4 years to adults.

As the variability of bar lengths increases, some bars are substantially longer, and in the universe of all possible stimuli the longer bars are more likely to be found in the set with the larger quantity. Results for the stimuli with intermediate levels of variability suggest that younger children's performance increased by selecting the stimuli with longer bars.

As bar length variability increases further, integrating the information represented by a set of bars becomes more challenging. In stimuli with greater variability, there are fewer bars of equal length in the two sets that can be excluded from consideration. In stimuli with very high variability, some bars in the comparison set are longer than any in the constant set and others are shorter, creating a conflict between the two simple rules adopted by many children. For these stimuli, children appear to have relied on the first rule, selecting the set with the longest bar. Adults also considered the location of the shortest bars, causing them to achieve the poorest performance of any age group when the set with the shortest bar contained the most chocolate.

Research on the development of quantity judgment - and more specifically on area quantity judgment - has explained children's performance in terms of the emerging ability to use basic algebraic rules (e.g., height + width rule; see Anderson & Cuneo, 1978; Cuneo, 1980), or to consider area independent of shape (Gigerenzer & Richter, 1990). However, these studies were based on relatively simple tasks, such as judging the area of a single rectangle (or a pair of rectangles), whereas in our case the task was more complex because children had to compare two sets of 5 bars each. We may hypothesize that the differences observed across age groups in our study are related to the development of children's cognitive capacities that occur over this age span (e.g., working memory capacity and executive functioning). Indeed, similar explanations have been offered for developmental changes in abstract number representation over this same age range (Halberda & Feigenson, 2008).

Chapter 5

Experiment 2. Age-related effects of stimulus variability: A fine-grained analysis

One of the most salient results emerging from Experiment 1 is the presence of an age-related effect of variability on participants' performance. In this chapter, we present a control experiment developed to further validate this finding by taking possible biases related to the specific experimental design into account. As we will see, the new results confirm what we found in Experiment 1.

5.1 Introduction

As shown in the previous chapter, the effect of stimulus variability on performance was different across age groups. Overall, we expected that performance would decrease as stimulus variability increased. However, this pattern held only among adults and 12-year-olds. In 4, 5, and 6-year-olds, children's performance was highest for intermediate levels of variability, thus resulting in a surprisingly inverted U-shaped effect of variability on performance. Finally, in 8-year-olds the effect of variability could be considered intermediate between that of younger children and adults.

A possible concern with our findings is related to the presence of stimuli including sets with equal amounts of chocolate. Indeed, in these situations participants

showed remarkable difficulties in providing the correct answer (i.e., "the same"). Such difficulties were especially evident in younger children (see Table 5.1), thus raising the possibility of an age-related judgment bias. In other words, could the observed age differences in the effects of stimulus variability on performance be due to age differences in response bias? There is evidence that the concept "same" develops in childhood, influencing children's understanding of quantity (Cantlon, Fink, Safford, & Brannon, 2007; Cowan, 1991). Hence, the effect of variability on performance observed in our data may partly be influenced by this decision bias.

To control for this potentially confounding effect, we ran an additional experiment in which: (i) the 5 stimuli with sets having the same quantity were eliminated; (ii) the response option "the same" was omitted. Furthermore, we decided to focus our attention on 6- and 8-year-olds because in the former group, the reversed U-shaped effect of variability was still present, whereas in the latter group this effect tended to disappear.

Table 5.1

Proportions of correct responses by age in sets with the same quantity in Experiment 1

4 years	**5 years**	**6 years**	**8 years**	**12 years**	**Adults**
.02	.18	.15	.13	.52	.61
(.01)	(.03)	(.03)	(.02)	(.03)	(.02)

Note. Standard errors are reported in parentheses.

5.2 Method

5.2.1 Participants

A total of 64 children (30 six-year-olds, 34 eight-year-olds) participated in Experiment 2. Children were recruited in a primary school following the same informed consent procedure described in Experiment 1.

5.2.2 Materials, design and procedure

The task in Experiment 2 was identical to the chocolate task used in the previous experiment, except for the following: (1) the 5 stimuli with sets having the same quantity were eliminated, thus resulting in a total number of 10 trials instead of 15; (2) the response option "the same" was not given to participants. As in Experiment 1, children completed two training trials in which the correct responses were left and right before proceeding to the 10 test trials.

5.2.3 Statistical analyses

Given that the data were repeated measurements of a binary response, we used a logistic mixed-effects model. In this model, accuracy (correct vs. incorrect response) was the dependent variable, age group (6- vs. 8-year-olds), mean chocolate bar length in the comparison set (2 levels: more vs. less than the constant set), and standard deviation of chocolate bars (quantitative variable with 5 increasing values) were considered as fixed effects. In addition, to evaluate the quadratic effect of variability emerging from Experiment 1, we included the quadratic term of the standard deviation of the chocolate

bars. To test our main hypothesis that the effect of variability differs as a function of age, the following interactions were also evaluated: (a) linear term of standard deviation by age; (b) quadratic term of standard deviation by age. Finally, subjects were included as random effect. In order to facilitate interpretation of results, the estimated effect of variability as a function of age will be presented graphically.

5.3 Results

Overall, children performed better in the current experiment compared to the prior one (mean proportion of correct responses = .68 vs. 49 in 6-year-olds, and .80 vs. 54 in 8-year olds). This finding reflects the difference in response options of the two experiments, since in the present study there was 1 correct answer out of 2 (i.e., probability of 50% to respond correctly by chance), while in the previous experiment the correct answer was 1 out of 3 (i.e., 33% of probability to provide a correct answer by chance).

Results of the logistic mixed-effects model are presented in Table 5.2.

Table 5.2

Results of the logistic mixed-effects model with accuracy as dependent variable

Fixed Effects	***B***	***SE***	***Odds Ratio***	$\chi^2(1)$	***p* value**
Age (8 year-olds)	2.86	.98	17.41	12.14	<.001
Comparison Set Mean (Less than the Constant set)	-.81	.20	.45	16.10	<.001
Comparison Set Standard Deviation (Linear)	2.08	.48	8.00	23.07	<.001
Comparison Set Standard Deviation (Quadratic)	-.37	.08	.69	29.63	<.001
Age X Comparison Set Standard Deviation (Linear)	-1.27	.73	.28	8.89	<.01
Age X Comparison Set Standard Deviation (Quadratic)	.15	.12	1.17	1.68	.194

Note. Baseline category for Age was "6 year-olds". Baseline category for Comparison Set Mean was "Greater than the Constant Set". For Comparison Set Standard Deviation, the degree of the estimated term (Linear or Quadratic) is reported in parentheses. Random effect was subject. Number of observations = 640. Number of subjects = 64.

We found a significant main effect of age, with older children (M = .80, SE = .02) performing better compared to younger children (M = .68, SE = .03). Mean amount of chocolate in the comparison set was also significant, with participants showing a better performance (M = .81, SE = .02) when the amount of chocolate was larger in the comparison set than in the constant set, and a worse performance in the opposite condition (M = .67, SE = .03).

With regard to the effect of variability, both the linear and the quadratic term of the standard deviation were statistically significant. However, this effect was moderated by child age, as there was a significant interaction between the linear term and age. As shown in Figure 5.1, the estimated effect of variability on performance by age indicates the presence of a reversed U-shaped effect in 6-year-olds, which attenuates in 8-year olds. This pattern clearly replicates the findings of Experiment 1, with the reversed U-

shaped effect in younger children being even more pronounced.

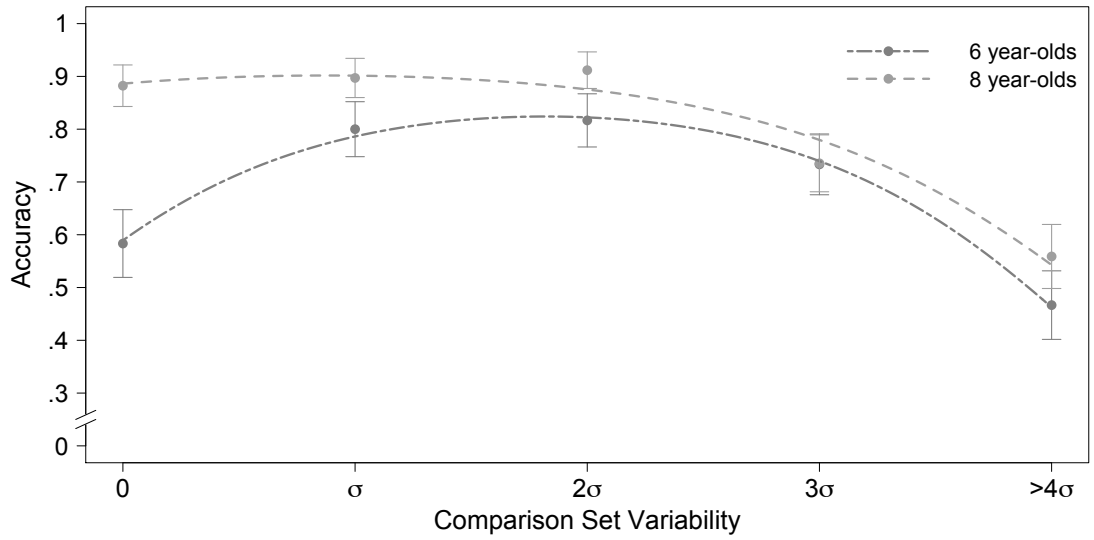

Figure 5.1. Mean proportions of correct responses by age and stimulus variability for Experiment 2. Error bars represent standard errors of the mean. Lines represent estimated effects of the model ($N = 64$).

5.4 Discussion

These findings confirm the non-monotonic effect of stimulus variability on younger children's performance. As in Experiment 1, the strongest evidence of a reversed U-shape effect was found for the youngest children, and this effect attenuates with age.

Of particular interest is the low performance observed for the stimulus characterized by the presence of null variability in the comparison set. In this situation, we expected that participants would show the best performance, since detecting the overall quantity of the comparison set – in which bars have all the same length - and

then comparing the results with the constant set should be relatively easy. While this was true for adults and 12-year-olds in Experiment 1, and to some extent for 8-year-olds in both Experiments 1 and 2, the data for the youngest children (6-year-olds in Experiment 1 and 2, and 4-and 5-year-olds in Experiment 1) did not confirm this expectation. Considering that the constant set has a low level of bar variability, a possible explanation is that in this particular case, young children make their judgments primarily based upon the perceived similarity in the overall shape of two sets. This may lead to poor performance when the amount of chocolate in the two sets is different. In contrast, older participants first focus their attention on the absence of variability in the comparison set, recognizing that is particularly easy to estimate the amount of chocolate, and then use this information to make their judgment.

Together, this pattern suggests the presence of two different cognitive mechanisms involved in quantity judgments among adults and 12-year-olds as compared to younger children.

Chapter 6

Conclusion

The concept of variability (i.e., dispersion of observed data) has a central role in statistics and in quantitative decisions in everyday life. In the educational literature, however, scholars have only recently directed their attention to the study of how reasoning about variability develops (Garfield & Ben-Zvi, 2005). From a developmental cognitive perspective, much research has been conducted on children's quantity judgments since Piaget's seminal work, but few studies have systematically investigated the development of children's reasoning about variability.

To address these gaps, the present study investigated development of the ability to make quantity judgments in the presence of variability.

In the first experiment, 241 children aged 4, 5, 6, 8, and 12 years and 82 university students were assessed using a computerized task in which they were asked to compare two sets of five chocolate bars. The mean and variability of the chocolate bar lengths were held constant in one set and manipulated in the other set. Participants indicated which set contained more chocolate or that the amounts of chocolate were equal.

Overall, the judgments were surprisingly difficult even for adults, who responded incorrectly to 29% of the stimuli. Recognizing that two sets represented

equal quantities proved especially difficult. The mean percentage correct when quantities were equal was only 12% for children and 61% for adults. There was a strong bias for responding that the sets represent different quantities.

As expected, we found that quantity judgment performance significantly increased with age, with mean performance increasing monotonically from 4 to 12 years. Despite this overall effect, there were also marked inter-individual differences indicating large overlap of performance between different age groups. Planned comparison analyses indicated that 8-year-olds performed significantly better than their younger counterparts; 12-year-olds performed significantly better compared to all the other children, but not compared to adults. In other words, the oldest children were more similar to adults in their performance than to the younger children, thus suggesting the presence of a developmental shift occurring between the ages of 8 and 12. This finding was supported by the qualitative analysis of children's difficulty judgments and use of strategies when faced with specific stimuli. Indeed, while most 12-year-olds recognized that task difficulty increased with increasing levels of stimulus variability, only less than half of 8-year-olds did so. Furthermore, when asked to make a quantity judgment of a stimulus containing an exceptionally long bar, the majority of older children reported to consider multiple aspects; in contrast, more than half of 8-year-olds reported to use a strategy involving centration (i.e., focusing on the longest bar). The stimulus feature analysis leads to the same conclusion: when confronted with stimuli having one bar that was substantially longer, the majority of children in the four youngest age groups consistently selected the set containing the longer bar, whereas the

12-year-old children and adults were much less likely to base their judgments only on the longest bar.

We also expected that participant performance would decrease as stimulus variability increased. However, this pattern held only among adults and 12-year-olds. In 4, 5, and 6-year-olds, children's performance was highest for intermediate levels of variability, thus resulting in a surprisingly inverted U-shaped effect of variability on performance. Finally, in 8-year-olds the effect of variability could be considered intermediate between that of younger children and adults. These findings suggest that increasing levels of stimulus variability are related to worse quantity judgments only in 12-year-olds and adults, whereas in younger children this relationship appears to be more complex.

The age-related effect of variability on participants' performance was further investigated in Experiment 2. Because younger children apparently showed a response bias against selecting the equal response, we re-administered the computerized task to six-year-old and eight-year-old children. The task in Experiment 2 was identical to the chocolate task used in the previous experiment, except for the following: (1) the 5 stimuli with sets having the same quantity were eliminated; (2) the response option "the same" was not given to participants. Analyses confirmed the non-monotonic effect of stimulus variability on younger children's performance. As in Experiment 1, the strongest evidence of a reversed U-shape effect was found for the youngest children, and this effect attenuated with age.

Overall, the differences observed across age groups in our study may be

explained in terms of the development of children's cognitive capacities that occur over this age span (e.g., working memory capacity and executive functioning). Indeed, similar explanations have been offered for developmental changes in abstract number representation over this same age range (Halberda & Feigenson, 2008). Future research may investigate the role of other cognitive abilities, such as mathematical and problem-solving skills, in children's quantity judgment performance. In addition, it would be useful to thoroughly analyze the reasoning strategies endorsed by children when making quantity judgments in the presence of variability through a microgenetic approach (Siegler & Crowley, 1991). Given the difficulties experienced by adults in responding correctly to the task, this method may also be applied to older ages as a means to gain insight into the development of cognitive reasoning biases.

From an educational perspective, the ability to reason about variability may be usefully improved by proposing simple quantity judgment tasks in the school setting to help children familiarize with basic statistical concepts, such as the mean and standard deviation, as well with graphical representations of data.

To conclude, judging quantity in the presence of variability is a relevant and difficult task. Understanding development of the ability to make quantity judgments in the presence of variability may suggest innovative teaching strategies and prevent possible reasoning biases in adults.

References

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Anderson, N. H., & Cuneo, D. O. (1978). The height + width rule in children's judgments of quantity. *Journal of Experimental Psychology: General, 107*, 335-378.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D.J., & Bates, D.M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*, 64-83.

Ben-Zvi, D. (2004a). Reasoning about data analysis. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 121-145). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D. (2004b). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*, 42-63.

Ben-Zvi, J., Garfield, D. (2004). The challenge of developing statistical literacy, reasoning, and thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of*

developing statistical literacy, reasoning, and thinking (pp. 3 -15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brunswik, E. (1934). *Wahrnehmung und Gegenstandswelt [Perception and the world of objects]*. Leipzig: Deuticke.

Cantlon, J., Fink, R., Safford, K., & Brannon, E. M. (2007). Heterogeneity impairs numerical matching but not numerical ordering in preschool children. *Developmental Science, 10,* 431-440.

Chance, B. L., DelMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, G. W. (1992). Report of the joint ASA/MAA committee on undergraduate statistics. In *The American Statistical Association 1992 Proceedings of the Section on Statistical Education* (pp. 281- 283). Alexandria, VA: American Statistical Association.

Cobb, P., McClain, K., & Gravemeijer, K. P.E (2003). Learning about statistical covariation. *Cognition and Instruction, 21*, 1-78.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum Associates.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Cowan, R. (1991). The same number. In D. Durkin & B. Shire (Eds.), *Language in mathematical education: Research and practice* (pp. 445-464). Hillsdale, NJ: Lawrence Erlbaum.

Cuneo, D. O. (1980). A general strategy for quantity judgments: The height + width rule. *Child Development, 51*, 299-301.

DelMas, R. C., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 87-116). New York: Lawrence Erlbaum.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123–129.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Model (2nd ed.)*. Thousand Oaks, CA: Sage.

Garfield, J., & Ben-Zvi, D. (2004). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*, 372-396.

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4,* 92–99.

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning. Research and Teaching Practice*. New York: Springer.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Gigerenzer, G., & Richter, H. R. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development, 5*, 235-264.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oaks, CA: Sage.

Goldstein, H., Rasbah, J,, Yang, M., Woodhouse, G., Pan, H., Nuttal, D., & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19,* 425-433.

Gould, R. (2004). Variability: One statistician's view. *Statistical Education Research Journal, 3*, 7-16.

Groth, R. E. (2005). An investigation of statistical thinking in two different contexts: Detecting a signal in a noisy process and determining a typical value. *Journal of Mathematical Behavior, 24*, 109-124.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology, 44,* 1457-1465.

Hammerman, J. K., & Rubin, A. (2003). Reasoning in the presence of variability. In C. Lee (Ed.), *Reasoning about variability: A collection of current research studies.*

Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3), July 23-28.

Hastie, T. J., & Tibshirani, R. J. (1986). Generalized Additive Models. *Statistical Science, 1*, 297-318.

Hawkins, A. (1996). *Can a mathematically-educated person be statistically illiterate?* Paper presented at the Nuffield Conference on 'Mathematics for the New Millennium - What Needs to be Changed, and Why?', October 1996.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kahneman, D., Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kahneman, D., Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*, 237-251.

Konold, C., & Pollatseck, A. (2002). Data analysis as the search for signals in noisy processes. *Journal of Research in Mathematics Education, 33,* 259-289.

Lehrer, R., & Schauble, L. (2002). Distribution: A resource for understanding error and natural variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on the Teaching of Statistics (ICOTS-6), Cape Town, South Africa [CD-ROM]*. Voorburg, The Netherlands: International Statistical Institute.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 149-176). Mahwah, NJ: Lawrence Erlbaum Associates.

Makar, K., & Confrey, J. (2005). Using distributions as statistical evidence in well-structured and ill-structured problems. In K. Makar (Ed.), *Reasoning about distribution: A collection of current research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-4)*, University of Auckland, New Zealand, 2-7 July.

Marangolo, P., Bonifazi, S., Tomaiuolo, F., Craighero, L., Coccia, M., Altoè, G., Provinciali, L., & Cantagallo, A. (2010). Improving language without words: First evidence from aphasia. *Neuropsychologia, 48,* 824-833.

Marceau, K., Ram, N., Houts, R., M., Grimm, K., J. & Susman, E. (2008). Individual differences in boys' and girls' timing and tempo of puberty: Modeling development with nonlinear growth models. *Developmental Psychology, 47*, 1389-1409.

Miller, S. A. (1976). Nonverbal assessment of Piagetian concepts. *Psychological Bulletin, 83*, 405-430.

Miller, S. A. (1982). Children's and adults' integration of information about noise and interest level in their judgments about learning. *Journal of Experimental Child Psychology, 33*, 536-546.

Moore, D. (1992). Teaching statistics as a respectable subject. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the Twenty-First Century* (pp. 14-25). Washington, DC: Mathematical Association of America.

Moore, D. (1998). Statistics among the liberal arts. *Journal of the American Statistical Association, 93*, 1253-1259.

Nisbett, R, E, Krantz, D, H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339-363.

Nisbett, R. E., & Ross, L. (1980). *Human Inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Opfer, J. E., & Siegler, R. S. (in press). Development of quantitative thinking. In K. Holyoak and R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning.* New York: Oxford University Press.

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*, 131-156.

Piaget, J. (1952). *The child's conception of number*. London: Routledge and Kegan Paul (Original work published in 1941).

Piaget, J. (1969). *The mechanisms of perception.* London : Routledge & Kegan Paul (Original work published in 1961).

Piaget, J., & Inhelder, B. (1967). *The child's conception of space*. New York: W.W. Norton (Original work published in 1948).

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal, 3(2)*, 84-105.

Reading, C., & Reid, J. (2010). Reasoning about variation: Rethinking theoretical frameworks to inform practice. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8),* Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201–226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist, 46*, 606-620.

Silverman J., W., & Paskewitz, S.L. (1988). Developmental and individual differences in children's area judgment rules. *Journal of Experimental Child Psychology, 46*, 74-87.

Tukey, J., W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tversky, A., Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Tversky, A., Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Tversky, A., Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804.

Watson, J. M, Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology, 34,* 1-29.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, 67, 223-265.

Wilkening, F. (1980). Development of dimensional integration in children's perceptual judgment: Experiments with area, volume, and velocity. In F. Wilkening, J. Becker & T. Trabasso (Eds.), *Information integration by children* (pp. 47-69). Hillsdale, NJ: Erlbaum.