anales de psicología, 2017, vol. 33, nº 3 (october), 450-460
http://dx.doi.org/10.6018/analesps.33.3.271021

ISSN print edition: 0212-9728. ISSN web edition (http://revistas.um.es/analesps): 1695-2294

Observational data analysis using generalizability theory and general and mixed linear models: an empirical study of infant learning and development

Angel Blanco-Villaseñor[1*] and Elena Escolano-Pérez[2]

1 Universidad de Barcelona, Barcelona (Spain).
2 Universidad de Zaragoza, Zaragoza (Spain).

Título: Análisis de datos observacionales mediante la Teoría de la Generalizabilidad y la utilización del Modelo Lineal General y Mixto: Un estudio empírico del desarrollo y aprendizaje infantil

Resumen: Una adecuada evaluación de las competencias infantiles tempranas es esencial para potenciar un desarrollo óptimo, pues los primeros años de vida son la base de todo el desarrollo y aprendizaje posterior. Sin embargo, todavía existen ciertas limitaciones y deficiencias en el ámbito de la medición del desarrollo y aprendizaje infantil. Con el objetivo último de contribuir a la mejora de esta situación, este trabajo presenta las posibilidades y ventajas que ofrecen nuevas técnicas de análisis de datos, tanto para controlar la calidad de los datos infantiles registrados a través de observación sistemática como para analizar su variabilidad. Se ha observado en tres edades diferentes (18, 21 y 24 meses) la actividad lógica y ejecutiva de 48 niños usando un diseño observacional nomotético, de seguimiento y multidimensional.

Dadas las particularidades de los datos del estudio que presentamos, desde el punto de vista metodológico y su análisis, realizamos análisis pormenorizados a través de la Teoría de la Generalizabilidad en tres vertientes posibles en un estudio observacional: Análisis de la fiabilidad intra e inter-observadores, Análisis de la validez del instrumento de observación y Estimación muestral de las facetas estudiadas (en concreto, la de participantes). De esta forma, se pretende optimizar el número de facetas y niveles necesarios para llevar a cabo un estudio de tales características.

Además, se utilizan otras técnicas analíticas para conocer la variabilidad del desarrollo y aprendizaje infantil, como son el Modelo Lineal General y el Modelo MIXED.

Los resultados indican cómo el uso de la Teoría de la Generalizabilidad permite controlar la calidad de los datos observacionales en una estructura única que integra la fiabilidad, validad y generalizabilidad.

Palabras clave: Observación sistemática; Modelo Lineal General; Teoría de la Generalizabilidad; desarrollo; aprendizaje; infancia.

Abstract: Accurate evaluation of early childhood competencies is essential for favoring optimal development, as the first years of life form the foundations for later learning and development. Nonetheless, there are still certain limitations and deficiencies related to how infant learning and development are measured. With the aim of helping to overcome some of the difficulties, in this article we describe the potential and advantages of new data analysis techniques for checking the quality of data collected by the systematic observation of infants and assessing variability. Logical and executive activity of 48 children was observed in three ages (18, 21 and 24 months) using a nomothetic, follow-up and multidimensional observational design.

Given the nature of the data analyzed, we provide a detailed methodological and analytical overview of generalizability theory from three perspectives linked to observational methodology: intra- and inter-observer reliability, instrument validity, and sample size estimation, with a particular focus on the participant facet. The aim was to identify the optimal number of facets and levels needed to perform a systematic observational study of very young children.

We also discuss the use of other techniques such as general and mixed linear models to analyze variability of learning and development.

Results show how the use of Generalizability Theory allows controlling the quality of observational data in a global structure integrating reliability, validity and generalizability.

Key words: Systematic observation; General Linear Model; Generalizability Theory; development; learning; childhood.

Introduction

Human development is a broad, complex phenomenon (Guralnick, 2015) characterized by a process of construction and continuous change arising from diverse dynamic interactions between numerous elements such as genes, neural activity, pre-, peri-, and post-natal behavior, physical environment, and social and cultural factors (Massand & Karmiloff-Smith, 2015). These changes, which start as early as conception, continue throughout a person's life and affect all dimensions of an individual, whether, physical, social, cognitive, linguistic, emotional, or personal. They do not, obviously, all occur simultaneously or with the same frequency or intensity. The first three years of a child's life is a critical period for development as it is a time of multiple, complex, and interacting changes affecting different dimensions that result in numerous gains that form the building blocks for even more complex gains in the future. What occurs during this period therefore lays the ground for lifelong development and learning (Scharf, Scharf, & Stroustrup, 2016).

Cognitive development is one of the numerous and complex changes that occur during early childhood, and it is a crucial part of development (Nelson & Luciana, 2008) as it involves the construction of highly diverse yet interdependent skills essentially involving processes linked to the acquisition, organization, retention, and use of information and knowledge that allow a person to adapt to a continuously changing environment (Goswami, 2010). This environment, in turn, affects the nature of the changes that occur, as, like the individual, it is an active participant in the development process. The construction of these cognitive skills, i.e., cognitive development, is closely related to cortical development, as the increasing structural complexity of the cortex gives rise to increasingly sophisticated cognitive capacities. In addition, neural complexity and organization is itself modified by its own functioning. The progressive specialization of neural structures is driven by environmental experiences that are expressly chosen by and participated in by the child. The acquisition of very basic cognitive skills therefore

*** Correspondence address [Dirección para correspondencia]:**
Angel Blanco-Villaseñor. Facultad de Psicología. Universidad de Barcelona. Campus Mundet. Pº Vall d'Hebrón, 171. 08035 Barcelona (Spain).
E-mail: ablanco@ub.edu

favors new neural connections that enable the acquisition of increasingly complex skills and learning (Karmiloff-Smith, Casey, Massand, Tomalski, & Thomas, 2014).

This study focuses on two types of cognitive skills that emerge in early childhood: logic and executive functions (EFs).

Logic is the ability to capture, elaborate, structure, and interiorize information; its origins lie in organized, mindful actions executed by a baby in its environment (Langer, 1986, 1990). Through some of these actions, the baby is focusing his attention on exploring and experimenting with the physical world around him, capturing information about the direct properties of objects. In short, he is constructing physical knowledge. Through others, he is focusing on the relationships between actions and objects. In this case, he is building logical-mathematical knowledge (Langer, 1986, 1990). Although action-based logic exists from the very first days of a child's life, important milestones are achieved in the second year of life. This is when basic action schemas, such as grasping or sucking, are replaced by differentiated actions, i.e., by actions adapted to the specific characteristics of the objects. Furthermore, these actions are coordinated, shaped, combined, and redefined, as the child tries out similar actions to obtain new information. He starts to become capable of bringing together different objects and then similar objects, and following an increasingly complex organization of actions, learns to establish relationships between elements that belong to different sets. One example is one-to-one correspondence, which involves sequentially pairing each element in one set to one and only one element in another set, such that the elements are equivalent. One-to-one correspondence tasks are essential for the learning and development of mathematical skills, which, in turn, are essentially for successfully participating in today's society (Izard, Streri, & Spelke, 2014). The ability to make one-to-one correspondences is influenced by the number and characteristics of the elements in each set and by the number of sets that have to be matched. Elements that are related in terms of shape and size (i.e., elements from one set that fit into elements in another) are easier to match (Langer, 1986, 1990; Sinclair, Stambak, Lézine, Rayna, & Verba, 1984). By the age of 18 months, almost all children (91.6%) are capable of establishing correspondences between two sets of two elements each. However, they cannot complete correspondences between three or more two-element sets or between two sets consisting of three or four elements each. By 21 months, all children can successfully pair elements from two sets featuring two elements each. By this age, some children will already have started to make correspondences between two sets composed of three elements (16.6%) and between three or four sets consisting of two elements (8.3% in both cases). Considerable progress is seen by the age of two. Most children at this age (58.3%) can now make correspondences between two sets of three elements, and for the first time, they start to successfully pair elements from two sets containing four elements. A surprisingly high percentage of children of this age (66%) can master this skill. There is also an increase in the percentage of children capable of making correspondences between three and four two-element sets (25% and 33.3%, respectively) (Langer, 1986).

Despite the importance of logic in cognitive development, tasks such as one-to-one correspondence during the first three years of life have received little attention from researchers. Most studies of one-to-one correspondence have targeted preschool and school-age children (i.e., those aged > 3 years) and have largely focused on studying mathematical skills in a formal education setting (Muldoon, Lewis, & Towse, 2005). Surprisingly little attention has been paid to the fact that these skills are built on knowledge gained in previous years.

EFs have been the subject of much research in the last decade, particularly in studies of early development (Carlson, Zelazo, & Faja, 2013). EFs are processes that allow an individual to control and self-regulate their behavior in order to achieve a goal in new or complex situations (Barkley, 2012; Guare, 2014). They are important not only for cognitive development but also for social, personal, and emotional development and have been identified as essential for school adaptation and success and even for health (Diamond, 2013; García, Rodríguez, González-Castro, Álvarez-García, & González-Pienda, 2016; Guare, 2014; Iglesias-Sarmiento, Carriedo, & Rodríguez, 2015). They are the building blocks for learning and adaptation, and allow children to pay attention, store information and not lose sight of objectives, refrain from not answering automatically, resist distractions, consider the consequences of actions, reflect on past experiences, and plan future ones. They are so important that several studies of preschool children have shown that the ability to remain seated, pay attention, and remember and follow rules (all aspects involving EFs) are more important for later adaptation, learning and success at school than early mathematical or linguistic skills, or even IQ (Viterbori, Usai, Traverso, & De Franchis, 2015; Wass, 2015). The role of early EFs in human learning and development, however, extends well beyond school years into later life, where the successful acquisition of EFs in early years is an important contributor to success at work and in one's personal, family, and social life. EFs have also been associated with better health and higher socioeconomic status (Miyake & Friedman, 2012). Some authors have even claimed that the study of the early development of EFs is essential in order to understand child learning and development (Moriguchi, Chevalier, & Zelazo, 2016) and even human learning and development in all its forms.

The most recent models of EFs include working memory (storing information for a short period of time while processing it mentally), inhibition (suppressing a predominant response or stimulus that is irrelevant to the task at hand), and cognitive flexibility (ability to quickly change and adapt one's course of thought or action to the demands of a continually changing situation) (Diamond, 2013; Miyake et al., 2000; Zelazo & Carlson, 2012). EFs start in early childhood (towards the end of the first year of life) and develop rapidly between the ages of 2 and 5. They continue to develop, albeit at a slower pace, into adolescence, when rapid improvements occur again. Following another period of more progressive development, EF skills reach their peak around the age of 20 years (Best & Miller, 2010; Brydges, Anderson, Reid, & Fox, 2013; Flores-Lázaro, Castillo Preciado, & Jiménez-Miramonte, 2014). EFs develop in tandem with the maturation of their main neuroanatomic structure: the dorsolateral prefrontal cortex (Funahashi & Andreau, 2013).

Inhibition is considered by numerous authors to be one of the most important cognitive EFs (Barkley, 2012; Miyake et al., 2000) and a key component of human behavior (Albert, López-Martí, & Carretié, 2010), intelligence (Duan, Wei, Wang, & Shi, 2010) and adaptive responses, which are essential for success in everyday life (Petersen, Hoyniak, McQuillan, Bates, & Staples, 2016). In childhood, inhibition is the best predictor of behavior and socio-emotional skills. In preschool children, it has been found to be a good predictor of later mathematical, reading, and writing skills and, as such, optimization of inhibitory control could help to prevent later learning difficulties (Stievano & Valeri, 2013).

Many authors believe inhibition to be a multidimensional construct, i.e., a family of separate yet related inhibitory processes (Brydges et al., 2013). Numerous hypotheses have been made about these processes (Dempster & Corkill, 1999; Friedman & Miyake, 2004; Harnishfeger, 1995; Howard, Johnson, & Pascual-Leone, 2014; Nigg, 2000; Nee & Jonides, 2008). In this study, we analyze what is known as resistance to distractor interference (Dempster, 1993; Friedman & Miyake, 2004). This basic process is the least studied of the inhibitory processes that affect cognitive development and very few longitudinal studies have been conducted in this area. Resistance to distractor interference is the ability to resist interference or distractions generated by external information or stimuli that are irrelevant to the task at hand and can interfere with its successful completion. It requires the person to select the information or stimulus he needs to resolve the task while ignoring competing distractions (Mishra, Anguera, Ziegler, & Gazzaley, 2014).

Resistance to interference, like other EFs, has been scarcely studied in children aged 0 to 3 years (Hendry, Jones, & Charman, 2016), despite its apparent importance for later development. As mentioned, most studies of EFs in childhood have focused on preschool and school-aged children, and the majority have been cross-sectional (Shanmugan & Satterthwaiter, 2015). Very few studies have examined the early development of EFs from a longitudinal perspective (Best & Miller, 2010; Willoughby, Holochwost, Blanton, & Blair, 2014; Willoughby, Wirth, & Blair, 2011). As a result, despite the extensive research into EFs that has emerged in recent years, our knowledge and understanding is skewed towards conceptual aspects, with much remaining to be learnt about how these functions develop over time and how they can be measured (Willoughby & Blair, 2016). One possible reason for the lack of longitudinal studies is the difficulty associated with actually capturing development and change (Isquith, Gioia, & Espy, 2004) and studying mental processes (García Molina, Tirapu Ustárroz, & Roig Rovira, 2007). These aspects cannot be studied by direct observation and researchers must therefore analyze the outputs of these processes and draw inferences (Willoughby et al., 2011). Additional obstacles include working with very young children (Clark, Flewitt, Hammersley, & Robb, 2013), as their behavior is irregular and they have a short attention span, highly fluctuating motivation (Aslin & Fiser, 2005), immaturity to cooperate consistently (Field & Behrman, 2004) and limited verbal abilities (Salley, Panneton, & Colombo, 2013). All these issues contribute to the difficulty and complexity of obtaining reliable and valid data related to child learning and development and, therefore, to the low number of studies about child cognitive development. In addition, we frequently find in these studies that the samples are small in size, due to ethical and legal issues and because sometimes parents may be reluctant to allow their children to participate in research (Alderson, 2004; Shaw, Brady, & Davey, 2011). Few methods are therefore suitable for studying this population. One of the most suitable options—and sometimes the only one—is observational methodology (Anguera, 2001, 2010; Bryce & Whitebread, 2012; Herrero, 1992; Whitebread & Coltman, 2010). Despite its numerous benefits and strengths, however, systematic observation also has disadvantages or difficulties, such as the time needed to collect and code the data and the extensive training needed for observers (Anguera, 2010). All the above factors probably explain, at least in part, why cognitive processes have been studied so little in very young children.

Nonetheless, early assessment of children's activities and skills is essential for favoring optimal learning and development, as it can identify possible barriers and permit the planning of suitable actions to overcome these barriers and prevent adverse consequences later in life. In this article, we explore the potential and advantages by recent data analysis techniques for analyzing observational data collected in natural settings with a particular focus on examining the quality of data collected by systematic observation in infants and analyzing variability. We investigate the use of new procedures for calculating intra-and inter-observer reliability and validity and for assessing the generalizability of results from a sample to a larger population with the same characteristics. We also discuss the advantages of general and mixed linear models for analyzing the variability of data.

In response to recent calls for solutions to overcome the limitations in this area, (Carlson, Faja, & Beck, 2016; Escolano-Pérez & Blanco-Villaseñor, 2015; Willoughby, Wirth, Blair, & Family Life Project Investigators, 2016), we hope that this study will contribute to identifying better approaches for measuring the acquisition and development of skills in children.

Method

Design

We employed a nomothetic, multidimensional, follow-up observational design (Anguera, Blanco-Villaseñor, Hernández-Mendo, & Losada, 2011; Anguera, Blanco-Villaseñor, & Losada, 2001). It was nomothetic because we studied several participants, multidimensional, because we observed various dimensions of children's behavior related to one-to-one correspondence and resistance to distractor interference, and follow-up because the participants were studied at three moments of their lives (ages 18, 21, and 24 months).

Participants

The sample consisted of 48 participants evaluated longitudinally at the ages of 18, 21, and 24 months. Development was considered to be normal in all children. They had had no congenital risk factors or diseases and there had been no pre-, peri-, or postnatal complications (Grupo de Atención Temprana, 2000). The socioeconomic status of their families was medium to high and all the children were enrolled at the same private education center.

The sample was a convenience sample selected by non-probability sampling. The children selected in this sample were extracted from a list encompassing all students enrolled in the education center who fulfilled the above mentioned characteristics (ages studied, normal development and absence of risk factors, diseases, and pre-, peri-, or postnatal complications) and whose parents signed the informed consent form authorizing their participation in the study.

All the children were treated with in compliance with international guidelines and ethical principles for scientific research. Informed consent was obtained from all parents.

Instruments

Different typologies of instruments were used:

1.- Three nonverbal recreational tasks were designed *ad hoc* to facilitate the establishment of one-to-one correspondences by the children, with no external intervention. The last of the tasks was designed to additionally test resistance to distractor interference.

The one-to-one correspondence tasks were facilitated by the use of two sets consisting of matching objects in terms of size (i.e., each object in one set fit into another object in the other set). One of the sets contained four cups of a different size and the other contained four balls with matching sizes. The fact that each of the balls fit into just one cup favored the successful completion of the task. According to the evidence on one-to-one correspondence abilities, both the number of sets (two) and the number of elements in each set (four) were adequate for capturing the process that the majority of very young children go through before they are able to successfully complete such a task by the age of 24 months (Langer, 1986). The colors of the cups and balls were modified to present increasing difficulty throughout the three tasks. In the first task, cups and balls of the same size were the same color. In the second task, all the cups and balls were white, meaning that the task had to be resolved based on size only, and finally, in the third task, there were four colored cups and four identically colored balls but matching cups and balls were a different color. In this last task, thus, color served as a distraction interfering with completion of the task, as reasoning based on color did not resolve the task (e.g., the biggest ball and the smallest cup were red). The children thus were required to resist the interference generated by color and focus on size only.

2.- The following instruments and equipment were used for the systematic observation of the tasks:

a) A digital video camera.

b) The Early Logical and Executive Development Assessment observation instrument (ELEDA) (Escolano-Pérez & Sastre-Riba, 2010), which combines a field format system and category systems designed to capture aspects of logic and EFs during children's activities, with a particular focus on one-to-one correspondences and resistance to distractor interference. Some examples of the dimensions and the category systems that comprise it are:

* *Content*: This dimension refers to the type of logical activities that are the previous, forthcoming and necessary activities in the development of the one-to-one correspondence in order to reach it (Langer, 1986; Sinclair et al., 1984). Since one-to-one correspondence tasks are essential in the study of infant logical activity development, and consequently in this research, the inclusion in the observation instrument of logical activities that indicate what is the course or degree of development of one-to-one correspondence is absolutely necessary. This dimension is formed by a category system of 8 exhaustive and mutually exclusive categories. Some of these categories are:

- «Grouping»: Putting together elements from different sets (Sinclair et al., 1984).
- «Collection»: Putting together elements from the same sets (Sinclair et al., 1984).
- «One-to-various/all distribution»: An element from one set is sequentially related with various/all of the elements in the other set.
- «Various/all-to-various/all distribution»: Various or all of the elements of one set are individually and sequentially related with various or all of the elements in the other set.

* *Adaptation*: This dimension informs about the existence or absence of agreement in the size and color of the related elements. Consequently, it informs about the facilitating role of the color in the resolution of the task 1 and about the interfering role of the color in the resolution of the task 3. In this previous case (task 3), it assesses the infant's ability to resist interference generated by the distracting stimulus (color). This dimension is formed by a category system of 6 ex-

haustive and mutually exclusive categories. Some of these categories are:

- «Adaptation of size and color»: All of the interrelated elements concur with one another in size and color. It is only possible in task 1 due to the characteristics of elements. This category assesses the facilitating role of the color in order to resolve task 1.
- «Adaptation of size but not color»: All the interrelated elements concur with one another in size but not in color. It is only possible in task 3 due to the characteristics of elements. It indicates the infant's ability to resist interference generated by the color.
- «Adaptation of color but not size»: All the interrelated elements concur with one another in color but not in size. It is only possible in task 3 due to the characteristics of elements. This category informs about no-resistance to interference generated by the color.

* *Scope*: This dimension indicates the number of elements used by children in their action or in its results. This dimension is formed by a category system of 2 exhaustive and mutually exclusive categories:

- «Exhaustive»: Participation of all of the elements in the action or in its results.
- «Nonexhaustive»: Participation of some elements in the action or in its results.

The Early Logical and Executive Development Assessment observation instrument (ELEDA) can be entirely consulted in Escolano-Pérez and Sastre-Riba (2010).

c) Match Video Studio v1.0 (Perea, Alday, & Castellano, 2006) was used to analyze and code the video recordings.

3.- The data were analyzed using SAS 9.1.3 3 (SAS Institute Inc., 2004; Schlotzhauer & Littell, 1997) and EduG 6.0-e (Cardinet, Johnson, & Pini, 2010).

Procedure

The children were video-recorded as they individually completed all three tasks, in order of difficulty, at the ages of 18, 21, and 24 months. Each child was allowed to play freely with the objects and did not receive any instructions until they voluntarily completed the activity, at which time the observation session was considered complete. All the sessions were recorded at the education center facilities. The children were each accompanied by an adult who provided them with the objects for the tasks but did not intervene. For each task, the cups and balls were randomly positioned on the floor by dropping them out of a bag, with care taken to ensure that none of the balls had accidentally rolled into a cup.

The video-recordings of the tasks were subsequently analyzed and coded using the *ad hoc* ELEDA instrument in Match Vision Studio v. 1.0. The same observer (an expert in both observational methodology and child logical and executive development, author of the observation instrument and co-author of this manuscript) coded each of the children's sessions. Furthermore, another observer (an expert in observational methodology and in child learning and development) was trained for the use of ELEDA. He registered 27 sessions belonging to participants of the three ages and in the three tasks. Some of his coded sessions were used for the inter-observer reliability analysis.

Statistical Analysis

As required by observational methodology, the quality of the resulting datasets was checked by statistical analyses. Data quality control is an essential part of any observational methodology study and can be analyzed from four perspectives: reliability, accuracy, validity, and estimation of sample size. These aspects can also be analyzed as a whole through a generalizability study. In this study we report on our analyses of these four aspects. In addition, we analyzed the variability of the observational data using the general linear model procedure (PROC GLM) and the mixed linear model (PROC MIXED) in SAS.

Data quality control: reliability, validity, sample size estimation, and generalizability

As we will see in the following paragraphs, data reliability can be estimated using different methods, each of which generates a different coefficient. For example, we can check ratings assigned to the same behaviors by a single observer on two different occasions (inter-observer reliability); ratings assigned by different observers at the same or a different time within a session or on different occasions separated by a short period of time (inter-observer reliability), or ratings assigned using different scales that measure the same behavior (parallel-forms reliability). However, these standard measures do not account for all possible sources of variation. One of the aims of this study was to apply a new method—based on the concepts of analysis of variance—to check the quality of data obtained from the systematic observation of very young children. We did this within the framework of generalizability theory (G theory), developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972). The use of G theory for assessing the reliability of measurements in observational studies was prompted by the work of Mitchell (1979), who clearly established that inter-observer agreement measures were inadequate in this setting.

The differences between agreement (concordance) and reliability (correlation) lie in how these measures are defined. As stated by Mitchell (1979, p. 382), "reliability coefficients partition the variance of a set of scores into a true score (individual differences) and an error component. Interobserver agreement percentages, on the other hand, carry no information at all about individual differences among subjects and contain information about only one of the possible sources of error—differences among observers." These measures, therefore, cannot be used to estimate variance components related to differences in observers, measurement tools, or moments of time, nor can they consider these sources of variation simultaneously. These limitations thus justify the need for a multivariate theory that takes into account all possible sources of error in addition to those contemplated by empirical validity tests (Blanco-Villaseñor, Castellano, Hernández Mendo, Sánchez-López, & Usabiaga, 2014). We agree that such an integrative approach is necessary for guaranteeing the quality of observational datasets.

Measurements used in observational methodology studies provide data that may be influenced not only by individual differences between study subjects but also by aspects related to the observation procedure itself (e.g., different observers, data collection times, recording methods, or observation instrument criteria). This is the perspective that defines G theory. G theory assumes the existence of sources of variation other than individual differences and integrates these within a global structure that contemplates not only the sources of variation in the above-mentioned reliability coefficients but also sources attributable to the observation instrument criteria and the study subjects. If the observer (intra-observer reliability) or observers (inter-observer reliability) were used as the instrumentation or generalization facets in the measurement design in G theory, we would be analyzing the reliability of the data (intraclass correlation coefficient). If, by contrast, these same facets were used as differentiation facets, we would be testing the validity of the observation instrument. Finally, if participants rather than observers was used as the instrumentation facet, we would be assessing whether the size of the sample is sufficient in order to generalize the results to the reference population.

Variability analysis

Variability analysis is important for numerous reasons. First, although used widely, conventional data analysis techniques frequently have little or nothing to do with situations studied in Educational and Developmental Psychology. In our opinion, these techniques are not appropriate for studying aspects related to human learning and development, particularly in its early stages, as the samples are neither adequate nor truly representative (i.e., they are not fully randomized). Another problem is that standard procedures for calculating variance, such as the least-squares method in PROC GLM do not take missing data into account, unlike PROC MIXED, which is based on maximum likelihood estimation.

When working with models focused on individual learning and development, it is appropriate to analyze data corresponding to characteristics or behaviors that are measured on two or more occasions. These studies are generally referred to as longitudinal repeated measures studies. Although certain aspects will necessary change (e.g., time, situation, session, age) what is being measured is not and we can therefore apply a repeated measures analysis that accounts for within-subject covariability over time.

Results

To calculate the intra- and inter-observer reliability coefficients for our study, we analyzed 16 of the 48 children (randomly selected) performing different tasks at the ages of 18, 21, and 24 months. We analyzed thus 16 observation sessions, of which 10 were used to measure intra-observer reliability and six were used to measure inter-observer reliability. Tables 1 and 2 show the results for one of the children, while Tables 3 and 4 show a summary of the results for all the children analyzed.

The generalizability coefficients for all the sessions were calculated in the EduG 6.0-e software program using a single three-facet measurement (observers [O], macro-categories [M], and categories [C]), where observer (intra-observer reliability) or observers (inter-observer reliability) were used in all cases as the instrumentation facet. The O x M x C measurement design therefore had two differentiation facets (MC) and one instrumentation facet (O). This formula actually coincides with the *ICC* as it reduces the bias that can be introduced by an observer who consistently assigns lower or higher ratings than the others.

Generally speaking, the results for both intra-and inter-observer reliability were very satisfactory (with *ICC*s in the range of .96-.99), particularly if we consider that these coefficients are not based on absolute error and are not percentages.

The *ICC*s for the 10 intra-observer reliability sessions were close to 1 (.98-.99) (Table 3). These results, which correspond to the same observation session viewed and coded by the same observer on two occasions, can be consisted excellent.

Note that the variability for the observer facet is 0% for 9 of the 10 sessions analyzed, indicating that the same observer coded the sessions almost identically on two separate occasions. The estimation of results for an infinite number of occasions shows that this observer would make very few rating errors.

The differences between one session and another are due to errors made by the observer when coding the macro-categories or criteria (O x M), but as seen, the variability did not exceed 1% on any of the occasions. As expected, more errors were made when coding the categories (O x C), with results varying from 0% to 3%.

Whatever the case, all the structures show that there are no additional sources of error resulting in serious bias or errors in the systematic observation system used. The residual error (O x M x C) for the 10 sessions i.e., the unknown source of variability, was zero.

Table 4 shows the excellent results obtained for inter-observer reliability (corresponding to two observers independently coding the same observation session at the same or at different times), although the *ICC*s were a little lower than those observed for intra-observer reliability (.96-.99 vs. .98-.99). This is logical as the perceptions of two different observers are more likely to contain more errors than those of a single observer. Indeed, it would not be possible to calculate inter-observer reliability if the intra-observer results were not close to 1, as was the case in our sample.

The results for the six sessions show that the errors made by the observers when coding the macro-categories (O x M) were similar to those made by the same observer when rating the sessions on two different occasions. We did, however, observe greater variability for the coding of categories, with values ranging from 1% to 5%. The session with a variability of 5% also had a lower *ICC*. With the exception of this case, the other results were similar to those observed for intra-observer reliability.

Table 1. Intra-Observer Reliability for Participant 28 During Task 3 Assessed by Generalizability Theory with Corresponding Relative and Absolute Generalizability Coefficients (*GC*, rel, abs).

Source	*SS*	*MS*	%	*GC* (rel, abs)
Observer	0.6621	0.0008	0	
Macro	4278.2820	6.4213	46	
O x M	1.6920	0.0050	0	
Categories	5324.3125	7.3863	53	.99 .99
O x C	6.1458	0.0170	0	
M x C	0	0	0	
O x M x C	0	0	0	

Table 2. Inter-Observer Reliability for Participant 28 During Task 2 Assessed by Generalizability Theory with Corresponding Relative and Absolute Generalizability Coefficients (*GC*, rel, abs).

Source	*SS*	*MS*	%	*GC* (rel, abs)
Observers	3.4594	0.0053	0	
Macro	1048.2167	1.5629	47	
O x M	7.3113	0.0219	1	
Categories	1250.7291	1.7034	51	.987 .986
O x C	24.2291	0.0673	2	
M x C	0	0	0	
O x M x C	0	0	0	

Table 3. Summary of 10 Intra-Observer Reliability Sessions, With Generalizability Coefficients (*GC*) Obtained by Generalized Theory and Intraclass Correlation Coefficient (*ICC*).

Intra-observer Sessions	1	2	3	4	5	6	7	8	9	10
Observer	0	0	0	0	1	0	0	0	0	0
O x Macro	1	1	0	1	1	1	1	0	2	1
O x Categories	0	2	0	3	3	0	3	3	3	3
O x M x C	0	0	0	0	0	0	0	0	0	0
ICC or *GC*	.993	.987	.99	.98	.98	.99	.979	.984	.987	.982

Table 4. Summary of 6 Inter-Observer Reliability Sessions, With Generalizability Coefficients (*GC*) Obtained by Generalized Theory and Intraclass Correlation Coefficient (*ICC*).

Inter-observer Sessions	1	2	3	4	5	6
Observers	0	0	0	0	0	0
O x Macro	3	2	1	3	0	1
O x Categories	1	5	2	4	1	1
O x M x C	0	0	0	0	0	0
ICC or *GC*	.981	.964	.983	.959	.996	.988

The design structures used to calculate reliability can also be used to assess the validity of the criteria and categories in the observation instrument, i.e., instrument validity.

As shown in Table 5, when the macro-category and category facets were considered together, variability was close to maximum levels (96%), with zero residual variance in all cases. The results for the macro-categories and categories both show that the maximum variability obtained for all the structures allows us to adequately determine what has been coded or assessed in one or other of the categories. Therefore, if for any of the sessions we were to apply G theory with a measurement design in which the macro-categories or categories (either separately or together) were used as the generalization or instrumentation facet, the coefficient would be close to 0 in all cases. We decided to omit all the results in Tables 2-4, and as the values are almost identical (high variability for macro-categories and categories and zero residual value). Our claim makes even more sense if we consider the real results observed in the different observation sessions. In other words, the observation instrument created for this study is valid for recording what it was designed to record based on the theoretical framework derived from the scientific literature and the corresponding hypothetical constructs. Obviously the macro-category and category facets have a different meaning depending on the measurement design used. When they were used as differentiation facets, the differences with the other facets were greatest, and when they are used as instrumentation facets, the coefficients were 0 in almost all cases. Our observation instrument is therefore valid.

Table 5. Summary of 10 Sessions Used to Calculate Intra-Observer Validity.

Intra-observer sessions	1	2	3	4	5	6	7	8	9	10
Macrocategories	61	47	46	37	46	39	28	52	29	33
Observer x M	1	1	0	1	1	1	1	0	2	1
Categories	38	51	53	59	49	60	68	45	66	63
O x C	0	2	0	3	3	0	3	3	3	3
M x C	0	0	0	0	0	0	0	0	0	0
O x M x C	0	0	0	0	0	0	0	0	0	0
GC	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0

In brief, the analysis of different observation sessions using G theory shows that a single three-facet design (observers, macro-categories, and categories, considered either separately or together) can be used to analyze the veracity of data using filters that are key to ensuring quality (intra- and inter-observer reliability and validity). The results of this initial data quality assessment attest to the quality of the data used in the subsequent analyses. Data quality control is particularly important in longitudinal studies involving very young children. As shown by our reliability and validity tests, the data obtained by our systematic observation system offers more than sufficient guarantees of quality and G theory allowed us to structure all the corresponding measures within a single analysis unit.

Regarding the adequacy of sample size, the model (Age x Task x Participant) had a high residual error (39%) (Table 6). In other words, 39% of the variance observed is due to unknown variables that were not contemplated in the model, indicating the need to include additional variables or facets to identify factors that will help to better explain the use and development of logic and executive functions in childhood.

Table 6. Estimation of Ideal Sample Size. Age x Task x Participants (A x T x P) Estimation Design by Generalizability Theory.

Source	Estimation of variance components	% total variance
Age	16.0436	0
Task	192.3609	9
Age x Task	51.5028	4
Participants	695.6675	11
Age x Participants	416.4789	12
Participants x Task	646.7291	25
A x T x P	440.4757	39

Table 7 shows the generalizability coefficient for the results for the 48 children in our sample (n_p = 48) and for two additional samples size tested in the optimization design: n_p = 60 and n_p = 70. The coefficient obtained for our sample was high, at .89, demonstrating that our results can be generalized to the study population with considerable confidence. The results obtained for the larger samples sizes were only slightly higher at .91 for 60 children and .92 for 70. In our opinion, this slight improvement would not compensate the additional costs of having to study more children, especially considering their young age.

Table 7. Optimization Design Age Task/Participant to Calculate Optimal Sample Size Using Generalizability Theory.

Facet	Levels	Sample size	Option 1	Option 2
Participants	n_p = 48	$N_p = \infty$	60	70
Generalizability coefficient = .89			.91	.92

One of the most interesting options for extending our research in the field of cognitive development in very young children is the use of growth curve modeling, designed to track individual development based on repeated measures over time (McArdle & Nesselroade, 2003). Growth curve models can be used to explore two levels of variability within the response variables: within-subject variability and between-subject variability. Table 8 and Table 9 respectively show the results of the GLM univariate and multivariate analyses of within-subject variability over time, although it should be noted that this procedure does not account for missing data, which is a common limitation of such analyses in Educational and Developmental Psychology. The results for all the facets and their interactions are significant in all cases except (surprisingly) Age and the interaction Participants x Actions.

Observations made at different times points in longitudinal studies are nested within the subject and therefore the study population has a two-level hierarchical structure, with within-subject longitudinal variability at the bottom and between-subject variability at the top. Estimation of population covariance matrices would provide percentages corresponding to the development curve. In this respect, the fact that recent software programs now offer more accurate maximum likelihood procedures is an important consideration. Table 10 shows the results for the same multivariate analysis using PROC MIXED in SAS, which is more suitable as it accounts for missing data.

Table 8. Univariate Analysis by General Linear Model.

Source	*DF*	*SS*	*MS*	*F* Value	*Pr* > *F*
Age	2	16.043689	8.021844	2.42	.0890
Task	2	192.360989	96.180495	29.03	< .0001
Participants	47	694.293450	14.772201	4.46	< .0001
Age x Task	4	52.876962	13.219240	3.99	.0031
Age x Participants	56	416.478932	7.437124	2.24	< .0001
Age x Actions	126	2945.405353	23.376233	7.06	< .0001
Participants x Actions	1974	6721.256565	3.404892	1.03	.2451

Table 9. Analysis of Multivariate Repeated Measures Data Using a General Linear Model that Does Not Account for Missing Data. Hypothesis Test for Between-Subject Effects and Univariate Hypothesis Test for Within-Subject Effects.

Source	*DF*	*SS*	*MS*	*F* Value	*Pr* > *F*
Participants	20	1286.46167	64.32308	9.65	< .0001
Task	2	397.10068	198.55034	29.79	< .0001
Error	2923	19480.68273	6.66462		

Source	*DF*	*SS*	*MS*	*F* Value	*Pr* > *F*
Age	2	13.78681	6.89341	2.25	.1060
Age x Participants	40	925.54614	23.13865	7.54	< .0001
Age x Task	4	110.36546	27.59137	8.99	< .0001
Error (age)	5846	17947.16869	3.06999		

Table 10. Multivariate analysis using the fixed-effects mixed linear model (PROC MIXED) in SAS that accounts for missing data.

Source	*DF*	*F* Value	*Pr* > *F*
Age x Task	6	2.89	.0082
Participants	47	4.37	< .0001
Task	2	27.10	< .0001
Actions	42	18.46	< .0001
Age x Participants	58	2.16	< .0001

Discussion

Certain individual cognitive and behavioral differences can be traced back to the first months of life (Bornstein, 2014), indicating that the risk of atypical learning and development in later years exists from a very young age. This highlights the importance of adequate and thorough evaluation of childhood learning and development as early as possible to permit the design and timely implementation of interventions at an age where the brain is most malleable and responsive (Karmiloff-Smith et al., 2014; Wass, 2015).

Observational methodology is the most suitable and perhaps the only option for capturing aspects of learning and development in infants, but it has been used in very few studies, despite its considerable advantages. Although the literature on childhood EFs has grown rapidly in recent years, most behavioral studies have measured behavior in clinical or laboratory settings or through questionnaires or surveys completed by third parties (e.g., parents or teachers). Both measurements systems have their limitations. Behaviors performed in an artificial, controlled setting, such as a laboratory, will necessarily differ from behaviors that occur in a natural everyday setting, and therefore any findings will have low ecological validity (Miranda, Colomer, Mercader, Fernández, & Presentación, 2016). On the other hand, while information supplied by third parties can provide insights into a greater number of situations, its reliability is questionable for numerous reasons related to, for instance, social desirability or recall bias, and even a lack of familiarity or sensitivity on the part of the observer to perceive and detect certain behaviors (Wertz, 2014). Systematic observation overcomes the above limitations in that it captures the spontaneous behavior of individuals in their natural environment and therefore has high ecological validity. Furthermore, the behaviors are rated or coded by one or more observers who are experts in both the "how" (the methodology) and the "what" (the subject being analyzed). Observers in this respect are "made not born" (Anguera, 2010). To ensure optimal results, observers participating in an observational methodology study should be provided with comprehensive training that ideally extends beyond the initial data collection phase.

These initial evaluation stages will determine subsequent stages and may lead to decisions that could have a determining impact on the child's learning and development. Accordingly, it is crucial to ensure the quality of the data used to make any decisions. Researchers should therefore take advantage of any relevant methodological advances that emerge to enhance the quality of data throughout all phases of a study. We have described some of these advances in this article

The repeated measures analysis used is available in standard software programs such as SAS and SPSS (Mushquash & O'Connor, 2006). The PROC GML procedure is also available in SAS, but it is valid only for traditional univariate and multivariate analyses. We believe that in future studies we might be able to use new procedures and structures in SAS (e.g., MIXED) that, through general covariance structures, will provide a better approximation to repeated measures modeling (Castellano, Blanco-Villaseñor, & Álvarez, 2011).

Future longitudinal studies of behaviors in young children will need to contemplate solutions that overcome the particularities of PROC GLM, which is limited by the dichotomy between within-subject and between-subject effects. One example is the repeated measures strategy offered by PROC MIXED, which has the additional advantage of accounting for missing data. It does this through maximum likelihood estimation (which requires full data) rather than through least squares estimation, which is used in PROC GLM (Schlotzhauer & Littell, 1997). We adopted such an approach in this study, although we believe that the best possible option would be to use multilevel growth models. These are typically referred to in the literature as longitudinal or repeated measures models (growth curves, life span curves, latent growth models) and they tend to simultaneously compare processes of stability and change in individuals and the groups they form. Such an approach would permit a more thorough and detailed analysis of the interactions underlying cognitive development in children.

Acknowledgements.- We gratefully acknowledge the support of both Spanish government projects (Secretaría de Estado de Investigación, Desarrollo e Innovación del Ministerio de Economía y Competitividad projects): [Grant PSI2015-71947-REDT; MINECO/FEDER, UE] and [Grant DEP2015-66069-P; MINECO/FEDER, UE]. We gratefully acknowledge the support of the Generalitat de Catalunya Research Group (*GRUP DE RECERCA E INNOVACIÓ EN DISSENYS [GRID]. Tecnología i aplicació multimedia i digital als dissenys observacionals),* [Grant 2014 SGR 971]. We gratefully acknowledge the support of Aragón Autonomous Community government to the research activity of *Grupo consolidado Educación y Diversidad* [S56].

References

Albert, J., López-Martín, S., & Carretié, L. (2010). Emotional context modulates response inhibition: Neural and behavioral data. *NeuroImage, 49*(1), 914-921. doi: 10.1016/j.neuroimage.2009.08.045

Alderson, P. (2004). Ethics. In S. Fraser, V. Lewis, S. Ding, M. Kellett, & C. Robinson (Eds.), *Doing research with children and young people* (pp. 97-111). London: Sage.

Anguera, M. T. (2001). Cómo apresar las competencias del bebé mediante una aplicación de la metodología observacional. *Contextos Educativos, 4*, 13-34.

Anguera, M. T. (2010). Posibilidades y relevancia de la observación sistemática por el profesional de la Psicología. *Papeles del Psicólogo*, *31*(1), 122-130. Recuperado de http://www.papelesdelpsicologo.es/pdf/1802.pdf

Anguera, M. T., Blanco-Villaseñor, A., Hernández-Mendo, A., & Losada, J. L. (2011). Diseños observacionales: ajuste y aplicación en psicología del deporte. *Cuadernos de Psicología del Deporte, 11*(2), 63-76.

Anguera, M. T., Blanco-Villaseñor, A., & Losada, J. L. (2001). Diseños observacionales, cuestión clave en el proceso de la metodología observacional. *Metodología de las Ciencias del Comportamiento, 3*(2), 135-160.

Aslin, R. N., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *TRENDS in Cognitive Sciences, 9*(3), 92-98.

Barkley, R. A. (2012). *Executive Functions: What They Are, How They Work, and Why They Evolved.* New York: Guilford.

Best, J. R., & Miller, P. H. (2010). A Developmental Perspective on Executive Function. *Child Development, 81*(6), 1641-1660. doi: 10.1111/j.1467-8624.2010.01499.x

Blanco-Villaseñor, A., Castellano, J., Hernández Mendo, A., Sánchez-López, C. R., & Usabiaga, O. (2014). Aplicación de la TG en el deporte para el estudio de la fiabilidad, validez y estimación de la muestra. *Revista de Psicología del Deporte, 23*(1), 131-137.

Bornstein, M. H. (2014). Human infancy … and the rest of the lifespan. *Annual Review of Psychology, 65*, 121-158. doi: 10.1146/annurev-psych-120710-100359

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: evidence from observational analysis of young children's behaviour during problem-solving. *Metacognition and Learning, 7*(3), 197-217.

Brydges, C. R., Anderson, M., Reid, C. L., & Fox, A. M. (2013). Maturation of cognitive control: delineating response inhibition and interference suppression. *PLoS ONE, 8*, e69826.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability Theory using EduG*. Londres: Routledge.

Carlson, S. M., Faja, S., & Beck, D. M. (2016). Incorporating early development into the measurement of executive function: The need for a continuum of measures across development. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 45-64). Washington, DC, US: American Psychological Association.

Carlson, S. M., Zelazo, P. D., & Faja, S. (2013). Executive function. In P. D. Zelazo (Ed.), *The Oxford handbook of developmental psychology, Vol. 1: Body and mind* (pp. 706-743). New York: Oxford University Press.

Castellano, J., Blanco-Villaseñor, A., & Álvarez, D. (2011). Contextual variables and time-motion analysis in soccer. *International Journal of Sports Medicine, 32*(6), 415-421. doi: 0.1055/s-0031-1271771

Clark, A., Flewitt, R., Hammersley, M., & Robb, M. (2013). *Understanding Research with Children and Young People*. London: Sage.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: theory of generalizability for scores and profiles*. New York: John Wiley and Sons.

Dempster, F. N. (1993). Resistance to interference: Developmental changes in a basic processing mechanism. In M. L. Howe & R. Pasnak (Eds.), *Emerging themes in cognitive development* (Vol. 1, pp. 3-27). New York, NY: Springer-Verlag.

Dempster, F. N., & Corkill, A. J. (1999). Interference and inhibition in cognition and behavior: unifying themes for education psychology. *Educational Psychology Review*, *11*(2), 1-88.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*, 135-168. doi: 10.1146/annurev-psych-113011-143750

Duan, X., Wei, S., Wang, G., & Shi, J. (2010). The relationship between executive functions and intelligence on 11- to 12-year-old children. *Psychological Test and Assessment Modeling, 52*, 419-431.

Escolano-Pérez, E., & Blanco-Villaseñor, A. (2015). The longitudinal measurement of change: Intraindividual variability in behavior and interindividual differences observed in childhood. *Anales de Psicología, 31*(2), 545-551. doi: 10.6018.analesps. 31.2.166361

Escolano-Pérez, E., & Sastre-Riba, S. (2010). Early Infant Cognitive Assessment: Validity of an instrument. *Behavior Research Methods, 42*(4), 759-767.

Field, M. J., & Behrman, R. E. (Eds.), (2004). *Ethical Conduct of Clinical Research Involving Children*. Washington (DC): National Academies Press.

Flores-Lázaro, J. C., Castillo Preciado, R. E., & Jiménez-Miramonte, N. A. (2014). Desarrollo de funciones ejecutivas, de la niñez a la juventud. *Anales de Psicología, 30*(2), 463-473. doi: 10.6018.analesps.30.2.155471

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control function: a latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101-135.

Funahashi, S., & Andreau, J. M. (2013). Prefrontal cortex and neural mechanisms of executive function. *Journal of Physiology-Paris, 6*, 471-482.

García Molina, A., Tirapu Ustárroz, J., & Roig Rovira, T. (2007). Validez ecológica en la exploración de las funciones ejecutivas. *Anales de Psicología, 23*(2), 289-299.

García, T., Rodríguez, C., González-Castro, P., Álvarez-García, D., & González-Pienda, J. A. (2016). Metacognición y funcionamiento ejecutivo en Educación Primaria. *Anales de Psicología, 32*(2), 474-483. doi: 10.6018.analesps.32.2.202891

Goswami, U. (2010). *The Blackwell Handbook of Childhood Cognitive Development.* New York: John Wiley and Sons.

Grupo de Atención Temprana (2000). *Libro blanco de la Atención Temprana.* Madrid: Real Patronato de Prevención y de Atención a Personas con Minusvalía. Consultado el 28 de Septiembre, 2015, en http://www.gat-atenciontemprana.org/

Guare, R. (2014). Context in the Development of Executive Functions in Children. *Applied Neuropsychology Child, 3*(3), 226-232. doi: 10.1080/21622965.2013.870015.

Guralnick, M. J. (2015). Merging policy initiatives and developmental perspectives in early intervention. *Escritos de Psicología*, *8*(2), 6-13. doi: 10.5231/psy.writ.2015.1004

Harnishfeger, K. K. (1995). The development of cognitive inhibition: Theories, definitions, and research evidence. In F. N. Dempster & C. J. Brainerd (Eds.), *Interference and Inhibition in Cognition* (pp. 175-204). San Diego, CA: Academic Press.

Hendry, A., Jones, E. J. H., & Charman, T. (2016). Executive function in the first three years of life: Precursors, predictors and patterns. *Developmental Review, 42,* 1-33. doi: 10.1016/j.dr.2016.06.005

Herrero, M. L. (1992). Posibilidades de la metodología observacional en el estudio analítico de conductas en el aula: Aplicación en escolares con problemas de comportamiento. *Anales de Psicología, 8*(1-2), 149-155.

Howard, S. J., Johnson, J., & Pascual-Leone, J. (2014). Clarifying inhibitory control: Diversity and development of attentional inhibition. *Cognitive Development, 31*(1), 1-21. doi: 10.1016/j.cogdev.2014.03.

Iglesias-Sarmiento, V., Carriedo, N., & Rodríguez, J. L. (2015). Updating executive function and performance in reading comprehension and problem solving. *Anales de Psicología, 31*(1), 298-309. doi: 10.6018.analesps.31.1.158111

Isquith, P. K., Gioia, G. A., & Espy, K. A. (2004). Executive function in preschool children: examination through everyday behavior. *Developmental Neuropsychology, 26*(1), 403-422. doi: 10.1207/s15326942dn2601_3

Izard, V., Streri, A., & Spelke, E. S. (2014). Toward Exact number: Young children use one-to-one correspondence to measure set identity but not numerical equality. *Cognitive Psychologyst, 72*, 27-53. doi: 10.1016/j.cogpschy.2014.01.004

Karmiloff-Smith, A., Casey, B. J., Massand, E., Tomalski, P., & Thomas, M. S. C. (2014). Environmental and genetic influences on neurocognitive development: The importance of multiple methodologies and time-dependent intervention. *Clinical Psychological Science, 2*, 628-632. doi: 10.1177/2167702614521188

Langer, J. (1986). *The origins of logic: One to two years.* New York: Academic Press.

Langer, J. (1990). Early cognitive development: Basic funtions. In C.A. Hauert (Ed.), *Developmental Psychology: Cognitive, perceptuo-motor and neuropsychological perspectives* (pp. 19-42). Amsterdam: North Holland.

Massand, E., & Karmiloff-Smith, A. (2015). Cascading genetic and environmental effects on development: implications for intervention. In K. Mitchell (Ed.), *The Genetics of Neurodevelopmental Disorders* (pp. 275-288). Hoboken, U.S.: Wiley-Blackwell.

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology: Research methods in psychology* (Vol. 2, pp. 447-480). Nueva York: Wiley.

Miranda, A., Colomer, C., Mercader, J., Fernández, I., & Presentación, M. J. (2016). Performance-based tests versus behavioral ratings in the assessment of executive functioning in preschoolers: associations with ADHD symptoms and reading achievement. *Frontiers in Psychology, 29.* doi: 10.3389/fpsyg.2015.00545

Mishra, J., Anguera, J. A., Ziegler, D. A., & Gazzaley, A. (2014). A Cognitive Framework for Understanding and Improving Interference Resolution in the Brain. *Progress in Brain Research, 207*, 351-377. doi: 10.1016/B978-0-444-63327-9.00013-8

Mitchell, S. K. (1979). The interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86,* 376-390.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: four general conclusions. *Cur-*

rent Directions in Psychological Science, 21(1), 8-14. doi: 10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex 'Frontal Lobe' tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49-100.

Moriguchi, Y., Chevalier, N., & Zelazo, P. D. (2016). Editorial: Development of Executive Functions during Childhood. *Frontiers in Psychology, 7*(6). doi: 10.3389/fpsyg.2016.00006

Muldoon, K., Lewis, C., & Towse, J. (2005). Because it's there! Why some children count, rather than infer numerical relationships. *Cognitive Development, 20*, 472-491.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*, 542-557.

Nee, D. E., & Jonides, J. (2008). Dissociable interference-control processes in perception and memory. *Psychological Science, 19*(5), 490-500. doi: 10.1111/j.1467-9280.2008.02114.x

Nelson, C. A., & Luciana, M. (2008). *Handbook of Developmental Cognitive Neuroscience* (2nd Ed.). London: The MIT Press.

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin, 126*(2), 220-246.

Perea, A. E., Alday, L., & Castellano, J. (2006). Registro de datos observacionales a partir del MATCH VISION STUDIO v1.0. In J. Castellano, L. M. Sautu, A. Blanco-Villaseñor, A. Hernández Mendo, A. Goñi, & F. Martínez (Eds.), *Socialización y Deporte: Revisión crítica* (pp. 135-152). Vitoria-Gasteiz: Arabako Foru Aldundia-Diputación Foral de Álava.

Petersen, I. T., Hoyniak, C. P., McQuillan, M. E., Bates, J. E., & Staples, A. D. (2016). Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review, 40*, 25-71. doi: 10.1016/j.dr.2016.02.001

Salley, B., Panneton, R. K., & Colombo, J. (2013). Separable Attentional Predictors of Language Outcome. *Infancy, 18*(4), 462-489.

SAS Institute Inc. (2004). *SAS 9.1.3 Help and documentation*. Cary, NC: SAS Institute Inc.

Scharf, R. J., Scharf, G. J., & Stroustrup, A. (2016). Developmental Milestones. *Pediatrics in Review, 37*(1), 25-37. doi:10.1542/pir.376266

Schlotzhauer, S. D., & Littell, R. C. (1997). *SAS system for elementary statistical analysis*. Cary, NC: SAS Institute Inc.

Shanmugan, S., & Satterthwaiter, T. (2015). Neural markers of the development of executive function: relevance for education. *Behavioral Sciences, 10*, 7-13. doi: 10.1016./j.cobeha.2016.04.007

Shaw, C., Brady, L. M., & Davey, C. (2011). *Guidelines for Research with Children and Young People*. London: National Children's Bureau Research Centre.

Sinclair, H., Stambak, M., Lezine, I., Rayna, S., & Verba, M. (1984). *Los bebés y las cosas*. Barcelona: Gedisa.

Stievano, P., & Valeri, G. (2013). Executive functions in early childhood: Interrelations and structural development of inhibition, set-shifting and working memory. *Neuropsychological Trends, 13*(1), 27-45. doi: 10.7358/neur-2013-013-stie

Viterbori, P., Usai, M. C., Traverso, L., & De Franchis, V. (2015). How preschool executive functioning predicts several aspects of math achievement in Grades 1 and 3: longitudinal study. *Journal Experimental of Child Psychology, 140*, 38-55. doi: 10.1016/j.jecp.2015.06.014

Wass, S. V. (2015). Applying cognitive training to target executive functions during early development. *Child Neuropsychology, 21*(2), 150-166. doi: 10.180/09297049.2014.882888

Wertz, F. J. (2014). Qualitative Inquiry in the History of Psychology. *Qualitative Psychology, 1*, 4-16.

Whitebread, D., & Coltman, P. (2010). Aspects of pedagogy supporting metacognition and self-regulation in mathematical learning of young children: evidence from an observational study. *ZDM Mathematics Education, 42*(2), 163-178. doi:10.1007/s11858-009-0233-1

Willoughby, M. T., & Blair, C. B. (2016). Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment, 28*(3), 319-330. doi: 10.1037/pas0000152

Willoughby, M. T., Holochwost, S. J., Blanton, Z. E., & Blair, C. B. (2014). Executive Functions: Formative versus Reflective measurement. *Measurement: Interdisciplinary Research and Perspectives, 12*(3), 69-95. doi: 10.1080/15366367.2014.929.453

Willoughby, M. T., Wirth, R. J., & Blair, C. B. (2011). Contributions of modern measurement theory to measuring executive functions in early childhood: An empirical demostration. *Journal of Experimental Child Psychology, 108*, 414-435. doi: 10.1016/j.jecp.2010.04.007

Willoughby, M. T., Wirth, R. J., Blair, C. B., & Family Life Project Investigators (2016). Executive Function in Early Childhood: Longitudinal Measurement Invariance and Developmental Change. *Psychological Assessment, 28*(3), 319-330. doi: 10.1037/pas0000152

Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, *6*(4), 354-360. doi: 10.1111/j.1750-8606.2012.00246.x

(Article received: 12-10-2016; revised: 29-11-2016; accepted: 28-02-2017)