MASTERARBEIT | MASTER'S THESIS

Titel | Title

The Relationship Between Child Temperament and Scores on the Cognitive Scale of the Bayley-III at 24 Months

verfasst von | submitted by

Noah Brandfellner BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Psych. Dr. Stefanie Höhl

Acknowledgements

First of all, I want to thank my supervisor, Univ.-Prof. Dipl.-Psych. Dr. Stefanie Höhl, for her support and guidance during the creation of this thesis. Furthermore, this could not have been done without the support, kind words, and helpful tips from my co-supervisor, Mag. Alicja Brzozowska, PhD, either. Thank you for taking the time to answer every little question, no matter how trivial it may have been. I am very grateful that I have been able to join this study and work with these two brilliant scientists.

Furthermore, I want to thank Valerie, Regina, and Elisa for all the fun testing sessions we had, for all the early Saturday and Sunday mornings we shared in the lab and the support we gave each other. It was a great time! Also, I want to thank Sophie, Theresa, Mae, Coni, and Valie, for their friendship, love, and support, for being there, when I need them, for all the countless hours spent together – I could not imagine better friends than you. Also, I want to thank Eva, no words would suffice to say how grateful I am to have you in my life.

Last but not least, I would like to thank Doris, for being my mentor and for taking me under her wing. It means so much.

Table of Contents

1 Introduction

Every person is unique. One aspect that majorly contributes to one's individuality is temperament. Temperament influences how we experience the world and interact with our environment, impacting various aspects of our lives (Rothbart, 2011). One of those aspects is cognitive development, which is a complex process that occurs throughout the entire life span (Gauvain, 2022). Cognitive abilities enable us to experience the world in the first place (Sternberg & Sternberg, 2017). Consequently, it follows that temperament and cognition interact with each other, which has also been examined in previous studies (Fagen et al., 1987; Fenstermacher & Saudino, 2016; Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019).

There are two possible explanations for why we may observe correlations between temperament dimensions and cognitive development indices: Actual influences of temperament on cognition and influences of temperament on test-taking behavior (Rothbart & Bates, 2006). Actual influences of temperament on cognition would mean that temperament directly influences the cognitive abilities of a child, hence, children who have certain temperamental characteristics also exhibit higher or lower cognitive abilities (Rothbart & Bates, 2006). On the other hand, influences of temperament on test-taking behavior would mean that temperament solely affects the way a child acts and interacts in a situation where cognitive abilities are tested (Rothbart & Bates, 2006). For instance, a child may be very shy and therefore might struggle to engage in such situations, possibly resulting in lower scores on those tests. Another child might be good at focusing their attention on tasks and games that are presented to them, leading to higher scores in tests concerning cognitive skills. Although it is not possible to discern these two possible influences in correlational studies, knowing more about this relationship can help us understand the nature of temperament-cognition interactions better. Thus, the present study will investigate the relationship between temperament and scores on tests capturing cognitive abilities.

2 Theoretical Background

2.1 Temperament

Temperament research has a long history, dating back to Hippokrates and Galen. Over the course of several centuries, various theories of temperament have been proposed (Eckardt, 2017). In the present day, still numerous theories of temperament coexist. However, according to Strelau (2020), there are some aspects on which most researchers in this field agree:

> Temperament (a) is a component of personality, (b) is present since early infancy, (c) has a biological background, (d) is partially genetically determined, (e) is considerably stable as compared to other personality characteristics, and (f) is present also in animals. (p. 5388)

It is important to note that not all aspects of temperament are regarded as stable in early childhood, however, they gain more stability as the child ages and develops, being considerably stable by preschool age (Shiner, 2015). The stabilization of temperament around preschool age is influenced by the development of the lateral prefrontal cortex as well as the anterior singular cortex, which are both linked to the development of the executive attention system in the brain (Gartstein et al., 2016). This development results in increased attentional control, which in turn increases regulation. Regulation alters the expression of temperament by decreasing the influence of its reactive and emotional aspects, subsequently leading to more stability (Gartstein et al., 2016). Moreover, in the process of development, temperament and its stability are progressively affected by contextual factors, experiences, and genetic influences (Gartstein et al., 2016; Shiner, 2015).

Temperament can be measured in various ways, such as self- and other-report measures and questionnaires but also through observation, behavioral tasks, and peer nominations (Shiner, 2015). Behavioral tasks are settings that are specifically constructed to elicit certain reactions or behaviors (Shiner, 2015). Peer nominations require that participants rate or assess their peers in regard to certain characteristics (Del Vecchio, 2011). The most common measure for children are other-report questionnaires, especially parent-report questionnaires (Shiner, 2015).

2.1.1 Temperament vs. Personality

Temperament and personality are related, yet they are not identical. As already stated by Strelau (2020), temperament is seen as a part of personality, sometimes temperament is also considered to be the core of personality (Chen & Schmidt, 2015; Rothbart, 2011). Therefore, they share some similarities, such as that they both describe interindividual differences which are consistent over different situations and time, entail reactivity and modulation of that reactivity, are impacted by biological influences and interplay with the environment (Chen & Schmidt, 2015). Nevertheless, they also exhibit some differences. One of the most distinguishing aspects between these two constructs is that temperament is understood to become apparent early in life, while personality traits develop later on and become more apparent over the course of childhood, partly being shaped by the experiences a

child has (Chen & Schmidt, 2015). As a person ages, more personality traits emerge, which have a foundation in more advanced social and cognitive abilities, coping mechanisms, and values (Chen & Schmidt, 2015). In general, personality traits include one's way of thinking and coping mechanisms, attitudes, self-perception and emotions related to self, like pride or shame, beliefs, perception of others and the environment, values, and morals. On the other hand, temperament entails "emotional, motor, and attentional tendencies and regulative capacities" (Rothbart, 2011, p. 12). Furthermore, temperament is considered to have stronger biological foundations than personality and is associated with broader emotional and behavioral processes (Chen & Schmidt, 2015).

2.1.2 Psychobiological Model of Temperament

This thesis assumes the psychobiological model of temperament (Rothbart & Derryberry, 1981) as a theoretical framework which can be applied to humans across the life span, from infancy to adulthood (Gartstein et al., 2016). The authors of the model defined temperament as "constitutional differences in reactivity and self-regulation" (p. 37), which are consistent across situations and stable over time (Rothbart & Derryberry, 1981). "Constitutional" refers to temperament being based on biological factors, such as genes and neurobehavioral mechanisms. "Reactivity" can concern an individual's attention, emotions, motor activity, behavior, and physiological responses and describes the way an individual reacts to changes in their surroundings, for instance whether a person is inclined to move towards or away from a certain stimulus, whether they are easily angered, or how their heart rate changes in response to a certain trigger (Gartstein et al., 2016; Rothbart, 2011; Rothbart & Derryberry, 1981). Thus, the concept of reactivity implies that those aspects of temperament are not omnipresent, rather, they have to be elicited (Rothbart & Bates, 2006). Rothbart and Bates (2007) give the example that children who have a low frustration tolerance are not frustrated or angry persistently, only in situations in which they cannot achieve their goals or when their expectations are not fulfilled (Rothbart & Bates, 2006).

The elicited reactions can be evaluated by examining their threshold, latency, rise time, intensity, and recovery (Rothbart & Derryberry, 1981). Threshold means the sensitivity to stimulation, latency describes the time between the stimulus and the response, intensity reflects how strong a response is, rise time describes how long it takes to until the response reaches maximum intensity, and recovery states the time an individual takes to recover from the reaction (Rothbart & Derryberry, 1981).

Another component of the psychobiological model is “self-regulation” (Rothbart & Derryberry, 1981). The term “self-regulation” reflects the ability to regulate reactivity, either by increasing or decreasing it, through processes like inhibition, approach and avoidance, attention as well as self-stimulation and self-soothing (Rothbart, 2011; Rothbart & Derryberry, 1981; Strelau, 2020).

Based on the psychobiological model, Rothbart and colleagues have constructed several questionnaires to capture temperament, such as the Infant Behavior Questionnaire (IBQ; Gartstein & Rothbart, 2003), the Early Childhood Behavior Questionnaire (ECBQ; Putnam et al., 2006), or the Children’s Behavior Questionnaire (CBQ; Rothbart et al., 2001). The questionnaires were designed to measure temperament at different ages: The IBQ is suited for use with children aged 3 to 12 months (Gartstein & Rothbart, 2003), the ECBQ for children aged 1.5 years to 3 years (Putnam et al., 2006), and the CBQ for children aged 3 to 7 years of age (Rothbart et al., 2001). The reason for this distinction is that temperament manifests differently in behavior depending on the age, and the circumstances that trigger different behaviors change as a person develops as well (Goldsmith & Rothbart, 1991).

The questionnaires consist of 14 to 20 dimensions, depending on the specific questionnaire (Gartstein et al., 2016). Factor analyses have identified three to four temperament factors: Negative Affectivity (sometimes also referred to as Negative Emotionality), Surgency, Effortful Control, and Orienting Sensitivity, with Orienting Sensitivity being found only in adults (Gartstein et al., 2016).

2.1.2.1 Negative Affectivity

The dimension of Negative Affectivity describes inclinations to perceive and express feelings like anxiety, anger, or physical discomfort (Gartstein et al., 2016). Negative Affectivity is relatively stable over time, however, intrapersonal variations can occur, as it is subject to external influences, like stress (Paulus & Zvolensky, 2020).

2.1.2.2 Surgency

Surgency reflects a person’s tendencies to enjoy strong stimuli, approach new experiences, or display laughter (Gartstein et al., 2016). Surgency is regarded to be considerably stable over time (Rothbart, 2011). It becomes apparent when infants are 2-3 months old, as exhibited through behaviors such as smiling and laughing or general activity (Rothbart, 2011). During development, the expression of surgency changes, with sociability and impulsivity emerging in childhood (Holmboe, 2016). Furthermore, in children as well as in adults, high levels of Surgency are related to lower inhibitory control (Holmboe, 2016).

2.1.2.3 Effortful Control

Effortful Control describes abilities related to attention, taking pleasure in calm activities, inhibitory control, and perceptual sensitivity (Gartstein et al., 2016; Rothbart, 2011). Effortful Control also involves being able to manage emotions, suppressing dominant responses and subsequently being in charge of one's actions, as well as shifting and focusing attention deliberately (Rothbart, 2011).

Effortful Control emerges between the ages of 6 and 12 months and develops fast between 2 and 7 years of age. It is assumed that this is paralleled by the development of the brain's executive attention system (Kochanska et al., 2000; Rothbart, 2011). Over the course of infancy, toddlerhood, and preschool years, infants' ability to intentionally focus their attention increases, with the ability to focus attention more voluntarily improving between 9 and 18 months of age (Eisenberg et al., 2016). Moreover, inhibition skills continue to progress from toddler to preschool age (Eisenberg et al., 2016). In general, Effortful Control is considered to be quite stable, with multiple studies examining its stability (Kochanska et al., 2000; Kochanska & Knaack, 2003; Li-Grining, 2007). Studies found that Effortful Control at 22 months predicted Effortful Control at 33 and 45 months and that attention focusing early in life predicted Effortful Control later in life. Moreover, a stability of 4 to 6 years was observed for certain parts of Effortful Control during early school age (Eisenberg et al., 2016).

2.1.2.4 Orienting Sensitivity

Orienting Sensitivity, the temperamental dimension only found in adults, encompasses the ability to notice small details from internal or external sources (perceptual sensitivity), the frequency of automatic connections between different cognitive concepts (associative sensitivity), and being conscious of feelings triggered by weak stimuli (affective perceptual sensitivity; Evans & Rothbart, 2007).

3 Cognitive Development

Cognitive development is a complex and lifelong process. Over the course of development, various cognitive functions continue to develop, such as language skills, attention, working memory, cognitive control, or processing speed (Casey et al., 2005; Luna et al., 2004).

Multiple factors contribute to cognitive development, internal as well as external (Gauvain, 2022). Internal factors are, for instance, biological factors, such as temperament, fundamental sensory capabilities enabling a person to perceive information in their environment, the ability to regulate one's emotions, or the ability to learn. Also belonging to

the internal factors are psychological factors, like previous learning experiences, the way one expresses themselves, and areas of interest (Gauvain, 2022). Considering external factors, various social, cultural, physical, and material elements play a part in influencing cognitive development, such as interactions with other humans or cultural influences (Gauvain, 2022). Children play an active role in their development as they interact with their environment and look for information that appeals to and engages them. Thus, engagement opportunities need to be available to the child (Gauvain, 2022). Developmental tests, such as the Bayley Scales of Infant and Toddler Development III, are based on the idea that children of a certain age should possess certain skills, with these skills advancing as the children develop (Bayley et al., 2014b). However, it is not only a pre-programmed neurobehavioral maturation that is the cause of progress in cognitive development, but rather a result of an interplay of multiple factors, such as possibilities offered to the child (Gauvain, 2022).

3.1 Cognitive Development and Temperament

It becomes apparent that various factors are at play when it comes to a child's cognitive development. Since temperament is considered to contribute as well, a literature search for studies that have examined the relationship between cognitive capacities and temperament was conducted. The following studies were the ones I considered to be most relevant to the topic of the current thesis, ordered by increasing relevance. To the best of my knowledge, no meta-analyses concerning the relationship between cognitive abilities and temperament in children exist.

Smith-Watts and colleagues (2019) examined, among other variables, the relationship between different temperament dimensions and general intelligence at multiple testing points in 412 twin pairs aged 14, 20, 24, and 36 months in the USA. They assessed concepts related to self-restraint as well as self-restraint by itself. The concepts of self-restraint and attentional control are related to the factor of Effortful Control in the psychobiological model of temperament (Rothbart, 2011; Rothbart & Derryberry, 1981). Furthermore, they assessed negative emotionality. Negative emotionality is associated with the dimension of Negative Affectivity in the psychobiological model (Gartstein et al., 2016; Rothbart & Derryberry, 1981). In addition, IQ was measured using a Stanford-Binet test (Smith-Watts et al., 2019). Parent-reported self-restraint was measured using certain items from the Colorado Childhood Temperament Inventory (Rowe & Plomin, 1977), the Toddler Temperament Scale (Fullard et al., 1984), and the Differential Emotions Scale (Izard et al., 1980), whereas observed self-restraint was assessed through a prohibition task in which the children were not allowed to touch an attractive toy. Results showed that IQ was significantly correlated with observed and

parent-reported self-restraint in both boys ($r_{observed}$ = .5, $r_{reported}$ = .53) and girls ($r_{observed}$ = .46, $r_{reported}$ = .47). Furthermore, IQ had a significant positive association with observed attentional control in girls (r = .6) as well as a significant positive correlation with both observed (r = .65) and parent-reported (r = .24) attentional control in boys. Additionally, it was found that IQ was negatively associated with observed negative emotionality (measured with a coding scheme developed by Smith-Watts et al. (2019)) in both girls (r = -.34) and boys (r = -.24). Moreover, observed negative emotionality was significantly negatively related to self-restraint for boys (r = .33) and girls (r = .41; Smith-Watts et al., 2019). Concluding, the study by Smith-Watts and colleagues (2019) yielded a broad range of results concerning various measures related to self-restraint, negative emotionality, and IQ, indicating that those dimensions are related to each other and worth investigating further. However, it is hard to discern whether those results reflect actual differences in IQ, implying that children with higher self-restraint also have a higher IQ, whether self-restraint simply influences test-taking behavior in a positive way, so that these tests result in higher scores, or whether it is a mixture of these two.

Moffitt et al. (2011) conducted a longitudinal study over the course of 32 years with a sample of 1037 children in New Zealand. The authors assessed a wide range of variables, with one of them being the association between self-control and IQ in childhood. They used the term “self-control” as an overarching concept including various constructs from multiple scientific branches, such as the concept of self-regulation or of executive function (Moffitt et al., 2011). Self-control was assessed at 3, 5, 7, 9, and 11 years of age through multiple measures: Observations as well as parent-, teacher- and self-reports. IQ was assessed at ages 7, 9, and 11 years, using the Wechsler Intelligence Scale for Children – Revised (Wechsler, 1974). A positive association between the concept of self-control and intelligence was found (r = .44), supporting the notion that children who exhibit higher levels of self-control also score higher on scales capturing intelligence and thus, cognitive performance (Moffitt et al., 2011).

A study conducted by Fenstermacher and Saudino (2016) examined the connection between development, temperament, and imitation in 311 24-month-old healthy twin pairs in the USA. Imitation is sometimes used an indicator of a young child’s cognitive development (Fenstermacher & Saudino, 2016), however, Fenstermacher & Saudino (2016) also used the Mental Development Index (MDI) of the English version of the Bayley Scales of Infant Development II (BSID-II; Bayley, 1994). The Mental Development Index score is comprised of items targeting cognitive and language development (Månsson et al., 2019). It is not quite

clear why they used the BSID-II, as the Bayley-III, which is an advancement of the BSID-II, had already been available for some years at the time of the study (Bayley, 2006a). Temperament was assessed via the Infant Behavior Record (Bayley, 1994), which is an observer-rated measure of temperament with three dimensions: Task Orientation, Activity, and Affect/Extraversion. Task Orientation concerns the children's attention, goal orientation, and persistence, thus capturing similar aspects to the dimension of Effortful Control in the psychobiological model (Fenstermacher & Saudino, 2016; Rothbart, 2011). Activity measures a child's energy and movement, whereas Affect/Extraversion captures cooperation, social responsiveness, and emotional tone (Fenstermacher & Saudino, 2016). Both dimensions partly assess aspects similar to the dimensions of Surgency as well as Negative Affectivity in the psychobiological model of temperament (Fenstermacher & Saudino, 2016; Gartstein et al., 2016; Rothbart, 2011). The study found that children who scored higher in Affect/Extraversion also exhibited higher levels of imitation (r = .33). Imitation, in turn, was associated with higher MDI scores (r = .39). The authors did not examine the relationship between Affect/Extraversion and MDI scores directly (Fenstermacher & Saudino, 2016). Concluding, the results suggest that children who are more outgoing score higher on cognitive tasks such as those used in the Bayley Scales, indicating a relationship between temperament and performance in the Bayley scales.

Furthermore, studies (e.g. Fagen et al., 1987; Sajaniemi et al., 1997) have specifically assessed the relationship between temperament and scores on the Bayley Scales. One of these studies was conducted by Sajaniemi and colleagues (1998). They compared 80 preterm born children (born between 23 and 34 weeks of pregnancy) and 80 full-term born children at 24 months of age in Finland. As a measure of temperament, the Toddler Temperament Scale (Fullard et al., 1984) was utilized, which is a parent-report questionnaire that captures temperament on nine dimensions: activity, adaptability, approach, distractibility, intensity, mood, persistence, rhythmicity, and threshold (Fullard et al., 1984). These dimensions relate to the psychobiological model (Rothbart & Derryberry, 1981) in various ways: Activity, approach, and mood are aspects associated with the dimension of Surgency in the psychobiological model (Gartstein et al., 2016). Additionally, mood also relates to the dimension of Negative Affectivity (Gartstein et al., 2016). Furthermore, distractibility and persistence play a role in the factor of Effortful Control (Gartstein et al., 2016; Rothbart, 2011). The dimensions of adaptability, intensity, and threshold of the Toddler Temperament Scale are not directly considered dimensions of temperament in the psychobiological model, however, they serve as important factors for evaluating elicited reactivity, as temperament was

defined as “constitutional differences in reactivity and self-regulation” (Rothbart & Derryberry, 1981, p. 37). Cognitive skills were assessed by using the MDI of the BSID (Bayley, 1969). Findings concerning the correlations between the BSID (Bayley, 1969) and the Toddler Temperament Scale (Fullard et al., 1984) included a weak negative correlation between Bayley scores and rhythmicity (r = -.24; Sajaniemi et al., 1998). Rhythmicity in this context refers to the rhythmicity of bodily functions, such as sleep and hunger as well as bowel movements (Fullard et al., 1984; Rothbart, 1981). Moreover, results showed again weak but significant negative relationships between Bayley scores and positive mood (r = -.21) as well as between Bayley scores and response threshold (r = -.29). Finally, a significant positive relationship between Bayley scores and persistence was also present (r = .33; Sajaniemi et al., 1998). Similar to the findings by Fagen et al. (1987), the negative correlation between MDI scores and positive affect seems counterintuitive, whereas the relationship between MDI scores and persistence could be explained by the possibility that toddlers who struggle to engage with a task for longer times may perform worse than children who persist despite potential obstacles. In addition, it may make sense that children with a high response threshold score lower on the BSID, since a high response threshold could lead to delayed or missing responsiveness to items, resulting in lower scores and a possible underestimation of their abilities.

Another study that has assessed the relationship between temperament and scores on the Bayley Scales was conducted by Fagen and colleagues (1987). They examined the relationship between infant temperament, measured with the Infant Behavior Questionnaire (Rothbart, 1981) and performance on the Bayley Scales of Infant Development (BSID; Bayley, 1969) at 4, 8, and 12 months in 42 infants born healthy and full-term. The study took place in the USA. The temperament dimensions used were activity level, smiling and laughter, duration of orienting, and fear. Activity level captures an infant’s gross motor activity, whereas smiling and laughter assesses the amount of positive affect exhibited by the child (Rothbart, 1981). Both relate to the factor of Surgency in the psychobiological model of temperament, as Surgency, among other things, reflects a person’s tendencies to display laughter (Gartstein et al., 2016). Furthermore, activity level has been categorized as being a component of Surgency in later publications, such as in the Early Childhood Behavior Questionnaire (Putnam et al., 2006). Duration of orienting describes the child focusing on a single object for a longer time period (Rothbart, 1981). As this ability is related to attention, it relates to the dimension of Effortful Control in the psychobiological model (Gartstein et al., 2016). The dimension of fear assesses an infant’s distress or hesitation to approach in

response to new stimuli or stimuli of high intensity (Rothbart, 1981). This relates to the dimension of Negative Affectivity in the psychobiological model (Gartstein et al., 2016). They excluded the dimensions distress to limitations and soothability, arguing that they were not relevant to performance on the Bayley Scales (Fagen et al., 1987). However, it seems possible that especially the dimension of distress to limitations may play a role in test taking behavior as well, as infants may become frustrated when faced with a task in the Bayley Scales that they cannot complete. Nonetheless, Fagen et al. (1987) found a significant positive association between activity level and MDI scores at 4 months (r = .26). Additionally, at 8 months, a significant negative association between smiling and laughter with MDI scores (r = -.26) was found. Finally, at 12 months, they found significant positive associations between MDI scores and both activity level (r = .32) and fear (r = .29) and a significant negative association between MDI scores and smiling/laughter (r = -.29; Fagen et al., 1987). The positive association between activity level and MDI scores at 4 months can easily be explained, since this could imply that children participated more actively and therefore were able to receive higher scores. However, the negative correlation between smiling and laughter and MDI scores at 8 months as well as 12 months seems counterintuitive, as one would assume that more positive affect would also influence test taking behavior, and therefore scores, positively. However, the authors of the study (Fagen et al., 1987) commented that based on the results of prior studies (Field et al., 1978; Peters-Martin & Wachs, 1984; Roth et al., 1984; Worobey, 1987), who either found a positive, a negative, or no relationship between various measures of mood and BSID scores, it may be that the true correlation between mood and BSID scores is zero (Fagen et al., 1987). Similarly, the positive correlation between MDI scores and fear at 12 months found by Fagen et al. (1987) is unexpected, since fear hindering performance would seem logical. Nevertheless, these results indicate that temperament dimensions may interact with performance in the Bayley Scales.

4 Present Study

Previous research has suggested that an association between temperament and cognitive abilities exists. Moreover, research indicated that temperament is also associated with performance in tests that aim to capture cognitive abilities, such as the Bayley Scales (Fagen et al., 1987; Fenstermacher & Saudino, 2016; Moffitt et al., 2011; Smith-Watts et al., 2019). This observation is relevant as it indicates that scores on tests assessing cognitive skills may not solely reflect a child's cognitive development but may be influenced by aspects beyond that, such as temperamental features like shyness or negative emotionality. Thus, the

relationship between temperament and scores on a standardized measure of cognitive development warrants further exploration. Prior studies either used an older version of the Bayley scales (either BSID or BSID-II) and along with that, the Mental Development Index score, used a different measure of cognitive abilities altogether, had no German speaking samples, tested children younger than 24 months, or preterm children (Fagen et al., 1987; Fenstermacher & Saudino, 2016; Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019). Hence, to the best of my knowledge, the relationship between temperament dimensions of full-term children at 24 months and cognitive scores of the Bayley Scales III in a German speaking sample has not been investigated yet. Consequently, in this thesis, the focus lies on the relationship between temperament and scores on a standardized test of development in healthy, full-term children at 24 months of age.

To this aim, I am going to examine the relation between scores on the Early Childhood Behavior Questionnaire (Putnam et al., 2006), a measure of temperament grounded in the psychobiological model, and the cognitive scale of the Bayley Scales of Infant and Toddler Development III (Bayley et al., 2014b). Using the Bayley-III enables me to target cognitive abilities more effectively, on the grounds that cognitive and language skills are assessed separately in the Bayley-III, compared to the previous versions where these two abilities were united in the Mental Development Index score (Bayley et al., 2014b; Månsson et al., 2019).

4.1 Hypotheses

The hypotheses are structured according to the dimensions of the psychobiological model of temperament (Effortful Control, Negative Affectivity, Surgency), as the Early Childhood Behavior Questionnaire, one of the measures used in this study, reflects its structure (Putnam et al., 2006; Rothbart & Derryberry, 1981). In general, I hypothesize that temperament affects test-taking behavior and, therefore, cognitive scores. However, as previously noted, differentiating between whether temperament is related to cognitive scores because it directly affects cognitive abilities or whether temperament is related to cognitive scores because it affects test-taking behavior, is not possible. My hypotheses are as follows:

H1: Children with higher levels of Effortful Control also show higher levels of cognitive development (in a standardized test of development).

As the concept of Effortful Control is associated with abilities related to attention and inhibitory control (Kirchhoff, 2015; Putnam et al., 2010), it may be possible that children who exhibit higher levels of Effortful Control can focus better on the task and consequently reach higher scores in said task.

Furthermore, multiple studies, such as those conducted by Smith-Watts et al. (2019) and Moffitt et al. (2011) found that constructs related to self-restraint, which is a concept associated with Effortful Control, were positively connected with cognitive performance.

H2: Children with higher levels of Negative Affectivity show lower levels of cognitive development (in a standardized test of development).

Negative Affectivity, entailing factors like discomfort, frustration, or fear (Kirchhoff, 2015; Putnam et al., 2010), may prevent a child from fully engaging in tests targeting cognitive skills, thus possibly resulting in lower scores.

Studies examining this relationship led to mixed results. In the study conducted by Smith-Watts and colleagues (2019), IQ was negatively associated with observed negative emotionality, whereas Fagen et al. (1987) found that smiling and laughter were negatively correlated with MDI scores. Sajaniemi and colleagues (1998) came to a similar conclusion, finding a significant correlation between low MDI scores and positive mood.

H3: Children with higher levels of Surgency show higher levels of cognitive development (in a standardized test of development).

The dimension of Surgency involves factors such as sociability and positive anticipation (Kirchhoff, 2015; Putnam et al., 2010), hence, it may be possible that children who are more extroverted are more engaged during developmental tests and are less shy, which in turn could lead to higher scores. Furthermore, Fenstermacher & Saudino (2016) found an indirect association between extroverted behavior and MDI scores.

5 Methods

The present research was part of the longitudinal study “Neural entrainment and dynamic attention in infants” of the Vienna Children Studies (“Wiener Kinderstudien”) at the Department of Developmental and Educational Psychology at the Faculty of Psychology of the University of Vienna. It consisted of three testing points in total: at 6 months, at 12 months, and at 24 months of age. However, the current thesis solely focuses on the third testing point. The study was supervised by Univ.-Prof. Dipl.-Psych. Dr. Stefanie Höhl and Mag. Alicja Brzozowska, PhD.

5.1 Sample

The total sample at the third testing point was comprised of 122 children, however, due to the timeline of the present thesis, the current sample only consists of 73 children, 35

boys and 38 girls. The children were recruited from the Vienna Children Studies database, a registry where caregivers can voluntarily enroll their children for potential participation in studies. When meeting the inclusion criteria for a certain study, caregivers were contacted by student assistants, receiving information about the study and being invited to partake, either via email or via phone call. Inclusion criteria for the present study were that the children were born full term, being aged 6 months at the first testing point, and having no known neurodevelopmental conditions.

Similar procedure was followed for the subsequent testing points. Parents whose children had already participated in the first testing point were contacted and asked whether they would still be available and open to participate. If they consented, a testing date was scheduled. The initial sample enrolled at the first testing point consisted of 140 infants, of which 13 % dropped out at the third testing point. The final sample for the present thesis consisted primarily of higher-income families. Data about the caregiver's income was available for 64 of the 73 children, with 52 families (81.2 %) stating a yearly income of more than 45 000 €. 10 families (15.6 %) reported an income between 26 000 € and 45 000 €, whereas only two families (3.1 %) reported an income below 26 000 €.

Following participation, caregivers received a compensation of 6 € for travel expenses. In addition, they were gifted a participation certificate with a photo of them and their child taken before/after the testing, a bag showcasing a logo of the Vienna Children Studies as well as either a voucher of 15 € for a drug store or a book about child development (dependent on their choice). The toddlers were offered various toys from which they could select one.

5.2 Research Design

Data collection for the third testing point began in October 2023 and ended in September 2024. It was carried out by two researchers per session. The testing sessions took place at the Faculty of Psychology of the University of Vienna and lasted approximately 2.5 hours. During the testing session, the child was accompanied by their caregiver(s). At the beginning of each session, the child and their caregiver(s) arrived, were offered drinks and were given time to settle in and get comfortable. Toys were offered to the child and the experimenters initiated interactions with the child in an attempt to engage as well as decrease potential shyness and wariness. The caregiver(s) received several forms to fill in, such as informed consent, consent regarding the use of video footage, a form enquiring socioeconomic information, and a questionnaire targeting the language development of the child ("ELFRA"; Grimm et al., 2019). The parents were instructed to fill in the ELFRA

(Grimm et al., 2019) during the testing. Additionally, they were informed that they would receive another questionnaire, the Early Childhood Behavior Questionnaire (Putnam et al., 2006), via email after the testing, including a code to access it.

Following the initial warming-up phase and the caregivers' consent, video cameras were installed and turned on, and the first two tests were carried out: A drumming task as well as a visual search task. The drumming task captured sensorimotor synchronization skills and the visual search task measured selective attention. However, as those tasks are not relevant to the present thesis, they will not be discussed further. Between all tasks and when necessary, short breaks were offered including small snacks and drinks.

After the first two tasks, the Bayley-III cognitive and gross motor scale were administered (Bayley et al., 2014a; Bayley et al., 2014b). However, the current thesis only uses data obtained from the cognitive scale. During the first few testing sessions, the administration of the Bayley-III was filmed as well, however, it soon proved to be impractical and unnecessary and was discontinued subsequently. For the cognitive scale, the child was sat at a children-sized table across from the two researchers. Optionally, if necessary and desired, the caregiver could sit next to the child, either on the floor or with a chair. The caregiver was instructed not to interfere with the administration of the test and not to give clues or support in solving the items. Going forward, one researcher presented the items, whereas the other one assisted, measuring the time the child took to complete the timed items and taking notes. Additionally, they took care of storing the material after items had been completed. The gross motor scale was administered standing and moving, given that movement was essential for the items of that scale. The administration of the Bayley-III came to an end either when the stopping criterion was reached or when the child refused to cooperate, usually due to fussiness based on factors such as tiredness.

5.3 Bayley Scales of Infant and Toddler Development III (Bayley-III)

The Bayley Scales of Infant and Toddler Development III (also called Bayley-III; Bayley, 2006) are a test developed by Nancy Bayley to assess a child's developmental status between the ages of 1 and 42 months, with some norms even allowing for testing infants as young as 16 days (Albers & Grieve, 2007). They are regarded as the gold standard assessment for assessing babies and toddlers (Armstrong & Agazzi, 2010). The Bayley Scales are not based on one singular developmental theory, rather, they are influenced by the general state of research in developmental psychology (Macha & Petermann, 2015).

The Bayley Scales-III are an advancement of the Bayley Scales of Infant Development II (BSID-II) (Albers & Grieve, 2007). From the BSID-II to the Bayley-III, relevant changes were made. The most notable one is that the Bayley-III scales examine language and cognitive development independently in two scales, whereas earlier versions of the Bayley used the Mental Development Index score, which was comprised of both language and cognitive skills (Albers & Grieve, 2007; Månsson, 2019). This is to disentangle language skills from cognitive skills, as most items on the cognitive scale of the Bayley-III now do not require a verbal response (Armstrong & Agazzi, 2010). However, it is important to note that the child still needs to understand the task, so a certain level of receptive language skills is necessary and does not make the cognitive scale completely independent from language skills. Moreover, expressive language skills are required to a certain degree as well, for instance when the child is asked to count toy blocks (Albers & Grieve, 2007; Bayley et al., 2014b). Further changes from the BSID-II to the Bayley-III are updated normative data, improved psychometric properties, easier test administration, and revised materials (Albers & Grieve, 2007). A novelty of the German version of the Bayley-III is that normative data for a German population was created, using a representative sample of 1009 children (Schultheiss, 2015).

Still, the Bayley-III is not the most recent version as the Bayley Scales of Infant and Toddler Development 4 (Bayley-4) was published in 2019 with further revisions (Bayley et al., 2019). However, this version is currently not yet available in German.

5.3.1 Scale Structure and Psychometric Properties

The English version of the Bayley-III consists of five subscales: the cognitive scale, the language scale, the motor scale, the social-emotional scale, and the adaptive behavior scale, with the latter two being questionnaires that are filled in by the child's caregiver (Albers & Grieve, 2007). The German scale only consists of three scales: the cognitive, language, and motor scale (Bayley et al., 2014b). In both languages, the language scale is divided into receptive and expressive language skills and the motor scale is divided into fine and gross motor skills. Thus, the total number of items across the cognitive, language, and motor scales and across all ages is 324 for the German version and 326 for the English version (Albers & Grieve, 2007; Schultheiss, 2015). These numbers differ because two items from the expressive language scale had to be removed as a translation to German was not possible. Apart from that, all other items in their respective sequence, the starting points, as well as rules regarding the administration remained the same as in the English version (Bayley et al., 2014b).

5.3.2 Normative Data

The Bayley Scales of Infant and Toddler Development III use various values to describe a child's performance and, more importantly, to interpret it by comparing their score to other children: Raw scores, scaled scores, composite scores and composite score equivalents, percentile ranks, confidence intervals, developmental age equivalents, and growth scores (Bayley et al., 2014a). However, for the current thesis, only raw scores and composite scores are of relevance. Consequently, only these two will be discussed in more detail.

5.3.2.1 Raw Scores

The most basic scores are the raw scores, those are obtained in two steps: First, by simply adding up the items the child successfully solved. Second, by adding to that the number of all items before the child's starting point. However, the raw score does not have any interpretative value yet (Bayley et al., 2014a).

5.3.2.2 Composite Scores and Composite Score Equivalents

Composite scores and composite score equivalents are obtained by transforming scaled scores, which are calculated by transforming raw scores, using tables in the administration manual (Bayley et al., 2014a). They can be utilized to compare a child's performance across different scales. In addition, comparison with other children is possible as well (Bayley et al., 2014a). The scaled scores of the language and the motor scale each consist of their two respective subscales, i.e. the receptive and expressive language scales as well as the gross motor and fine motor scales, respectively (Bayley et al., 2014a). Since the cognitive scale does not have any subscales, a composite score equivalent is calculated. The composite scores and the composite score equivalents span from 40 to 160, with a mean of 100 and a standard deviation of 15 (Bayley et al., 2014a). See *Table 1* for information about the interpretation (Bayley et al., 2014b). It is important to note that this table specifies the interpretation of the German version of the Bayley-III, as the English version differs in the interpretation of the scores (Bayley, 2006b; Bayley et al., 2014b).

Table 1: Interpretation of Composite Scores and Composite Score Equivalents

Score	Interpretation
≥ 145	Extremely above average
130 – 144	Considerably above average
115 – 129	Above Average
85 – 114	Average

70 – 84	Below average
55 – 69	Considerably below average
≤ 54	Extremely below average

5.3.3 Reliability

The internal consistency of the German Bayley-III sample was evaluated using split-half reliability (Bayley et al., 2014b). The mean reliability of the subscales spans from .77 (receptive language skills) to .89 (gross motor skills), indicating medium to high internal consistency (Bayley et al., 2014b; Moosbrugger & Kelava, 2020). Retest reliability was examined using a sample of 54 premature children, which were tested at 7 and 24 months of age. The retest reliability coefficients span from .20 (expressive language skills) to .46 (gross motor skills), exhibiting small to medium retest reliability (Bayley et al., 2014b; Moosbrugger & Kelava, 2020).

5.3.4 Validity

5.3.4.1 Construct Validity

In order to assess construct validity, intercorrelations between subtests and scales were calculated. For correlations between subtests, the values ranged from .22 (correlation between gross motor scale and receptive language scale) to .41 (correlation between receptive and expressive language scales; Bayley et al., 2014b). Concerning the correlations between the scales, correlations were between .36 (correlation between motor scale and language scale) and .43 (correlation between motor scale and cognitive scale; Bayley et al., 2014b). Overall, the highest correlation with .69 was found between the language scale and the expressive language subtest, the lowest with .22 between the gross motor scale and the receptive language scale (Bayley et al., 2014b).

5.3.4.2 Criterion Validity

Criterion validity was determined by examining the correlation between the subscales and external measures in a sample of 54 preterm children. For cognitive performance, the German version of the parent-report questionnaire “Parent-based Assessment of the Cognitive Abilities” (PARCA; Reuner et al., 2010) was utilized. It consists of items asking about visual-spatial skills, play behaviors, and pre-linguistic skills. Moreover, language abilities were assessed using a short questionnaire (SBE-2-KT; Suchodoletz, 2008), in which parents report the child’s usage of certain words and phrases (Bayley et al., 2014b).

Correlations of the scales and subtests with the PARCA ranged from .18 (subtest receptive language skills) to .33 (cognitive scale), with the latter one being significant (Bayley

et al., 2014b). Although only exhibiting a moderate correlation, the finding that the PARCA displays the highest correlation with the cognitive scale makes sense and supports the validity of that scale. Furthermore, correlations of the Bayley scales and subtests with the SBE-2-KT were higher, spanning from .24 (subtest gross motor skills) to .69 (subtest expressive language skills), with all correlations between the Bayley and the SBE-2-KT being significant (Bayley et al., 2014b). Thus, in general, the results support the validity of the Bayley-III, as the highest correlations between the questionnaires and the Bayley scales were found in the subscales designed to measure those specific skills (Bayley et al., 2014b).

5.3.5 Factorial Structure

The factorial structure of the Bayley-III in the German sample was assessed by conducting confirmatory factor analyses. Models with one, two, and three factors were compared across two age groups: 16 days to 6 months 22 days and 6 months 23 days to 43 months 15 days (Bayley et al., 2014b). Results indicated that in both age groups as well as in the total sample, the model with three factors (consisting of the motor scale, the language scale, and the cognitive scale) was the most fitting (Bayley et al., 2014b).

5.3.6 Bayley-III Cognitive Scale

The cognitive scale is, like all scales of the Bayley-III, not based on a singular theory. Rather, it measures a broad range of cognitive abilities that are considered relevant to a child's development, such as preference for novel stimuli, memory, attentional focusing, habituation, exploration, object manipulation, processing and processing speed, problem-solving, and conceptual reasoning (Armstrong & Agazzi, 2010; Schultheiss, 2015). In addition, during the construction of the items, the vital role of play in children's development was taken into account and items targeting various forms of play, such as imaginative play or pretend play, were added (Armstrong & Agazzi, 2010; Bayley et al., 2014b).

When evaluating the cognitive scale, it is important to consider its psychometric properties to understand its effectiveness in assessing cognitive abilities. Analogous to the psychometric properties of the whole scale, split-half reliability, test-retest reliability, construct validity, and criterion validity have been examined (Bayley et al., 2014b).

The split-half reliability of the scale is, across all age groups, .82, indicating good internal consistency. Additionally, test-retest reliability between 7 and 24 months was found to be .23, showing a small correlation (Bayley et al., 2014b; Moosbrugger & Kelava, 2020). Construct validity was examined calculating correlations between the cognitive scale and the other scales and subtests, which ranged from .31 (gross motor skills) to .43 (motor scale),

demonstrating small to medium relationships (Bayley et al., 2014b; Moosbrugger & Kelava, 2020). This implies that the various scales measure different constructs, which is desirable. Finally, criterion validity was established by evaluating the correlation between the cognitive scale and two questionnaires, PARCA (Reuner et al., 2010) and SBE-2-KT (Suchodoletz, 2008). The correlation between PARCA and the cognitive scale was .33 and between SBE-2-KT and the cognitive scale .45, with both correlations being significant (Bayley et al., 2014b; Moosbrugger & Kelava, 2020). It is important to note that the rather low correlations between the PARCA and the cognitive scale may partly stem from the different format of the tests, on the grounds that the Bayley is a developmental test, whereas the PARCA is an other-report questionnaire. In general, the psychometric properties of the cognitive scale are satisfactory (Bayley et al., 2014b; Moosbrugger & Kelava, 2020).

5.3.7 Test Administration

When administering the Bayley-III, different starting points exist, depending on the age of the child – the older a child is, the more advanced the items become. The starting points are assigned letters, which range from A (16 days to 1 month 15 days) to Q (39 months to 42 months 15 days) and always cover an age range. For instance, when a child is 24 months old, the correct starting point would be “M”. It is also possible to correct the age for children that were born preterm (Albers & Grieve, 2007; Bayley et al., 2014b).

If a child successfully completes an item, they receive a score of one point, whereas incorrect items are scored with zero. The child needs to pass the first three items of the scale, otherwise the test instructor has to revert to the prior starting point. Furthermore, if the child fails to complete five items in a row, the stopping criterion is reached and the subscale is finished (Albers & Grieve, 2007; Bayley et al., 2014b).

In the current study, test administration was carried out by trained researchers that mainly consisted of student assistants and master's students. Prior to the start of the data collection, the researchers completed a training at the Vienna General Hospital (Allgemeines Krankenhaus der Stadt Wien) in July 2023, which was held by two clinical psychologists proficient in the administration of the Bayley-III. The training consisted of a theoretical introduction followed by an observation of a testing session with patients at the General Hospital. In addition, the researchers were required to practice the administration of the Bayley-III with each other as well as having the opportunity to pose questions, until they felt confident using the tool.

5.3.8 Data Processing

Subsequent to the testing sessions, the Bayley-III scores were calculated and analyzed. In a first step, the scores of the items were summed up, thus constituting the subscale scores. In the current study, scores for the cognitive and the gross motor subscale were determined. Correctly solved items were allocated a score of 1, whereas incorrectly solved items were scored with 0 points. According to the Bayley-III manual, items prior to the starting point should be treated like solved items and therefore should be added to the score (Bayley et al., 2014b). Going forward, scaled and composite scores were calculated and assigned to the raw scores, in order to simplify interpretation. Finally, scores were recorded in an Excel spreadsheet, in conjunction with the participant numbers.

5.4 Early Childhood Behavior Questionnaire

The Early Childhood Behavior Questionnaire (ECBQ) is a parent-report questionnaire capturing the temperament of children between the ages of 1.5 and 3 years. It consists of 201 items, 18 scales, and 3 factors, in English as well as in German (Kirchhoff, 2015; Putnam et al., 2006). It is based on the psychobiological model of temperament by Rothbart and Derryberry (1981). Thus, it focuses not only on reactivity but also on self-regulation to modulate that reactivity (Putnam et al., 2006; Rothbart & Derryberry, 1981). The ECBQ is an advancement of the Toddler Behavior Assessment Questionnaire (TBAQ), allowing a more fine-grained assessment of temperament than before (Putnam et al., 2006).

5.4.1 Factorial Structure

The three main factors that were found by factor-analysis are Surgency, Negative Affectivity, and Effortful Control, as discussed earlier (Kirchhoff, 2015; Putnam et al., 2006). Surgency consists of the components/scales Activity Level, High Intensity Pleasure, Impulsivity, Positive Anticipation, and Sociability (Kirchhoff, 2015). Activity Level rates the level of gross motor activity, High Intensity Pleasure describes the level of enjoyment regarding circumstances in which there is a stimulus of high intensity, Impulsivity assesses how fast a response is initiated, Positive Anticipation captures the anticipation concerning expected events, and Sociability is concerned with enjoying social engagements with others (Putnam et al., 2010).

Negative Affectivity is comprised of Discomfort, Fear, Frustration, Motor Activation, Perceptual Sensitivity, Sadness, Shyness, and Soothability (Kirchhoff, 2015). Discomfort describes the level of negative emotions in response to sensory aspects of stimuli, Fear assesses negative emotions associated with expected pain or frightening situations, such as

anxiety, Frustration entails negative reactions in response to being hindered in achieving one's goals or being interrupted during a task, Motor Activation is concerned with repeated fine-motor movements, Perceptual Sensitivity describes the perception of subtle stimuli of low intensity from one's surroundings, Sadness captures decreased mood associated with being confronted with one's own or other's sorrow, loss of objects or approval, or being disappointed, Shyness consists of feeling uncomfortable in social settings that involve unfamiliarity and/or decreased approach behaviors in such situations, and Soothability describes the speed of recovering from excitation (Putnam et al., 2010).

Finally, Effortful Control contains Attention Focusing, Attention Shifting, Cuddliness, Inhibitory Control, and Low Intensity Pleasure (Kirchhoff, 2015). Attention Focusing describes the continuous direction of attention towards a particular object and not being distracted, Attention Shifting consists of being able to move one's attention from one point to another, Cuddliness assesses the child's display of pleasure and physical adjustment of the body when being held, Inhibitory Control involves the ability to disengage from a behavior when being told to do so, and Low Intensity Pleasure describes the level of enjoyment regarding circumstances in which there is a stimulus of low intensity (Putnam et al., 2010).

5.4.2 Psychometric Properties

The German version of the ECBQ shows adequate internal consistency, with Cronbach's α for the three main factors varying between .81 and .87 (Kirchhoff, 2015). The internal consistencies for the subscales range from .61 to .84, with the exception of Activity Level, which has a Cronbach's α of .56 (Kirchhoff, 2015). In a German sample, Sieber and Zmyj (2022) found that stability for parental reports concerning temperament in a 6-month span was high, with all correlations except for Positive Anticipation being significant in the time period of the children's age between 12 and 18 months, as well as correlations being above $r = .5$ from 18 to 24 months. Additionally, the factor of Effortful Control only became apparent in the sample at 18 and 24 months, whereas Surgency and Negative Affectivity were present at all testing points (Sieber & Zmyj, 2022). It is important to note that the psychometric properties described by Kirchhoff (2015) and Sieber and Zmyj (2022) have only been examined in one sample each. Thus, the results of these studies should be regarded with this consideration in mind. Ideally, more studies should examine the psychometric properties of the German version of the ECBQ.

The items of the ECBQ consist of commonly experienced situations and are designed in a stem-and-leaf format, meaning that a single statement is succeeded by multiple responses.

For instance, a statement would be “When visiting a new place, how often did your child”, which is followed by “not want to enter?” and “go right in?” (Putnam et al., 2006). Both questions are rated separately on a 7-point Likert scale ranging from “never” to “always”, with the option to check “does not apply” as well. Some items are reverse scored (Putnam et al., 2006).

5.4.3 Administration

In the current study, the ECBQ was administered online via the platform SoSci Survey (https://sosci.univie.ac.at/). Caregivers received an individualized code via email that allowed them to access the questionnaire at any time as well as making it possible to pause it and continue later.

5.4.4 Data Processing

To process ECBQ data, the collected data was downloaded from SoSci Survey. Furthermore, the information was transferred to an Excel spreadsheet that had been programmed to calculate the scores of all subscales and factors. In addition, the scores of each participant were linked to their access code and their participant numbers, so that Bayley and ECBQ data from each person could be matched and merged without problems.

5.5 Statistical Methods

Statistical analyses were performed using JASP Version 0.19.0 (JASP Team, 2024). The final sample consisted of the children who had already completed the testing sessions at the time of the data analysis (May 2024). Furthermore, only cases where valid data existed for both the Bayley cognitive scale and the ECBQ were included. Due to issues during the test situation or parents not filling in the ECBQ, multiple cases had to be removed beforehand. After removal, the final sample for the main analysis was comprised of 70 children, whose data was available for both Bayley and ECBQ.

6 Results

The method chosen for the statistical analyses is multiple linear regression. Consequently, prior to analyzing the data, tests examining the fulfillment of the assumptions of this method were performed.

The first assumption is that the independent variable should be metric or dichotomous and the dependent variable needs to be metric (Field, 2017). The independent variables in this study are the dimensions of the ECBQ (Surgency, Negative Affectivity, Effortful Control;

Putnam et al., 2006) and the dependent variable is the score on the Bayley-III cognitive scale (Bayley et al., 2014a). Due to the structure of those measures, both can be considered metric.

Further requirements are that all predictors should be included in the model (Field, 2017), which is the case. Moreover, independent observations are required (Field, 2017), which can also be confirmed due to the structure of the testing sessions. In addition, the variance of the predictors must not be 0 (Field, 2017). This too is the case. Another prerequisite is that the outcome variable should be linearly related to the predictors (Field, 2017). To test this, partial regression plots between the independent variable and the respective dependent variables were created, which did not indicate any other non-linear pattern and, although not very strong, showed a linear relationship between the outcome and the predictors. The plots are displayed in *Figure 1* in section 6.2 Main Analysis.

Homoscedasticity was checked using a residual plot, which did not display any pattern, indicating homoscedasticity. The plot can be seen in *Appendix, Figure 2*. Next, the normality of residuals was examined (Field, 2017). A standardized residuals histogram was created, see *Appendix, Figure 3*. The histogram indicated a close-to-normal distribution.

In a last step, multicollinearity was considered and collinearity statistics for all three predictors were calculated. Variance Inflation Factors of all three predictors were around 1 ($VIF_{Surgency}$ = 1.056, $VIF_{NegativeAffectivity}$ = 1.056, $VIF_{EffortfulControl}$ = 1.012), which indicated no multicollinearity (Field, 2017).

Concluding, all assumptions of a multiple linear regression are fulfilled, thus, it is possible to proceed with the main analysis.

6.1 Descriptive Statistics

Descriptive statistics for the three predictors as well as for the outcome were calculated. There were 73 valid cases for the ECBQ, whereas only 70 were valid for the Bayley Scales. Means and standard deviations of each measure were calculated and a Shapiro-Wilk test of normality was conducted. Surgency and Effortful Control were normally distributed ($p > 0.05$), however, the Shapiro-Wilk test of normality for the variable Negative Affectivity reached significance with $p = .05$ as well as for the variable Bayley Cognitive Scale Composite Scores, with $p = .03$. For more detailed information about the descriptive statistics, see *Table 2*.

Table 2: Descriptive Statistics for All Measures

	Surgency	Negative Affectivity	Effortful Control	Bayley Cognitive Scale Composite Scores
Valid cases	73	73	73	70
Mean	235.18	272.29	258.01	104.57
Std. Deviation	30.22	39.90	29.41	17.30
P-value of Shapiro-Wilk	0.19	0.05	0.67	0.03
Minimum	140	204	194	60
Maximum	316	394	323	135

6.1.1 Bayley Composite Score Equivalents

Additionally, frequency tables for the Bayley composite score equivalents were created and the respective interpretations assigned. The majority of children tested (n = 38, 54.29%) scored average on the Bayley cognitive scale. Furthermore, 21 (30%) scored either above average or considerably above average, however, none were extremely above average. Further 11 children (15.71%) scored either below average or considerably below average (Bayley et al., 2014b). See *Table 3* for all frequencies.

Table 3: Frequencies of Composite Score Equivalents

Composite scores	Interpretation	Frequency (n = 70)
≥ 145	Extremely above average	0
130 – 144	Considerably above average	8
115 – 129	Above average	13
85 – 114	Average	38
70 – 84	Below average	10
55 – 69	Considerably below average	1
≤ 54	Extremely below average	0

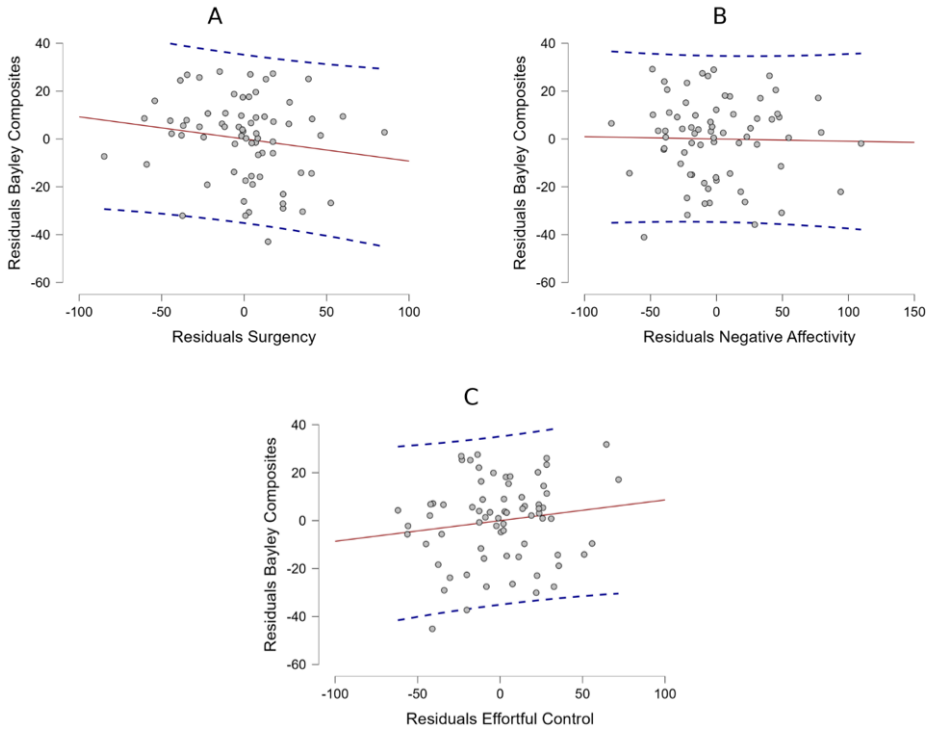
6.2 Main Analysis

A linear regression was used to examine the relationship between the temperamental dimensions of Effortful Control (H1), Negative Affectivity (H2), Surgency (H3) and the cognitive scores on the Bayley-III. For that, Effortful Control, Negative Affectivity, and Surgency were used as the predictors and the scores on the cognitive scale as the outcome. This method was chosen because regression allows for determining the influence of a predictor while controlling for the influence of the other predictors. All predictors were entered simultaneously.

The model indicated an $R^2 = .05$ and an adjusted $R^2 = .004$. However, the ANOVA yielded a significance level of $p = .36$, indicating that the model was not significant. None of the individual predictors was significant either, indicating no significant relation of either predictor with the scores of the Bayley-III. Though very small, the unstandardized coefficients of Surgency ($B = -0.09$) and Negative Affectivity ($B = -0.01$) indicated a negative relationship with the scores of the cognitive scale of the Bayley-III. See *Panel A* in *Figure 1* for a graphic representation of the relationship between Surgency and the Bayley Composite Scores and *Panel B* for a representation of the relationship between Negative Affectivity and the Bayley Composite Scores. Effortful Control, on the other hand, displayed a positive relationship with the cognitive scores ($B = 0.09$), which can be seen in *Panel C* of *Figure 1*. However, it is important to note that none of these relationships were significant.

Figure 1

Partial Regression Plots



The red lines in the plots represent the regression line, whereas the blue lines depict the prediction intervals, which aid in identifying outliers. All three plots displayed two outliers each, otherwise, all values lay between the prediction intervals.

Detailed information about the p-levels, t-values, and unstandardized coefficients is noted in *Table 4*.

Table 4: Coefficients in Linear Regression

	t-value	**Unstandardized regression coefficient**	**p-value**
Negative Affectivity	-0.17	-0.01	0.86
Effortful Control	1.20	0.09	0.24
Surgency	-1.33	-0.09	0.19

As the evidence uncovered in the statistical analysis indicates, no predictor had a significant relationship with the scores of the cognitive scale of the Bayley-III. Consequently, I fail to reject the H0 for all three hypotheses.

In addition to a linear regression model that includes all three predictors, a linear regression model containing only Effortful Control was calculated, as the most consistent findings (Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019) in the literature reviewed for the current thesis were that there is an association between measures related to Effortful Control and scores on scales measuring cognitive abilities. The model resulted in an $R^2 = .02$ and an adjusted $R^2 = .004$, not indicating a better fit than the previous model. The t-value for Effortful Control was 1.13, the unstandardized regression coefficient $B = 0.08$, and $p = 0.27$, again not suggesting that a model only containing Effortful Control explains scores on the cognitive scale of the Bayley-III better than a model containing all three predictors.

6.3 Exploratory Analysis

In addition to the main analysis, exploratory analyses were conducted to shed more light on the topic of temperament and cognitive development with the data available. Moreover, literature (e.g. Else-Quest et al., 2006; Gagne et al., 2013; Krogh & Væver, 2019) suggests that there are sex differences in both cognitive development and temperament:

Krogh and Væver (2019) found that in a sample of 33 children at 24 months, girls scored significantly higher on the Bayley-III cognitive subscale than boys (Cohen's $d = 0.95$). In addition, a study conducted by Gagne and colleagues (2013) that examined 714 three-year-old children showed that boys exhibited significantly higher levels of activity than girls (d ranging from 0.26 to 0.36, depending on the measure; Gagne et al., 2013). Activity is associated with Surgency in the psychobiological model of temperament (Gartstein et al., 2016). Furthermore, boys exhibited significantly lower levels of inhibitory control than girls (d ranging from -0.23 to -0.5, depending on the measure; Gagne et al., 2013). Inhibitory control is related to the dimension of Effortful Control (Gartstein et al., 2016; Rothbart, 2011). Additionally, in a meta-analysis by Else-Quest and colleagues (2006) that involved 189 studies, results indicated that boys displayed higher levels of Surgency than girls ($d = 0.5$), whereas girls exhibited higher levels of Effortful Control ($d = -1.01$).

6.3.1 Sex as an Additional Predictor

Sex was added as a dichotomous variable, with 0 indicating male sex and 1 indicating female sex, to check whether sex was a significant predictor of the scores on the Bayley Cognitive Scale. The adjusted R^2 decreased (adjusted $R^2 = -.01$ as compared to adjusted $R^2 = .004$ in the former model) and the model did not reach significance ($p = .53$). Moreover, sex as a predictor itself was not significant, either ($p = 0.98$). In addition, the test yielded a $t = 0.03$ and an unstandardized regression coefficient $B = 0.11$. The latter suggests that on average, boys and girls in this sample merely differed by 0.11 points in their cognitive scores. As this difference is very small and not statistically significant, it indicates that both sexes performed very similarly on the Bayley-III in this sample. Therefore, the present data does not indicate that there is a relationship between sex and the scores on the Bayley cognitive scale.

6.3.2 Independent Sample t-Tests

Furthermore, independent sample t-tests ($n_{total} = 73$, $n_{boys} = 35$, $n_{girls} = 38$) were calculated to examine whether boys and girls differed in any of the three temperament dimensions.

Girls (labeled "1") scored slightly higher in Effortful Control ($m_{girls} = 259.55$, $m_{boys} = 256.34$) as well as in Negative Affectivity ($m_{girls} = 273.00$, $m_{boys} = 271.51$). However, boys (labeled "0") scored minimally higher in Surgency ($m_{girls} = 235.08$, $m_{boys} = 235.29$).

Nonetheless, as depicted in *Table 5*, the differences between the two sexes at group level were not significant. Thus, it can be concluded that the current data does not indicate a difference in temperament dimensions in boys and girls.

Table 5: Independent Sample t-Tests Examining Sex Differences in Temperament

	t-value	df	p-value
Surgency	0.029	71	0.977
Effortful Control	-0.463	71	0.645
Negative Affectivity	-0.158	71	0.875

6.3.3 Interactions Between the Temperamental Dimensions

In this part of the exploratory analysis, interactions between the temperamental dimensions (Effortful Control, Negative Affectivity, and Surgency) were considered, as temperament dimensions have been shown to interact with each other (Gartstein et al., 2016; Rothbart, 2011).

6.3.3.1 Surgency and Negative Affectivity

In a first step, Surgency and Negative Affectivity were centered by subtracting their respective means. Then, a linear regression model containing the interaction term between Surgency and Negative Affectivity, the centered term of Surgency, the centered term of Negative Affectivity, and Effortful Control was established.

The significance for the interaction term resulted in $p = .5$, not indicating an interaction between Surgency and Negative Affectivity. Furthermore, the interaction term displayed an unstandardized regression coefficient $B = -0.0008$, suggesting a very small negative relationship between the interaction term and cognitive development. The t-value for the interaction term was -0.426.

6.3.3.2 Surgency and Effortful Control

Analogous to the prior model, Surgency and Effortful Control were centered and a linear regression model consisting of the interaction term, the centered term of Surgency and Effortful Control, respectively, and Negative Affectivity was calculated.

The interaction term between Surgency and Effortful Control yielded $p = .18$, which suggests that Surgency and Effortful Control did not interact. Additionally, the unstandardized regression coefficient for this interaction resulted in $B = -0.004$, again indicating a very weak relationship between the interaction term and cognitive development. In addition, the interaction term displayed a t-value of -1.75.

6.3.3.3 Negative Affectivity and Effortful Control

In this model, Negative Affectivity and Effortful Control were centered and a linear regression model with the interaction term of those two, Negative Affectivity centered, Effortful Control centered, and Surgency was conducted.

The p-value of the interaction term yielded $p = .51$, indicating no significant interaction, and the unstandardized regression coefficient resulted in $B = -0.0004$. Furthermore, the interaction term displayed a t-value of -0.21. Hence, this model only indicates a very small and non-significant relationship between the interaction term and cognitive development.

7 Discussion

Temperament affects various parts of our lives and is vital to interindividual differences (Chen & Schmidt, 2015; Rothbart, 2011). Moreover, literature suggests that temperament influences cognitive abilities and performance on scales capturing cognitive skills (Fagen et al., 1987; Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019). Consequently, the present study examined the relationship between temperamental dimensions using the Early Childhood Behavior Questionnaire (Putnam et al., 2006) and scores on the cognitive scale of the Bayley Scales of Infant and Toddler Development III (Bayley et al., 2014b), with the aim to aid in disentangling these relationships.

The hypotheses H1-H3 investigated the relationship between each temperament dimension (Surgency, Negative Affectivity, and Effortful Control) of the ECBQ (Putnam et al., 2006) and the scores on the cognitive scale of the German version of the Bayley-III (Bayley et al., 2014a; Bayley et al., 2014b). However, in the present study, no significant relationships were found. This is not in line with previous research.

H1 hypothesized that children with higher levels of Effortful Control also display higher levels of cognitive development, as it was expected that these children may be able to focus better on a task due to advantages in their attentional skills and inhibitory control. This was not found in the current study. This is contrary to prior studies, such as the ones by Smith-Watts and colleagues (2019) as well as by Moffitt and colleagues (2011), which found that abilities associated with self-restraint were positively related to cognitive performance. A possible reason for the differing results may be the use of different scales – neither Smith-Watts et al. (2019) nor Moffitt et al. (2011) used the Bayley Scales or the ECBQ. As a measure for cognitive abilities, both studies used the IQ. A study by Månsson and colleagues

(2019) that examined the relationship between scores on the cognitive scale of the Swedish version of the Bayley-III (Bayley, 2009) and the Full Scale Intelligence Quotient of the Swedish version of the Wechsler Intelligence Scale for Children – Fourth Edition (Wechsler, 2007) found that while Bayley-III cognitive scores and the Full Scale Intelligence Quotient showed a correlation of .41, they were also significantly different (Månsson et al., 2019). Moreover, the measurements are based on different concepts (Månsson et al., 2019). Although this study was conducted in Swedish, it indicates that it is possible that the differing results between the studies conducted by Smith-Watts et al. (2019) as well as Moffitt et al. (2011) and the current study may partly reflect the differences in measurement.

H2 assumed that children with higher levels of Negative Affectivity show lower levels of cognitive development on the grounds that aspects of negative emotions, such as fear or frustration, would hinder adequate engagement in the tasks. Again, no significant relationship could be found with the current data. This does not align with findings by Fagen et al. (1987), who found a negative relationship between smiling and laughter and MDI scores or with findings by Sajaniemi et al. (1998), whose results suggested a significant positive correlation between low MDI scores and positive mood. Neither does it align with the results found by Smith-Watts et al. (2019). They found that IQ had a negative relationship with observed negative emotionality (Smith-Watts et al., 2019). However, a possible explanation for the results found in the current study could be that children who exhibit higher levels of fearfulness are more prone to following rules, which may have resulted in adequate engagement with the test (Kochanska et al., 2001). Nonetheless, this explanation does not account for other aspects of Negative Affectivity, such as frustration or discomfort.

H3 hypothesized that children who exhibit higher levels of Surgency also show higher levels of cognitive development. This was based on the idea that children who are more outgoing tend to seize opportunities to engage more readily. This could not be seen in the current data, which is contrary to the findings by Fenstermacher & Saudino (2020), who found an indirect relationship between MDI scores and extroverted behavior. A possible reason for non-significant results in the current study may be that due to a different composition of the sample, different components of Surgency may have prevailed, such as Impulsivity or Activity Level (Kirchhoff, 2015), which could have influenced engagement during the testing situation negatively.

Furthermore, an additional regression model was fitted with only Effortful Control as a predictor of cognitive development, as in the literature reviewed for the current thesis, the

most consistent evidence for relationships between cognitive development and temperamental dimensions was found for temperament dimensions related to Effortful Control (Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019). However, this regression model did not indicate a better fit than the model containing all three predictors, nor was it significant, suggesting that a model with only one predictor does not provide a better explanation for the present data. The unstandardized regression coefficient, although not significant, displayed a positive relationship between Effortful Control and cognitive development, which is in line with previous research (Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019). Still, the significant relationships found in previous literature could not be replicated with the current data.

In addition to testing the hypotheses, exploratory analyses were conducted. In an attempt to examine whether sex was associated with scores on the cognitive scale of the Bayley-III, sex was added as a predictor. This, however, did not yield significant results, either and the analysis indicated a decrease in model fit compared to the previous model. Consequently, it can be concluded that boys' and girls' performances were very similar in the current study and that sex did not significantly contribute to explaining the variance in cognitive scores. In 2019, a study by Krogh & Væver examined the effects of gender on scores of the Bayley-III (Bayley, 2006a) and test-taking behavior. They attempted to improve limitations found in previous studies by using a longitudinal design, testing full-term children, and using the Bayley-III (Bayley, 2006a). The results displayed significant differences at various time points for the different scales of the Bayley-III, however, a significant difference between boys and girls in the cognitive scale was found at 24 months (Krogh & Væver, 2019). Thus, the current results do not align with their findings.

In another analysis, gender differences in the temperament dimensions of the ECBQ were explored. Again, no significant gender differences were identified, indicating that both sexes did not differ significantly in temperament. Moreover, correlations between the temperament dimensions were examined, yielding no significant relationships. Finally, interactions between the temperament dimensions were examined, indicating no significant interactions, either. Overall, the findings indicate that neither temperament nor sex were significantly associated with the performance on the Bayley-III.

There are multiple possible explanations why the current results differ from previous studies. First, there is the possibility of publication bias, meaning that there may have been similar studies that did not find any significant relationships that simply were not published.

However, since various studies that found effects exist, alternative explanations need to be considered as well. It may be partly due to the fact that the current sample differed in their characteristics to the samples in the studies cited above, especially because none of those studies used a German speaking sample or measures in German language. Additionally, some of the studies on which the hypotheses are based were published multiple decades ago (Fagen et al., 1987; Sajaniemi et al., 1998) and the characteristics examined may be subjected to temporal changes and cohort effects. Also, results may differ due to a different gestational age of the children examined, as some of the children in the study conducted by Sajaniemi et al. (1998) were born pre-term. Furthermore, not all studies used the Bayley-III or the ECBQ as measures, possibly leading to different results because their measures did not exactly capture the same concepts as the present study. Finally, the current study has several limitations, which also could have affected the results.

7.1 Limitations and Implications for Future Research

The first limitation is that the sample used for the study is not based on a power analysis, rather, the inclusion criterion was the number of children who already had participated at the time of the data analysis. This led to a smaller number of participants than in previous studies (Fenstermacher & Saudino, 2016; Moffitt et al., 2011; Sajaniemi et al., 1998; Smith-Watts et al., 2019) whose numbers of participants ranged from 160 (Sajaniemi et al., 1998) to 1037 (Moffitt et al., 2011), with the exception of the study conducted by Fagen and colleagues (1987) that only included 42 infants, which is fewer than in the current study. The fact that the majority of these previous studies had larger sample sizes than the current study suggests that the current study may have been underpowered. Moreover, the sample is not representative because the participants were recruited through the data bank of the university, which only contains participants who willingly signed up for studies. In addition, the majority of the children's caregivers reported a household income of more than 45 000 € per year, indicating an overrepresentation of higher-income families. Another important aspect is that the present research was part of a longitudinal study with three testing points, therefore, the possibility of systematic dropouts between the testing points exists. Moreover, consistently participating in all three testing points possibly requires a certain level of caregiver's perseverance. As Rothbart and Derryberry (1981) assume a biological basis of temperament, the participating children may also have exhibited higher levels of perseverance or similar traits than the average population, possibly resulting in too little variance concerning the children's temperament. All these factors could affect the composition of the

sample and, in turn, the results. Moreover, it decreases external validity. Future research could examine a larger, representative sample that is based on a power analysis.

Furthermore, although the participants' well-being was a priority, external factors, such as tiredness, having a cold, or being overwhelmed by the new environment could not be completely ruled out, resulting in unsystematic influences on the data. Also, a small team conducted the testing sessions, thus, different sessions were carried out by different people. Hence, despite a standardized procedure, small differences between the testing sessions could occur. This was amplified by the participants' varying temperaments, which influenced interactions with the researchers and, therefore, the whole testing.

Another limitation is, as already mentioned earlier, the inability to distinguish between the effects of temperament on cognitive development per se and the effects of temperament on test-taking behavior. It may also be possible that scores on measures capturing cognitive abilities result from an interplay of those two. Additionally, this study is merely correlational and the statistical methods applied do not allow for causal interpretations.

In addition, the measures employed have limitations. Even though the Bayley-III (Bayley, 2006) is considered gold standard for assessing the development of infants and toddlers (Armstrong & Agazzi, 2010) and exhibits satisfactory psychometric properties (Bayley et al., 2014b), it does not specify which specific cognitive skills it captures, complicating the interpretation of its results (Armstrong & Agazzi, 2010). Moreover, by abolishing the MDI, the Bayley-III attempted to disentangle cognitive skills from language skills, however, receptive language skills are still necessary to understand the tasks and a few items require verbal responses from the children as well (Albers & Grieve, 2007; Bayley et al., 2014b). As language skills were not captured in the present study, they were not controlled for. In future studies, it may make sense to apply the language scale in addition to the cognitive scale to have a deeper understanding of the children's abilities.

Regarding the ECBQ, one of its limitations is that it is an other-report questionnaire (Putnam et al., 2006). Compared to other measures, such as behavioral observations, other-report questionnaires are more likely to be subject to biases (Chen & Schmidt, 2015). Examples are memory biases, where some information may be remembered incorrectly, or biases such as raters not paying attention to the instructions of the questionnaire, raters answering the questions influenced by social desirability, or contrast effects, where certain characteristics of a child appear more pronounced or less noticeable when compared to other children (Chen & Schmidt, 2015). Also, data created by observational measures of

temperament and other-report questionnaires does not always align (Smith-Watts et al., 2019). Furthermore, although satisfactory, the psychometric properties of the German version of the ECBQ have only been examined in two samples (Kirchhoff, 2015; Sieber & Zmyj, 2022). Future research could apply a second measure of temperament in addition to the ECBQ, such as an observational measure, to support the data collected by the ECBQ.

7.2 Conclusion

The present study examined the relationship between temperament, assessed by the ECBQ, and cognitive development, captured by the Bayley-III, in children aged 24 months. It was hypothesized that temperament influences test-taking behavior and, in turn, scores on scales capturing cognitive skills. However, statistical analyses did not find evidence supporting these hypotheses. In further explorations, the relationship between sex and the scores of the Bayley-III was investigated as well, again indicating no significant relationship. Moreover, boys and girls did not significantly differ in temperament. Future research should focus on examining bigger and more representative samples, while taking into account and controlling for the limitations associated with the measures. The study of temperament and cognitive abilities offers exciting insights into how interindividual differences may influence cognitive performance. Knowledge about this relationship could aid in creating and establishing test situations that capture cognitive abilities as exactly as possible.

8 References

Albers, C. A., & Grieve, A. J. (2007). Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development– Third Edition. San Antonio, TX: Harcourt Assessment. *Journal of Psychoeducational Assessment*, *25*(2), 180–190. https://doi.org/10.1177/0734282906297199

Armstrong, K. H., & Agazzi, H. C. (2010). The Bayley-III Cognitive Scale. In L. Weiss, T. Oakland, & G. Ayland (Eds.), *Bayley-III Clinical Use and Interpretation* (pp. 29–45). Elsevier Inc. https://doi.org/10.1016/B978-0-12-374177-6.10002-9

Bayley, N. (1969). *Bayley Scales of Infant Development*. Psychological Corporation.

Bayley, N. (1994). *Bayley Scales of Infant Development* (2nd ed.). Psychological Corporation.

Bayley, N. (2006a). *Bayley Scales of Infant and Toddler Development* (3rd ed.). Pearson.

Bayley, N. (2006b). *Bayley Scales of Infant and Toddler Development: Technical manual* (3rd ed.). PsychCorp.

Bayley, N. (2009). *Bayley Scales of Infant and Toddler Development: Svenskt manual supplement* (3rd ed.). Pearson.

Bayley, N. & Aylward, G. P. (2019). *Bayley Scales of Infant and Toddler Development* (4th ed.). Pearson.

Bayley, N., Reuner, G., & Rosenkranz, J. (2014a). *Manual zur Durchführung und Auswertung der Bayley-III Skalen* (3rd ed.). Pearson.

Bayley, N., Reuner, G., & Rosenkranz, J. (2014b). *Technisches Manual zu den Bayley-III Skalen und dem Bayley-III Screening Test* (3rd ed.). Pearson.

Casey, B., Tottenham, N., Liston, C., Durston, S., & Durston, S. (2005). Imaging the developing brain: what have we learned about cognitive development?. *Trends in Cognitive Sciences*, 9, 104-110. https://doi.org/10.1016/j.tics.2005.01.011.

Chen, X., & Schmidt, L. A. (2015). Temperament and personality. *Handbook of Child Psychology and Developmental Science*, 1-49.

Del Vecchio, T. (2011). Peer Nomination Technique. In S. Goldstein & J. A. Naglieri (Eds.), *Encyclopedia of Child Behavior and Development* (p. 1074). Springer US. https://doi.org/10.1007/978-0-387-79061-9_2097

Eckardt, G. (2017). *Persönlichkeits- und Differentielle Psychologie: Quellen zu ihrer Entstehung und Entwicklung*. Springer.

Eisenberg, N., Smith, C. L., & Spinrad, T. L. (2016). Effortful Control: Relations with Emotion Regulation, Adjustment, and Socialization in Childhood. In K.D. Vohs & R.F. Baumeister (Eds.), *Handbook of Self-Regulation: Research, Theory, and Applications* (pp. 458-478). The Guilford Press.

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, *132*(1), 33–72. https://doi.org/10.1037/0033-2909.132.1.33

Evans, D. E., & Rothbart, M. K. (2007). Developing a model for adult temperament. *Journal of Research in Personality*, *41*(4), 868–888. https://doi.org/10.1016/j.jrp.2006.11.002

Fagen, J. W., Singer, J., Ohr, P. S. & Fleckenstein, L. K. (1987). Infant temperament and performance on the Bayley Scales of Infant Development at 4, 8, and 12 months of age. *Infant Behavior & Development*, *10*(4), 505–512. https://doi.org/10.1016/0163-6383(87)90047-6

Fenstermacher, S. K., & Saudino, K. J. (2016). Exploring links among imitation, mental development, and temperament. *Infancy*, 21(5), 536–559. https://doi.org/10.1111/infa.12132

Field, A. P. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Field, T., Hallock, N., Ting, G., Dempsey, J., Dabiri, C., & Shuman, H. H. (1978). A first-year follow-up of high-risk infants: formulating a cumulative risk index. *Child development*, *49*(1), 119–131.

Fullard, W., Mcdevitt, S. C., & Carey, W. B. (1984). Assessing temperament in one-to three-year-old children. In *Journal of Pediatric Psychology* (Vol. 9, Issue 2). https://academic.oup.com/jpepsy/article/9/2/205/962227

Gagne, J. R., Miller, M. M., & Goldsmith, H. H. (2013). Early-but modest-gender differences in focal aspects of childhood temperament. *Personality and Individual Differences*, *55*(2), 95–100. https://doi.org/10.1016/j.paid.2013.02.006

Gartstein, M. A., Putnam, S. P., Aron, E. N. & Rothbart, M. K. (2016). Temperament and personality. In S. Maltzman (Eds.), *The Oxford Handbook of Treatment Processes and Outcomes in Psychology: A Multidisciplinary, Biopsychosocial Approach* (pp. 11-41). https://doi.org/10.1093/oxfordhb/9780199739134.013.2

Gartstein, M. A., & Rothbart, M. K. (2003). Studying infant temperament via the revised infant behavior questionnaire. *Infant Behavior & Development, 26*(1), 64–86. https://doi.org/10.1016/S0163-6383(02)00169-8

Gauvain, M. (2022). *Cognitive development in infancy and childhood*. Cambridge University Press. https://doi.org/10.1017/9781108955676

Goldsmith, H. H., & Rothbart, M. K. (1991). Contemporary instruments for assessing early temperament by questionnaire and in the laboratory. In J. Strelau & A. Angleitner (Eds.), *Explorations in temperament: International perspectives on theory and measurement* (pp. 249–272). Plenum Press. https://doi .org/10.1007/978-1-4899-0643-4_16

Grimm, H., Doil, H., Aktaş, M. & Frevert, S. (2019). *ELFRA. Elternfragebögen für die Früherkennung von Risikokindern.* Hogrefe.

Holmboe, K. (2020). Surgency. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 5326-5331). Springer.

Izard, C. E., Huebner, R., Risser, D., McGinnes, G. C., & Dougherty, L. (1980). The young infant's ability to produce discrete emotional expressions. *Developmental Psychology, 16*, 132–140. https://doi.org/10.1037/0012-1649.16.2.132

JASP Team. (2024). *JASP* (Version 0.19.0) [Computer software]. https://jasp-stats.org/download/

Kirchhoff, M.C. (2015). *Temperament and behavioral difficulties across cultures: A comparison between German and U.S. toddlers*. [Doctoral dissertation, Washington State University]. Research Exchange. https://hdl.handle.net/2376/6209

Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development*, *72*(4), 1091–1111. https://doi.org/10.1111/1467-8624.00336

Kochanska, G., & Knaack, A. (2003). Effortful control as a personality characteristic of young children: antecedents, correlates, and consequences. *Journal of personality*, *71*(6), 1087–1112. https://doi.org/10.1111/1467-6494.7106008

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36*(2), 220–232. https://doi.org/10.1037/0012-1649.36.2.220

Krogh, M. T., & Væver, M. S. (2019). Does gender affect Bayley-III scores and test-taking behavior? *Infant Behavior and Development*, *57*. https://doi.org/10.1016/j.infbeh.2019.101352

Li-Grining, C. P. (2007). Effortful control among low-income preschoolers in three cities: Stability, change, and individual differences. *Developmental Psychology, 43*(1), 208–221. https://doi.org/10.1037/0012-1649.43.1.208

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A. & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Development*, *75*(5), 1357–1372. https://doi.org/10.1111/j.1467-8624.2004.00745.x

Macha, T., & Petermann, F. (2015). Bayley Scales of Infant and Toddler Development, Third Edition – Deutsche Fassung. *Zeitschrift Für Psychiatrie, Psychologie Und Psychotherapie*, *63*(2), 139–143. https://doi.org/10.1024/1661-4747/a000232

Månsson, J., Stjernqvist, K., Serenius, F., Ådén, U., & Källén, K. (2019). Agreement between Bayley-III measurements and WISC-IV measurements in typically developing children. *Journal of Psychoeducational Assessment*, *37*(5), 603-616. https://doi.org/10.1177/0734282918781431

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H. L., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Moosbrugger, H., & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion*. Springer.

Paulus, D. J. & Zvolensky, M. J. (2020). Negative Affectivity. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 3146-3150). Springer.

Peters-Martin, P., & Wachs, T. D. (1984). A longitudinal study of temperament and its correlates in the first 12 months. *Infant Behavior & Development, 7*(3), 285–298. https://doi.org/10.1016/S0163-6383(84)80044-2

Putnam, S. P., Gartstein, M. A., & Rothbart, M. K. (2006). Measurement of fine-grained aspects of toddler temperament: The Early Childhood Behavior Questionnaire. *Infant Behavior and Development*, *29*(3), 386–401. https://doi.org/10.1016/j.infbeh.2006.01.004

Putnam, S. P., Jacobs, J., Gartstein, M. A., & Rothbart, M. K. (2010, March). *Development and assessment of short and very short forms of the Early Childhood Behavior Questionnaire*. [Poster Presentation]. ICIS 2010, Baltimore, MD, United States. https://research.bowdoin.edu/rothbart-temperament-questionnaires/files/2016/08/ICIS_2010_ECBQ_sf_poster.pdf

Reuner, G., Voigt, B., & Pietz, J. (2010). *Validierung eines Elternfragebogens zur kognitiven Entwicklung an (korrigiert) zweijährigen frühgeborenen Kindern* [Conference Presentation]. Neurowoche Mannheim, Germany.

Roth, K., Eisenberg, N., & Sell, E. R. (1984). The relation of preterm and full-term infants' temperament to test-taking behaviors and developmental status. *Infant Behavior & Development, 7*(4), 495–505. https://doi.org/10.1016/S0163-6383(84)80008-9

Rothbart, M. K. (1981). Measurement of Temperament in Infancy. *Child Development*, *52*(2), 569–578. https://doi.org/10.2307/1129176

Rothbart, M. K. (2011). *Becoming who we are: Temperament and personality in development.* Guilford Press.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: the Children's Behavior Questionnaire. *Child development*, *72*(5), 1394–1408. https://doi.org/10.1111/1467-8624.00355

Rothbart, M. K., & Bates, J. E. (2006). Temperament. In N. Eisenberg (Ed.), *Handbook of Child Psychology* (6th ed., Vol. 3, pp. 99–166). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0303

Rothbart, M. K., & Derryberry, D. (1981). Development of individual differences in temperament. In M. L. Lamb & A.L. Brown (Eds.*), Advances in developmental psychology* (Vol. 1, pp. 37–86). Guilford.

Rowe, D. C., & Plomin, R. (1977). Temperament in early childhood. *Journal of Personality Assessment, 41*, 150–156. https://doi.org/10.1207/s15327752jpa4102_5

Sajaniemi, N., Salokorpi, T., & von Wendt, L. (1998). Temperament profiles and their role in neurodevelopmental assessed preterm children at two years of age. *European child & adolescent psychiatry*, *7*(3), 145–152. https://doi.org/10.1007/s007870050060

Shiner, R. L. (2015). The development of temperament and personality traits in childhood and adolescence. In M. Mikulincer, P. R. Shaver, M. L. Cooper, & R. J. Larsen (Eds.), *APA Handbook of Personality and Social Psychology, Vol. 4. Personality processes and individual differences* (pp. 85–105). American Psychological Association. https://doi.org/10.1037/14343-004

Sieber, F., & Zmyj, N. (2022). Stability and structure of infant and toddler temperament in two longitudinal studies in Germany. *Infant Behavior and Development*, *67*. https://doi.org/10.1016/j.infbeh.2022.101714

Sternberg, R. J., & Sternberg, K. (2017). *Cognitive psychology* (7th ed.). Cengage Learning.

Strelau, J. (2020). Temperament. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 5388-5407). Springer.

Suchodoletz, W.v. (2008). Der SBE-2-KT – ein neuer Kurztest zur Früherkennung von Late Talkers bei der U7. *Kinderärztliche Praxis, 79*, 308.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children Revised: Manual*. Psychological Corporation.

Wechsler, D. (2007). *Wechsler Intelligence Scale for Children, Svensk version* (4th ed.). Pearson.

Worobey, J. (1987). Temperament and sensorimotor intelligence. *Early Child Development and Care*, *28*(1), 1–11. https://doi.org/10.1080/0300443870280101

9 List of Figures

10 List of Tables

11 Appendix

Figure 2

Residual Plot

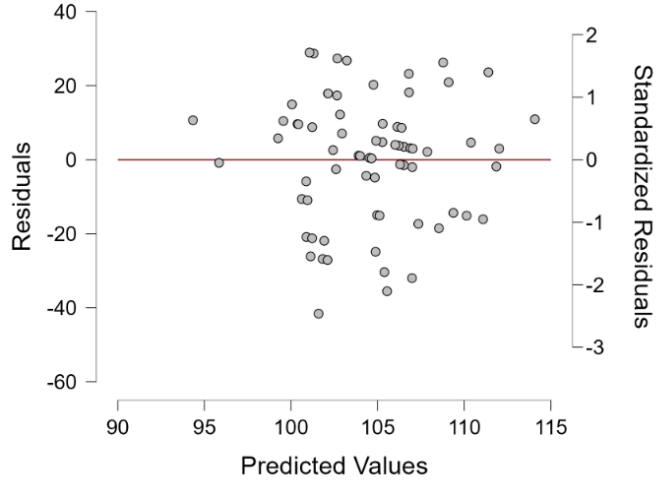
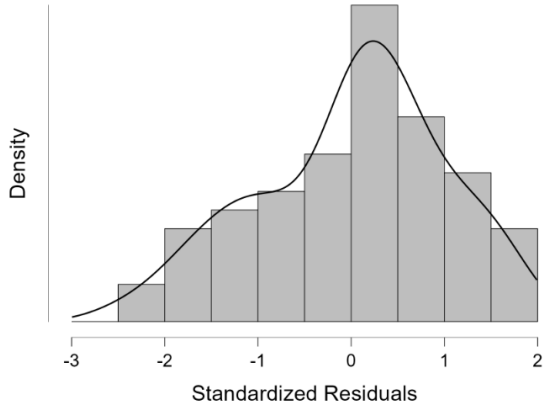


Figure 3

Standardized Residuals Histogram



12 Abstract (English)

Temperament is of high importance to humans' individuality. Cognition, on the other hand, is essential to our perception of and interaction with the world. Thus, it follows that temperament and cognitive abilities interact with each other in development, influencing how we perceive and engage with our environment. The present study examined the relationship between temperament, measured with the Early Childhood Behavior Questionnaire, and cognitive development, assessed with the cognitive scale of the Bayley Scales of Infant and Toddler Development III, in a sample of 73 toddlers aged approximately 24 months. I expected that the temperament dimensions of Effortful Control and Surgency would display a positive relationship with the scores on the cognitive scale, as higher levels of those dimensions may impact test taking behavior positively, whereas the dimension of Negative Affectivity would be negatively correlated with the scores on the cognitive scale. However, no evidence supporting these hypotheses was found. Moreover, scores on the cognitive scale of the Bayley-III were not related to the children's sex, nor did boys and girls differ significantly in temperament dimensions. The results indicate that temperament and scores on the cognitive scale of the Bayley-III were not associated with each other in the present sample.

13 Abstract (German)

Das Temperament eines Menschen ist für seine Individualität von hoher Relevanz. Kognitive Fähigkeiten hingegen spielen bei der Wahrnehmung von und Interaktion mit der Umwelt eine essenzielle Rolle. Daher wäre es naheliegend, dass Temperament und kognitive Fähigkeiten in der Entwicklung miteinander interagieren und sich dies darauf auswirkt, wie wir unsere Umgebung wahrnehmen und mit ihr in Kontakt treten. Die vorliegende Studie hat den Zusammenhang zwischen Temperament, erfasst mit dem Early Childhood Behavior Questionnaire, und kognitiver Entwicklung, erhoben mit der kognitiven Skala der Bayley Scales of Infant and Toddler Development III, in einer Stichprobe von 73 Kleinkindern im Alter von circa 24 Monaten untersucht. Erwartet wurde einerseits ein positiver Zusammenhang zwischen den Temperamentdimensionen Effortful Control sowie Surgency mit den Testwerten der kognitiven Skala des Bayley-III, da angenommen wurde, dass höhere Ausprägungen dieser Dimensionen sich positiv auf das Testverhalten auswirken würden. Andererseits wurde erwartet, dass die Temperamentdimension Negative Affectivity eine negative Korrelation mit den Testwerten der kognitiven Skala aufweist. In der statistischen Analyse konnten jedoch keinerlei Hinweise auf derartige Zusammenhänge gefunden werden. Darüber hinaus waren die Ergebnisse in der kognitiven Skala des Bayley-III unabhängig vom Geschlecht der Kinder, außerdem haben sich Mädchen und Buben nicht signifikant in der Ausprägung ihrer Temperamente unterschieden. Die Ergebnisse deuten darauf hin, dass in der gegenwärtigen Studie Temperament und die Testwerte der kognitiven Skala des Bayley-III nicht miteinander zusammenhängen.