SENNA: Inventory for the Assessment of Social and Emotional Skills: Technical Manual

Ricardo Primi,

Universidade São Francisco, and *EduLab21, Ayrton Senna Institute*, Brazil,

Daniel Santos,

Universidade de São Paulo and *EduLab21, Ayrton Senna Institute*, Brazil,

Oliver P. John,

University of California, Berkeley, and *EduLab21, Ayrton Senna Institute*, Brazil, &

Filip De Fruyt

Ghent University, Belgium and *EduLab21, Ayrton Senna Institute*, Brazil

16/08/2021

Summary

Summary

Chapter 1: Background and Conceptual Foundations of Social-emotional Skills

Over the past two decades, education sciences and policy-making has witnessed an increased attention for the assessment and learning of Social-Emotional Skills, also called 21st century skills (from here onwards abbreviated as SEMS). This interest shift built on the notion that learning cognitive skills (like reading and math) is not sufficient for a successful and happy life. This insight was already recognized in 1952 by Albert Einstein, the preeminent physicist and astute observer of human beings. In a letter to a friend, he wrote: "I have no special intellectual gifts. I am only passionately curious." Einstein was, of course, superbly intelligent (and also modest), but he also recognized that noncognitive attributes, like being curious and working hard, were not only important in life, but also formed key ingredients for learning in the classroom and beyond.

Today, SEMS are considered both as means and end-products of education processes, in addition to more traditional indicators of scholastic performance, such as language and mathematics' achievement. The amplified attention for SEMS raised questions on their conceptual status, their underlying structure, and how SEMS develop either naturally (normative development), socially influenced by formal, and informal experiences, and through scholastic intervention (targeted malleability); these issues have been addressed in several recent reviews. Parallel to these scientific investigations, intensive policy debates started in many countries, including Brazil, on how to represent SEMS learning in the educational curriculum (OECD, 2015). One significant initiative in Brazil is the Brazilian Common Core for Teaching Fundamentals (Base Nacional Comum Curricular – BNCC) (MEC, 2017).

Since the BNCC was approved in December 2017, educators around the country are mobilizing themselves to actualize its guidelines in the curriculum, given that they need to meet the 2020 deadline defined by the Ministry of Education for these guidelines to be fully implemented in schools. In order to provide these educators with some support on that matter, Instituto Ayrton Senna has built a digital guide entitled "*BNCC: construindo um currículo de educação integral*" which aims at assisting school administrators and their teams when elaborating curriculums for their municipalities

and states assuring their commitment with integral education as indicated in the BNCC text, a commitment Instituto shares. As part of this digital guide, public has access to an exclusive analysis associating the set of general competences contained in BNCC, covering several aspects of SEMS, with the most robust SEMS taxonomy available in the scientific community.

The BNCC is a consensus-based framework, developed through intensive discussions among the various stakeholders and eventually reaching sufficient consensus and political approval. Such consensus-based, political frameworks need to be distinguished from scientific taxonomies that are based on systematic research and findings. In the academic and public policy discussions regarding the BNCC, the need became apparent for a clearly articulated, science-based taxonomy of SEMS that describes comprehensively and compactly those SEMS that have been demonstrated to be important for learning and functioning in the 21st century. The present manual will describe the development of such a taxonomy, the Senna SEMS-framework, and the construction of an accompanying assessment inventory, Senna, to assess SEMS in school settings in Brazil.

Initial evidence for the importance of SEMS

In the early 1960s, teachers, education activists, and researchers in the United States set out to prove that education could make a big difference for improving the lives of people from poor and disadvantaged backgrounds. A few years later, a similar motion occurred in Brazilian settings. The Perry Longitudinal Study focused on early preschool education and started in the Detroit area in the US state of Michigan. This extensive intervention program provided regular preschool education as well as home visits to children from severely disadvantaged families. To allow for a formal assessment of the success of the intervention, the researchers set up a randomized control trial—that is, half of the kids in the study were randomly assigned to receive the intervention (the intervention group), and the other half did not (the control group) (Schweinhart e cols., 2005).

What makes this study so unique is that the researchers conducted regular follow-up assessments of these kids all the way into their 40s. They kept records of the kids'

school attendance and degree completion; they gave the kids tests of their intelligence; they visited them at home and kept track of their employment, their marriages, and many other important outcomes as they matured into adults. Careful re-analyses by economist James Heckman of the University of Chicago and his team in 2010 came to a clear conclusion: "We find statistically significant and economically important program effects for both males and females" (Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010). Compared to the kids in the control condition, kids who had been in the early intervention group had been more successful in school, showed fewer arrests and imprisonment as adults, were more likely to be employed and have better economic outcomes, including greater incomes by age 40. Compared to the substantial costs of this preschool intervention, Heckman and his team nonetheless estimated a real and important 6-10% societal return on the investment.

These results from a randomized control trial speak directly to the issue question of malleability (or plasticity): the kids in the intervention group learned skills at their school but the kids in the control group did not learn those skills as much. That is, the kids in the intervention group were not born with these skills, and their parents did not teach them these skills at home. The Perry School findings demonstrate that there are important life skills that help kids get more from their education and stay out of jail and hold on to jobs and earn a living-- and these skills are not fixed or in-born but are learned and can be learned in school. This is the lesson of plasticity, and the Perry School study teaches that lesson in a particularly powerful way because the study started out with the most impoverished and disadvantaged families. The Perry School findings were important because they provided direct and causal evidence for the beneficial long-term effects that excellent preschool education and support can have: it helps disadvantaged kids be more successful in life. These findings have been central to the argument for early childhood preschool programs throughout the world.

Initially researchers thought the intervention would work by making the kids "smarter"—that is, raise the scores the kids in the intervention group achieved on IQ tests. However, by the time the kids reached adolescence, the intervention group did not score higher than the control group on any of the "cognitive skills" tests the researchers had so carefully administered. To be clear, they found their intervention had

no long-term effects on kids' cognitive skills! Thus, Heckman's team concluded in 2010 that the impressive effects of the Perry Preschool program arise "primarily from boosts in non-cognitive traits" (Heckman et al., 2010).

A definition of SEMS

Although Heckman initially labelled these causal agents "non-cognitive traits", the bifurcation between 'cognitive' and 'non-cognitive' is rather artificial. The term SEMS is to be preferred over "non-cognitive skills" because many of the skills subsumed under the former header require specific cognitive abilities (De Fruyt, Wille, & John, 2015). Collaborative problem-solving, for example, entails both cognitive and non-cognitive abilities. Likewise, planning and organization skills also implicate cognitive capacity. Furthermore, we suggest using 'skills' instead of 'traits', because the trait term is strongly associated with the personality trait concept, often misunderstood as immutable and chiefly genetically determined. Although trait psychologists today acknowledge that traits have both stable and developing features, the skill construct is more dynamic, and is comparable in nature to the competency construct used to denote characteristics that individuals require on the labor market for successfully filling job vacancies or performing in jobs. Cognitive abilities and personality traits are to be considered as (interactive) building blocks of SEMS/competencies (De Fruyt, Bockstaele, Taris, & Van Hiel, 2006). For the educational domain, we prefer however the term SEMS.

There are probably as many definitions of SEMS as there are models structuring these skills (John, & De Fruyt, 2015). Relying on an extensive review of the literature and on inputs from various experts and policy makers panels from OECD countries, SEMS have been defined as: "Individual capacities that (a) are manifested in consistent patterns of thoughts, feelings and behaviors, (b) can be developed through formal and informal learning experiences, and (c) influence important socioeconomic outcomes throughout the individual's life" (OECD; 2015, p. 4). Such definition is comprehensive enough to accommodate a wide range of skills, highlighting at the same time their malleability and consequential effects for individuals and society.

Towards an agreed upon taxonomy of SEMS

There is a plethora of different SEMS' taxonomies, with some advocating only a few and others proposing hundred or more skills. One influential group (Elias et al., 1997) proposed six major domains of social and emotional learning, including (1) recognizing and managing emotions, (2) setting and achieving positive goals, (3) appreciating the perspectives of others, (4) establishing and maintaining positive relationships, (5) making responsible decisions, and (6) handling interpersonal situations constructively. Durlak, Weisberg, Dymnicki, Taylor and Schellinger (2011) advocate that social-emotional learning programs should foster five broad competency sets but they are different from those proposed by Elias et al. (1997): self-awareness, self-management, social awareness, relationship skills and responsible decision-making.

From a different angle, Saarni (1999, 2011) focused on what a child needs to learn to become an emotionally and socially competent adult. She distinguished eight affect-oriented behavioral, cognitive, and regulatory skills that she assumed were prerequisites for emotional competence. These skills are: (1) Awareness of one's own emotional state; (2) Skills in discerning and understanding the emotions of others; (3) Skill in using the common vocabulary of emotion and expression; (4) Capacity for empathic and sympathetic involvement in others' emotional experiences; (5) Skill in realizing that inner emotional states need not correspond to outer expression; (6) Capacity for adaptive coping with aversive or distressing emotions by using self-regulatory strategies that ameliorate the intensity or temporal duration of such emotional states; (7) Awareness that relationships are defined by emotional genuineness of expressive display and reciprocity; and (8) Capacity for emotional self-efficacy (i.e., individuals can accept their own emotional experience and view themselves as generally feeling the way they want to feel). Trilling and Fadel (2009), however, distinguish among a set of more than 160 different skill terms, including concepts like abnegation and altruism, engagement and enthusiasm, innovation and inquisitiveness, self-discipline and self-control, stability and tranquility, and many more. Conceptually, in an attempt to simplify, these 21st century skills have been grouped into the broader themes labeled as "the four **C**s": Collaboration, Communication, Critical thinking, and Creativity-innovation (e.g., Trilling & Fadel, 2009.

Although there may be obvious reasons why these frameworks are very different, this mixture of models and diversity of vocabularies hampers an integrative and in-depth debate among the various stakeholders, including how to operationalize the construct. The field further suffers from the jingle-jangle fallacy, where similarly named constructs across frameworks refer to different skills (jingle), whereas nearly identical skills are labelled differently (jangle). The construct of Grit (Duckworth, Peterson, Matthews, & Kelly, 2007), for example, refers to perseverance and passion for achieving long-term goals and has received considerable attention in education over the past decade. Recent behavior-genetic (Rimfeld, Kovas, Dale, & Plomin, 2016) and meta-analytic work (Crede, Tynan, & Harms, 2016), however, showed that this 'new' construct is very similar to the personality construct of conscientiousness, with a well-documented history as an important antecedent for learning within the personality field (e.g., John, Caspi, Robins, et al., 1994; for reviews, see Poropat, 2009, 2014a, 2014b). Recently, Dumfart and Neubauer (2016) showed that conscientiousness and intelligence independently predicted language performance and overall GPA in 8th grade students in Austria but that Duckworth Grit dimension showed no incremental validity beyond conscientiousness, consistent with the evidence provided by Rimfeld et al. (2016). In other words, Grit does not seem to add anything new beyond the well-known effects of conscientiousness. The next section (1.4) of this manuscript will examine how this jingle-jangle fallacy can be overcome when constructs are properly operationalized in terms of convergent and discriminant validity.

The social-emotional "Big Five"

The challenge to bring order in the chaos of (overlapping) SEMS' terms closely resembles personality psychologists' efforts to structure the hundreds of personality descriptive terms available, and this similarity is what reasons this topic being brought at this point of this text. These terms have been finally summarized in the Big Five personality dimensions (John, 1990). Today, personality psychologists agree that the five dimensions of openness (or open-mindedness), conscientiousness, extraversion, agreeableness, and negative emotionality form the largest common denominator to describe personality differences observable in various age and cultural groups (De Fruyt & Van Leeuwen, 2014; McCrae & Terracciano, 2005). The availability of this empirical

framework helped to solve the disputes caused by using different labels for similar constructs and to examine the overlap among presumably distinct constructs. This break-through greatly advanced the personality field leading to improved knowledge on how these individual differences are best assessed and how personality changes and develops across the life-course. A similar kind of structuring and trimming exercise seems recommended for the field of SEMS.

Although the Big Five were initially introduced as a personality descriptive taxonomy, John and Srivastava (1999) described five broad psychosocial systems that help to explain why the big five dimensions are manifested across cultures, societies, and time. This more functional angle on the key dimensions of personality is extremely helpful to think about structuring SEMS and how to connect SEMS with outcomes of education, as it is further explored.

A first mechanism distinguished by John and Srivastava (1999) is the "*Approach*" system that helps students getting into contact with others and moving ahead. Its neurobiological substrate is the Behavioral Activation System (BAS) that is activated through the anticipation of rewards and gains. Its equivalent in the Big Five is the interpersonal dimension of extraversion, chiefly describing the frequency of social interaction with others. A second evolutionary based system is "*Belonging,*" which describes humans' need to live in groups and form bonds with others in order to guarantee survival, ensure social support and well-being of the individual and the group. Agreeableness is the dimension reflecting individual differences in the quality of the social relationships that students and adults form. The *approach* and *belonging* systems together describe much of the interaction patterns observable among humans in terms of up versus down (status hierarchy) versus horizontal (belonging).

Students not only achieve successes and gains, but they also have to deal with failures, losses and punishments. The "*Coping*" system has evolved to help people deal with negative events and to regulate the negative affect and emotions associated with the disappointments of life. The corresponding dimension in the Big Five is negative emotionality (or neuroticism) versus emotional stability.

Two additional systems help to implement and learn new skills and knowledge, so students can execute assigned tasks efficiently and effectively and learn from experiences. The "Self-management" system helps students perform tasks and complete assignments, and meet standards, goals, and expectations. Skills in this domain are associated with the Big Five conscientiousness domain.

The final system focuses on exploration, challenging the status quo and helping the individual to flexibly navigate and explore different interests, available options and possibilities, and learn from experiences. Its Big Five equivalent is the openness to experiences dimension. Overall, this functional and dynamic account of the Big Five dimensions provides a deeper elaboration of what Heckman (2010) had simply labeled "non-cognitive traits".

Outlining the contours of the Senna SEMS framework

In Brazil, interest in the assessment and development of SEMS has been growing over the past decade, as a means to improve general welfare and prepare youth for upcoming challenges via general education and intervention programs. The Ayrton Senna Institute has played a key role in Brazil in raising awareness for this challenge and initiated and supported research to list and examine the underlying structure of measures used to examine social and emotional development and school learning in Brazil (Primi & Santos, 2014).

Primi and Santos (2014) started a systematic search and identified eight SEMS instruments which constructs showed power to predict consequential outcomes of education, were feasible to administer (low cost, little time to administer, formulated in age-appropriate language, useful for self-description without assistance), assessed malleable constructs, and showed robust psychometric characteristics. Primi and colleagues (2016) demonstrated that the items and scales of these eight frequently used measures to evaluate SEMS-learning could be easily mapped within the Big Five scheme. A joint factor analysis of the Nowicki-Strickland Locus of Control Scale (Nowicki & Strickland, 1973), Rosenberg Self-Esteem Scale (Rosenberg, 1979), Strengths and Difficulties Questionnaire (Goodman, 1997), Big Five Inventory (BFI; John, Donahue, & Kentle, 1991), Self-Efficacy Questionnaire for Children (Muris, 2001), Core Self-

evaluations (Judge, Erez, Bono, & Thoresen, 2003), and the Grit Scale (A. L. Duckworth & Quinn, 2009), showed that all their items could be easily structured under the umbrella of the five major dimensions of personality. Also, a sixth factor popped up, tentatively labelled as External Locus of Control/Negative Valence, though it was unclear at that point whether this factor reflected substance or was more an artefact or method factor (Primi et al., 2016).

Additional empirical evidence for the comprehensiveness and utility of the Big Five framework to structure SEMS was provided by John and Mauskopf (2015) who used factor-analysis to examine the large set of more than 160 SEMS that were listed initially by Trilling and Fadel (2009) and further expanded by so-called "character" skills Fadel and his collaborators (2015). When the resulting SEMS factors from the Fadel list were empirically compared with the Big Five Inventory (BFI-2; Soto & John, 2016), John and Mauskopf found that they converged very closely. In other words, the major conceptual domains defined by the 21st century character skills confirmed the well-known Big Five domains. To illustrate, we summarize the kinds of skills included in each of the five domains in Tables 1 – 5, with about 10 specific skills as examples. We begin with the two skill domains that are known to be most closely linked to learning outcomes: Open-mindedness and Self-Management.

Let us consider Open-Mindedness first. Table 1 shows the 21st century skills most strongly associated with Open-Mindedness in the John and Mauskopf (2015) study. Here we find inquisitiveness and exploration, mental flexibility and critical thinking, and the ability to recognize how things are interconnected. These are skills that are critical for developing curiosity, and passion for learning and understanding, which Einstein was so concerned about (see Introduction). Moreover, these kinds of skills are important not only for kids learning in our schools. They also serve critical functions for today's adults that need to adjust to the changing needs of the 21st century, promoting flexible adaptation to changing environments and innovation through learning. Open-mindedness is also an essential ingredient for personal change and growth—it has been linked to unconventional thinking and creativity, helping us to think differently and do things in new and better ways. Finally, open-mindedness can also be applied to oneself and to others: To be *open to the self* means self-reflection, to pay attention to our inner

thoughts and feelings, and to examine our beliefs and actions; these skills correspond to "Self-awareness" that we have seen mentioned in many conceptions of socio-emotional learning, such as Durlak et al.'s (2011) that have been discussed above. To be *open to others* means to appreciate how other people are unique and different, to be able to see their diversity, and to learn to respect differences in backgrounds, perspectives, and cultures; these skills have often been referred to as "Tolerance"—being able and willing to overcome stereotypes and prejudice against individuals who are different in some way (e.g., kids with disabilities).

Table 1

21st Century Skills Associated with Open-Mindedness in the Big Five: Einstein's "Open Your Mind" Domain

_____

Inquisitiveness: Is inquisitive, inquiring; interested in learning new things

Exploration: Likes to explore, discover new things

Mental flexibility: Is intellectually flexible; thinks about things from many different perspectives

Critical thinking: Thinks critically, is not satisfied with superficial, "quick and easy" explanations

Interconnectedness: Recognizes that things are interconnected; what I do affects others and vice versa

Wonder: Has a sense of wonder (for nature, history, or ideas)

Creativity: Is creative; comes up with new ways of doing things

Individualism: Is individualistic, unconventional, unique; doesn't mind being different

Self-reflection: Is self-reflective; examines their motives, actions, and feelings

Diversity: Appreciates diversity; respects differences in backgrounds, perspectives, and cultures

_____

Would the skills summarized under Open-mindedness be sufficient, by themselves, for learning in school and college, and for building a successful and satisfying life and work? Skills from the Open-mindedness domain are highly developed

in scientists (like Einstein) and in thinkers, writers, and artists. However, they are also found in dreamers and idealists who never get anything done. The kinds of skills that help us plan a task or job, get organized, stay focused, and persevere until the task is done are rather different from Open-mindedness—they all involve *Self-management.* They form their own, separate factor, and that factor is illustrated in Table 2. Again, we give examples of the major 21st century skills that have been linked to Self-management.

Table 2

21st Century Skills Associated with Self-Management in the Big Five: Even Geniuses Need to Work Hard

_____

Productivity: Is productive; works hard and gets a lot done

Reliability: Is reliable, dependable; can be relied upon

Execution: Executes plans, follows through

Persistence: Sticks with things; is steadfast and single-minded in getting things done; persistent

Self-control: Shows self-control at work or school; keeps impulses and distractions at bay when things need to get done

Results orientation: Is results-oriented; prioritizes getting things done

Focus: Has a laser-like focus; can keep full attention on a single task, goal, or project

Effort: Puts great effort into their work; pushes self to achieve excellence

Grit: Shows a lot of grit, passion and perseverance for long-term goals

High standards: Has high standards for themselves; always wants to do their best

_____

Consider the 21st century skills shown in Table 2: they illustrate that Self-management involves what psychologists call "the self-regulation system", which helps us focus our attention and behavior on the important things when we need to meet standards. This function is similar to what neuroscientists call "executive control"— these are skills that help the individual *initiate, coordinate, monitor, and complete complex, long-term, and goal-directed behavior.* Finishing a university education (e.g., a bachelor degree) is a good example, as it requires students to perform all these

executive functions in a coordinated way over several years in order to reach their goal of earning a diploma.

The 3rd domain involves skills that help people successfully regulate their emotions, especially negative emotions. Negative emotions occur when we are experiencing "bad" events: when we fail at an exam or a task and feel anxious and upset; when we have a fight with our best friend and feel frustrated and angry; when a close friends moves to another state and we are left behind, feeling sad and lonely. But human beings are not passive victims of their "passions" (as we used to call emotions) but can actively regulate their feelings. Well-developed emotion regulation skills are critical for the individual to maintain an equilibrium of life experience and evaluation of the self. With good emotion-regulation skills, stress, frustration, and upsetting circumstances do not lead to extreme responses; after a negative experience, the individual is able to quickly bounce back and return to emotional balance and equanimity. As shown in Table 3, these skills include self-confidence (defined as having positive thoughts about oneself, and not ruminating or obsessing about failures, disappointments, or set-backs) and also optimism (having positive expectations about the future outcomes in one's life). These skills will help buffer the individual from negative experiences and thus lead to a more positive general self-evaluation.

Table 3

21st Century Skills Associated with Positive Self-Evaluation: Successful Regulation of Stress and Negative Emotion

_____

Equanimity: Shows equanimity, emotional composure, calm

Optimism: Has optimistic expectations about own future outcomes; has a "can-do" mind-set

Tranquility: Is tranquil; centered, not easily flustered

Self-confidence: Has positive thoughts about self; does not ruminate or obsess about failures, disappointments, or set-backs

Balance: Is a very stable person; returns to their center quickly, even after intense emotion

Serenity: Is serene, at peace

Composure: Keeps things in perspective; able to maintain composure and even temper in difficult situations

Patience: Is patient; accepts or tolerates delays and setbacks calmly, without becoming annoyed or anxious

Happiness: Is a happy person; generally in good spirits

Flexibility: Is flexible, not rigid; readily adjusts to changing circumstances

Resilience: Is resilient; able to recover quickly from challenging or difficult experiences

_____

Finally, schools are social places where children need to learn to engage with new people, both teachers and peers. But engaging with others is not enough: They need additional skills to turn these strangers into long-term friends with whom they can share their thoughts and feelings and who will provide them with crucial social support when things are are not going well. These social or interpesonal skills are organized into the 4th and 5th domain, Engagement with Others and Amabilidade.

Table 4

21st Century Skills Associated with Engagement with Others

_____

Engagement: Is engaged in a lot of activities/projects; finds ways to get involved and committed

Social Presence: Has social presence; commands attention and respect

Zest: Has zest for life; approaches things with liveliness and great energy

Passion: Is passionate; is alive with intense emotions; has strong feelings about things

Inspiration: Inspires others; motivates them to take action

Leadership: Is a leader

Courage: Is courageous; dauntless, not easily intimidated

Assertiveness: Is assertive; speaks up for their own rights and for the rights of others

Charisma: Is charismatic (magnetic, charming); captivates the attention of others

Vigor: Is vigorous; strong, robust, and healthy

_____

The examples in Table 4 show that Engagement with Others involves not only being present in the social world but also captures engaging with zest and passion, inspiring others and taking on leadership roles, and using assertiveness not only (or necessarily) in the service of one's own agenda but also to speak up and advocated for the rights and needs of others.

In addition to Engaging with Others, the interpersonal skill domain includes a second dimension (called Amabilidade) that helps students turn their engagements with others into stable, long-term friendships that provide the individual with closeness, connection, and mutual support. Examples of these skills are shown in Table 5.

Table 5

21st Century Skills Associated with Amabilidade

_____

Empathy: Has a lot of empathy; understands other people's feelings well

Team work: Is a team player; works well with others

Citizenship: Is a good citizen; cares about the well-being of their community

Harmony: Values harmony with others

Respect: Shows respect for others

Inclusiveness: Values inclusiveness; makes sure nobody is excluded

Gratitude: Feels a lot of gratitude to others and shows it to them

Generosity: Is generous; happily gives more than is expected or necessary

Humility: Is humble; acts with humility; does not think they are more important than others

Camaraderie: Values camaraderie (spirit of familiarity and trust between friends)

_____

In this interpersonal domain, we see all the core skills for building long-term trust and friendship, including empathy, respect, humility, and gratitude. Also included are skills important for collaboration and working in teams. These skills are important not

only for the individual building good relationships but also for building a community out of many individuals, as shown by citizenship, social harmony, and inclusion.

The importance of these SEMS dimensions in everyday life—during the school years and into adulthood

Back to the Perry Preschool kids in the intervention group: They had better educational outcomes, were more likely to be employed as adults, and less likely to be in prison. What can we now tell about the major socio-emotional skills they learned? Open-Mindedness and conscientious self-management are prime candidates as we know that each of them predicts better performance in school (Poropat, 2009; Poropat, 2014). An early study from 1994 studied kids at age 12, living in poverty in the inner city of Pittsburgh in the United States (John et al., 1994). Those kids who had learned greater self-management skills (such as organization and persistence skills; see Table 2) scored higher on achievement tests for reading, writing, and math; improved achievement was also observed for kids who had mastered higher levels of open-mindedness (e.g., curiosity and interest for learning and understanding things). Over the past 20 years, these findings have been replicated in hundreds of subsequent studies and extended to both younger kids and older adults, and in samples from diverse countries (see Poropat, 2009), including recently in Brazil (Primi & Santos, 2014).

A second important finding is that conscientious self-management is important not only for getting things done as a student in school but also, later on, in work or career. Research shows that successful salespeople, accountants, entrepeneurs, and janitors have one thing in common: they all need to finish their tasks and projects in an efficient, careful, and timely way. Thus, self-management skills like planning, organizing, and concentration are generally useful skills for getting a job, keeping it, and getting paid over longer stretches of time (Duckworth, 2011).

Conscientious self-management also helps keep adolescents and young people out of jail: research has shown the effects of having skills that help us plan and focus on important goals, rather than act impulsively. They are particularly consequential for adolescents and young adults who need to invest energy, effort, and time to build a family and work or career.

Finally, there is reason to suspect that their experiences at the Perry School taught the kids ways to successfully regulate their emotions, and that their positive experiences at school translated into a generally positive, optimistic, and self-confident self-evaluation. This Big Five dimension has been found a strong socio-emotional predictor of satisfaction with life and work, personal well-being, and general psychological health (Hutz & Nunes, 2001).

Table 6 provides a summary of the research on important outcomes that each of the Big Five domains have been shown to predict, both in school and in later life. The first column shows which kinds of tasks and behaviors this Big Five domain helps kids do well. For example, kids with highly developed Openness do particularly well in innovation and creativity tasks at school. The second column shows which kinds of problems and behaviors this Big Five domain helps prevent. For example, kids with highly developed Conscientiousness have fewer problems with paying attention in class, planning their homework, and organizing their time; they also show less truancy (absence from school) and delinquent behavior. The third column shows which kinds of tasks and work this Big Five domain helps people do well after school, during early and middle adulthood. For example, having developed high levels of Engagegment with others helps adults attain success in social jobs (e.g., customer relations; managing people), in entrepeneurial contexts (that require personal initiative and energy), and in terms of having a large social network. Overall, the empirical findings summarized in Table 4 are quite consistent with the results of the Perry School Study.

Table 6

Examples of Important Outcomes that Each of the Big Five Predict in School and in Later Life

Big Five Domain	(1) Do well in	(2) Have fewer problems with	(3) Have particular life success in
Openness	Innovation and creativity tasks; school performance	Prejudice and negative attitudes to outgroups	Investigative and artistic work; political-involved citizenship
Self-Management	School performance and achievement tests; grades	Attention, planning, and organization;	Productive and successful at work;

		truancy; juvenile delinquency	less unemployment; less crime
E	Leaderships tasks; Working & learning in groups	Making new friends, giving presentations, public speaking	Enterprising and social jobs; making social connections
A	Social knowledge and empathy tasks	Aggressive behavior; close relationships; social support	Care-giving jobs; collaboration; stable family life
Emotional Resilience	Confidence and psychological well-being	Worrying, test anxiety, and depression; somatic issues (like headaches)	Stable and satisfying work, marriage, and relationships

Sources: Ozer & Benet-Martinez (2006); George, Helson, & John (2011); Nigg et al. (2002); John et al. (2008).

On the plasticity and malleability of these SEMS dimensions

As mentioned earlier, the results from the Perry study teach us one other important lesson: the kids in the intervention group learned socio-emotional skills at their school but the kids in the control group *did not* learn those skills as much. That is, the kids were not born with these skills, and their parents did not teach them these skills at home. No, they learned these socio-emotional characteristics *in their school program*. In other words, the Perry School results illustrate that there is a considerable amount of plasticity in socio-emotional skills and that schools can be an important agent in teaching these skills.

That insight contrasts starkly with an intuition we often hear from people who do not follow the research. Many people seem to have what is called a "fixed mind-set": they believe that socio-emotional characteristics are "set like plaster"—fixed and unchanging and there isn't much we can do about that.

This belief, however, has been thoroughly disproven by research over the past 25 years. Figure 2 is a good example; it shows how levels of self-management change with age in the United States. These data reflect the natural course of development, with no interventions being administered. The figure shows mean level differences by year of age, in a large sample of adolescents and adults from age 10 to 65. The data show the expected turmoil during adolescence—note the substantial low point in the middle teenage years, around age 15, and the notable increase after that. But change continues, even after age 20! Both women and men continue to increase steadily, even

though men never quite catch up with women. This increase is linear at first and steep, then slows down a bit by age 50 as people enter later adulthood and retirement age when self-management is becoming less important. But, clearly, we still see considerable change in adulthood, even after age 30.

The overall finding shows plasticity, rather than plaster: Individuals are learning and changing throughout life (Soto, & John, 2012; Soto, John, Gosling, & Potter, 2011; Srivastava, John, Gosling, & Potter, 2003). Such findings hold in all kinds of research studies, and it holds up when all the available research findings are combined and analyzed together, a data-summary technique scientists call *meta-analysis*. In fact, the degree of plasticity is much higher for these socio-emotional skills than it is for measures of IQ or for performance on math and verbal achievement tests in school, which are two other constructs of interest, and well studied in educational settings. In other words, consistent with the Perry Preschool Study results, we have a much greater chance to teach and change these socio-emotional skills than IQ scores and that can also be observed from the fact that psychologists have long track records at teaching socio-emotional skills through various kinds of programs, such as teaching organizational skills to kids and adults that have attention problems, and teaching assertiveness and other social skills to kids and adults who suffer from shyness and social withdrawal.

In a more recent meta-analysis, Taylor et al. (2017) reviewed 82 school-based, universal social and emotional learning (SEL) interventions involving 97,406 kindergarten to high school students aiming at evaluating whether SEL interventions that encourage the development of social and emotional assets through school-based interventions yield significant effects at follow-up on multiple positive and negative indicators of well-being. Follow-up outcomes (collected 6 months to 18 years postintervention) demonstrated SEL's enhancement of positive youth development and the benefits were similar regardless of student's race, socioeconomic background, or school location. The positive effects of SEL interventions were found at follow-up for each of the seven outcome categories, namely SEL skills, attitudes, positive social behavior, academic performance, conduct problems, emotional distress, and drug use, with means ESs ranging from .13 to .33.

Requirements for a SEMS assessment tool in Brazil

Retaining the best items loading the core dimensions of the Big Five framework, Primi and colleagues (2016) initially developed an early "Beta" version of the Senna instrument, a 54-item set assessing social-emotional skills in students. As these 54 items were derived from a variety of instruments, they referred to a mixture of trait-related and self-efficacy items describing an individual' standing on five key dimensions of social-emotional functioning.

Feedback from educators, teachers, policy makers and researchers made clear that it was feasible to distinguish between how children *typically* manifest a social-emotional skill from how well children actually think they *can master* a particular skill when required to do so, basically requesting to distinguish between an identity and self-efficacy description of SEMS. How well these two different manifestation modes of SEMS are distinguishable is an empirical question, though worth investigating, given their conceptually distinct nature. In addition, bottom-up feedback also pointed to the necessity to assess SEMS at a more fine-grained level, distinguishing more specific SEMS beyond the broader SEMS dimensions. For both formative and summative assessment purposes, educational practitioners, researchers and policy makers require information of educational level of specific SEMS related to learning, such as for example 'presentation skills', 'planning and organization' or 'self-confidence'.

These two requests formed the starting point of a new series of development and research activities to construct a comprehensive and at the same time manageable set of constructs to assess SEMS in Brazilian students. To make the instrument useful for large-scale assessments, it had to satisfy the following requirements:

Comprehensiveness. Given the variety of educational curricula and the multidimensional nature of several of the objectives stipulated therein, it is necessary to have a broad set of skills represented in the instrument. Education in the 21st century and in Brazil aims to develop the integral student preparing her/him not only for professional challenges but also as an individual and as a contributing member of society. Enclosed skills hence needs to reflect this complexity and assess hybrid curriculum objectives in their constituting parts. The general competences in 3rd version

of the Brazilian common core are examples of how multiple skills need to be merged into the content so integral education is a reality for the students. Although comprehensive, the final skill set needs to be manageable at the same time, because both attention span of respondents and assessment time are usually constrained, so the tool assesses exactly those skills that are (a) represented across a range of curricula, (b) are a common denominator of 21st century skill taxonomies, and (c) are necessary to explain the consequential outcomes envisaged in different curricula.

Specificity. Although the five dimensions assessed by Senna Beta version broadly capture underlying core SEMS qualities providing a parsimonious account of an individual's SEMS, the bandwidth-fidelity trade-off suggests that it is unlikely that this broad level is most appropriate for applied research and evaluative purposes. Senna Beta version dimensions are by necessity broad and may have less fidelity to predict specific outcomes. To better serve formative and summative assessment purposes, we need further specification distinguishing among specific skills structured under these broader functional dimensions. The final version of Senna needs to describe students' skills also at a more fine-grained level of analysis, preferably at both the identity and the self-efficacy levels. More about these levels will be explored when we describe Senna final version in further detail in the next chapter. A key question is hence to define this optimal number of lower-level of 21st century skill constructs, that is comprehensive enough though still manageable in number. The addition of lower-level differentiation brings in extra challenges to have separate SEMS scales that share variance with one or more broader dimensions, though have specificity enough to emerge as a distinct SEMS.

Specific skills can be further combined into skill compounds, that probably better reflect the complex nature of more multidimensional socio-emotional skill constructs reflected in educational curricula. A construct like Grit (Duckworth, Peterson, Matthews, & Kelly, 2007), for example, can be probably best conceptualized as an aggregate of 'achievement striving' and 'concentration/focus' (both related to conscientiousness) with 'energy/drive' that is more associated with the trait of extraversion. Alternative ways to translate complex constructs in basic trait/skill building blocks are described by Shoss and Witt (2013).

Potential to measure change. The final version of Senna scales needs to be used in both cross-sectional and longitudinal formative and summative designs. The instrument hence needs to adequately describe students' standing (identity) and mastery (self-efficacy) of SEMS as well as changes therein, including normative (Roberts, Walton, & Viechtbauer, 2006; Roberts, Wood, & Smith, 2005) and intentional development patterns. The instrument also needs to be able to chart linear and non-linear forms of growth, and to distinguish among latent groups of students following specific developmental trajectories (in response to interventions) beyond normative change patterns (De Fruyt & Van Leeuwen, 2014).

Broad applicability. The final version of Senna needs to be useful for both formative and summative assessment where students representing culturally diverse and socio-economic different classes will have to describe themselves. These requirements necessitate grammatically streamlined and short items, an easily understandable scale format, and clear instructions. Measurement equivalence across these different cultural, age and social-economic groups has to be demonstrated so meaningful comparisons can be made across different groups of students.

Ease of administration and large-scale applicability. Finally, the final version of Senna needs to be easy and flexible to administer, both paper and pencil as well as electronically. The application in large-scale surveys require specific indices to check on the quality and internal validity of the data, and specific guidelines can filter data retaining valid protocols. The electronic platform needs to respect specific requirements to ensure confidentiality of the data and create a stable assessment platform even when internet capacity is poor and volatile.

Chapter 2. Senna Developmental History and Steps

Most of the work on assessing socio-emotional variables has been done in the United States, Northern Europe, and other English-speaking countries, like Australia. Thus, the vast majority of instruments has been developed in English, with a focus on applying the measures to participants who live in Western industrialized nations and are generally well-educated. In contrast, here we aim to develop a measure that can be applied to youth starting at age 10 years in Brazil, including the rich and the poor, the more educated and the less educated. That meant, an initial round of research was needed to test out Brazilian adaptations of English-language instruments and then use the results of that initial round of research as a jumping-off point for the construction of the measure we call Senna. Fortunately, work on socio-emotional characteristics had been under way for a few years, and the developmental work undertaken here could be based on this foundational work (Primi & Santos, 2014; see also Primi et al., 2016).

Initial review and analysis of existing measures of socio-emotional skills

With the support of the Ayrton Senna Institute, our team of researchers (see Primi, Santos, John, & DeFruyt, 2016) set out to address the need for measures that would be appropriate for the Brazilian context and to identify core qualities to be included in a new and comprehensive measure specifically for Brazil. They began with a careful review of measures that assess social and emotional characteristics and either were available in Brazil or could be translated into Portuguese. To guide their systematic search for relevant instruments, they used four key criteria:

(1) *Evidence of predictive power*: means that research has shown that the measures had been shown to be empirically related to at least one measure of an important future outcome, such as well-being, school achievement, drop-out, wages, etc.;

(2) *Feasibility*: Given the constraints in Brazil, the instrument should be low cost, take relatively little time to administer, be easy to understand and respond to, and be usable for self-descriptions without the assistance of a professional psychologist;

(3) *Malleability*: The measured constructs should be malleable during school age, preferably with evidence that formal education can impact their course; and

(4) *Robustness*: The instrument should have good psychometric properties.

Adopting these criteria, the literature review identified eight instruments. They are listed, by year of publication, in Table 5, along with their abbreviations, number and names of their scales, and their total number of items. This overview shows a wide range of scales measuring diverse constructs that are popular in educational, psychological, and economic research on socio-emotional skills, such as self-concept, self-esteem and self-efficacy, motivation, attitudes and control beliefs (locus of control, referring to attributing control internally or externally), social and emotional behavior that may be adaptive or maladaptive, and personality characteristics.

Note that Table 7 shows that the eight instruments included 21 scales, which are measured with *more than 200 unique items*! Clearly, we cannot (and need not) measure each of the individual concepts that are implied by all these scales. As in the work on the structural models in other fields, like intelligence, it seems unlikely that these scales all capture unique information. Instead, they are likely to show considerable overlap because they measure very similar constructs, if not outright redundancy. In general, we expected a small set of core themes to be represented across many of these instruments and their scales.

Table 7

Instruments selected and studied by Primi et al. (2016): Names, Abbreviations, Number of Items and Scales, and Scale Names

Instrument	Items	Scales	Names of scales
Rosenberg Self-Esteem Scale Rosenberg (1965)	10	1	Self-esteem
Locus of Control Scale Nowicki & Strickland (1973)	21	1	Locus of control
Strengths and Difficulties Questionnaire (SDQ) Goodman (1997)	25	5	Emotional symptoms, Conduct problems, Hyperactivity, Peer problems, Prosocial skills

Big Five Inventory (BFI) John et al. (1991)	44	5	Extraversion, Agreeable-ness, Conscientiousness, Negative emotionality, and Open-mindedness
Self-Efficacy Questionnaire for Children (SEQ-C) Muris (2001)	24	3	Academic, Social, and Emotional self-efficacy
Big Five for Children (BF-C) Barbaranelli et al. (2003)	65	5	Negative Affect, Extra-version, Openness, Agreeableness, Conscientiousness
Core Self-Evaluations (CORE) Judge et al. (2003)	12	1	Positive vs. negative self-evaluation (combined self-esteem, generalized self-efficacy, locus of control, low neuroticism)
Grit Scale Duckworth & Quinn (2009)	8	1	Grit
Total	209	21	16

Background on Senna Beta version: Key development steps and foundation

To make these instruments available for use in Brazil, a five-step approach was taken to translate, adapt, and review the content of the instruments, to make sure they were suitable for the Brazilian context. For example, after the initial translation process, a content analysis was performed: Two groups of school administrators from the Rio de Janeiro State and the municipal education systems examined the items and their proposed translations to check their suitability for the target population and evaluate possible negative reactions to particular items by teachers, school directors, and parents (Step 1). To verify that less educated individuals could use the instruments, researchers checked whether participants understood each item and properly used the 5-point rating scale (Step 2). They also evaluated whether the item content was suitable for the relevant age groups and understandable within their sociocultural experience of Brazilians (Step 3).

This careful pilot work was also important for establishing how many items the least-educated respondents could answer without losing interest or attention (Step 4); that research informed our decision to limit the number of items administered in the final version of Senna to less than 170 items; for kids ages 10-12, the number of items should not exceed 90 items. After this extensive preparatory work, the items were

administered to more than 3.000 students who attended one of 86 classes within 16 schools selected to represent typical levels of performance on standardized achievement tests in Rio de Janeiro State (Step 5). Statistical analyses yielded clear evidence that all of the Big Five dimensions could be recovered from these Brazilian school data (Primi et al., 2016; Primi, Santos, Hauck, De Fruyt, & John, 2019). Specifically, the majority of the 209 items could be empirically aligned and interpreted within the Big Five framework, and all Big Five dimensions were clearly represented within this comprehensive item set.

The findings from this first study served as the starting point to develop a new instrument with groups of internally consistent items that would assess each dimension well. The goals were to achieve both bandwidth and parsimony: providing a comprehensive assessment of the variance identified in the 8 instruments selected by the literature review but to do so with a much smaller, manageable set of items. The final item selection process was conducted by four independent reviewers taking into account all the available information. This process resulted in an instrument of 92 items we called Senna beta version 1 (Primi et al., 2016) To ensure replication and confirm the previous results, the resulting set of candidate items was tested one more time. The final item set was administered to a new sample of more than 24.000 public school students. Confirmatory analyses showed excellent fit of the new data with the structure established in the first study, and the resulting initial Senna scales had good psychometric properties.

Santos and Primi (2014) conducted an important validity check on these initial versions of the Senna scales: Portuguese and Math scores on the state-wide achievement test were available for most of the students, and so they could test whether the previous findings (see Table 8) would also hold in the Brazilian school context. Indeed, both open-mindedness and self-management predicted how much the children had advanced in their learning, closely replicating the international findings.

Table 8.

Correlations of SENNA v1 scales with standardized achievement in Portuguese and

Math from Rio Study (Santos & Primi, 2014)

	portu	math
Open-mindedness	***0.19	***0.10
Conscientiousness	***0.12	***0.09
Extraversion	*0.02	0.00
Agreeable-ness	***0.10	***0.06
Negative emotionality	-0.01	***-0.05
Locus of Control Negative Beliefs	***-0.16	***-0.07

Note. * p < .05, *** p < .001

Three new directions for Senna final version

Given this successful scale construction project and the demonstration of the expected link with school performance, we decided to further extend the Senna beta version approach in several different directions:

(a) Measuring at two conceptual levels, namely both *Self-Identity* and *Self-Efficacy*;

(b) Measuring multiple, more specific facets within each of the five broad socio-emotional domains; and

(c) Measuring in novel ways to control for potential response tendencies and biases in the ways children report about their self-identity and self-efficacy.

First, we had found that most of the existing instruments measured the five initial versions of the Senna dimensions in terms of *Self-Identity* statements—items that express possible answers to the "Who am I" question that students are beginning to ask themselves in middle and late childhood. For example, the identity statement "**I am** a dedicated and hardworking student" indicates that, in the domain of Conscientious Self-management, this student has developed and internalized a general view of himself or herself as "dedicated and hardworking" in the way they typically approach their school work.

However, we also found that some *self-efficacy* statements had been used to assess this domain (often labeled as "academic self-efficacy"); these statements

included "I am able to complete all my homework" or "I am good at concentrating on the project I am working on and not getting distracted by other things," which assess the *self-perceived capability* of the student in the Conscientious Self-management domain. Self-efficacy items did exist in the literature also for Emotional Resilience (called Emotional Self-efficacy) and for Engagement with Others (called Social Self-efficacy). However, such Self-efficacy items had never been developed for the Open-mindedness and Amity domains; we set out to develop Self-efficacy items for all the constructs we considered, as the examples in the next chart show.

Table 8

Examples of Identity and Self-efficacy items for Open-mindedness and Amity domains

Domain	Identity items	Self-efficacy items
Amity	I get worried with what happens to others	How well do you manage to understand what others are feeling?
	I apologise to people I hurt	How well and respectfully do you treat people you don't like?
	I trust to leave my things with others	How well can you trust people to what over your things?
Open-Mindedness	I like artistic activities	How much are you able to feel interest for literature and poetry, or expression of feelings?
	I'm original, I have new ideas	How easily do yo u find it to creat artistic things, like a poem?
	I like to learn how things work.	How well can you manage yourself to find out how something works?

Second, for school context, measurement at the broad Big Five level is not specific enough. Soto and John (2017) reviewed the literature on more narrowly defined "facet" concepts that make up each of the Big Five domains, and found evidence for at least 3 important distinctions within each domain. For example, the broad Big Five domain of Open-mindedness consists of three major facets or subdimensions, namely Intellectual Curiosity, Creative Imagination, and Aesthetic Sensitivity. On the basis of that literature review, we developed 17 facets for the final version of Senna. Specifically, we differentiated 3 facets for each of the domains, except for the domain of Conscientious Self-management where we differentiated 5 facets because this domain

has been found so important in the school context (see Table 4). The 17 facets are shown in the next figure below.
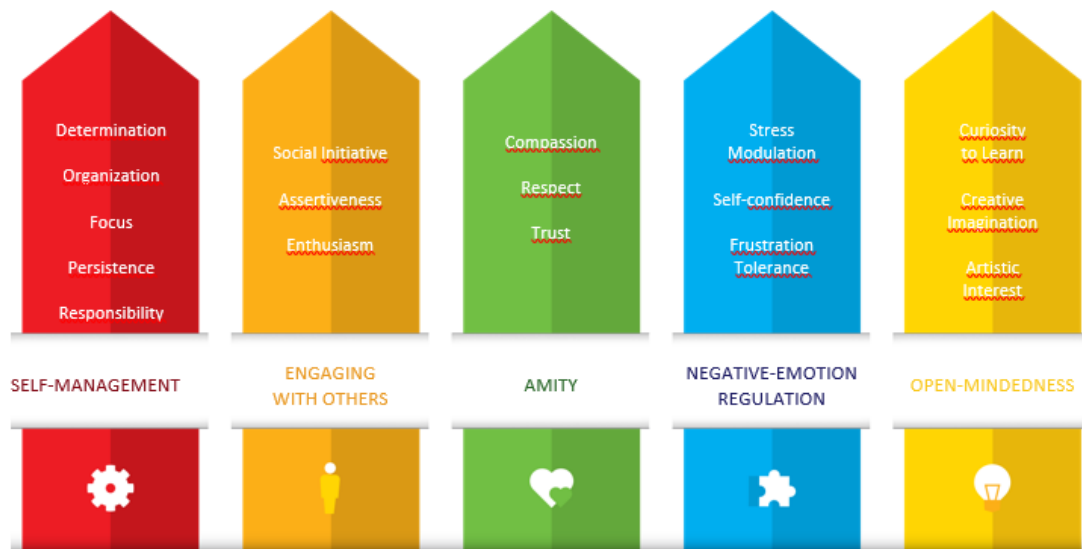
Figure 1. Facets and domains of Senna framework

Third, we developed and tested a series of measurement innovations because self-reports by children and adolescents may show a number of response styles and biases. For example, Soto, John, Gosling and Potter (2008) addressed the concerns that analyses of socio-emotional survey items rarely show the structural validity found in adults and well- educated samples even in the U.S. Examining a large sample of youth ranging from age 10 to 20, they found the underlying cause is great variability in how children and adolescents use the numerical rating scale to indicate how well a particular item describes their identity or self-efficacy. Specifically, on the 1-5 rating scale, some youth would preferentially use the upper or right-hand side of the scale (i.e., use 4 and 5 more often than 1 and 2), whereas others would preferentially use the lower or left-hand side of the scale (i.e., use 1 and 2 more often than 4 and 5). The first response bias is called high acquiescence because these youths are agreeing with more of the items (yeah-saying), where the second response bias is called low acquiescence (nay-saying). Soto et al. (2008) also showed that individual differences in acquiescence bias were most pronounced at age 10 and then decreased substantially (by 50%) all the way to age 18, at which point they reached stable adult levels. Primi, Fruyt, Santos, Antonoplis, & John (2019) replicated these findings with public school students in Brazil and found that the

acquiescence corrections proposed by Soto et al. (2008) improved substantially the structural validity of the responses to our Senna items as well as criterion validity (see also: Primi, Santos, De Fruyt, & John, 2019; Primi, Hauck-Filho, Valentini, Santos, & Falk, 2019, Primi, Hauck-Filho, Valentini & Santos, in press). We therefore carefully developed items sets that would allow us to assess individual differences in acquiescence adopted, and provide scoring and analysis procedures that correct for this important response bias.

In addition, when the Senna survey is used in large school contexts, it is critical to verify that the youth completing the surveys are motivated and engaged with the self-report task and pay close attention to the content of each item (i.e., what they are being asked). If they do not pay attention or even answer the items randomly, their results are not reliable and valid. This response bias has been called *lack of attention* or *random responding*. To assess this response bias, we used a number of attention checks where students are told "This is just check. Simply respond by ticking 4 for this item." Across multiple such attention check items, deviations from the instructed response (here 4) provide a reliable indicator of random responding. Finally, when students complete the Senna survey online, we also assess the response times; implausibly fast response times also indicate problems with random responding and can be used to identify invalid response protocols.

Descriptions of the final version of the Senna facet scales

The names and definitions of the 17 facet scales are shown in Table 8 below. The facets scales are grouped by the broad Big Five domain, starting with Open-mindedness and ending with Emotional Resilience. At the end of each definition, we give some abbreviated item examples in parentheses.

Table 9

Proposed Names for Domains and Facets in Senna-2: Brief Definitions and Item Examples (Nov. 2016)

E: Engaging with Others (vs. Withdrawal and Avoidance) [3 Facets]
1 **Social initiative (or approach/connection)**: Able to approach and connect with others, both friends and strangers, initiating, maintaining, and enjoying social contact and connections; skilled at teamwork,

including expressive communication skills, such as public speaking skills (Is outgoing, comfortable around people)

2 **Assertiveness (or courage/finding your voice)**: Able to speak up, voice opinions, needs, and feelings, and exert social influence; capacity to assert own will to accomplish goals in the face of opposition, such as speaking out, taking a stand, and confronting others if needed; courage (Takes on leadership roles)

3 **Enthusiasm (energy; positive attitude)**: Able to show passion and zest for life; to approach daily tasks with energy, excitement, and a positive attitude (Is full of energy, shows enthusiasm)

A: Amity (vs. Enmity): "Tending and Befriending" Others [4 Facets]

1 **Compassionate caring (empathy)**: Able to use empathy and perspective taking skills to understand the needs and feelings of others, act on that understanding with kindness and consideration of others, and investing in close relationships by helping and providing support and assistance, both material and emotional; is rewarding and easy to deal/live/work with (Considerate and kind to everyone)

2 **Respect for others (politeness)**: Able to treat others with respect and politeness, the way oneself would like to be treated, according to notions of fairness, justice, and tolerance, and keeping aggressive and selfish impulses in check (Is respectful; treats others with respect vs. breaking rules; known for defying teachers)

3 **Trust and forgiveness (appreciation of others)**: Able to assume that others generally have good intentions and forgiving those that have done wrong; avoid being harsh and judgmental, giving people another chance (Assumes the best about people)

4 **Modesty/Gratitude (humility, self-effacement)**: Able to be modest about one's abilities and status in the world, feel gratitude for what we have, and stay in the background, rather than thinking of oneself as extraordinary, better than others, deserving special treatment, and claiming attention for self (I avoid calling attention to myself, vs. I put myself first because I am very special).

N: Negative-Emotion Regulation (or: Emotional Resilience = ER) [3 Facets]

1 **Stress resistance**: Is effective in modulating anxiety and response to stress; untroubled by excessive worry and able to calmly solve problems (Is relaxed, handles stress well)

2 **Self-confidence/Optimism**: Is able to feel satisfied with self and current life, think positive thoughts, and maintain optimistic expectations; anticipates success in actions undertaken; has a "can-do" mind-set; does not ruminate about failures, disappointments, or set-backs (Feels secure, comfortable with self)

3 **Emotional (Temper) control**: Has effective strategies for regulating temper, anger, and irritation; able to maintain tranquillity and equanimity in the face of frustrations; not moody or volatile (Keeps their emotions and temper under control)

C: Conscientious Task Performance (or: Goal Orientation) [5 Facets]

1 **Organization (Orderliness)**: Has organizational skills and meticulous attention to detail that are useful for planning and executing plans to reach longer-term goals (Keeps their school things neat and tidy; not disorganized or messy)

2 **Goal Striving (Ambition; High standards)**: Is able to set goals and high standards, motivate themselves, work very hard (in terms of time and effort), and apply themselves fully to the task, work, or project at hand. This is the pro-active side of **C** (I do more than what is expected of me; I do my work as well as I possibly can; vs. I only need to be in the average; I find it difficult to motivate myself to excel)

3 **Concentration**: Is able to focus attention and concentrate on the current task, and avoid distractions even while performing repetitive tasks (I manage to concentrate on things I do, vs. I don't pay close attention during class and end up forgetting things)

4 **Self-discipline/Persistence**: Is able to overcome obstacles in order to reach important goals; "implement, persist, and finish." The emphasis here is on completing tasks and finishing whatever one has undertaken, rather than procrastinating or giving up. Related concepts are grit, perseverance, and effortful control (I finish my work by the time I have planned to, vs. I leave everything until the last minute)

5 **Sense of Responsibility (Reliability, Dependabilty)**: Has self-management skills needed for doing one's **duty**, meet commitments, act in reliable and consistent ways, and engender trustworthiness; this facet has a secondary link to **A** and should be important for predicting civic involvement and commitment (Is reliable, can always be counted on)

O: Open-Mindedness: Interest and devotion to matters of the mind [3 Facets]
1 **Intellectual curiosity**: Able to muster interest in ideas and a passion for learning, understanding, and intellectual exploration; an *inquisitive* mind-set that facilitates critical thinking and problem solving (Likes to think, play with ideas) 2 **Creative imagination**: Is able to generate novel ways to think about or do things through experimenting, tinkering, learning from failure, insight, and vision (Is original, comes up with new ideas) 3 **Aesthetic sensitivity (Appreciation of aesthetics)**: Valuing, appreciating, and enjoying design, art, and beauty, which may be experienced or expressed in writing, visual and performing arts, music, and other forms of self-actualization (Is fascinated by art, music, or literature)

Outline of the development objectives of Senna final version

One of the advantages of this project was that we already had available knowledge based on the earlier work on the initial Senna scales. First, we had an initial pool of both Identity items and Self-efficacy items. Second, we already had some items for most of the proposed 17 facets; for example, for Engaging with others, we already had some items for the Social Initiative and the Enthusiasm facets in the initial scales but no Assertiveness items. Third, because we needed to develop an acquiescence index and had evidence for the value of false-keyed (or reversed-keyed) items, we knew we needed to generate an item pool with equal numbers of true and false-keyed items. We also knew this would be challenge especially for Open-mindedness and Emotional Resilience because we had learned that writing pairs of true and false-keyed items for these two domains is particularly difficult. For example, such a pair of items for Creative Imagination from Open-mindedness is "Sou original, tenho ideias novas." (true) and "Dificilmente tenho ideias originais." (false). For Frustration Tolerance from Emotional Resilience, we have for instace, "Sou calmo(a) e controlo bem meu estresse" (true) and "Fico muito bravo e costumo perder a paciência" (false).

Study 1 (Ceará 2014 BIB): development of large database psychometric studies facets measures for SENNA

Our goal was to use empirical methods to identify 9 items for each facet scale, that is, a total of more than 150 items. Using the definitions of the 17 facets shown in Table 7. We assembled a large pool of more than 527 candidate items – 92 items from previous version of SENNA and 435 new created items, holding multiple sessions across several days and using the input from with research psychologists, education experts, and former teachers. We created items in two forms – identity and self-efficacy items –

and also from high and low pole of the construct (so called false-keyed items). In appendix A we show shows the distribution of item per a preliminary classification onto domain (five plus negative beliefs and locus of control) and facets, layer and pole (negatively keyed or positively keyed items)

Because we could not administer so many items to the same students, we used a planned missing block design, where each student only completed a subset of the items. This design is also called Balanced Incomplete Block Design (BIBD, van der Linden, Veldkamp, & Carlson, 2004).

We selected the best items from Senna beta version 1 and composed an anchor test form of 54 items based on internal structure and external validity. We called "anchor" meaning that we wanted to investigate the empirical links of the newly created items with the conceptual space anchored on the big five we developed on the previous two studies done in Rio de Janeiro. This anchor version will serve as a conceptual and empirical link of the Rio items database and the ne new items.

We split the 435 new items into two sets of 220 and 215 items. Then each item set was split into eight subtests of 27 or 28 items each. In the end we have a database of 17 subtests. Subtest one is the anchor test and the other 16 subtests splits the new 435 new items evenly. We tried to have items representing all domains and pole in each of these subtests. We then combined these subtests using an optimum BIB design. We used two BIBs one with each big set of items (220/215 items).

BIB is a combinatoric design of tree dimensions. Let's consider the first big set of 220 items subdivided into 8 subtests of 27/28 items plus anchor test of 54 items. One dimension is exactly how many subtests we have, in this case, nine subtests. The second dimension is the number of subtests each subject will respond, in our case three out of the nine possible. The third and last dimension is the number of test booklets, that is, number of test forms assembled from a particular combination of subtests (for instance, 1, 7 and 9 or 4, 5 and 7, and so forth). In our case it was twelve test forms. We could have lot of combinations or test forms that can be made from nine testes picking up tree each time. There are actually 84 combinations possible. A BIB is an optimal group of combinations so as to contain all pairwise combinations of two tests equally balanced.

This condition is the one that determined picking up 12 particular combinations out of 84 combinations possible because they met this condition. With these particular 12 combinations, each subtest appears one time with the other seven subtests in one booklet and each subtest appears four times in the twelve booklets. A detailed view of this design is given in Appendix B.

With this design the number of items in each booklet varied from 77 to 112 with an average of 90 items. So, a student will answer 90 items in average. Booklets were distributed randomly to students in each classroom (spiraling distribution) such as to split our main sample into 24 random subsamples. Because of BIB's optimal combination and a random distribution of subjects we can estimate the full inter-item variance-covariance matrix for the entire set of 527 items.

We administered our first item pool to a large sample of public-school students in Ceará in a collaboration with the Ceará's State Secretary of Education. Our questionnaires were administered in paper and pencil format as part of standardized tests of language and math of Ceará Permanent System of Evaluation of Basic Education [Sistema Permanente de Avaliação da Educação Básica do Ceará – SPAECE]. Sample was comprised of 33,766 students from 5th, 9th and 10th grades. These students were randomly assigned to the 24 booklets. In average 1,406 students answered to each of the 24 booklets (min 1,268, max 1,427).

Because of BIB design we were able to estimate the full inter-item correlation matrix for the 220 items of BIB set 1 and 215 items of BIB set 1. Across BIBs we have a set of 54 common anchor items that provided markers of the six factors from previous version of Senna beta version 1. We examined item psychometric prosperities to select the best items from this database using three procedures related to internal structure validity: (a) item exploratory bi-factor analysis by facets, (b) facets exploratory factor analysis (c) item-total correlation (item-rest correlation) with facet scores.

First, we run a series of exploratory bi-factor analysis by groups of items that theoretically measures the same facet. Our goal was to identify items that were measuring more strongly their intended construct – the general factor – relative to secondary factors - group factors - that eventually could emerge because of specific

nuances shared with other items, or similar words used or because they had strong specific factor than the general factor.

Simultaneously, we run exploratory factor analysis of facet scores. Since each student answered to a sample of the items, calculation of facet raw scores is limited to particular item selection each student was administered. With this method we could not calculate scores for any possible combination of items since we don't have data on full item set for all students. Fortunately, Revelle (2018) produced a function (*scoreItems*) in his *psych* R package that uses basic principle of algebra related to the covariance of sums[1] to calculate reliability and correlations among scales based on inter-item correlation matrix of items that compose each scale. Therefore, since we estimated a full inter-item correlation matrix, we could compose two scales based on any combination of items and calculate the correlation between scales. Using this method, we calculated inter-facet correlation matrices and then run exploratory factor analysis of facet level to investigate if they group according to big five. We run this analysis twice, each time with a big set of items that composed each BIB, with a separate group of items as indicators but same facets, therefore, served as a replication design.

As a third method we also looked at item total correlations of each items with all facets built both convergent and discriminant validity considerations into the item selection process: for an item to be selected for a facet scale, it had to correlate with the sum of the other items from their scale and correlate more strongly with its own scale than with the other facet scales within the Big Five domain.

Using these three procedures we selected 306 items from this first item pool that met the criteria of internal structure validity, that is, items that were a homogeneous measure of the intended facet, correlated more strongly with other facets of the same domain, that is, convergent validity and had relatively smaller correlations with facets of other domains of the big five, that is, had discriminant validity. We also selected items measuring the positive and negative pole of the construct as well as identity and self-

---

[1] The principle of covariance of sums (or correlation of sums) asserts that, for instance, the covariance of two scales A and B, for example, composed of two variables each: $S_A = X_1 + X_2$ with $S_B = X_3 + X_4$ is equal to the sum of covariances among items that compose the two scales: $COV(S_A, S_B) = COV(X_1, X_3) + COV(X_1, X_4) + COV(X_2, X_3) + COV(X_2, X_4)$. See Ferguson and Takane (1989) for a very simple and clear explanation (p. 492).

efficacy items for all facets. Among these items we kept the 54 anchor items that wer

also part of Senna version 1.

Study 2 (Ceará 2015 Dress-rehearsal test-retest): Psychometric study of a balanced set of 306 items database for measuring 18 facets of SENNA

For study 2 we had the best 306 items selected from Study 1 for measuring 18 facets of SENNA described in Table 8. We had 12 identity items (12 X 18 = 216) and 5 self-efficacy items (5 X 12 = 90) per facet. We subdivided 216 identity items into five subtests of 43/44 items and 90 self-efficacy items into two subtests of 45 items each totaling seven subtests. We organized subtests using a BIB (7 X 3 X 7) of seven booklets of three subtests each that balances the pairwise combination of seven subtests. With this design each student answered about 131/133 items and we could estimate the full inter-item correlation matrices.

At this time, we developed a computerized web application that administered SENNA test to students. In this second study 5,485 high school students answered to one of the seven booklets (average of 783 students answered each booklet). They were attending 10^{Th} to 12^{Th} grades (N= 1,740/826/508 respectively), aged 14 to 21 years old (M=16.3, SD=1.1) and 53.9% were girls. One week later 3,012 students answered the same booklet a second time. This database was used to estimate test-retest reliabilities for the final SENNA scales. The design also make it possible to calculate the full inter item correlation within and between time points. A detailed view of this design and item database is given in Appendix C.

We followed the same rational as before when selecting items. We looked to (a) exploratory factor analysis of facets (b) item-total correlation (item-rest correlation) with facet scores. We then selected the best nine items for each the eighteen facet scales of SENNA. As described above, we included pairs of true and false-keyed Identity items (thus an equal number of both); for the Self-efficacy items, we could only use true-keyed items asked students "how well" they could perform or execute each socio-emotional skill (i.e., it doesn't make sense to ask them about the lack or absence of the skill). Thus, because we needed pairs of true and false-keyed items to assess individual differences in response scale use, the acquiescence index is based only on the Identity

items. Six items were for identity scales (3 true keyed items and 3 false keyed items) and three items for self-efficacy scales. To measure random responding, we also included 3-6 attention check items. This composed the 162-final version of SENNA v2.0 that we will be the focus of validity and reliability studies of chapter three.

Chapter 3: Reliability, validity and Norms for SENNA v2.0

This chapter presents a series of studies of the final version of SENNA 162 items selected form dress rehearsal study 2 presented in chapter 2. This third study was run in November of 2015 and administered SENNA via web platform in a sample of 50,209 students who were attending grades 6th to 12th (6th N=8,117, 7th N=5,370, 8th N=6,672, 9th N=7,664, 10th N=8,216, 11th N=7,124, 12th N=6,986), 52.7% were girls, ranged in age from 11 to 28 years (M=14.9, SD=2.1). This sample were well distributed on the state of São Paulo, they were from 234 cities and 501 schools within these cities.

All data were collected in the course of social-emotional skill assessments conducted by researchers from Edulab21 (Ayrton Senna Institute) as part of a reading program run by educational department of the institute. We also had information on measured students' academic achievement with a standardized test (SARESP) for language and for math earlier in 2015. The scores for each student were provided by the Secretariat of Education of the State of São Paulo, which administers these measures of school performance as part of their regular assessment cycle. As expected, the language and math scores were positively correlated (r = .67). Details can be found here: http://file.fde.sp.gov.br/saresp/saresp2015/Arquivos/SE_2015_online.pdf.

Most studies presented here were run on this São Paulo sample. But some studies were based in different samples. We will inform when it is the case.

Methodological Innovations Included in the SENNA Questionnaire: Acquiescence and quality control indicators

The approach taken in developing the SENNA questionnaire was informed by recent methodological advances in research on large-scale socio-emotional assessments conducted in public schools. When students are asked to self-report on their own socio-emotional characteristics and efficacy in a classroom setting, not all of them will be motivated to pay attention to the task at hand and some may struggle to follow instructions, due to problems with reading and understanding the verbal materials presented. Moreover, students differ in the way they use the multi-step Likert-type response scale they have to use to record their responses. These issues of motivation,

capability, and response styles in self-reports have been addressed by vigorous reasearch during the last decade, and some of this research has been conducted by the developers of the SENNA questionnaire (see Abrahams et al., 2019).

Indeed, large-scale international assessments, such as the Programme for International Student Assessment (PISA), have started to pay attention to these measurement issues as they have started to include students' social-emotional skills among their assessment batteries (OECD, 2013). This is an important trend in educational research and assessment, shifting from a primary focus on purely *cognitive* abilities to a growing concern about students' *social-emotional* skills.

In developing the SENNA questionnaire, we have included three major recent methodological advances. The first addresses a long-standing and well-known bias in all self-report measures: namely, in reporting on their socio-emotional skills, students differ substantially from each other in how they use the low and high sides of a typical Likert rating scale, an issue captured by the concept of *individual differences in acquiescence*. All the analyses reported in this chapter have been corrected for individual differences in acquiescence.

Correction for acquiescence. Acquiescent responding is an individual's general tendency to consistently agree (yea-saying) or disagree (nay-saying) with questionnaire items, regardless of their content, and its effects have long been recognized in adults (e.g., Cronbach, 1946; Jackson & Messick, 1958). When left uncontrolled, individual differences in acquiescence tend to positively bias interitem correlations, increasing the (positive) correlations between two items keyed in the same direction but decreasing the (negative) correlations between items keyed in the opposite direction (one true and one false keyed).

Whereas the existence of acquiescence effects has long been recognized in adults, Soto, John, Gosling, and Potter (2008) studied acquiescence effects in children, expecting acquiescence to be one of the culprits responsible for the much less clear factor structures typically found in younger respondents. Indeed, individual differences in acquiescence were much more pronounced in children, decreasing linearly and substantially (by about half) from age 10 to the typical adult level at age 19. Even more

important, acquiescence variance seriously distorted the factor structures of personality self-ratings in the youngest children; at age 10, for example, the standard Big Five structure could not be recovered in the raw data but did emerge when individual differences in acquiescence were controlled. These findings for children have been extended to adults with lower educational attainment; for example, Rammstedt and Farmer (2013) found that, like the earlier findings for children, the expected Big Five structure could be obtained for low educational-attainment groups only after correcting for individual differences in acquiescence. These findings have led to a renewed interest in understanding the causes and consequences of acquiescence, and researchers have suggested different approaches to correcting individuals' responses to rating scales, such as within-person centering around an acquiescence index (Soto et al., 2008; see also Rammstedt, Danner, & Bosnjak, 2017), using anchoring vignettes (see below; Mõttus et al., 2012), and applying statistical corrections (Buckley, 2009). Here we have used within-person centering around the acquiescence index formed by the True and False scored SENNA items, as recommended originally by Soto et al. (2008) and further elaborated by Primi et al. (2020).

Anchoring vignettes. In our initial studies developing SENNA, we also made use of a new technique called anchoring vignettes (e.g., Primi, Zanon, Santos, De Fruyt, & John, 2016). This work is based on the recognition that respondents from different cultural backgrounds differ in the way they define the high and low points in Likert-type rating scales. Using simple Likert scales, studies using the PISA 2003 data found a counterintuitive pattern of findings in which correlations between noncognitive factors (such as conscientiousness) and actual cognitive achievement measures were positive at the within-country level (as expected), but unexpectedly negative at the between-country level (Bertling, Borgonovi, & Almonte, 2016). Kyllonen and Bertling (2013) and He and Van de Vijver (2016) suggested that the cross-country comparability of students' self-descriptions on Likert scales could be improved by rescaling students' ratings based on their responses to so-called anchoring vignettes.

Anchoring vignettes serve the goal of taking individuals' differential interpretation and use of rating scales into account, and may help correct various response styles (Primi, Zanon, Santos, De Fruyt, & John, 2016). Using the anchoring

vignette technique, individuals rate the behavior or traits of a person in a hypothetical scenario using the same response format and on the same rating scale as they later rate their own behavior or traits (for an example see Primi, Zanon, et al., 2016). Because all individuals rate the same vignette, variability among individuals in their responses can be attributed to interindividual variability in scale use (Vonkova, Bendl, & Papajoanu, 2017). Individuals' self-ratings can then be rescaled based on their ratings on the anchoring vignettes.

Following Primi, Zanon, et al.'s (2016) findings with students in Brazilian public schools, we used anchoring vignettes to identify students whose ratings of the vignettes did not follow the expected normative pattern, suggesting these students did not understand or pay close attention to the task at hand. Indeed, when the responses of these students were omitted, the reliability and factorial validity of the responses improved markedly. In other words, in the SENNA questionnaire, we advocate using anchoring vignettes as a way to identify random or otherwise inconsistent response protocols.

Response Consistency Checks. We have also developed a direct approach to identify random or otherwise inconsistent response protocols: we typically add to the SENNA questionnaire 8-10 questions that have pre-determined answers. These items do not ask a substantive question about socio-emotional skills (e.g., How well can you tell a funny story to your class colleagues?) but instead instruct the student to give a particular response. For example, one of the items to check response consistency reads "This is a consistency check. For this item, please check the number 5." These items allow us to compute an overall consistency index; students who don't actually read the item content but respond randomly to the SENNA items will score low on this consistency index (i.e., get a lot of these consistency check items wrong). Our research has shown that these students tend to also show other problems with reliability and their responses should be set aside. The analyses reported in this chapter have been screened for response consistency.

Validity evidence based on Internal structure: factor analysis

Validity is the extent to which the measures truly capture what we are trying to measure (see AERA, APA, & NCME, 2014). It refers to the extent that the interpretation of the test results find empirical support. There are several ways to gatther vailidy evidence abou the interpretations intended from test scores results. This section adresses a core validity question about the final item set that composed SENNA. It addresses whether relationships among facets organizes according to the theoretical expectation that each group of facets measure primarily one of the big five domains. In order to test this structure, we fitted a five-factor model with target rotation using Exploratory Structural Equation Modeling (ESEM) using MPLUS. This rotation aims at maximizing the confirmatory approach in ESEM and allows us to define all the relationships among the factors and the items, providing greater control when specifying the model (Marsh, Nagengast, & Morin, 2014). Indicators for the five factors were 54 sub facet scores calculated from 3 items each of: 18 facet scores X 3 keying directions plus self-efficacy combination (3 identity true-keyed items, 3 identity false-keyed items and 3 self-efficacy items). Before calculating sub-facet scores, we recoded items raw scores controlling for acquiescence.

Table 10 shows the results of this analysis. The first three columns show domains, facets, layer (identity and self-efficacy) and pole (T: true-keyed and F: false-keyed). The last five columns show the factor loadings. Note that for this analysis, false-keyed sub facets were reversed before the analysis.

We can see 50 out of 54 scales have its highest loadings on the expected domain and 40 out of 54 were higher than .40. Factor loadings formed a five-factor structure, as expected. Mainly, all facets loaded on the expected domain, mostly above .40. When analyzing secondary loadings, we can find meaningful patterns. Those can be reasonably explained when looking for the item contents and this pattern was replicated previously studies (Soto & John, 2016). The only facet that were not recovered was trust.

Table 10.

Factor loadings of Exploratory Structural Equation Modeling of 54 sub facet scores of 3 items each from 18 facet scores X 3 layer plus self-efficacy).

Domain	facet	Layer	Pole	O	C	E	A	N
O: Openness to the new	**Artistic interest**	Ident.	F	**0.426**	0.079	0.046	*0.316*	-0.132
		Ident.	T	**0.589**	-0.062	-0.038	0.123	-0.016
		Self-eff.	T	*0.724*	0	-0.175	-0.033	0.128
	Creative imagination	Ident.	F	**0.473**	0.029	0.17	0.106	-0.137
		Ident.	T	**0.703**	-0.04	0.15	-0.1	-0.05
		Self-eff.	T	*0.812*	0.021	-0.006	-0.11	0.121
	Curiosity to learn	Ident.	F	0.351	0.021	0.107	*0.263*	-0.177
		Ident.	T	**0.448**	0.066	0.143	0.183	-0.114
		Self-eff.	T	*0.562*	0.169	0.049	0.07	0.123
C: Self Management	**Determination**	Ident.	F	0.149	0.36	0.106	0.086	-0.132
		Ident.	T	0.018	**0.797**	-0.011	0.006	-0.079
		Self-eff.	T	0.168	**0.493**	0.148	0.069	0.132
	Focus	Ident.	F	0.205	**0.524**	-0.103	-0.047	0.022
		Ident.	T	0.078	**0.56**	-0.046	0.032	0.046
		Self-eff.	T	0.182	**0.574**	-0.11	-0.059	0.253
	Organization	Ident.	F	-0.149	**0.727**	-0.027	0.01	0.003
		Ident.	T	-0.139	**0.867**	-0.031	-0.014	-0.02
		Self-eff.	T	-0.051	**0.75**	-0.063	-0.047	0.183
	Persistence	Ident.	F	0.013	**0.716**	0.016	0.103	-0.191
		Ident.	T	0.007	**0.673**	0.124	0.033	-0.046
		Self-eff.	T	0.165	**0.68**	-0.074	-0.063	0.099
	Responsibility	Ident.	F	-0.008	**0.459**	0.173	*0.214*	-0.126
		Ident.	T	-0.042	**0.53**	0.123	*0.265*	-0.1
		Self-eff.	T	0.029	**0.456**	0.037	*0.268*	0.085
E: Engaging with Others	**Enthusiasm**	Ident.	F	0.079	-0.074	**0.534**	0.052	-0.041
		Ident.	T	-0.017	0.063	**0.584**	0.06	*0.221*
		Self-eff.	T	0.08	0.236	0.3	0.137	*0.322*
	Assertiveness	Ident.	F	0.195	0.126	**0.445**	-0.207	0.01
		Ident.	T	0.261	0.139	0.351	-0.188	-0.123
		Self-eff.	T	0.189	*0.381*	0.191	-0.13	0.179
	Social Iniciative	Ident.	F	0.052	-0.181	**0.607**	-0.073	0.023
		Ident.	T	-0.05	-0.078	**0.649**	0.223	0.014
		Self-eff.	T	0.059	0.072	0.316	0.05	*0.259*

Domain	facet	Layer	Pole	O	C	E	A	N
A: Amity	**Empathy**	Ident.	F	0.148	0.074	0.215	**0.451**	-0.141
		Ident.	T	0.053	-0.022	0.217	**0.54**	-0.013
		Self-eff.	T	0.245	0.039	0.184	0.388	-0.006
	Modesty	Ident.	F	0.036	0.064	-0.109	**0.615**	-0.083
		Ident.	T	0.05	-0.093	-0.199	0.346	-0.045
		Self-eff.	T	0.187	0.124	-0.175	0.318	0.249
	Respect	Ident.	F	0.031	*0.299*	-0.216	**0.457**	0.138
		Ident.	T	-0.016	*0.366*	-0.025	**0.508**	-0.017
		Self-eff.	T	0.136	0.166	-0.168	**0.437**	*0.353*
	Trust	Ident.	F	-0.011	-0.007	0.128	0.087	*0.267*
		Ident.	T	-0.002	-0.117	0.202	0.326	0.187
		Self-eff.	T	0.069	-0.129	0.123	0.173	*0.347*
N: Negative Emotional Regulation	**Tolerance to frustration**	Ident.	F	0.046	-0.041	-0.023	*0.176*	**0.586**
		Ident.	T	-0.036	-0.06	-0.081	*0.137*	**0.527**
		Self-eff.	T	0.071	-0.074	-0.124	*0.205*	***0.747***
	Tolerance to stress	Ident.	F	0.031	0.085	0.193	-0.189	**0.444**
		Ident.	T	0.046	0.102	0.214	-0.033	0.397
		Self-eff.	T	0.11	0.034	-0.028	0.033	***0.739***
	Self-confidence	Ident.	F	-0.127	0.238	*0.294*	0.009	0.349
		Ident.	T	-0.068	0.134	*0.291*	0.126	0.3
		Self-eff.	T	0.068	0.169	*0.254*	0.09	***0.496***

Note. Fit indices were $\chi^2(1171) = 78{,}888$; *CFI* = 0,74; TLI = 0.68; *RMSEA* = 0,07 and *SRMR* = .04.

We computed congruency coefficients where loadings from factor analysis were correlated with vectors of theoretically perfect loadings (i.e., each facet loads entirely and exclusively on its expected factor with loading of 1 and 0 on the other factors). Each row of Table 11 shows results of empirical loadings. Each column of Table 11 contains the theoretical vectors. Each cell of the table shows the correlation of empirical loading vector with the theoretical perfect vector. For instance, there was a congruency of .89 of the factor loadings of an empirical factor we called O and theoretical vector of O. This same empirical loadings vectors had a congruency of .08 with theoretical vector of C and so on.

Each empirically observed factor shows substantial and clear congruence with its expected theoretical factor, and the congruence coefficients were above .80 for O, C and N and above .78 for A and E. Appendix D shows additional exploratory factor analysis of identity versus self-efficacy scales.

Table 11.

Congruence Coefficients of Five Domains of ESEM loadings versus theoretically perfect loadings

Empirical loadings	Expected perfect loadings				
	O	**C**	**E**	**A**	**N**
O	**0.89**	0.08	0.15	0.14	0.02
C	0.04	**0.92**	0.09	0.09	0.08
E	0.09	0.04	**0.78**	0.03	0.19
A	0.16	0.13	-0.01	**0.79**	0.11
N	-0.04	0.02	0.15	0.19	**0.80**

Reliability of SENNA facet and domain scales: Internal Consistency and Stability

Reliability refers to the degree of which the score on a measure is affected by error of measurement. A reliability index will tell us how much a person's scores will change when we have another opportunity to measure it presuming that persons haven't changed in the measured attribute. One way to estimate reliability index assess how consistent responses are to sligtly different questions measuring the same underlying factor. This index is called Cronbach Alpha (α) coefficient. In an unlike situation with no error internal consistency index would be 1; in another unlike situation if there is only error of measurementm the value would be zero. Of course, in real life situations values fall somewhere in between. Another way to estimate reliability indexes is when we measure the same people twice in the same test with a short interval such as that thay don't have had a chance to change in the atribute measured. When we estimate reliability index in this way it is called test-retest reliability or stability coefficient.

According to the classical benchmarks, values of .90 or greather are excelent, between .80 and .90 are very good, between .70 and .80 are acceptable and values between .60 and .70 are considered adequate for very short scales (John & Benet-Martínez, 2000). Table 12 shows internal consistency for SENNA facet and domain scales broken down by the age of students from 11 to 18 years old. Indexes were calculated from items scores after controling for acquiescence.

We expected SENNA´s scales to demonstrate adequate to very good internal consistency although this may vary through ages. Since within-domain coherence improves with age we expected they will be relatively lower in younger samples and increase with age (see developmental psychometrics Soto, John, Gosling, & Potter, 2008).

Table 12.

Internal consistency for acquiescence-controlled scores of SENNA by age.

Domain (in bold) and facet scales	Ages (years)							
	11	12	13	14	15	16	17	18
Open-mindedness (O)	**.87**	**.88**	**.89**	**.90**	**.90**	**.91**	**.90**	**.91**
Artistic interest	.77	.79	.81	.82	.83	.84	.85	.85
Creative imagination	.65	.67	.70	.73	.75	.76	.76	.77
Curiosity to learn	.74	.76	.79	.79	.80	.81	.80	.82
Self-management (C)	**.94**	**.94**	**.95**	**.95**	**.95**	**.95**	**.95**	**.94**
Determination	.80	.80	.82	.82	.82	.82	.83	.83
Focus	.68	.70	.74	.75	.78	.78	.78	.77
Order	.87	.88	.89	.89	.89	.89	.89	.89
Persistence	.78	.79	.81	.81	.82	.81	.82	.80
Responsibility	.79	.81	.81	.82	.82	.82	.83	.82
Engaging with others (E)	**.71**	**.76**	**.79**	**.81**	**.83**	**.84**	**.85**	**.86**
Enthusiasm	.51	.56	.62	.66	.69	.71	.72	.73
Assertiveness	.63	.65	.71	.74	.76	.78	.78	.79
Social initiative	.48	.53	.60	.63	.65	.67	.69	.71
Amity (A)	**.85**	**.83**	**.85**	**.85**	**.85**	**.85**	**.86**	**.86**
Compassion	.80	.79	.81	.81	.81	.80	.80	.80
Modesty	.64	.66	.70	.71	.71	.72	.74	.74
Respect	.76	.75	.78	.80	.80	.81	.82	.83
Trust	.55	.51	.54	.58	.58	.59	.61	.60
Negativa Emotion regulation (N)	**.83**	**.86**	**.88**	**.89**	**.89**	**.90**	**.90**	**.90**
Frustration tolerance	.74	.76	.80	.82	.82	.84	.84	.85
Stress modulation	.55	.66	.72	.74	.75	.77	.78	.78

Self-confidence	.63	.66	.70	.72	.73	.76	.77	.78

Most of internal consistency indexes were in the expected range of very good to acceptable values. Some scales had excelent reliability indexes. Domain scales values ranged from .71 to .95. Lowest values were seen in the lower ages as expected. Specifically, the domain Engaging with others have shown the highest improvement as the age increases, whereas the other domains did not present much variation among ages. Similar results can be seen with some facets Modesty, Social initiative and Enthusiasm/Energy Self-confidence.

Table 13 shows the stability coefficients (test-retest reliability index). These correlations were calculated based on the database of Study 2 Ceará 2015 where 3,012 students took SENNA twice within 15 days interval. Remember that this study had a large item database presented to students via booklets using BIB design. Therefore, the correlations presented in Table 3 were calculated from 162 final selected SENNA items inter-item correlations matrix across two time points. We used the principle of correlations of sums to calculate correlation among scales (facets and domains) between time 1 and 2.

All five domains approached very good levels of reliability (r ≥ .85). Facets had lower values since they were composed of less items but still achieved acceptable to very good values (the only exception was curiosity to learn).

Table 13.

Stability coefficients (test-retest reliability) of SENNA scales (Ceará Study 2 2015 sample)

Domain (in bold) and facet scales	Test-retest
Open-mindedness	**.85**
Artistic interest	.79
Creative imagination	.82
Curiosity to learn	.67
Self-management	**.89**
Determination	.83
Focus	.78
Order	.82
Persistence	.81

Responsibility	.77
Engaging with others	**.90**
Enthusiasm	.75
Assertiveness	.87
Social initiative	.85
Amity	**.85**
Compassion	.77
Modesty	.69
Respect	.78
Trust	.81
Negative Emotion regulation	**.86**
Frustration tolerance	.79
Stress resistance	.76
Self-confidence	.76

Validity evidence based on the relationship with external variables: standardized achievement

One important question related to the view of socio emotional skills as influencing important socioeconomic outcomes throughout the individual's life is whether they are associated to learning at school. So, another core validity question we addressed in this section is whether SENNA socio-emotional skills measures predict standardized achievement scores for Portuguese language and for Math.

For this study we merged data standardized achievement test (SARESP) for language and for math in 2015. The scores for each student were provided by the Secretariat of Education of the State of São Paulo, which administers these measures of school performance as part of their regular assessment cycle[2]. Test were administered one months earlier of the SENNA data collection. As expected, the language and math scores were positively correlated (r = .67). We matched scores from 12,987 students (52.7% female) from 425 public schools located in 216 cities in the State of São Paulo in Brazil. They attended grades 7 (N=840), 9 (N=6,474) or 10 (N=5,673) and ranged in age from 12 to 20 years (M=16, SD=1.85). We present descriptive statistics of domain and facet scores in Appendix F.

---

[2] Details can be found here: http://file.fde.sp.gov.br/saresp/saresp2015/Arquivos/SE_2015_online.pdf.

Poropat's (2009) meta-analysis shows that self-management (conscientiousness) is the construct most related to academic achievement across all levels of education (ρ = .24 and ρ = .28 in primary education). It indicates personal capabilities such as perseverance, discipline, endeavor and responsibility, that is, important characteristics in activity that involves medium and long-term commitment, such as study or work. Openness also shows relatively high correlation with achievement (ρ = .12 and ρ = .24 in primary education). It indicates students' desire to learn, imagination, and also a tendency to have a more creative process for daily activities (see also Kaufman, 2013). Finally, Amity is related with achievement specially in primary education (ρ = .07 and ρ = .30 in primary education). It refers to personal characteristic of cooperation with teachers and principals' instructions and quality of interpersonal relationships with other students. Our first question with this study was: Can we replicate the links found by this meta analysis?

Poropat's (2009) meta-analysis investigated broad big five domains. With SENNA scales we could move a step forward by analyzing what facets are related to academic achievement in addition of broad domains. A high score is self-management, for instance, can be a result of very different profiles in terms of its constituents' facets. Our main goal from evolving SENNA v1 to v2 is that facet scales will provide more nuanced information about students or larger amounts of specific personality information (Soto & John, 2017), therefore, it should be a better predictor of external variables such as achievement. So, our additional questions in this study were: Does measuring facets improve predictive validity over using only domain scores? Are facets of socioemotional skills a better way to understand how academic achievement is affected compared to the role of socioemotional domains? And which of these facets may play more decisive role?

We run two analysis: (a) first we calculated zero order correlations among SENNA scales (domain /facets) and standardized scores on Language and Math; (b) second, we run multiple regressions by domain, one time predicting language, and a second time predicting math from facet scores. Table 14 shows a synthetic result of this analysis.

Table 14.

Criterion validities: correlations, and unique contributions from multiple regression of domain and facets predicting language and math.

	Language			Math		
	r	*β*	*R²*	*r*	*β*	*R²*
Open-mindedness	**0,24**	**0,24**	**0,057**	**0,18**	**0,18**	**0,032**
Artistic interest	0,15	-0,03	0,087	0,09	-0,06	0,057
Creative imagination	0,18	*0,01*		0,14	0	
Curiosity to learn	0,29	0,30		0,23	0,25	
Self-management	**0,21**	**0,21**	**0,040**	**0,17**	**0,17**	**0,030**
Determination	0,27	0,29	0,114	0,22	0,23	0,072
Focus	0,18	0,07		0,16	0,07	
Order	0,05	-0,24		0,05	-0,19	
Persistence	0,17	-0,04		0,14	-0,04	
Responsibility	0,24	0,19		0,19	0,14	
Engaging with others	**0,14**	**0,14**	**0,019**	**0,12**	**0,12**	**0,015**
Enthusiasm	0,09	0,04	0,043	0,07	0,03	0,036
Assertiveness	0,20	0,21		0,18	0,20	
Social initiative	0,03	-0,07		0,03	-0,07	
Amity	**0,24**	**0,24**	**0,058**	**0,19**	**0,19**	**0,035**
Compassion	0,23	0,15	0,090	0,14	0,06	0,049
Modesty	0,26	0,18		0,21	0,16	
Respect	0,20	0,05		0,15	0,05	
Trust	*0,01*	-0,07		0,04	-0,01	
Negative emotion regulation	**0,05**	**0,05**	**0,002**	**0,09**	**0,09**	**0,009**
Frustration tolerance	0,04	0,04	0,007	0,09	0,08	0,010
Stress modulation	*0,01*	-0,06		0,07	0,01	
Self-confidence	0,07	0,08		0,07	0,03	

Note. r shows simple bivariate correlations of SENNA scales (domains and facets) with achievement. *β* are the regression coefficient of facets of multiple regression analysis (unique contribution of each facet controlling for the others in the model). *R²* shows the percent of variance accounted for. For domain scores they are simply *r²*. Values in italics are not significant at p =.05.

All domain scales were significantly and moderately related to achievement and values were similar to those reported by Poropat's (2009). Interestingly the magnitude of Openness was higher than Self-management. This is similar of Poropat's data from primary education where magnitudes were equal. Amity was also moderately related to achievement again as was found for primary education by Poropat. This confirmed our first expectation about the replicability of results found in literature with SENNA scales corroborating its validity.

By taking a closer look at the facets level, we found the predictive validity increases. For all domains explained variance increased in most cases being close to twice as higher. For instance the facet *Curiosity to learn* has a unique contribution of β = 0,30 for Language and β = 0,25 for Math. Also, the facets *Determination* and *Responsibility* presented unique effects of β = 0,29 / 0,23 for Language and β = 0,19 / 0,14 for Math, respectively. Meanwhile, facet *Order* presented a unique negative effect of β = -0,24 for Language and β = -0,19 for Math. Finally, the facets *Compassion* and *Modesty* presented, respectively, unique effects of β = 0,15 / 0,06 for Language and β = 0,18 / 0,16 for Math.

We also run multiple regressions predicting language and math using all SENNA scales. One time we included domain scores as predictors. This resulted a R^2 = .08 for language and R^2 = .05 for math. We ran the same regressions but now with facets as predictors. This resulted an increase in prediction to R^2 = .18 for language and R^2 = .12 for math. Transforming these values to correlations (multiple correlations), to be close to the metric of validity coefficients, we see coefficients of R = .28 /.22 for domains and R =.42/.34 for facets. So, we took a big step at increasing criterion validity going from domains to facets. Note that we have the same set of 162 items. In the first analysis, while scoring items by domains, we have more reliable measures with broad conceptual breadth representing the sum of facets. In the second analysis, while scoring items by facets, we had less reliable scores but a narrower conceptual definition of each of the big five domains. This made possible to see that some unique contributions tuned to negative (the effect of a skill considering constant the level of other skills). This indicates a suppression effect. Therefore, when we aggregate all facets together, we lose predictive power. Results of these analysis presented in Appendix F.

From these results, it is possible to say not only socioemotional domains replicate what is found in the literature for the Big Five model in education but at the facets levels, the power of explanation is higher. As hypothesized, the difference seen at both levels falls in the perspective that "facet-level subscales provide the descriptive and predictive precision of high fidelity" (Soto & John, 2017, p.118), allowing nuances of each trait to be more or less related to the prediction of academic achievement.

NTPPS experimental study: Validity evidence based Impact of intervention program in SENNA scales of socio emotional skills

Another important question about validity asks whether SENNA scales measures traits that are potentially malleable and whose development is shaped by environmental factors, that is, formal and informal learning experiences? In this section we address the question: Can SENNA scales pick up changes intentiaonaly promoted in a school-based intervention program?

This section reports a study of impact evaluation that admnistered SENNA as dependent varibels aiming to gauge if intervantion has an impact on socio emotional skills. This stuydy followed a randomized experimental design comparing control versus tretment group that answered SENNA at previous (pre) and after the intervention (post).

The program is a innovative social and emotional curriculum-based intervention for high school students called "Work, Research and Social Practices" (Núcleo de Trabalho, Pesquisa e Práticas Sociais – NTPPS in Portuguese). In this program students in normal secondary schools in the State of Ceará, Brazil (Ensino Medio, a 3-year period phase of education) receive training to develop personal, social and labor skills over the three years of this educational level. The program is developed by the Department of Education of Ceará (SEDUC-CE) in partnership with Instituto Aliança (IA) and consists of a new curriculum component inducing innovative teaching practices in school, based on methodologies that highlight: (i) student involvement, (ii) interdisciplinarity, (iii) project-based learning, (iv) the use of research as a learning tool, and (v) preparation for the world of work.

NTPPS shares the same general characteristics of the Positive Youth Development (PYD) programs that have become popular in the US. Studies on PYD interventions (Durlak et al., 2011, Sklad et al, 2012) reveal that these programs typically succeed in improving feelings of self-confidence and self-esteem, positive social behaviors and school bond, as well as academic achievement. They also reduce problem behaviors and drug use. In the most recent and broad meta-analysis (Taylor et al, 2017),

the average effect size of SEL interventions was 0,14 standard deviations[3] for conduct problems and 0,16 for emotional distress, typical markers of emotional stability. They also find an important impact of 0,13 for positive social behavior. These authors, as well as Durlak et al (2011), stressed that four features seem to be present in successful SEL interventions, summarized in the acronym SAFE (Sequenced, Active, Focused and Explicit), all of which presented in NTPPS. They also report that most of the PYD interventions contain structured material and lessons, lasting between 30 and 45 minutes. Programs run by the school staff (such as NTPPS) also display higher impacts than those whose activities are conducted by external personnel (Durlak et al, 2011). Zins et al (2004) summarizes the proposed theory of change of PYD programs, which suggests that PYD interventions affect mainly Positive Social Behavior, Conduct Problems and Emotional Distress, through two mechanisms: social and emotional skill acquisition and improved attitudes about self, others and school.

It is worth to notice that the first year of the NTPPS curriculum emphasizes precisely such attitudes, and as shown further in this study, has its most visible impacts exactly on interpersonal dimensions (related to positive social behavior) and emotional resilience. In the second year, the focus is on citizenship and the relationship between the individual and the community; and in the last year of high school the central concern is the relation of the individual to the world of work.

Regarding the way the NTPPS sessions are organized, they consist of five hour lessons per week divided into three sessions (two sessions of 2 hours and one of 1 hour). The two longer sessions focus on building a life project, while the shorter session is dedicated to learning Information and Communication Technologies (ICT). The pedagogy is based on participatory methodologies that value students' opinions and experiences, and provides an association between disciplinary contents, real life experiences and practices.

Th sample of this study were 5,561 students from the 72 schools that: (a) offer regular high school (not professional education) or normal school (magisterio) and at

---

[3] Here and thereafts, in this session, effects will be presented using unities of standard deviation metric. Therefore, an effect of .14 indicates that treatment group differs .14 standard deviations from control group. We will omit the word "standard deviations thereafter in this session to avoid repetition.

least one class takes place in the morning or afternoon; (b) had not yet received the NTPPS program (or his predecessor the CDD program) in previous years; and (c) had information regarding the proficiency of students coming from the Permanent System of Evaluation of Basic Education of Ceará (SPAECE) for the years 2013, 2012 and 2011. The intervention took place in 2015 and the 72 schools that volunteered to implement the program were randomly assigned into 2 groups: those who immediatly started in the program (treatment group) and those who would not participate in the program (controle group). These 72 were organized into 36 pairs of schools with the lowest Mahalanobis distance using the following characteristics: Portuguese and math proficiency for the cohort enrolled in 1st and 3rd year of high school, a variable indicating whether the school is located in Fortaleza, an indicator of whether the school participates in the Project called Jovem de Futuro, the student teacher ratio in the first year of high school and the total number of students enrolled in the school.

In the beginning of the academic year students answered SENNA v1 beta version with 90 items. In the follow-ups they answered SENNA v2 with 162 items and a questionnaire designed to collect data on socio-economic characteristics of the students' households, the school life of students, including student perceptions of their teachers, and students' expectations about their post-secondary school future. It is important to highlight that the program was designed as a 3-year curriculum throughout high school. Students were followed for three years, and we were then able to compare the social and emotional changes between the treated and control groups at the end of first and last grades of high school.

The results indicated that there is evidence of a positive impact of NTPPS on students' social-emotional characteristics, especially in self-efficacy in dimensions associated with intrapersonal and inter-personal behavior. The most significant impacts at the end of the first grade were in "Engagement with others" (0,16), "Amity" (0,16) and "Negative Emotional Regulations" (0,17). The socioemotional impacts were particularly high for students with a history of school repetition, Learning difficulties in Portuguese and boys. This result is particularly important given that this program had bigger effects of student that were "more vulnerable" and thus, had important effects on equity. No negative effect of NTPPS on student proficiency was observed. It is

interesting to note that the dimensions on which the program had effect have a great overlap with the contents prioritized in the first year of the NTPPS, such as the individual's relationship with him/herself, with the school and the family, and the domains on which the program had the greatest impact are those related to interpersonal characteristics and emotional resilience. The medium-term impacts of NTPPS, measured at the very end of the program, when students were about to graduate high school, were highest in Openness to New Experiences (0,18), Engagement with others (0,18) and Emotional Resilience (0,15).

These results show that SENNA scales are able to capture change over time once an especific intervention is put into place aiming at developing social-emotional skills. The content of intervention were well mached with the content of the specific scales we noted significant changes. These results suggest that the atributes that SENNA scales measures are maleable aspects of traits. It is a strong evidence since it was based aon a randomized experimental design. More on the study and its results can be found on Ortiz, Borges, and Santos (2017).

Validity evidence based on the relationship with external variables: correlations of SENNA with alternative measures of socio emotional skills

The examination of whether tests that are intended to measure similar constructs exhibits high correlations between them is another important consideration about validity. Since the beta version of SENNA we have been proposing and testing the use the big five's empirical taxonomy to map alternative frameworks and measures developed by social cognitive and educational researchers. Because SENNA scales sample a comprehensive set of dimensions of personal characteristics instantiated from the big five, we expect they will shed light on what primary dimensions of personality underlies measures developed by other social cognitive and educational researchers. We hope to understand and link the constructs used in these fields to an empirical taxonomy of personality theory. We expected that SENNA scales will correlate more strongly with similar constructs in other frameworks. We hypothesized several links between big five these major frameworks elsewhere (see Primi, Santos, Hauck, De Fruyt,

& John, 2019; Primi, Alves & Marino, no prelo). We will test these links in a validity evidence exercise in this section.

The present study occurred in the city of Chapecó in 2016 where elementary and high school students answered several questionnaires and tests as a part of bigger joint project of Ayrton Senna Institute, OECD-coordinated international study on Critical Reasoning and Creativity and Chapecó City Department of Education. A sample 1,041 students from 5th to 12th grades, 54% girls, answered to SENNA and several other instruments of important frameworks on socio emotional skills. The instruments were:

Mindsets, Essential Skills, and Habits (MESH) This instrument is a self-report questionnaire consisting of 25 items measuring: growth mindset, self-efficacy, self-management (grit) and social awareness. MESH is designed for use in the group of California schools that are developing a new, more holistic system for teaching quality monitoring. The scales were derived from the very influential work on the socio-emotional assessment of Angela Duckworth (Duckworth & Quinn, 2009), Clancy Blair, Carol Dweck and the CASEL organizations, Consortium on Chicago School Research (CCSR).

Student's Approach to Learning (SAL, Marsh e cols, 2006): This instrument is a self-report questionnaire composed of 52 items distributed in 14 scales measuring relevant constructs in psychoeducational assessment: Instrumental Motivation, Interest in Reading, Interest in Mathematics, Control Strategies, Memorization, Elaboration, Effort/Persistence, Cooperative Learning, Competitive Learning, Verbal Self-Concept, Math Self-Concept, Academic Self-Concept, Self-Efficacy and Control Expectation. SAL was used in PISA 2000 and was the result of a systematic review of the most relevant socio-emotional constructs in psychoeducational assessment.

National Assessment of Educational Progress (NAEP) contextual questionnaires (https://nces.ed.gov/nationsreportcard/). NAEP is a similar instrument to SAEB consisting of standardized tests to assess student achievement and monitor the education system. The student questionnaire includes questions about socio-emotional characteristics. In this study, 25 self-report questions were used, distributed into four scales: desire to learn, grit, growth mindset and self-control (shcool work).

Character Growth Card (CGC, Park, Tsukayama, Goodwin, Patrick, & Duckworth, 2017): CGC is a 30-item self-report questionnaire distributed across eight scales: curiosity, gratitude, grit, optimism, self-control (interpersonal), self-control (schoolwork) , social intelligence and enthusiasm. It was developed by Angela Duckworth (www.characterlab.org) and is based on the concept of positive psychology of character strengths. The instrument was developed to be a formative assessment tool for teachers and students.

Students answered tests via SENNA's computerized web system in regular school activities in their classes. In total there was 294 items organized in 47 scales. Of course, it is not feasible to administer these many scales to each one student. It will cause fatigue and increase the chances of demotivation and inattention. Therefore, we used an Incomplete Balanced Blocks (BIB) design, as in previous studies, and administered selected subtests to each student and still obtained a reliable estimation of a complete correlation matrix. Details of the design can be found in Appendix G.

Table 15 presents correlations of SENNA scales with four MESH scales. Examples of MESH items are (a) growth mindset: "*My intelligence is something that I can't change very much*"; (b) self-efficacy: "*I can master the hardest topics in my classes*"; (c) Self-management (Grit): "*I got my work done right away instead of waiting until the last minute*" and (d) social awareness: "*How well did you get along with students who are different from you?*".

We found a strong link between SENNA self-management facets and MESH scales of grit and self-efficacy. We found positive, moderate to strong correlations with Determination, Persistence and Responsibility (from .42 to .62). Note that MESH scales of Grit and Self-Efficacy are also highly correlated (r = .54). We also found a strong link between SENNA facet of Compassion and Social awareness (r = .56) and SENNA facet of Respect and Grit. Grit scale contains aspects of interpersonal relationships – ex. "*I was polite to adults and peers*" - and emotional regulation – ex: "*I kept my temper in check*". Therefore, it is coherent a to see a positive association with facets of Amity and Emotional Resilience.

Also a link between and between The scales self-efficacy and had Also, these two scales presented correlations with facets from Emotional resilience and Amity, and self-management had a moderate result with Curiosity to learn (r = 0.42). Growth mindset had a positive and moderate correlation with Assertiveness (r = 0.45), while social awareness presented similar results with Compassion (r = 0.56).

Table 15.

SENNA's scales correlations with Mindset Essential Skills and Habits scales (MESH)

SENNA scales	Growth mindset	Self-efficacy	Self-management (Grit)	Social awareness
Open-mindedness				
Artistic interest	0,14	0,35	0,38	0,32
Creative imagination	0,08	0,20	0,20	0,22
Curiosity to learn	0,22	0,29	0,42	0,28
Self-management				
Determination	0,14	**0,50**	**0,50**	0,27
Focus	0,10	**0,40**	**0,47**	0,19
Order	-0,02	**0,45**	**0,54**	0,25
Persistence	0,01	**0,47**	**0,62**	0,28
Responsibility	0,18	**0,42**	**0,56**	0,35
Engaging with others				
Enthusiasm	0,09	0,34	0,24	0,36
Assertiveness	**0,45**	0,36	0,20	0,12
Social initiative	0,20	0,12	0,06	0,13
Amity				
Compassion	0,15	0,30	**0,44**	**0,56**
Respect	0,07	0,34	**0,50**	0,37
Trust	0,11	0,18	0,22	0,20
Gratitude	0,16	0,03	0,19	0,27
Emotion resilience				
Frustration tolerance	0,06	0,19	0,33*	0,28
Stress modulation	0,09	0,18	0,19	0,00
Self-confidence	0,10	0,28	0,29*	0,17
MESH				
Self-efficacy	0,12			
Self-manag. (Grit)	0,04	0,54		
Social-awareness	0,07	0,40	0,58	

Marsh et al. (2006) organized these scales using a Boekaert's model of self-regulated learning into four conceptual groups: (a) metacognitive learning strategies (elaboration strategies, memorization strategies, control strategies); (b) motivational preferences (interest in reading, interest in mathematics, instrumental motivation, effort and persistence in learning); (c) self-related cognitions and beliefs (verbal self-concept, math self-concept, academic self-concept, self-efficacy, and control expectations) and finally (d) preferences for learning situations (cooperative learning and competitive learning).

Table 16 presents correlations of SENNA scales with the first two SAL scales scale groups: metacognitive learning strategies and motivational preferences. Examples of items are: elaboration strategies: "*When I study, I try to relate new material to things I have learned in other subjects*", memorisation: "*When I study, I memorize as much as possible*", control strategies: "*When I study, I start by figuring out exactly what I need to learn*", interest in Math: "*When I do mathematics, I sometimes get totally absorbed*", interest in Reading: "*When I read, I sometimes get totally absorbed*", instrumental motivation: "*I study to increase my job opportunities*", effort and persistence to learn: "*When studying, I keep working even if the material is difficult*".

Table 16.

Correlations of SENNA scales with Students Approach to Learning scales (SAL) scales of metacognitive learning strategies and motivational preferences

	cstrat	elab	memor	effper	insmot	intmat	intrea
Aes	0.20	0.21	0.14	0.13	0.20	0.29	0.46
CrImg	0.31	0.29	0.18	0.04	0.16	0.28	0.35
IntCur	0.31	0.37	0.25	0.30	0.30	0.34	0.52
Achv	0.44	0.38	0.42	0.55	0.26	0.53	0.52
Conc	0.56	0.49	0.42	0.47	0.30	0.50	0.42
Ord	0.54	0.47	0.50	0.61	0.31	0.52	0.44
SD	0.51	0.43	0.40	0.57	0.31	0.38	0.39
SofR	0.42	0.51	0.46	0.55	0.43	0.30	0.30
Act	0.22	0.16	0.31	0.33	0.26	0.34	0.16
Assr	0.33	0.24	0.33	0.35	0.21	0.23	0.18

Soc	0.19	0.24	0.21	0.26	0.15	0.18	0.08
Cmp	0.31	0.34	0.34	0.37	0.24	0.22	0.16
Mod	0.14	0.22	-0.05	0.04	0.09	0.24	0.20
Resp	0.55	0.45	0.50	0.54	0.35	0.59	0.34
Tru	0.06	0.14	0.20	0.18	0.00	0.06	-0.01
LAngrVol	0.14	0.18	0.07	-0.03	0.07	0.29	0.15
LAnx	0.24	0.21	0.06	0.06	0.23	0.26	0.16
LDep	0.32	0.35	0.31	0.44	0.24	0.27	0.15

Self-management shows the most moderate to high correlations r > .40 specially with effortful persistence, memorisation strategies, instrumental motivation. Effortful persistence is conceptually linked with self-management scales. Instrumental motivation items mention studying as a way to get a job and financial security; therefore, it can be interpreted as planning skills that is also conceptually linked to self-management. Meta cognitive strategies can be interpreted as top down organization skill in the context of school activities.

Open-mindedness scales tend to correlate highly with motivational aspects like interest in math and instrumental motivation. Amity facet respect has high correlations with memorisation, interest to read and effortful persistence. Respect items refers to good behaviour at the school (respect authority, avoid discussions, listen respectfully, apologize) and are dependent on top down regulation skills to prevent socially undesirable behaviour. Because of that, respect is highly related with self-management facets (around *r* = .50), consequently, they tend to show a similar profile of correlations with external variables as self-management scales do.

Table 17 presents correlations of SENNA scales with the last two groups of SAL scales <u>beliefs</u> and <u>preferences for learning</u>. Examples of items are: verbal self-concept "*I learn things quickly in Portuguese class*", math self-concept "*Mathematics is one of my best subjects*", academic self-concept "*I'm good at most school subjects*", self-efficacy "*I'm certain I can understand the most difficult material presented in texts*", control expectation "*When I sit myself down to learn something really difficult, I can learn it*", preference for competitive learning "*Trying to be better than others makes me work well*" and preference for co-operative learning "*I like to work with other students*".

Table 17.

Correlations of SENNA scales with Students Approach to Learning scales (SAL) of <u>beliefs</u>
and <u>preferences for learning</u>.

	cexp	scacad	scmath	scverb	selfef	comlrn	coplrn
Aes	0.10	0.42	0.18	0.40	0.26	-0.04	0.17
CrImg	0.28	0.47	0.29	0.34	0.27	-0.18	0.19
IntCur	0.25	0.34	0.18	0.32	0.31	0.03	0.27
Achv	0.30	0.48	0.40	0.41	0.32	-0.03	0.23
Conc	0.35	0.46	0.38	0.34	0.40	-0.07	0.23
Ord	0.40	0.41	0.35	0.40	0.47	-0.08	0.30
SD	0.30	0.42	0.29	0.37	0.37	-0.08	0.24
SofR	0.50	0.30	0.17	0.32	0.42	-0.06	0.32
Act	0.19	0.20	0.27	0.26	0.24	0.03	0.21
Assr	0.21	0.45	0.29	0.15	0.25	0.01	-0.07
Soc	0.22	0.12	0.15	0.13	0.23	-0.04	0.31
Cmp	0.19	0.26	0.06	0.30	0.33	-0.14	0.27
Mod	0.10	0.09	0.13	0.12	0.09	-0.36	0.18
Resp	0.37	0.34	0.36	0.33	0.43	-0.13	0.30
Tru	-0.04	0.12	-0.02	0.09	0.11	-0.16	0.07
LAngrVol	0.05	0.14	0.21	0.09	0.20	0.00	0.06
LAnx	0.29	0.28	0.21	0.04	0.28	0.03	0.18
LDep	0.25	0.20	0.19	0.27	0.38	0.00	0.21

Beliefs of being able to understand complex topics (self-efficacy), learn quickly, learn most subjects (academic self-concept) and doing well in math and language (math and verbal self-concept) have high correlations with self-management and open-mindedness. Self-management skills are associated with a self-image of being good at schoolwork such as learning, mastering complex tasks and doing well on tests.

It is interesting to note that open-mindedness is related to SAL scales that refers to aspects of intelligence such as the capacity to learn quickly, learn all subjects and understand complex topics. It also correlates relatively high with language domain. This is coherent with correlations of openness with standardized reading achievement.

We observed a strong correlation of assertiveness with academic self-concept. SENNA's assertiveness items talks about leadership and interaction with teachers about learning. Therefore, it concerns to the social interaction behaviour of a good student.

Preferences for cooperative learning is positively associated with facets of self-management – order and sense of responsibility, amity – compassion and respect - and extroversion - sociability. It is related to a mixture of characteristics such as being organized, reliable, caring and respectful with others and gregarious. Preference for competitive learning tends to be negatively associated with all scales but specially with modesty. This indicates that students that prefers competitive learning are prone to value their individuality and specialness before the group.

We found correlations of respect with all SAL beliefs scales. This also can a be reflection of the "good student" factor underlying both domains of self-management and amity. We observed small correlations of stress modulation and self-confidence with beliefs scales. This indicates that skills of regulating negative emotions – specially anxiety and sadness - is associated with positive self-image as being capable to learn, understand and to master topics even when they are difficult.

Table 18 presents correlations of SENNA scales with NAEP scales. Examples of items are: grit "*I try very hard even after making mistakes*", mindset "*In school, how possible is it for you to change …Behaving well in class"),* desire to learn "*I enjoy situations where I will have to think about something*", and self-control (school work) "*I paid attention and resisted distractions*".

We found very strong correlations of self-management facets – notably responsibility, persistence and order - with grit and self-control. The magnitude of these correlations suggest that all these scales measure the same underlying latent construct although they have different names and were developed from different theoretical backgrounds. Item content represent behaviours of finishing things even when facing difficulties, fulfilling commitments, staying long hours to finish your duties, paying attention, resisting distractions, being organized and being prepared in advance for school activities.

Open-mindedness scales are correlated with desire to learn referring to interest in thinking in intellectual challenges. Facet curiosity to learn has a strong correlation with grit scale. One shared aspect between scales is a prolonged work on complex intellectual tasks. Curiosity to learn share this motivational drive with grit scales.

Table 18.

SENNA's scales correlations with National Assessment of Educational Progress scales (NAEP)

	Desire to learn	Grit	Growth Mindset	Self control (school)
Open-mindedness				
Artistic interest	0,40	0,28	0,21	0,36
Creative imagination	**0,45**	0,37	0,18	0,18
Curiosity to learn	**0,47**	**0,66**	0,38	0,33
Self-management				
Determination	**0,51**	**0,48**	0,28	**0,47**
Focus	0,34	**0,44**	0,31	**0,46**
Order	0,31	**0,62**	0,39	**0,69**
Persistence	**0,40**	**0,68**	**0,51**	**0,69**
Responsibility	0,34	**0,85**	**0,47**	**0,65**
Engaging with others				
Enthusiasm	0,24	0,39	0,19	0,20
Assertiveness	0,30	0,38	0,10	0,18
Social initiative	0,26	0,32	-0,01	0,18
Amity				
Compassion	0,30	**0,40**	0,17	0,33
Respect	0,26	**0,61**	**0,40**	**0,54**
Trust	0,13	0,07	0,07	0,15
Gratitude	0,01	0,21	0,14	0,14
Emotion regulation				
Frustration tolerance	0,15	0,18	0,08	0,15
Stress modulation	0,26	0,25	0,01	0,14
Self-confidence	0,30	0,46	0,17	0,26

Growth mindset correlated highly with persistence, responsibility and respect. An important detail is that growth mindset items refers to beliefs that you can change your persistence (putting effort into a work) and your bad behaviors (behaving well in class) and not the usual beliefs that you can change your intelligence. Therefore,

correlations show this dual shared aspect with persistence and responsibility (facets self-management) and respect (a facet of amity).

Table 19 presents correlations of SENNA with eight scales of Character Growth Card (CGC). Example of CGC items are: curiosity "*I was eager to explore new things*", gratitude "*I showed appreciation for the good things I have in my life*", grit "*I tried very hard even after experiencing failure*", optimism "*I believed that effort would improve my future*", self-control - interpersonal "*I remained calm even when criticized or otherwise provoked*", self-control - school work "*I got my work done right away instead of waiting until the last minute*", social intelligence "*I was able to find solutions during conflicts with others*", and zest "*I showed enthusiasm*".

Curiosity to learn correlated with CGC's curiosity scale as expected and with social intelligence that have items that refers to finding solutions when in conflict with others and adapting to different social situations that has a shared aspect of openness. Creative imagination correlated with five CGC scales: social intelligence, curiosity, grit (persistence when facing difficulties) optimism (belief that effort on trying different ways will make you improve) and zest (approaching new situations with excitement and energy). Looking close to item content it is possible to see they shared aspects of openness such as trying new ways to do thinks and persist on intellectually challenging tasks.

Table 19.

SENNA's scales correlations with Character Growth Card (CGC).

	Soc. Intell.	Curios.	Gratit.	Grit	Optim.	Self-control (interp.)	Self-control (school work)	Zest
Open-mindedness								
Artistic interest	0,22	0,24	0,25	0,38	0,30	0,19	0,34	0,24
Creative imagination	**0,40**	**0,58**	0,34	**0,44**	**0,40**	0,29	0,23	**0,47**
Curiosity to learn	**0,41**	**0,45**	0,30	0,34	0,35	0,21	0,29	0,33
Self-management								
Determination	0,34	0,47	0,39	**0,54**	**0,42**	0,20	**0,48**	**0,52**
Focus	0,37	0,36	0,22	**0,40**	**0,43**	**0,48**	**0,51**	0,36
Order	0,30	0,36	0,29	**0,45**	0,33	0,37	**0,56**	**0,44**
Persistence	**0,49**	**0,45**	**0,42**	**0,59**	**0,47**	0,33	**0,62**	**0,44**
Responsibility	0,44	**0,40**	**0,42**	0,39	0,36	0,32	0,38	0,37
Engaging with others								
Enthusiasm	0,35	0,30	0,29	0,40	0,32	0,17	0,25	**0,40**

Assertiveness	0,23	**0,48**	0,14	0,22	0,38	0,10	0,13	**0,41**
Social initiative	0,27	0,30	0,36	0,27	0,16	0,07	0,22	0,32
Amity								
Compassion	**0,49**	0,34	**0,45**	**0,42**	0,37	0,36	0,37	0,33
Respect	**0,43**	0,39	**0,56**	0,32	0,31	**0,65**	**0,51**	0,33
Trust	0,20	0,18	0,23	0,14	0,15	0,30	0,19	0,14
Gratitude	0,40	0,24	0,31	0,44	0,37	0,38	0,22	0,28
Emotion regulation								
Frustration tolerance	0,19	0,14	0,09	0,25	0,24	**0,45**	0,15	0,26
Stress modulation	0,21	0,30	0,08	0,20	0,29	0,35	0,31	0,31
Self-confidence	**0,55**	**0,55**	**0,57**	**0,51**	**0,48**	**0,45**	**0,51**	**0,56**

Self-management scales have the high number of positive correlations with several CGC scales. As expected, facets correlate with grit, self-control (school-work) and zest that mentions the enthusiasm on school activities. Enthusiasm and assertiveness correlated with zest. These scales have a shared content of positive energy and active participation in social activities at school. Amity facets of compassion and respect showed expected correlations with CGC scales of self-control (interpersonal), gratitude and social intelligence.

Frustration tolerance facet correlated with self-control (interpersonal). This indicates that a shared aspect of this skill is the management of negative emotions of anger and irritability. One intriguing finding is a general high correlation of self-confidence with all CGC scales. We suggest that this may be related to the type of question that CGC asks to students: "In the past four months, how often have you done the following items? During that period … "I was able to find solutions during conflicts with others" and so forth. CGC asks students to respond on frequency scales: almost never, rarely, sometimes, often, almost always. Therefore, CGC is asking students to say how often they experienced positive experiences or positive emotions at school. Each item has a specific content that refers to experiences associated to the specific strength measured but they all have a general content referring to positive emotions. Senna self-confidence scale is meant to measure this aspect too with items like "*I am happy with few negative thoughts*", "*I can overcome difficulties*" (true keyed), or ": "*I can't stop thinking about negative events*" (false keyed). Therefore, this might be the shared aspect driving this positive correlations.

Validity evidence based on the relationship with external variables: Violence

A major challenge in educational contexts nowadays is to deal with youth aggression. Violence seems to be increasing and starting earlier in recent years. In 2018 we collected data from all middle school students in the city of Sobral, at the Brazil Northeast region. Together with SENNA, we applied a rich questionnaire about early involvement with violence and gang belonging. Results show that *Amity* and *Self-Management* predict this involvement, even after controlling for socioeconomic variables. This gradient tends to be bigger for younger ages. In the case of *Amity*, for instance, the mean difference between students involved and not involved with gangs is 0,22 sd, whereas the same calculation for *Self-Management* is 0,15 sd. This finding matches what has been found in the literature (e.g. John et al, 1994; Nigg et al, 2002) relating the Big Five domains with youth delinquency.

Besides the intervention studies, sensitiveness of the SENNA constructs to contextual variables have been explored in correlational investigations. Santos et al (2016) used an earlier version of SENNA and shows that students attending schools in violent neighborhoods have, on average, lower scores on *Amity, Engagement with Others* and *Emotional Resilience*. In the recent Sobral data collection mentioned before, we see that middle school students who experience or witness domestic violence display lower scores of *Amity, Self-Management* and *Emotional Resilience*.

Norms: benchmarks for interpreting SENNA scales

Classic score metric of SENNA's scales

Each SENNA scale is comprised of 9 items: 3 true keyed identity items, 3 false keyed identity item and 3 true keyed self-efficacy items. Each item asks subjects to respond in a 5-categories Likert scale. For identity items we instructed students: '*Each item shows some personal characteristic that may or may not have to do with you. To answer the questions, think about how you are/feel/behave in most situations*', and after each item, like, '*I adapt easily to new situations without worrying too much*' (true keyed) or '*I have trouble controlling my anxiety in difficult situations*' (false keyed) we presented five categories: '1' (not at all like me), '2' (little like me), '3' (moderately like me), '4' (a

lot like me) and '5' (completely like me). For self-efficacy items we instructed students to '*Evaluate on the scale below how well can you …*' and each item completed the sentence such as: '*Cheer up when you're sad*' or '*Create new things*' and were followed by the response categories: '1' (Nothing), '2' (A little), '3' (Moderately), '4' (Very) and '5' (Totally).

Therefore, item scores are on a 1-5 metric indicating the general degree of similarity to self or degree of confidence. Note that we reverse false keyed items (5 to 1, 4 to 2, 3 = 3, 2 to 4 and 1 to 5) making item scores toward 5 always indicating high level on the construct, for instance, a student will have a high score on engaging with others when he/she simultaneously agreed with an item '*I like to talk with others*' (answered '5') and s disagreed to items like '*I don't speak to my classmates at school much*' (answered '1' that is then reversed to '5').

A simple way to create a score for SENNA scale is to calculate an average of item scores indicating an average endorsement. This will produce e a metric of 1 to 5 and those numbers will be associated to their category labels. Although this procedure can provide one way to interpret the scale it has some shortcomings.

Let's consider the short version of SENNA scales comprised by three items, one true keyed identity item, one false keyed identity item and one true keyed self-efficacy item. Let's calculate the average level of social initiative of given sample twice one with the full scale (9 items) and another with the short scale (3 items). Are these averages equal? They should be since we are measuring the same sample. But they may differ because of error of measurement. More importantly they may differ due to systematic differences in "item easiness/difficulty to endorse". Imagine that short version has the easiest to endorse items of the full scale. Now the short version's average score will be systematically higher. It will suggest higher level of social initiate when compared with scores calculated with full version. Classical scores metrics (sum or average of item scores) of different tests are not directly comparable. Metrics will depend on the test/items used.

Psychometric methods for constructing standardized metric

There are four kinds of psychometric scaling methods that attribute a standardized meaning of classical scores: (a) domain referenced scales, (b) norm referenced scales, (c) criterion referenced scales, and (d) Item Response Theory scaling and item referenced meaning.

If items on a test are randomly selected from the universe of possible items that assess a given psychological concept or domain, classical scores would have a substantive meaning. Tests composed of different items will have equivalent difficulties and represent the same underlying construct because of the random process of assembling items to a test. Under this circumstance the metric will be comparable. Scales with this characteristic are called domain referenced scales but are hard to construct.

Normative reference scales atribute standardized meaning comparing subject's score with the distribution of all scores in a given normative sample and indicates the relative position of a given person in the distribution. It express this result with percentiles, that is, a number of 1 to 100 indicating the percentage of scores/people in the distributioin that a given score is above at. For instance a P75 will indicate that a given score is above 75% of the scores of the normative sample. Becasue of that it indicates a high score. For instance P25, P50 and P75 will indicate normatively low, averge and high scores in a statistical sense. If we order all scores on the sample from lowest to the highest P25 will separate the 25% lowest scores and P75 will separate the 25% highest scores. Scores between P25 and P75 will be scores arround averge/median.

Criterion reference scales atribute standardized meaning to a subject's making a prediction about what will be the score on a external variable usually a criterion variable. Criterion variables are outcomes of primary interest becuase of their inherent relavance. For instance consider college graduation as 1 if an adult has finished college and 0 if he/she has stoped at high school. If intelectual curiosity significantly correlates with graduation on college we could break the scale into four levels (using the percentiles 25, 50 and 75 producing four groups low: 1-25, low-mid 26-50, mid-high 51-75 and high 76-100) and for each group calculate the percentage of students that has finished college.

This procedure conects the metric of the scale with the metric of the oucome attributing a relevant meaning to the scale.

Normative scores don't explain what a person has communicated to us via his/her responses to the specific items we asked them. Item Response Theory (IRT) methods attribute meaning to scores via construct maps and item referenced meaning. In adition IRT scaling methods can in theory produce comparable metrics of different tests made of different items measuring the same construct.

IRT models what item's category of response persons with different levels on the latent factor will answer. It predicts which response (1 to 5) a person will be most likely to choose to particular items given their latent factor score. From this model we can draw a construct map that will informs us which responses will be most likely to be chosen as we move progressively from low to high score on the latent factor scale. Construct map relates scale scores to the specific response patterns that originated that score. Content analysis of reponse patterns cam give more substance meaning to scale metrics.

IRT metric: Rasch-Masters Partial Credit Model

To construct norm benchmarks to interpret SENNA scales we used a combination of norm referenced interpretation and IRT scaling. We calibrated item parameters according to the Rasch-Masters Partial Credit Model (PCM; Wright, & Masters, 1982). With items like we used with five response categories, PCM break them down into five equations each one predicting each response of 1 to 5: $P(X_{ni} = 1)$, $P(X_{ni} = 2)$, $P(X_{ni} = 3)$, $P(X_{ni} = 4)$ $P(X_{ni} = 5)$. PCM model is called divide-by-total because it predicts the relative probability of choosing a category, for instance 1, over all possible response categories 1 to 5.

Following the notation from Wu, Tam and Jen (2016), PCM will model the probability of person n answering each of the response categories 1 to 5 for item i called X_{ni} as:

$$P(X_{ni} = 1)$$
$$= \frac{1}{1 + e^{(\theta_n - \delta_{i1})} + e^{(2\theta_n - \delta_{i1} - \delta_{i2})} + e^{(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})} + e^{(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})}}$$

$$P(X_{ni} = 2)$$

$$= \frac{e^{(\theta_n - \delta_{i1})}}{1 + e^{(\theta_n - \delta_{i1})} + e^{(2\theta_n - \delta_{i1} - \delta_{i2})} + e^{(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})} + e^{(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})}}$$

$$P(X_{ni} = 3)$$

$$= \frac{e^{(2\theta_n - \delta_{i1} - \delta_{i2})}}{1 + e^{(\theta_n - \delta_{i1})} + e^{(2\theta_n - \delta_{i1} - \delta_{i2})} + e^{(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})} + e^{(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})}}$$

$$P(X_{ni} = 4)$$

$$= \frac{e^{(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})}}{1 + e^{(\theta_n - \delta_{i1})} + e^{(2\theta_n - \delta_{i1} - \delta_{i2})} + e^{(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})} + e^{(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})}}$$

$$P(X_{ni} = 5)$$

$$= \frac{e^{(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})}}{1 + e^{(\theta_n - \delta_{i1})} + e^{(2\theta_n - \delta_{i1} - \delta_{i2})} + e^{(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})} + e^{(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})}}$$

The more general and compact PCM formula of a person *n* scoring *x* on item *i* is:

$$P(X_{ni} = x) = \frac{e^{\sum_{k=1}^{x}(\theta_n - \delta_{ik})}}{\sum_{h=1}^{m_i} e^{\sum_{k=1}^{x}(\theta_n - \delta_{ik})}}$$

Where for the first response category "1" the exponent $\sum_{k=1}^{1}(\theta_n - \delta_{ik}) = 0$ is defined as 0, therefore the numerator at this score category will be $e^0 = 1$. PCM has the parameter theta θ_n indicating the level of peron *n* on the construct measured, and four delta parameters δ_{i1} to δ_{i4} that are item parameters indicating how easy or difficult is to endorse response categories. The four delta parameters δ_{ik} indicate a location in the theta/delta latent continuum where two adjacent categories have the same probability of endorsement, that is, the transition between score categories 1 to 2, 2 to 3, 3 to 4 and 4 to 5. For instance, δ_{i1} will indicate the location where score categories 1 and 2 are equally likely, δ_{i2} indicate the location where score categories 2 and 3 are equally likely and so forth. Therefore, these values point to levels on latent continuum separating response categories and mark regions where they are more likely to be observed[4]. For instance, from theta values from -∞ to δ_{i1} PCM models expect response category 1, from δ_{i1} to δ_{i2} model expects category 2 and so on. For instance, for persons whose $\theta_n > \delta_{i1}$ value model predicts response 2 or higher. If $\theta_n < \delta_{i1}$ model will predict a

---

[4] More precisely this will be true only when delta parameters are ordered

response 1. These predictions will permit to create a construct map; a plot where we can see what the more likely response pattern associated with theta values.

Note that PCM includes information of item difficulty/facility via delta parameters. If an item is very difficult to endorse it will have deltas shifted to the right of the scale meaning that if a person answered '5' on those items, it is because they have high theta values. PCM predicts item scores to be high either because a person has a high theta or because an item is too easy. The essence of PCM model is to predict a response via these two components (θ_n and δ_{ik}). The inclusion of difficulty in the model makes the metric of theta disentangled from item difficulty. So IRT scaling will produce comparable metrics even when tests have different items provided items measure the same construct and are proper linked.

Calibrating IRT parameters

We first selected a sample from Sao Paulo study to calibrate item parameters. We selected students attending grades 6th to 12th, that have passed attention checks - that is, who answered the three control questions correctly - that had spent at least ten minutes answering the test and had relatively consistent responses. We measured consistency calculating within-subject correlation between antonym pairs, that is, antonym index. We have 54 pairs of semantic opposites, that is, 54 pairs of true keyed *vs* false keyed items. So, each student we have two vectors of 54 responses that should be negatively related if he/she was responding consistently. We calculated this index and selected only students that have *r* < .10 (that is mostly zero or negative, the mean *r* = -.33, min = -.98 and max .10). The final calibration sample contained 31,735 students from grades 6th to 12th, 60.8% girls with age raging from 11 to 24 (M=15.05, SD=2.08).

We run five calibrations, one for each domain (**O**: Open-mindedness, **C**: Self-management, **E**: Engaging with others, **A**: Amity and **N**: Emotional resilience) on Winsteps software (Linacre, 2020). Winsteps implements Joint Maximum Joint Maximum Likelihood Estimation (JMLE) to calculate theta and delta parameters simultaneously. We examined item fit indices and found that majority of them fitted the PCM model (only 6 out of 162 had infit > 1.3, **A**: 2, **C**: 2, **O**: 2). All detailed results are provided in Appendix I (see infits/outfits on Table I6 on the appendix).

We fixed theta metric to have a mean of 0 and standard deviation of 1 (M=0, SD=1) to satisfy estimation's algorithm identification requirements. This implies that on the SENNA metric, for each domain, theta = 0 is equal the average of the students of this first calibration sample from São Paulo. We saved the delta parameters of each item of these five calibrations (see Table I7 on Appendix I) and defined them as the basic set to use in all estimations of SENA scores forward in future datasets. We have been using them to estimate theta scores in all subsequent SENNA datasets since then.

After calibration we run analysis again estimating theta for all students on the São Paulo sample (N=50,209). We fixed item delta parameters using the values discovered in the calibration phase. We estimated 5 domain scores and 18 facets scores. When estimating facet scores, we fixed delta parameters on the values discovered in the calibration phase where we run analysis by domains. Therefore, item difficulty parameters can be compared across facets within a domain. This is also consistent with the unidimensionality assumption of IRT models. But of course, on top of this domain factors there may be secondary dimensions related to facets. We will explore the influence of these secondary factors on IRT estimates in the future. Because this run involve more subjects then the calibration sample the final statistics of theta scores may not necessarily be centered at M=0 and SD=1.

Finally, we studied different R packages that performs IRT calibrations to check their estimation algorithms and compare their results with those produced by Winsteps. Since we defined the metric of SENNA with parameters calculated on Winsteps we wanted to reproduce it's results using R and then write an R package for scoring SENNA. We calibrated theta parameters using TAM (Robitzsch, Kiefer & Wu, 2019), ltm (Rizopoulos, 2018), catR (Magis, Raiche & Zaragoza, 2018) and mirt (Chalmers, e cols. 2019) packages. Although all packages produced equivalent theta estimates on the same data (usually correlations across results of packages were r = .98-1.0) they weren't exactly the same even when we chose estimation method as JMLE or the closest available. We found that mirt and catR produced estimates that are linearly equivalent to Winsteps differing just by intercept/scaling constants. Using a linear transformation with these constants we could recover the exact theta scores Winsteps produced. Therefore, we decided to use catR function 'thetaEst' with PCM model and Maximum

Likelihood (ML) method of estimation. We wrote an R package for scoring SENNA that wraps 'thetaEst' function and uses our database of delta parameters and linear constants to produce theta scores on the original metric we calibrated the items. This package can be found on github: https://github.com/curso-r/IAStri[5].
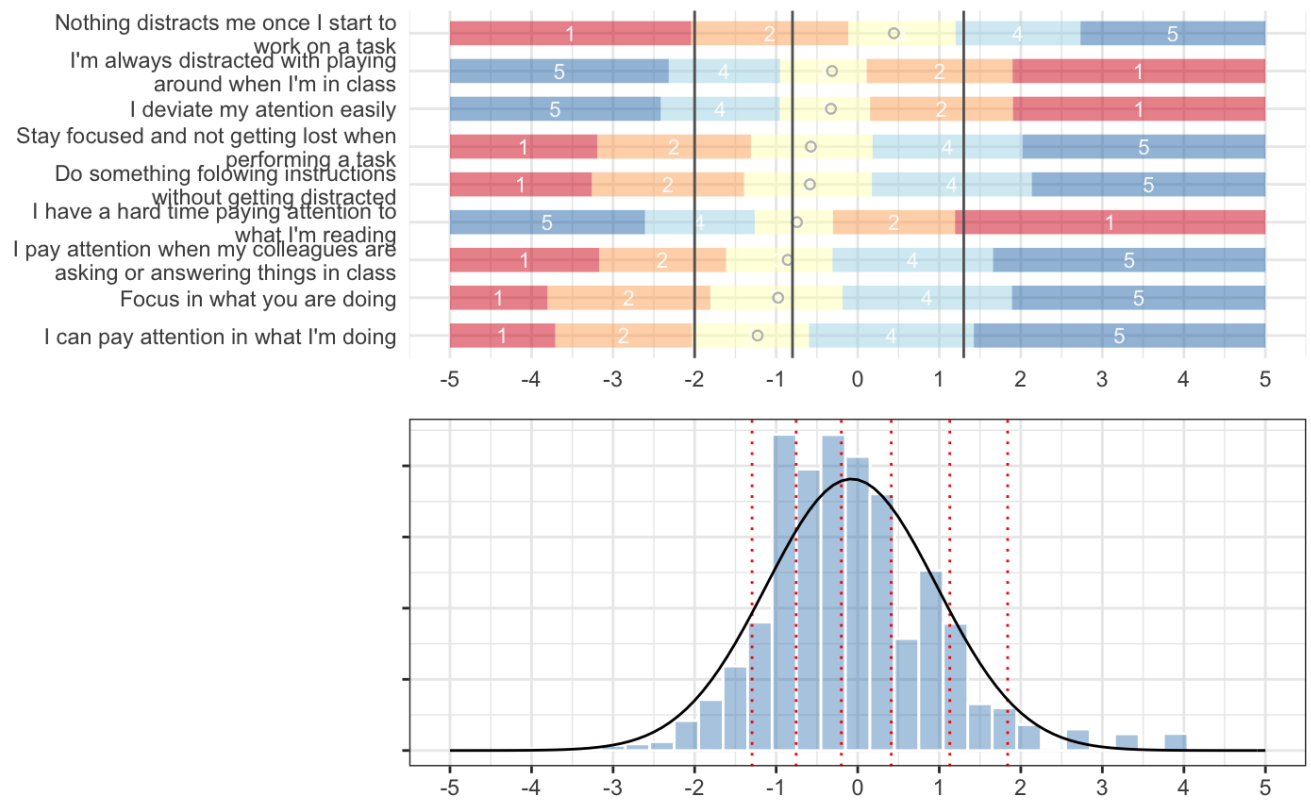
Construct maps

On the next step we created create construct maps to define substantive benchmarks to interpret SENNA scales. Construct maps (also known as person-item maps, wright maps, Wilson, 2005, 2011) are visual representations of the IRT model parameters and other qualitative information about the items that enhance the interpretation of the scale. Construct maps devise "item referenced meaning" a more modern psychometric method of standardizing interpretations of classic scales (Embretson, 2006). Usually construct maps shows information of theta (persons) and delta (items) parameters. Because these parameters are in the same metric, we can compare them directly and predict who (persons n with a given theta level) is going to answer what (answer categories of item i with a given delta difficulty parameter). Common metric creates a connection between theta scale and item difficulty allowing one to interpret theta scores using qualitative content of the items. We also added normative information to enrich the interpretation with statistical frequency of the scores.

Figure 2 shows the construct map of Focus facet of Self-Management domain. Lower half of this figure shows a histogram plotting the distribution of the theta scores of the students in São Paulo sample. X-axis shows the theta scores (same metric as delta parameters) or the logit metric as is known among Rasch community. Red dashed lines show percentiles 5, 10, 25, 50, 75, 90, and 95. This part of the figure allow us to interpret what is a low, medium and high theta score from their statistical frequency. We also show a density function of a theoretical normal curve with the parameters μ and σ set at the at sample values of mean (M) and standard deviation (SD).

---

[5] This package was wrote by Daniel Falbel, Wiliam Amorim from curso-r (https://www.curso-r.com) and Ricardo Primi

Figure 2.

Construct map of the facet Concentration of Self management domain.

Upper half of Figure 2 shows the item map. X-axis represents the delta meric (same metric as theta parameters). The horizontal bars represent what areas along the scale we expect observe each response category (1: red, 2: orange, 3: yellow, 4: light blue, 5: blue). Boundaries for these categories were obtained from half-score points from PCM model, that is, transition between 1 and 2, that is, the level where we PCM expects a response "1.5"; transition between 2 and 3 where it expects score 2.5 and so on. With this figure we can reconstruct expected responses patterns for a given theta score. For instcance, consider a percentile 10, θ = -1.34 (very low score) and trace a vertical line at this point. This line cross eight yellow bars (items) corresponding to category '3' ('*moderately like me*') and one orange bar '2' ('*little like me*') on last item 'Nothing distracts me once I start to work on a task'. Hence, PCM model predicts the response pattern 333333332 for this level. We can do the same exercice sistematicaly from low, mid to high level on the theta scale and check response patterns for each level

understanding qualitatively what subjects comunicate about their self image on the questions we asked them.

On item maps, items are ordered according to the average of their delta parameters (a global item location). These averages are plotted on item maps as a open circle. They invariantly appeared over yellow bars (category 3). Therefore difficult items (high global item location) are located on top; and easy items (low global item location) are located on he bottom. This global average is a index of general difficulty to endorse item. We can see that to score '4' and '5' in the most difficult item '*Nothing distracts me once I start to work on a task*' requires a higher theta than the easiest item '*I can pay attention on what I am doing*' (a self-efficacy item). This order also informs how constrruct are defined becasue increased levels of theta is represented by agreeing more strongly with items on the top in addition to the bottom ones.

Another feature to note is that flase keyed item are ploted before reversing their scores. This feature makes item map to show the exact response given. This allolw us to quickly recover the expeted response a student comunicated via our questionnaire. It is also interesting to see how false keyed items make the meaning of the low end of the construct more objective and direct. Students on the low end are comunication they have 'hard time paying attention'. Without false keyed items, the low end of the comnstruc will be ill defined, or at best defined indirectly.

Simple scale for reporting: 1: Low/to be developed, 2: emergent, 3: capable and 4: very capable

Figure 2 item map has another detail depicted in trhee vertical lines that divide the scale into four levels. Those lines are cut points to develop a simple 4-level straightfoward scale to interpret SENNA metric aiming to simplify comunication of results. SENNA authores worked in two-day session analyzing construct maps of each facet as the one presented in Figure 1 to produce tree cutpoints to create four levels 1: low/to be developed, 2: emergent, 3: capable and 4: very capable. They analized the response patterns and item comntent associated to increasing levels of theta to decide where to set these cutpoints. They also considered normative references to adjust those cutpoints. The decision where to set cut points were based on consensus among

authors. After defining the cutpoints authors wrote verbal descriptions for each level based on typical response pattern and item content. Apendix J contains all construct maps, cut points and descriptions that were written for each level.

Descriptive statistcs of the normative sample

We selected a sample from São Palulo to be the first reference group to calculate initial norm reference points. We selected students attending grades 6th to 12th, that have passed attention checks - that is, who answered the three control questions correctly, that had spent at least ten minutes answering the test and had age range consistent with grades they are enrolled. The final norm sample contained 29,976 students from grades 6th to 12th, 61.7% girls with age raging from 11 to 18 (M=14.8, SD=1.97). Theta domain and facet scores descriptive statistics and percentiles for this sample are presented in Tables 20, 21 and 22.

Table 20.

Desciptive statistics and percentiles of SP normative group of SENNA domains

	O	C	E	A	N
N	29976	29976	29976	29976	29976
Mean	-0.18	-0.154	-0.302	-0.407	-0.113
Median	-0.305	-0.266	-0.338	-0.465	-0.121
Standard deviation	0.849	0.879	0.673	0.761	0.789
Minimum	-3.94	-4	-3.7	-4	-4
Maximum	4	4	3.81	4	4
Skewness	1.06	0.711	0.175	0.463	0.0597
Std. error skewness	0.0141	0.0141	0.0141	0.0141	0.0141
Kurtosis	2.45	1.2	0.938	0.892	1.37
Std. error kurtosis	0.0283	0.0283	0.0283	0.0283	0.0283
10th percentile	-1.08	-1.15	-1.11	-1.33	-1.03
20th percentile	-0.862	-0.88	-0.838	-1.03	-0.685
30th percentile	-0.671	-0.649	-0.658	-0.822	-0.47
40th percentile	-0.513	-0.461	-0.476	-0.647	-0.297
50th percentile	-0.305	-0.266	-0.338	-0.465	-0.121
60th percentile	-0.0834	-0.0283	-0.148	-0.281	0.0593
70th percentile	0.107	0.195	-0.009	-0.0474	0.244
80th percentile	0.42	0.514	0.203	0.198	0.484
90th percentile	0.928	0.98	0.537	0.532	0.843

Table 21.

Desciptive statistics and percentiles of SP normative group of SENNA facets of O and C

	Aes	CrImg	IntCur	Achv	Conc	Ord	SD	SofR
N	29976	29976	29976	29976	29976	29976	29976	29976
Mean	-0.105	-0.087	-0.169	-0.120	-0.077	-0.001	-0.078	-0.072
Median	-0.177	-0.331	-0.422	-0.294	-0.201	-0.285	-0.214	-0.344
Standard deviation	1	1.17	1.09	0.979	1.05	1.38	1.18	1.27
Minimum	-4	-4	-4	-4	-4	-4	-4	-4
Maximum	4	4	4	4	4	4	4	4
Skewness	0.979	1.150	1.610	1.080	0.633	0.886	1.220	1.250
Std. error skewness	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014
Kurtosis	2.520	2.160	3.750	2.800	1.750	1.310	2.390	2.050
Std. error kurtosis	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028
10th percentile	-1.180	-1.320	-1.200	-1.220	-1.300	-1.410	-1.290	-1.400
20th percentile	-0.863	-0.953	-0.992	-0.828	-0.890	-1.060	-0.955	-1.030
30th percentile	-0.642	-0.709	-0.779	-0.697	-0.620	-0.812	-0.724	-0.774
40th percentile	-0.414	-0.585	-0.548	-0.430	-0.343	-0.557	-0.479	-0.494
50th percentile	-0.177	-0.331	-0.422	-0.294	-0.201	-0.285	-0.214	-0.344
60th percentile	-0.054	-0.067	-0.140	-0.010	0.096	0.169	-0.071	-0.012
70th percentile	0.203	0.216	0.021	0.294	0.411	0.521	0.242	0.380
80th percentile	0.632	0.703	0.400	0.635	0.752	0.947	0.610	0.614
90th percentile	1.170	1.360	1.250	1.040	1.130	1.530	1.400	1.680

Table 22.

Desciptive statistics of SP normative group of SENNA facets of E, A and N

	Act	Assr	Soc	Cmp	Mod	Resp	Tru	LAngrVol	LAnx	LDep
N	29976	29976	29976	29976	29976	29976	29976	29976	29976	29976
Mean	-0.341	-0.214	-0.265	-0.461	-0.269	-0.483	-0.006	-0.058	-0.050	-0.205
Median	-0.391	-0.311	-0.325	-0.684	-0.308	-0.847	-0.080	-0.101	-0.085	-0.183
Standard deviation	0.925	0.912	0.915	1.23	0.836	1.35	1.26	1	0.969	0.966
Minimum	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
Maximum	4	4	4	4	4	4	4	4	4	4
Skewness	0.428	0.611	0.543	1.210	0.917	1.360	-0.167	0.182	0.054	0.209
Std. error skewness	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.0141
Kurtosis	1.840	2.310	2.330	2.120	4.410	2.470	0.318	2.700	1.240	2.13
Std. error kurtosis	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.0283
10th percentile	-1.490	-1.260	-1.260	-1.680	-1.200	-1.850	-1.640	-1.180	-1.200	-1.260

20th percentile	-1.040	-0.898	-0.997	-1.440	-0.867	-1.470	-1.070	-0.726	-0.719	-0.871
30th percentile	-0.883	-0.664	-0.733	-1.180	-0.710	-1.250	-0.689	-0.517	-0.559	-0.605
40th percentile	-0.558	-0.429	-0.464	-1.030	-0.431	-0.997	-0.288	-0.210	-0.242	-0.468
50th percentile	-0.391	-0.311	-0.325	-0.684	-0.308	-0.847	-0.080	-0.101	-0.085	-0.183
60th percentile	-0.221	-0.066	-0.182	-0.480	-0.078	-0.478	0.351	0.118	0.224	-0.033
70th percentile	0.130	0.192	0.121	-0.004	0.035	-0.247	0.568	0.349	0.376	0.123
80th percentile	0.313	0.474	0.458	0.272	0.272	0.352	0.992	0.596	0.678	0.455
90th percentile	0.710	0.794	0.853	0.925	0.560	1.270	1.580	1.020	1.150	1.030

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

California Office to Reform Education (2014). *Social-Emotional Skills.* Retrivied from http://www.ousd.org/cms/lib07/CA01001176/Centricity/Domain/3154/Social%20and%20Emotional%20Learning.pdf

Carroll, J. (1963). A model of school learning. *The Teachers College Record*, 64(8), 723-723.

Crede, M., Tynan, M. C., & Harms, P. D. (2016). Much Ado about Grit: A Meta-Analytic Synthesis of the Grit Literature. Journal of Personality and Social Psychology, 113(3), 492–511

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2019). mirt: Multidimensional Item Response Theory (Version 1.31) [Computer software]. https://CRAN.R-project.org/package=mirt

Collaborative for Academic, Social, and Emotional Learning. (2012). *2013 Casel Guide: Effective Social and Emotional Learning Programs. Preschool and Elementary School Edition.* Chicago: KSA-Plus Communications.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112(1,155-159).

De Fruyt, F., Bockstaele, M., Taris, R., & Van Hiel, A. (2006). Police interview competencies: Assessment and associated traits. European Journal of Personality, 20(7), 567-584.

De Fruyt, F., & Van Leeuwen, K. (2014). Advancements in the field of personality development. [Article]. Journal of Adolescence, 37(5), 763-769.

De Fruyt, F., Wille, B., & John, O. P. (2015). Employability in the 21st century: Complex (interactive) problem solving and other essential skills. *Industrial and Organizational Psychology*, 8(02), 276-281.

De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of Personality, 10,* 303–336.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. Journal of Personality and Social Psychology, 92(6), 1087-1101.

Duckworth, A. L., & Quinn, P. D. (2009). Development and Validation of the Short Grit Scale (Grit-S). [Article]. Journal of Personality Assessment, 91(2), 166-174.

Dumfart, B., & Neubauer, A. C. (2016). Conscientiousness Is the Most Powerful Noncognitive Predictor of School Achievement in Adolescents. [Article]. Journal of Individual Differences, 37(1), 8-15.

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412.doi:10.1111/bjop.12046

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. [Article]. Child Development, 82(1), 405-432.

Elias, M. J., Zins, J. E., Weissberg, R. P., Frey, K. S., Greenberg, M. T., Haynes, N. M., et al. (Eds.). (1997). Promoting social and emotional learning: Guideliness for educators. Alexandria, VA: Association for Supervision and Curriculum Development.

Ferguson, G. A. & Takane, Y. (1989). *Statistical analysis in psychology and education*. New York, NY, US: McGraw-Hill.

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N.O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review*. Chicago: University of Chicago Consortium on Chicago School Research.

Gabrieli, C. Ansel,D, & Krachman, S. B. (2015). Ready To Be Counted: The Research Case for Education Policy Action on Non-Cognitive Skills. Boston: Transforming Ed/CORE districts. (https://coredistricts.org/ready-to-be-counted-why-non-cognitive-skills-must-be-incorporated-into-ed-policy-practice/)

Geary, D. C. (2007). An evolutionary perspective on learning disability in mathematics. *Developmental neuropsychology*, 32(1), 471-519.Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. Journal of Child Psychology and Psychiatry, 38, 581-586.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. and Yavitz, A. (2010), Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. Quantitative Economics, 1: 1–46. doi:10.3982/QE8

Hutz, C. S., Nunes, C. H. S. S. (2001). Escala Fatorial de Ajustamento Emocional/Neuroticismo. São Paulo: Casa do Psicólogo.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In *Handbook of research methods in social and personality psychology* (pp. 339–369). New York, NY, US: Cambridge University Press.

John, O. P. & De Fruyt, F. (2016). *Framework for the Longitudinal Study of Socio and Emotional Skills in Cities*. Paris: OECD. http://www.oecd.org/callsfortenders/CfT%20100001311%20Longitudinal%20Study%20of%20Social%20and%20Emotional%20Skills%20in%20Cities.pdf

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The "Big Five" Inventory - Versions 4a and 54. Berkeley: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: Discovery, measurement, and conceptual issues. In O.P. John, R.W. Robins,

& L.A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114-158). New York: Guilford.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self-evaluations scale: Development of a measure. [Article]. Personnel Psychology, 56(2), 303-331.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). Personality, motivation, and college readiness: A prospectus for assessment and development. *ETS Research Report Series*, 1-48.

Lipnevich, A. A., Preckel, F., & Roberts, R. D. (Eds.). (2016). Psychosocial Skills and School Systems in the 21st Century: Theory, Research, and Practice. Springer International Publishing. Retrieved from https://www.springer.com/us/book/9783319286044

Linacre, J.M. (2020). Winsteps® (Version 4.5.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2020. Available from https://www.winsteps.com/

Magis, D., Raiche, G., & Zaragoza, J. R. B. (2018). catR: Generation of IRT Response Patterns under Computerized Adaptive Testing (Version 3.16) [Computer software]. https://CRAN.R-project.org/package=catR

Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6(4), 311-360.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, 22(3), 471.

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2012). Measurement invariance of Big-Five Factors over the life span: ESEM tests of gender, age, plasticity, maturity and la dolce vita effects. Developmental Psychology, 49(6), 1194-1218. doi: 10.1037/a0026913.

McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. Journal of Personality and Social Psychology, 88(3), 547-561.

Ministério da Educação (2017). Base Nacional Comun Curricular - 3ª versão.

Millimet, D. L., & Tchernis, R. (2013). Estimation of treatment effects without an exclusion restriction: With an application to the analysis of the school breakfast program. *Journal of Applied Economics, 28*(6), 82-1017. doi: 10.3386/w15539

Muris, P. (2001). A brief questionnaire for measuring self-efficacy in youths. Journal of Psychopathology and Behavioral Assessment, 23, 145-149.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

National Research Council. (2012).*Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.

Nowicki, S., & Strickland, B. O. (1973). A locus of control scale for children. Journal of Consulting and Clinical Psychology, 40(148-154).

OECD (2015), Skills for Social Progress: The Power of Social and Emotional Skills, OECD Skills Studies, OECD Publishing. http://dx.doi.org/10.1787/9789264226159-en

Organização de Cooperação e de Desenvolvimento Econômico [OECD], (2018). Social and Emocional Skills: Well-being, Connectedness, and Sucess. Paris, France: OECD. http://www.oecd.org/education/school/UPDATED%20Social%20and%20Emotional%20Skills%20-%20Well-being,%20connectedness%20and%20success.pdf%20(website).pdf

Organização de Cooperação e de Desenvolvimento Econômico OECD (2016) Longitudinal study of social and emotional skills in cities. Paris: OECD (http://www.oecd.org/callsfortenders/CfT%20100001311%20Longitudinal%20Study%20of%20Social%20and%20Emotional%20Skills%20in%20Cities.pdf)

Osher, D., Kidron, Y., Brackett, M., Dymnicki, A., Jones, S., & Weissberg, R. P. (2016). Advancing the Science and Practice of Social and Emotional Learning: Looking Back and Moving Forward. Review of Research in Education, 40(1), 644–681. https://doi.org/10.3102/0091732X16673595

Ortiz, Borges, and Santos (2017).

Park, D., Tsukayama, E., Goodwin, G. P., Patrick, S., & Duckworth, A. L. (2017). A tripartite taxonomy of character: Evidence for intrapersonal, interpersonal, and intellectual competencies in children. *Contemporary Educational Psychology*, 48, 16–27. https://doi.org/10.1016/j.cedpsych.2016.08.001

Partnership for 21st Century Skills (2015). www.p21.org.\

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. [Review]. Psychological Bulletin, 135(2), 322-338.

Poropat, A. E. (2014a). A meta-analysis of adult-rated child personality and academic performance in primary education. [Article]. British Journal of Educational Psychology, 84(2), 239-252.

Poropat, A. E. (2014b). Other-rated personality and academic performance: Evidence and implications. Learning and Individual Differences, *34*(2014), 24-32.

Primi, R., Alves, G. & Marino, R. (no prelo). O modelo dos cinco grandes fatores e as capacidades pessoais socioemocionais. Em M. Bueno e M Ricarte (orgs). Habilidades socioemocionais: Aspectos teóricos e práticos. São Paulo: Hogrefe.

Primi, R., De Fruyt, F. D., Santos, D., Antonoplis, S., & John, O. P. (2019). True or false? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. *International Journal of Testing*, 0(0), 1–25. https://doi.org/10.1080/15305058.2019.1673398

Primi, R., Hauck-Filho, N., Valentini, F., Santos, D., & Falk, C. F. (2019a). Controlling acquiescence bias with multidimensional IRT modeling. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 39–52). Cham: Springer International Publishing. 10.1007/978-3-030-01310-3_4

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019b). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, 0(0). https://doi.org/10.1111/bmsp.12168

Primi, R., Santos, D., John, O. P., & Fruyt, F. D. (2016a). Development of an Inventory Assessing Social and Emotional Skills in Brazilian Youth. *European Journal of Psychological Assessment*, 32(1), 5–16. https://doi.org/10.1027/1015-5759/a000343

Primi, R., Santos, D., Hauck, N., De Fruyt, F., & John, O. P. (2019c). Mapping self-report questionnaires for socio-emotional characteristics: What do they measure? *Estudos de*

Psicologia (Campinas), 36, e180138. http://dx.doi.org/10.1590/1982-0275201936e180138

Primi, R., Hauck-Filho, N., Valentini, F. & Santos, D. (in press). Classical perspectives of controlling acquiescence with balanced scales. In M. Wiberg, S., D. J. González, U. Bockenholt, & Jee-Seon (Eds.), *Quantitative psychology* (pp. ??–??). Cham: Springer International Publishing.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016b). Anchoring Vignettes: Can They Make Adolescent Self-Reports of Social-Emotional Skills More Reliable, Discriminant, and Criterion-Valid? European Journal of Psychological Assessment, 32(1), 39–51. https://doi.org/10.1027/1015-5759/a000336

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. Journal of Personality and Social Psychology, *111*(57), 780-789.

Rizopoulos, D. (2018). ltm: Latent Trait Models under IRT (Version 1.1-1) [Computer software]. https://CRAN.R-project.org/package=ltm

Roberts, R., Martin, J. E. & Olaru, G. (2015). *A Rosetta Stone for Noncognitive Skills: Understanding, Assessing, and Enhancing Noncognitive Skills in Primary and Secondary Education*. New York: Asia Society, Professional Examination Service.

Rosenberg, M. (1979). Conceiving the self. New York, NY: Basic Books.

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test Analysis Modules* (Version 3.3-10) [Computer software]. https://CRAN.R-project.org/package=TAM

Saarni, C. (1999). The development of emotional competence. New York: Guilford Press.

Saarni, C. (2011). Emotional development in childhood. In R. E. Tremblay, R. G. Barr, R. V. Peters & M. Boivin (Eds.), Encyclopedia on early childhood development: Emotions (pp. 1-7). Montreal: Quebec: Centre of Excellence for Early Childhood Development and Strategic Knowledge Cluster on Early Child Development.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). Lifetime effects: The High/Scope Perry Preschool study through age 40. (Monographs of the High/Scope Educational Research Foundation, 14). Ypsilanti, MI: High/Scope Press

Shoss, M. K., & Witt, L. A. (2013). Trait interactions and other configural approaches to personality. In N. D. Christiansen & R. P. Tett (Eds.), Handbook of Personality at Work (pp. 392-418). New York: Routledge.

Santos, D., & Primi, R. (2014). *Social and emotional development and school learning: a measurement proposal in support of public policy*. Technical report for Organization for Economic Cooperation and Development (OCDE) Rio de Janeiro State Education Department (SEEDUC) and Ayrton Senna Institute. São Paulo: Ayrton Senna Institute.

Sailer, O. (2005) crossdes: A package for design and randomization in crossover studies. Rnews 5/2, 24-27.

Sklad M., Diekstra, R., Ritter, M., Ben, J., & Gravesteijn, C. (2012). Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students'

development in the area of skill, behavior, and adjustment? *Psychology in the Schools, 49*(9), 892-909. doi: 10.1002/pits.21641

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117-143. doi: 10.1037/pspp0000096

Soto, C. J., & John, O. P. (2012). Development of big five domains and facets in adulthood: mean-level age trends and broadly versus narrowly acting mechanisms. Journal of Personality, 80(4), 881-914. doi:10.1111/j.1467-6494.2011.00752.x

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, 94(4), 718–737. https://doi.org/10.1037/0022-3514.94.4.718

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big-Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology, 100, 330–348.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? Journal of Personality and Social Psychology, 84(5), 1041-1053. Retrieved from https://search.proquest.com/docview/209804598?accountid=11077.

Tabachnick, B. G., & Fidel, L. S. (2013). *Using Multivariate Statistics* (6th ed., pp. 153-235). Essex: Pearson Education

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting Positive Youth Development Through School-Based Social and Emotional Learning Interventions: A Meta-Analysis of Follow-Up Effects. *Child Development, 88*(4), 1156-1171. doi: 10.1111/cdev.12864

Transforming Education (2016). Measuring MESH: Student and Teacher Surveys Curated for the CORE Districts. Boston: Transforming Ed/CORE districts. (https://www.transformingeducation.org/measuring-mesh/)

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times.* San Francisco, CA: Jossey-Bass.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, 28(5), 317–331. https://doi.org/10.1177/0146621604264870

Wright, B., & Masters, G. (1982). Rating scale analysis. Measurement and Statistics. Retrieved from https://research.acer.edu.au/measurement/2

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational Measurement for Applied Researchers. https://doi.org/10.1007/978-981-10-3302-5. Singapure: Springer

Zins, E. J., Weissberg, R. P., Wang, M. C. & Walberg, H. J. (2004), Building academic success on social and emotional learning: What does the research say? (pp.3–22). New York: Teachers College Press.

Appendix

See file: Senna_Tech_Manual_Appendix_v0.pdf