SKILLS FOR SUSTAINABLE DEVELOPMENT:
ESSAYS ON HOW CREATIVITY, ENTREPRENEURSHIP AND EMOTIONS
FOSTER HUMAN DEVELOPMENT

PABLO ANDRES EGAÑA DEL SOL

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2016
ABSTRACT
Skills For Sustainable Development: Essays On How Creativity, Entrepreneurship
And Emotions Foster Human Development
Pablo A. Egaña del Sol

This dissertation deals with the understanding of policies and interventions that foster human development, with a focus on children and youth, and to build a link between human development and a learning society by using an interdisciplinary approach. This thesis studies how individuals develop their learning capabilities as well as their creative, entrepreneurial, and socio-emotional skills.

The dissertation is structured in three main chapters in addition to this brief introduction. The first chapter, “Affective Neuroscience meets Labor Economics: Assessing Non-Cognitive skills on Late Stage Investment on at-Risk Youth,” studies the role of a program designed to foster entrepreneurial and self-confidence through learning by failure using insights from micro-econometric, behavioral economics and applied neuroscience. The second chapter, “How Much Should We Trust Self-reported Measures of Non-cognitive Skills?,” explores the relation between transient emotional states and self-ratings on self-reported measures of socio-emotional skills using a behavioral and a neuro-physiological experiment. This chapter also works as a “proof of concept” of the methods — e.g. emotion-detection theory and lab-in-the-field experiments implemented on chapter 1. The third chapter, “Can Art-based Programs Nurture Human Capital? Evidence From Public Schools in Chile,” studies the impacts of an art-based program in high school in Chile following a quasi-experimental design using propensity score matching techniques.
Contents

List of Figures iii

List of Tables v

1 Introduction 1

2 Affective Neuroscience Meets Labor Economics: Assessing Socio-emotional Skills on Late Stage Investment on at-Risk Youth 9

2.1 Introduction .................................................. 9

2.2 Background and Conceptual Framework .......................... 15

2.3 The Role of Emotions .................................. 29

2.4 Experimental Paradigm ....................................... 38

2.5 Results .......................................................... 45

2.6 Conclusions .................................................. 68

2.7 Appendix 1: Educational Outcomes and EEG features .......... 72

2.8 Appendix 2: Balance of Attrition ............................. 82

2.9 Appendix 3: Robustness Checks ................................ 83

3 How Much Should We Trust Self-reported Measures of Non-cognitive Skills? 87
# List of Figures

2.1 Arousal-Valence Model of Emotions ................................. 30
2.2 Geographic Location of Schools in the Field Experiment .......... 40
2.3 Field Experiment Timeline ........................................... 41
2.4 Field Experiment Paradigm .......................................... 42
2.5 Experimental Paradigm of Emotional Responsiveness .............. 43
2.6 Field Experiment Setting............................................... 44
2.7 Box Plots: EEG features and Educational Outcome ............... 75
2.8 Distribution of EEG Features by Registration to SAT-like test (PSU) status 76
2.9 Distribution of EEG Features by Treatment status ................. 77

3.1 Emotiv EPOC Device .................................................. 100
3.2 10-20 Location System of Electrodes ............................... 101
3.3 GAPED Negative Stimulus Examples .................................. 104
3.4 GAPED Positive Stimulus Examples ................................... 104
3.5 GAPED Neutral Stimulus Examples ................................... 105
3.6 Experimental Paradigm ................................................ 112
3.7 Average Accuracy of Emotion Detection (Multiple Classifiers) ..... 115
3.8 Emotionally-laden Stimuli and Valence .............................. 116
3.9 Proof of Concept: Grit Test and Emotional Valence ............... 118
3.10 Valence: Positive-Negative .......................................... 125
3.11 Component via PLR ................................................... 126
4.1 Examples of TTCT Dimensions: Fluency ........................................ 137
4.2 Examples of TTCT Dimensions: Elaboration ............................ 138
4.3 Examples of TTCT Dimensions: Criterion-referenced Measures . . . . 138
## List of Tables

2.1 Experimental Balance at Baseline .............................................. 46  
2.2 Attrition ...................................................................................... 47  
2.3 Registration to take PSU (SAT-like) test ....................................... 51  
2.4 Students Dropout Rates ............................................................... 51  
2.5 Difference Model on Non-cognitive Skills: Creativity, Locus Control and Grit ............................................................. 53  
2.6 Difference in Difference Model on Non-Cognitive Skills: Creativity, Locus Control and Grit ............................................................. 55  
2.7 Impact on Emotional Regulation: Emotional State and Responsiveness . 57  
2.8 Probit Model: Registration to SAT-like test (PSU) and EEG Features . . 78  
2.9 Variance Ratio Test: Arousal Baseline (no stimuli) by Treatment Status 79  
2.10 Variance Ratio Test: Valence Baseline (no stimuli) by Treatment Status . . 79  
2.11 Variance Ratio Test: Valence Positive stimulus by Treatment Status . . 80  
2.12 Variance Ratio Test: Valence Negative stimulus by Treatment Status . . 80  
2.13 Variance Ratio Test: Difference Valence Positive and Baseline by Treatment Status ................................................................. 81  
2.14 Variance Ratio Test: Difference Valence Negative and Baseline by Treatment Status ................................................................. 81  
2.15 Balance of Attrition (missed subjects compared at baseline) .......... 82  
2.16 Placebo Test - Raven-like Test and Grade Previous Year ............... 83
I want to thank first and foremost my main advisors, Miguel Urquiola and Paul Sajda, for their guidance, feedback, and friendship of incomparable quality. A special thanks to Miguel for his tireless work in reading dozens of versions of my PhD thesis and providing meticulous and assertive comments and suggestions. I would also like to extend particular thanks to my committee members Mark Dean and Elke Weber, who also gave me thoughtful feedback during the process of planning and undertaking setting this thesis. Additional thanks to John Mutter, who always supported me and encouraged me to take risks.

I thank the extraordinary faculty for whom I served as a teaching assistant during my time as a Ph.D. student. Guillermo Calvo, Victoria Murillo, and Andres Velasco all told me that excellence in research and teaching are important complements. Andres, in addition to his invaluable feedback and unconditional support in the job market, gave me the invaluable advice that good research has also to be relevant to policy.

I also thank Mike Woodford, Mark Dean, and Pietro Ortoleva for inviting me to participate in the Cognition and Decision Lab at the Department of Economics. The experience in the lab was extremely fruitful and interesting and allowed me to get to know Ambuj Dewan, Mel Win Khaw, Silvio Ravaioli, Evan Friedman, Felix Mauersberger, Adriano Fernandez, Ziang Li, and Gabriella Zacarias, who are now amazing colleagues and close friends.
I would like to thank Cristian Pop-Eleches, Joseph Stiglitz, and Chris Blattman for their frequent support and advice over the years I spent working on this project. For their invaluable comments and criticisms, I would also like to recognize Suresh Naidu, Angela Duckworth, Ricardo Hausmann, Cesar Hidalgo, Rosalind Picard, Karla Hoff, Sanjay Peters, Michael Platt, Douglas Almond, Eric Verhoogen, Douglas Almond, Ernesto Reuben, Alessandra Casella, Paulo Barraza, Michael Platt, Esteban Puentes, Martha Kluttig, Dante Contreras, Francisco Gallego, Paula Luengo, Hugo Nopo, and Claudia Martinez.

To my colleagues at Columbia University—Claudia Allende, Gonzalo Munoz, Gonzalo Mena, Piero Montebruno, Micaela Sviatschi, Riccardo Ciacci, Evan Riehl, Goran Lazarevski, Scott Weiner, Chris Hansman, and Nicolas de Roux—I thank you for your friendship and comments in early stages of this project.

Thank you Mona Khalidi, Tomara Aldrich, Douglas Almond, and John Mutter for successfully managing this interdisciplinary program, and supporting us with all administrative needs. Mona Khalidi deserves special thanks for her warm support and help. She embraced me with a welcoming attitude from the moment I began my program.

Nicole Moscowitz and Tao Tu provided me with invaluable research assistance along the neuroscience paradigm. Without them, the interdisciplinary goal of this thesis would never have been achieved. Moreover, Simon Ballesteros, Macarena Gonzalez, Nicolas Fernandez, and Jose Manuel Belmar provided effective research assistance in the field.

For the implementation of the impact evaluation of the “Mining Rockstars” program, I thank Benjamin Faivovich, Jose Miguel Barros, and Esteban Alvarez from Emprende Joven in Chile.

I would like to thank all my fellows at the Sustainable Development program:
Aaron Baum, Anthony D’Agostino, Habtamu Fuje, Kimberly Lai, Stephanie Lackner, and Xiaojie Zhang. Their friendships are the most precious gift that this program gave me.

I would like to thank Jan von der Goltz, Solomon Hsiang, Amir Jina, Gordon McCord, Prabhat Barnwal, Kyle Meng, Anna Tompsett, Belinda Archibong, James Rising, Jesse Anttila-Hughes, Johannes Castner, and Kayleigh Campbell, all of whom loved the kind of research that I was doing besides been outside the core of the program.

Last but not least, to the younger cohorts—Ana Varela, Mehdi Benatiya Andaloussi, Raimundo Atal, Anouch Missirian, Eyal Frank, Eugenie Dugoua, Francis Annan, Ruinan Liu, Tim Foreman, and Zihan Cui—for their friendship and in anticipation of sharing the excitement of doing meaningful research with them in the future.

Finally, I acknowledge the financial support provided by Columbia University, the Centre for Social Conflict and Cohesion Studies (CONICYT/FONDAP/15130009) at University of Chile, and the InterAmerican Development Bank.
This page intentionally left blank
Dedication

La pasión por la rigurosidad en el trabajo bien hecho y en el pensamiento se la debo por siempre a mi hermana M. Francisca. Sin la convicción y apoyo incondicional de mi hermana mayor, nunca hubiese tenido la fuerza suficiente para emprender estos desafíos.

Esta tesis se la dedicó a Stephanie Froimovich por su inmenso cariño, por creer en mí, y en este proyecto.

A la memoria de mis abuelos Fernando y Rafael Luis, quienes me enseñaron mucho diciendo poco.
Chapter 1

Introduction

Motivation

We are facing a great challenge as a society to create a learning society, as Stiglitz and Greenwald (2015) pointed out in their recent book, Creating a Learning Society. My doctoral thesis focuses on the role of creativity and emotional regulation in human development and how they can create such a learning society. The role of emotional regulation, which involves developing the ability to affect emotional response to a stimulus, and participation in artistic and creative activities on multiple behaviors and educational outcomes, are a potential contribution to the understanding of how to achieve this goal. The United Nations Development Program (UNDP) defines human development as the process of expanding the richness of human life; that is, developing capabilities in health, education, and income, rather than simply the richness of the economy in which human beings live. It is an approach that is focused on people, their opportunities, and their choices.

Another goal of this dissertation is to advance the understanding of policies to accomplish one of the United Nation’s Sustainable Development Goals (SDGs) that
claims to “Ensure Effective Learning for All Children and Youth for Life and Livelihood: Primary schooling outcomes, Secondary schooling outcomes and youth participation in the labor force.” Furthermore, indirect implications of my thesis would help to achieve crosscutting SDGs such as early childhood development programs, progress in science, technology and innovation, such as entrepreneurial skills, transition into labor markets, among others.

This dissertation deals with the understanding of policies and interventions that foster human development, with a focus on children and youth, and aims to build a link between human development and a learning society through an interdisciplinary approach. It studies how individuals develop their learning capabilities as well as their creative, entrepreneurial, and socio-emotional skills. In particular, I consider two interventions in public schools in Chile: a program designed to foster entrepreneurial and self-confidence through learning by failure, and an art-based program. Both programs targeted children and youth at risk who lives and attend school among the poorest neighborhoods in Chile.

This thesis aims to contribute to a deeper understanding of human development by combining behavioral economics and applied neuroscience. After many years working on these topics, I advocate for the opportunity and necessity to advance interdisciplinary research that can contribute to both general theory on human development and provide evidence that is more policy-relevant. In fact, there are several limitations on disciplinary research in order to understand how learning, socio-emotional skills, and creativity are fostered through public policy. First, the literature from education, while having the best qualitative evidence and in-situ understanding, rarely uses causal inference in quantitative research. Literature on economics, on the other hand, remains limited to quantitative approaches that are based on self-reported measures on skills or tautological relations between past and future behaviors. In applied neuroscience, the focus on well-designed experimental data has limitations on the ex-
tent to which the findings can be generalized to other populations. Finally, program evaluation research in the psychology literature relies on self-reported psychometric scales; for example, learning abilities and personality traits that are stable over time and less influenciable in the short or medium term through policy adaptations. When integrated, these fields have advanced our understanding of many dimensions of human development, but there is a need to make this research more applicable to the way public policy shapes education and labor markets. In my thesis you will find examples of such interdisciplinary work. In order to accomplish the work done along this thesis, I have had the pleasure to work with artists, economists, psychologists, neuro-engineers, among others.

All chapters in the present thesis apply field, lab and/or lab-in-the-field experiments and randomized controlled trials to test insights from behavioral and labor economics, and psychology in the contexts of human development—mostly formal educational settings—and to evaluate programs and policy effectiveness using methods from econometrics and applied neuroscience. I was motivated to gather data firsthand because of how important primary data is for overcoming methodological constraints. I complemented my experimental data with administrative data from governmental agencies. I received funding to do all this work from many sources, including governments, multinational institutions, and universities.

Chapter summaries

The dissertation is organized into three main chapters in addition to this brief introduction. The first chapter, “Affective Neuroscience meets Labor Economics: Assessing Non-Cognitive skills on Late Stage Investment on at-Risk Youth,” studies the role of a program designed to foster entrepreneurial and self-confidence through
learning by failure using insights from micro-econometric, behavioral economics and applied neuroscience. The second chapter, “How Much Should We Trust Self-reported Measures of Non-cognitive Skills?,” explores the relationship between transient emotional states and self-ratings on self-reported measures of socio-emotional skills using a behavioral and a neuro-physiological experiment. This chapter also works as a “proof of concept” of the methods used to measure them; that is, and includes, emotion-detection theory and lab-in-the-field experiments implemented on chapter 1. The third chapter, “Can Art-based Programs Nurture Human Capital? Evidence From Public Schools in Chile,” studies the impacts of an art-based program in high school in Chile following a quasi-experimental design using propensity score matching techniques.

The first chapter aims to understand the mechanisms through which programs designed to develop socio-emotional skills affect educational and labor market outcomes. This chapter uses evidence from chapter 2 (where I show that emotions also bias socio-emotional skill test scores). This bias implies that empirically detecting a program’s impact on socio-emotional skills and creativity is a non-trivial task. In particular, I conducted a randomized control trial (RCT) on a program designed to foster creative skills among poor students in vocational schools in Chile. I conducted a baseline survey in March 2015 and conducted a follow-up in August 2015, collecting more than 500 electroencephalograms (EEG) and the results of standard psychometric tests. To the best of my knowledge, this is the largest neuro-physiological recording ever done in the field. Through the results, I found that the relevant program does not improve self-reported measures of non-cognitive skills and creativity, which is consistent with the puzzle in related literature. Instead, I found an impact on emotional regulation using EEG measures as a proxy.

The second chapter serves as a proof of concept of the methodology that I designed using EEG recordings. Measures of cognitive skills—i.e., GPA, IQ, et cetera—
account for a small fraction of the variance in salaries and other economic outcomes (Bowles et al., 2001; Heckman et al., 2006). Therefore, there is an increasing interest in elucidating other factors that might explain that variance; in particular, the role played by socio-emotional skills. In various economics literature, empirical attempts to measure the aforementioned dimensions (cognition, socio-emotional skills, and creativity) have been insightful but are problematic due to the low reliability of the proxies used in their measurements (Cunha and Heckman, 2008; Cunha et al., 2010; Almlund et al., 2011; Calero et al., 2014; Attanasio et al., 2015a; West et al., 2015).

This chapter is intrinsically related to the impact evaluation done in chapter 1 and was designed as a deeper proof of concept of the relationship between emotions and self-reporting on measures of non-cognitive skills. Its purpose is to show a plausible correlation between self-reported psychometric tests and transient emotional states, with the latter estimated from EEG recordings from low-cost portable devices. To make this conclusions, several relationships had to be established: namely, (i) showing the behavioral correlation between self-reported psychometric tests and transient emotional states, (ii) detecting emotional state and responsiveness from scalp EEG recordings, and (iii) showing the correlation between self-reported psychometric tests and transient emotional states using data from the recordings.

I argue that there is a positive correlation between emotional state and test scores that rely on self-ratings. In particular, I claim that self-reported tests used to measure both cognitive and non-cognitive skills are usually biased because of transient emotions that arise during testing. This study’s results differ from the asymmetric relation found in Querengasser et al. (2014), which solely observed effects for negative transient emotions on specific dimensions of a personality test.

This chapter contemplates two experiments. The first, referred to as the Behavioral experiment, estimates the capacity of the GAPED database to elicit emotions
and generate a significant change on self-ratings on self-reported socio-emotional measures. The results demonstrated that a positive stimulus increases test results relative to a negative one by 0.31σ and 0.16σ on the Grit Scale, and for a non-cognitive skill index, which comprises both Locus of Control and Grit Scale, respectively. These results are economically relevant considering the evidence of effects size from educational policies and programs (e.g. Duflo et al., 2012; Murnane and Gaminian, 2014). The second, the neurophysiological experiment, enriches this study in two ways. First, it validates the capacity of a low-cost EEG headset—i.e. the Emotiv EPOC—to detect emotions through traditional methods of affective neuroscience, with a 79% average accuracy level across subjects and classifiers. Further, it establishes a positive correlation between EEG-derived emotional state and self-reported psychometric test scores. Specifically, a statistically significant positive correlation was found for a random effects model.

In addition, using the same experiment, I estimated the relationship between EEG-derived and self-reported emotional states. I estimated the correlation between emotional-state indices—arousal and valence—from EEG recordings and those from the self-rating in common psychometric tests. I found a positive correlation, which is significant at an 80

The third chapter of my dissertation is an example of my interest in studying how individuals in vulnerable contexts develop their learning capabilities as well as their creative, entrepreneurial, and socio-emotional skills. Specifically, this chapter explores the role that participation in an art-based program in Chilean public high schools has on these dimensions. This chapter provides one of the first plausible causal estimates of the role that intensive participation in art-based programs has on knowledge, learning and cognitive skills; for instance, GPA in different subjects as well as creativity among students with low socioeconomic status. I explore the effects on cognitive, socio-emotional, and creative skills using innovative psychometric in-
struments. For instance, I measured creativity using a written format inspired by the work of Guilford’s alternative uses test (1967), and the graphical format that requires subjects to make several drawings called Torrance’s Test of Creative Thinking (1966). Both tests measure creativity in three elemental dimensions —fluency, flexibility and originality— plus a multidimensional index that measures thirteen creative virtues. In order to produce accurate assessments on all dimensions, I recruited three research assistants who are artists in addition to a psychologist.

Finally, this chapter was motivated by work I did at the University of Chile where I studied the impact of a youth orchestra on knowledge and learning (academic achievement) as well as perseverance on educational attainment (SAT-like scores increases over years). Two main findings can be derived from my empirical findings. Firstly, the intensity of treatment is crucial. The impacts of an art based program on the aforementioned dimensions become statistically and economically significant when students had participated in least two workshops. Secondly, the program influences academic achievement by producing a more flexible and sharp understanding of novel knowledge or theoretical frameworks. An understanding of the link between art-based program participation and the development of cognitive and non-cognitive skills would allow for the design of educational policies that pursue a more fruitful outcome than traditional educational systems.
Chapter 2

Affective Neuroscience Meets Labor Economics: Assessing Socio-emotional Skills on Late Stage Investment on at-Risk Youth

2.1 Introduction

In developing countries, youth often face scarce employment opportunities. The literature generally cites a lack of appropriate skills as a key reason for this (Bassi and Urzua, 2010; Pierre et al., 2014). Since the work of Bowles and Gintis (1976); Gintis (1971), research holds that these skills can be both cognitive and non-cognitive in nature. Cognitive skills encompass the use of language, memory, and logical-mathematical knowledge, while non-cognitive skills —also referred to as socio-emotional or life skills— include self-confidence, self-control, internal locus control, and grit (perseverance).

Recent evidence shows that interventions aimed at the non-cognitive skills of youth
can have a significant impact on behavior and/or outcomes. For instance, Carrel and Sacerdote (2013) study a college coaching intervention in New Hampshire and find a positive effect on enrollment and continued attendance in college. Blattman et al. (2015), shows that behavioral therapy can reduce crime and violence among at-risk youth in Liberia. In addition, studies suggest that programs to improve socio-emotional skills are more effective among students who are still enrolled in secondary schools (Heckman and Kautz, 2012; Cunha et al., 2010).

The existence of such impacts is not surprising to the extent that the neuroscience literature suggests that it is possible to affect non-cognitive skills during adolescence.\(^1\) The prefrontal cortex, which is related to emotion and self-control, is malleable into the early 20s (Fuster, 2002, 2013; Sigman et al., 2014) because the brain is still developing, particularly those areas related to identity, moral and social consequences of actions, and emotions (Fuster, 2013). Indeed, there is more synaptic activity of new neurons during adolescence than in the first years of life (Koelsch, 2012; Fuster, 2002, 2013; Levitin, 2006).\(^2\) To summarize, both the labor economics and the neuroscience literature suggest that late-stage non-cognitive investments can have a positive impact on behavioral outcomes that are associated with non-cognitive skills.

In this context a puzzle arises in the labor economics literature. Specifically, the evidence indicates that educational or labor market programs can improve individuals’ labor market outcomes; yet, contrary to the expectation the above would suggest, they do not seem to affect individuals’ non-cognitive skills measurements (Calero et al., 2014; Card et al., 2011; Ibarraran et al., 2014; West et al., 2015). This

---

\(^1\) Respect to cognitive skills is still an open debate. On one hand, some authors argue that after age 10, IQ level—one of the main measures of cognitive skills—remains relatively stable (Almlund et al., 2011; Koelsch, 2012). On the other hand, there is also evidence that shows that even cognitive skills can be affected during adolescence or later (Sigman et al., 2014).

\(^2\) The brain’s synapses are programmed to grow for a number of years until they shift to pruning away unneeded connections. Myelination—a substance that coats the axons, speeding up the synaptic transmission—is boosted between the ages of 8 and 16 (Koelsch, 2012).
is despite the fact that the qualitative evidence arising from participant, employer, expert, and program manager interviews usually suggests that improvements in non-cognitive skills are a main channel for the labor market effects (Calero et al., 2014; Ibarraran et al., 2014; Fazio, 2011). This emerges, for instance, in work on active labor market policies (ALMP) (Calero et al., 2014; Card et al., 2011; Ibarraran et al., 2014) and in the literature covering in-school programs (Morrison and Shoon, 2013; West et al., 2015; Egana-delSol, 2016a).

In short, the theory and the qualitative evidence suggest that interventions’ impact on labor market outcomes should be mediated by improvements in non-cognitive skill, yet there is little evidence that they actually affect socio-emotional skills. There are two main candidate explanations for this. First, it may be that non-cognitive skills simply do not account for the labor market impacts that are found. Second, it may be that the impacts on socio-emotional skills are difficult to detect because they are measured with noise, in part because they are generally self-reported. Consistent with the latter possibility, recent meta-analyses conclude that studies using self-reported psychometric tests to proxy socio-emotional skills face potentially significant measurement error.3

This study provides further evidence consistent with the second possibility. My hypothesis is that social programs affect participants’ emotional regulation4 which in turn makes it difficult to measure their impact on non-cognitive skills.5

Emotional regulation matters because neuroscientists highlight the critical role

---

3 See for a review Almlund et al. (2011); Heckman and Kautz (2012).

4 Emotional regulation is a different concept than emotional intelligence. On one hand, emotional regulation can be define as a mixture of cognitive and emotional processes that shape a mental state—i.e. a disposition to act (Salzman and Fusi, 2010). In other words, emotional regulation involves developing the ability to affect the emotional response to a stimulus. On the other hand, emotional intelligence consists of four capabilities or competencies: self awareness, self management, social awareness and social skills (Goleman, 2010).

5 In Chapter 3 of this dissertation, I study the relation between transient emotional state and self-reporting on non-cognitive skills tests. See also Querengsser and Schindler (2014).
that emotions play in cognition, perception, attention, and memory (Damasio, 1994; Lakoff, 2008; Salzman and Fusi, 2010; Fuster, 2013). Furthermore, work that combines neuroscience, behavioral science, and economics supports the notion that emotions influence economic behavior and decision-making, as well as labor market performance (e.g., occupational choice, salaries, entrepreneurship, etc.).

Taken together, this suggests that emotion can influence cognition and behavior in powerful ways. Indeed, the behavioral economics literature suggests that even minor mood manipulations have a substantial impact on behavior (DellaVigna, 2009).

To explore these issues, I follow the James-Lange theory in which emotions are organized along the arousal and valence locus. Arousal is related to excitement. Valence could be interpreted as a positive or negative mood, as well as an attitude of either approach or withdrawal towards/from a stimulus (Harmon-Jones et al., 2010; Kassam et al., 2013). To proxy emotional regulation —i.e. emotional state and responsiveness— I rely on emotion-detection theory from the affective neuroscience literature that uses electroencephalogram (EEG) recordings to measure emotions. In particular, I use low-cost portable EEG headsets to obtain a proxy measure of subjects’ emotional states — i.e. pre-test resting state— as well as emotional responsiveness to both positive and negative stimuli, in the arousal-valence locus. This theory and its methods have been widely studied in the applied neuroscience literature.

In particular, I study the Chilean program “Mining’s Rockstars” ( “Rockstars para la Mineria” in Spanish), which aims to foster life skills with a focus on self-confidence and creativity, following principles that are summarized as: “learning by

---

6 For recent reviews see Weber and Johnson (2009); Lempert and Phelps (2014); Lerner et al. (2015).

7 For reviews see Loewenstein et al. (1992, 2001); Damasio (1994); Weber and Johnson (2009); Lempert and Phelps (2014); Lerner et al. (2015).

8 See Chapter 3 for details.

9 See for example Takahashi et al. (2004); Bos (2006); Brown et al. (2011); Ramirez and Vamvakousis (2012); Verma and Tiwary (2014).
failing, gaming, doing, and rethinking”. The program is implemented in the IV Co-
quimbo Region of Chile, where mining is the central economic activity. “Mining’s
Rockstars” creates didactic materials, namely: a student textbook, a teacher text-
book, and class-by-class guidelines together with videos for each activity, all based on
the aforementioned principles and a Harvard leadership model (McClelland, 1973).
The intervention consists of weekly workshops in which students participate in differ-
ent activities designed to improve their life skills. Program instructors work together
with school teachers to train them and enhance local capabilities at the schools.

The program was randomly assigned to technical/vocational high schools enrolling
students aged 16-18. I use this set-up, which is a small cluster-randomized con-
trolled experiment, to explore the effects that students’ participation (for one aca-
demic semester—around four months) has on educational outcomes —e.g. school
dropout rates and SAT-like test registration— and on their non-cognitive skills.10 I
measure these using three tests: Rotter’s locus control scale, Grit—perseverance—
Scale, and Torrance’s test of creative thinking.11 In addition, I explore how program
participation affects emotional regulation.

There are three main findings. First, I find that the program had significant
impact on educational outcomes —i.e. in both dropouts rates and SAT-like test
registration— yet no impact on the expected mechanism, such as socio-emotional
skills or creativity measures, which is consistent with the findings in the labor eco-
nomics literature referenced above. Second, I find significant impacts on emotional
state —in both arousal and valence indices— from neurophysiological recordings.
If emotional disposition can bias self-reported measures of non-cognitive skills, as

---

10 I also observe the effects on cognitive skills using the Raven’s similar test (Raven, 1936); I use
this score as a placebo. See the Appendix for details.

11 For details see the next section or Grit Scale (Duckworth et al., 2007), Locus Control Scale
(Rotter, 1966), and Torrance’s test of creative thinking (Kim et al., 2013; Kaufman and Sternberg,
2010).
suggested above (Querengsser and Schindler, 2014; Egana-delSol, 2016b), this may account for the lack of evidence of an impact on non-cognitive skills —the puzzle highlighted above. Third, program participation also reduces individuals’ emotional reaction to negative stimuli —one might say that it makes individuals more resilient. Together, the results may suggest that more cool-headed, and/or phlegmatic students exhibit better educational outcomes.

The present study involves different scientific disciplines, including the economics of education, labor economics, affective neuroscience, and applied neuroscience — i.e. brain-computer interface. Given the specificity, sensitivity and complexity of the required dimensions of this study, the data were directly collected by the author. To the best of my knowledge, this is the first research study to apply large-scale neurophysiological recordings from EEG in field experiments, and also in the context of a social program evaluation. In short, the purpose of this study is to contribute to the understanding of the human capital production function in terms of cognitive, socio-emotional and creative skills by combining neurophysiological, behavioral and demographic data in a late-stage intervention setting.\footnote{Regardless, since I am doing a program evaluation in the medium scale range, this study will not address equilibrium effects, which are more likely to occur when the program or policy are taken to a scale and sustained for a longer period of time (Banerjee and Duflo, 2008; Acemoglu, 2010; Deaton, 2010; Pop-Eleches and Urquiola, 2013).}

The methodology proposed in this paper has many benefits. First, it offers a way to incorporate emotion into the labor economics field. The importance of emotional stability in labor markets and overall life satisfaction has recently been highlighted for both developed and developing countries (Deming, 2015; OECD, 2015). Furthermore, international multilateral institutions have emphasized the need to improve cognitive and socio-emotional skill assessment.\footnote{For instance, the Organization of Economic Cooperation and Development (OECD, 2015), the World Bank (McKenzie et al., 2014; McKenzie, 2014), and the Inter American Development Bank (Ibarraran et al., 2014).}
Second, it also may aid evaluation of similar programs attempting to foster non-cognitive skills. In fact, the effects on emotional state and responsiveness also have implications for the experimental evaluation of educational interventions. Social programs for education and ALMP usually aim to impact socio-emotional or life-skills, such as perseverance, self-control, goal-oriented effort, and so forth. However, these factors suffer from measurement bias—i.e. reference or emotional bias—due to self-reporting. The methodology proposed here allows us to measure emotional state and responsiveness from EEG recordings, which is a non-invasive and low-cost method. Therefore, further research could incorporate physiological measures of emotional regulation to study the human capital production function, educational interventions, and the ALMP effectiveness.

The paper is organized as follows: section 2 defines the background and conceptual framework for education production function and skill measurement. Section 3 describes the theories of emotions, their implications on behavior, and their methods to be measured. Section 4 explains the methodology, experiment and data, and section 5 presents the results and a discussion about the main findings. Finally, conclusions are presented in section 6.

### 2.2 Background and Conceptual Framework

This section has four objectives: to explain and contextualize the intervention studied; to review the literature on late-stage interventions aiming to foster non-cognitive skills; to sketch a conceptual framework relative to the education production function; and to delineate and discuss the definitions and measurements of non-cognitive skills.
Mining’s Rockstars Program

Technical and Professional Education (TPE), which is similar to vocational-technical schools in the US, started in Chile in the late 60s as an educational alternative oriented to working life, but with the possibility of continuing to higher education. It began as a four year cycle after completion of elementary school education. In the early 80s, TPE was divided into two blocks: the first two years dedicated to normal high school programming, known as Scientific and Humanistic Education (SHE), and the next two years focused on TPE.

After the 1981 education liberalization reform, schools were given the ability to choose whether they wanted to offer SHE or TPE, and to manage the content of each track (Cox, 2006). The goal of allowing high school managers —i.e. the principal and the municipality education department (DAEM) officer— to choose among tracks was to improve the quality and pertinence of the education offered, though evidence shows that this was not necessarily the result (Albarran and Gonzalez, 2015).

Nowadays, 42% of students in TPE programs continue on to higher education, compared to 66% of the students in SHE programs (Bassi and Urzua, 2010). Only 39% of those TPE students complete their degree. This suggests that students of TPE are disadvantaged in terms of access to and completion of higher education, which is reflected in the achievements of these students in several dimensions. In fact, after controlling for demographics and institutional characteristics, TPE schools underperform SHE schools in grades, access to higher education, completion rates, among others (Valenzuela et al., 2013). In addition, employers, in particular those in the mining sector, argue that TPE students lack employability skills, such as self-initiative, ability to finish a task well, ability to work as a team, tendency to honor

---

14DAEM is the Spanish acronym of Municipal Education Department (Departamento de Administracion de Educacion Municipal). The DAEM is the institution within municipalities in charge of provide public education in Chile after the 1981 reform.
commitments, and so forth (Albarran and Gonzalez, 2015). The National Mining Council also established a Competencies Framework for the mining sector, with its first priority is the development of life skills among applicants for jobs in the industry (Albarran and Gonzalez, 2015).

The Mining’s Rockstars program was created to tackle these challenges. The program aims to foster ten different skills through a didactic weekly intervention following a recognized Harvard leadership model by McClelland (1973) with the principles of “learning by failing, gaming, doing and rethinking.” These skills include: networking and mutual support, information searching, calculated risk taking, accomplishment of work-related commitments, systematic planning and monitoring, persuasion, demand for high quality and efficient work, and self-confidence. The intervention has a particular focus on self-confidence and creativity.

Accordingly, the program created didactic materials to be used by students and teachers, including a student textbook, which is hardcover, full color, and has more than 200 pages, and a teacher textbook, which has class-by-class guidelines and videos for each activity. The program’s coaches work together with teachers in order to generate local capabilities. The program also includes three training seminars and pilot/simulated training workshops before the program starts, as well as an online web support platform for teachers to use during implementation.

As a part of the intervention, the classroom is transformed into a friendly game room. All school desks are put aside, and chairs are arranged in a half-moon form, so that each student sits the same distance from the center. Then, through games and reflections/deliberations, working on teams under pressure, addressing complex

---

15 McClelland (1973) tested more than 500 businesses, sports, religious and political leaders of the world with the intention to see what skills they addressed made them different from normal people. He found 30 non-cognitive skills that were relevant, where 10 are developable in the short term. These 10 skills are: 1. Search for the opportunity and take initiative; 2. Persistence (grit); 3. Work in support networks; 5. Take calculated risk; 6. Comply work commitments; 7. Systematic planning and monitoring; 8. Persuasion; 9. Demanding efficiency and quality 10. Self-Confidence.
tasks and setting their own goals to achieve through these tasks, students are taught to recognize their own strengths and weaknesses.

The intervention evaluated here was a cluster-level —i.e. school level— randomization that targeted young students who were in technical/vocational high school for one academic semester, lasting around four months. Since there is high regional heterogeneity among schools in Chile, only TPE schools belonging to the 4th Region of Coquimbo were considered. The program’s managers agreed with municipal education department (DAEM) authorities on a potential subset of schools where it was feasible and desirable to implement the program. Restrictions included having more than 50 students per cohort and the guaranteed participation of the school’s principal in the program. Finally, “Mining’s Rockstars” managers randomly selected a group of four schools to be treated, as there were not sufficient resources to apply the program to all schools in the region. The annual estimated budget of the program was around USD$100,000, and the Ministry of Education agreed to fund only a pilot.

Within the treatment schools, the program was open to those students in their 4th and last year of high school. Therefore, I collected data on 4th grade students in both treatment and control group schools. As an additional exercise, I considered another control group consisted of students who attended same schools and were in their 3rd year. Since the intervention targeted only students in 4th grade, it was natural to create a within school control group as a way to perform additional robustness check on the results. Finally, the sample with valid EEG measurements is comprised of 296 students in total, of which 140 belong to the treatment group. The empirical

---

16 Valenzuela et al. (2013).
17 The fixed cost —methodology design, experts, professionals, book design and videos— was 64%, while the variable cost —books, workshops, monitoring, etc— account for the rest. The estimated cost of the book was USD$ 24.
18 I collected almost 500 EEG recordings in students. Filtering for artifacts and data quality gives us the final sample. See Attrition subsection for details.
strategy is explained in a following section.

**Brief Review of Literature on Late Stage Interventions**

This subsection reviews recent studies on late-stage interventions that are relevant to the present study.\(^\text{19}\) Recent evidence shows that late-stage non-cognitive interventions on youth can have significant impacts in terms of behavior and/or outcomes (Blattman et al., 2015; Calero et al., 2014; Carrel and Sacerdote, 2013). However, evidence concerning the returns from late-stage non-cognitive investments on academic and/or labor market outcomes is scarce.\(^\text{20}\) There are only a few rigorous evaluations of interventions of this type in the context of developing countries.\(^\text{21}\)

A couple of experimental studies conducted during the last few years in Latin America are closely related to the present study. Specifically, Card et al. (2011) conducted one of the first rigorous studies of the effectiveness of a labor training program that fosters vocational and socio-emotional skills, in the context of the Dominican Republic’s Youth and Employment —Juventud y Empleo— program. Attanasio et al. (2015a) studied a similar program in Colombia. Both studies show similar results:

---

\(^{19}\) In addition, since I am doing a program evaluation in the medium scale range, this study will not address equilibrium effects, which are more likely to occur when the program or policy are taken to a scale and sustained for a longer period of time (Banerjee and Duflo, 2008; Acemoglu, 2010; Deaton, 2010; Pop-Eleches and Urquiola, 2013).

\(^{20}\) There is a literature focused on Active Labor Market Policies (ALMP) focused in industrialized countries, with few studies related to developing countries, in particular, in the context of Latin America. Examples of ALMPs are: demand incentives, such us targeted wage subsidies or tax incentives, public employment, flexible employment arrangement such us remote work or internship, and training and vocational programs. For examples on these ALMP see the extensive review of Puentes and Urzua (2010) for Latin American countries, or Card et al. (2010), Kluve (2010), and Card et al. (2015) for industrialized countries context. In particular, in the industrialized countries context, the main conclusion is that ALMPs have low effectiveness in term of both earnings and employment in the developed world, with classroom and on-the-job training being the most effective settings (Card et al., 2010). Nonetheless, compared to industrialized countries, developing countries, in particular those in Latin American, unveil greater effectiveness, especially among women, although the absolute magnitude is yet small (Puentes and Urzua, 2010).

\(^{21}\) As mentioned in the introduction, there is recent evidence in a developed country context. In particular, Carrel and Sacerdote, 2013 study a college coaching intervention in New Hampshire. The authors find a positive and significant impact of coaching on enrollment and continuing in college.
no impact on employment and only a modest impact on earnings and quality of employment, conditional on working.

**Calero et al. (2014)** study the effectiveness of an art-based program to improve employability among at-risk youth in Brazil’s slums. The program jointly employs expressive art and theater training with mainstream vocational and academic instruction, aiming to improve labor market outcomes such as earnings and obtainment of paid formal employment, among others. The authors do not find robust and statistically significant impacts on non-cognitive measures, although their own qualitative data shows that there are impacts on valuable labor market skills such as punctuality, responsibility, and dedication to work. They point out that this lack of quantitative impact could be due to selection problems as well as to the self-reported nature of the psychometric tests employed.\(^{22}\)

In a follow up study of the same Youth and Employment program, **Ibarraran et al. (2014)** found similar results. Interestingly, they incorporated measurements of life skills, including personality, grit and self-esteem. Using customized instruments to measure those skills, the authors find a positive but moderate impact on some dimensions. They also remark that improvements in the instruments to measure life skills are needed. Other authors have similar conclusions (**Almlund et al., 2011; Calero et al., 2014; Heckman and Kautz, 2012**). For instance, recent meta-analysis conducted by **Heckman and Kautz (2012)** concludes that studies using self-reported psychometric tests to proxy socio-emotional skills face potentially significant measurement error with respect to the tests used. Moreover, **Cunha and Heckman (2007)** point out that measurement error is high in all of the proxies that they are using to measure cognitive and non-cognitive skills in a latent factor model.\(^{23}\) Indeed, the same authors argue

\(^{22}\) The Social and Personal Competencies Scale, an adaptation of the Big Five Inventory done by **Brea (2010)**; and the Grit Scale, developed by **Duckworth et al. (2007)**.

\(^{23}\) See Chapter 3 for details on this type of model.
that slight changes in those measures substantially affect the direction and significance of their findings.

The most recent study is conducted by Blattman et al. (2015), who evaluates a late-stage non-cognitive intervention in at-risk youth in Liberia. In his assessment of a youth population recruited in the streets, the authors find significant results on reducing violence, crime and poverty indicators from a cognitive behavioral therapy intervention. The main conclusion is that orthodox policies to increase labor participation of at-risk youth by increasing police control and low-skill emergency job creation had failed, and that there is a room for late investment in non-cognitive skills in order to change youth behavior.

**Education Production Function**

I used a production function to describe the process through which the skills of students evolve between the beginning and the end of the program. The baseline period is indexed as 1, at the beginning of the academic year, in March of 2014, while the post-intervention period is identified as 2, when students start the second academic semester, in August of 2014. Students’ outcomes or skills at time 2 are assumed to be a function of previous investments and some other shocks.

Moreover, relevant inputs in the production and investment functions will depend on the intervention. In this sense, the model allows for changes in both the parameters —i.e. elasticities or relative prices— and parent/student behavior due to participation in the program.24

The main contrastable groups are the following: the treated group, that comprises those students who participated in the program for one semester —i.e. attend a high-school that was selected to participate, and were enrolled in 4th grade; and the

---

24 For details see Todd and Wolpin (2003).
control, or non-treated, group, that includes students who did not participate in the program —i.e. were enrolled in neighboring schools or were pursuing 3rd grade in the selected into program’s schools.

This type of model allows the exploration of general equilibrium effects, or behavioral responses from students, principals and families. I did not collect information about parents’ investment or school adjustments, therefore, the estimation of the model will contemplates only individual investments and behaviors. Moreover, the notion of individual behavioral response is slightly changed for a emotional response because of the program. The conceptual framework will follow Todd and Wolpin (2003) and Pop-Eleches and Urquiola (2013), where achievement depends on the school and previous family investment. This conceptual framework will be later adapted to the context of this study. In fact, I consider that the student’s achievement —e.g. on non-cognitive tests— at the beginning of time 1 is:

\[ A_1 = g_0(F_0, \mu), \]  

(2.1)

where \( F \) and \( \mu \) index the her family’s investments and her innate ability, respectively.

Then, in the second period we have:

\[ A_2 = g_1(S_1, F_1, F_0, \mu) \]  

(2.2)

where \( S \) accounts for school’s inputs in the previous period.

Now, define the parent’s expected school’s inputs as:

\[ \bar{S}_1 = \theta(A_1, W, \mu) \]  

(2.3)

where \( W \) denotes family wealth. The actual school’s input at period 1 is defined as:

\[ S_1 = \psi(A_1, \mu) \]  

(2.4)

Finally, family’s input is defined by:

\[ F_1 = \phi(A_1, W, \mu, S_1 - \bar{S}_1) \]  

(2.5)
where the family observed \((S_1 - \overline{S}_1)\) before set the household investment \(F_1\).

The Production Function Effect (i.e. \textit{ceteris paribus}) of a change in \(S_1\) will be:

\[
\frac{\partial A_2}{\partial (S_1 - \overline{S}_1)} = \frac{\partial A_2}{\partial S_1} = \frac{\partial g_1}{\partial S_1} 
\]  \hspace{1cm} (2.6)

While the aggregate Policy Effect (with behavioral adjustment):

\[
\frac{\partial A_2}{\partial (S_1 - \overline{S}_1)} = \frac{\partial A_2}{\partial S_1} = \frac{\partial g_1}{\partial S_1} + \frac{\partial F_1}{\partial (S_1 - \overline{S}_1)} 
\]  \hspace{1cm} (2.7)

Finally, Pop-Eleches and Urquiola (2013) identify a specific change in school’s inputs (i.e. change in X while Y remains constant)

\[
\frac{\partial A_2}{\partial (S_1^x - \overline{S}_1^x)} = \frac{\partial A_2}{\partial S_1^x} = \frac{\partial g_1}{\partial S_1^x} + \frac{\partial g_1}{\partial S_1^y} \frac{\partial S_1^y}{\partial S_1^x} + \frac{\partial F_1}{\partial (S_1^x - \overline{S}_1^x)} + \frac{\partial F_1}{\partial (S_1^y - \overline{S}_1^y)} 
\]  \hspace{1cm} (2.8)

Equation 2.8 indicates the expected program’s impact on student’s achievement relative to non-cognitive skills.

This conceptual model illustrates some parameters that this study aims to identify. In particular, it is expected that the program will generate effects on the outcomes as well as emotional responses. Educational outcomes, which are obtained from administrative records, are SAT-like test registration and school dropout rates. Socio-emotional skills and creative abilities, which are proxied using psychometric tests, are considered outcomes as well. The student’s emotional response, which are measured through neurophysiological recordings, are considered, in the context of this model, as a "behavioral response". As mentioned, the interpretation of the behavioral response differs from used by Todd and Wolpin (2003) and Pop-Eleches and Urquiola (2013), because here I only consider the individual behavioral response, which is understood
as an emotional response, instead of a behavioral response from parents’ or school’s principals.

From a policy viewpoint, it is important to understand how life skills can be fostered by young people, especially for those at-risk. In order to fully understand the impact of a public policy it is necessary to study potential behavioral responses. In the case of students who participate in a program to promote socio-emotional skills, it is reasonable to expect behavioral responses in all different life spheres, such as interactions with peers, self-perception, emotional regulation and behaviors at school and home. Likewise, behavioral responses appear to be context dependent, and thus vary according to each public policy and to changes over time (Pop-Eleches and Urquiola, 2013).

Definition and Measurement of Skills

Skill Definitions

Almlund et al. (2011) and Heckman and Kautz (2012) have recently surveyed different approaches and definitions of cognitive and non-cognitive skills. Gintis (1971) defines cognitive skills as the process of logically combining, analyzing, interpreting and applying informational symbols. Common examples of these types of skills are the use of language, memory and logical-mathematical knowledge. Non-cognitive skills, also known as socio-emotional skills, life skills or personality traits are harder to define, as they are a combination of behaviors, feelings, values and thoughts. Examples of these skills are self-confidence, self-control, perseverance, grit, and locus control.

For example, Pop-Eleches and Urquiola (2013) argue that parents might react to their children going to a better school by lowering their own efforts; or that at-risk students who make it into better schools might feel inferior or stigmatized beside their more privileged peers. Regardless of the relevance of these behavioral responses, developments in the literature have only happened recently (e.g. Cullen et al., 2006; Pop-Eleches and Urquiola, 2013)

See for example, Borghans et al. (2008); Cunha et al. (2010); Almlund et al. (2011).
large body of evidence reviewed in Almlund et al. (2011) suggests that stable character skills exist and are predictive of many behaviors. Moreover, Borghans et al. (2008) highlight the existence of quasi-cognitive skills which are a mix between both cognitive and non-cognitive skills, such as creativity or innovation, emotional intelligence, and practical intelligence, among others. The definition and measurement of cognitive and non-cognitive skills are challenging. Any type of skill is difficult to measure, even when identifiable or separable, but it is even more difficult to define and separate non-cognitive skills from cognitive ones (Heckman and Kautz, 2012).

Consistent with the labor economics literature, the present study uses self-reported psychometric tests to measure skills. The battery of tests includes psychometric tests that intended to measure non-cognitive, creative and cognitive skills. Regarding socio-emotional skills, the Grit Scale (Duckworth et al., 2007), Locus Control Scale (Rotter, 1966) and Big Five Inventory (BFI, John and Srivastava, 1999a) were included in the battery of tests. The Big Five Inventory is a widely accepted taxonomy of personality traits that describes personality as consisting of the following five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (John and Srivastava, 1999a). The Grit Scale delves deeper into the measures of the Big Five’s Conscientiousness, defined as “perseverance and passion for long-term goals.” The Locus Control Scale measures the degree to which individuals attribute their fate to external or internal factors. There is an open debate about the malleability of personality in the medium term\(^\text{27}\), and thus the analysis on the present study will be focused on Grit and Locus Control scales instead of the Big Five Inventory (BFI).

In order to test creativity, the written version of the 1976 Torrance’s Test of Creative Thinking [TTCT] was also used (Kim et al., 2013; Kaufman and Sternberg, 2010). It is worth mentioning that in the psychology literature on creativity, a distinction is regularly made between convergent thinking, which is referred to here as

\(^{27}\) See for example Almlund et al. (2011).
cognitive ability, and divergent thinking, which is denoted here as creative thinking.\textsuperscript{28} The written form of the TTCT, is an assessment tool of creativity that was adapted to Chile by Egana-delSol (2016a).\textsuperscript{29} Following Torrance (1966), I shaped an instrument to measure four dimensions of creativity: fluency, flexibility, originality and the “13 creative forces.”\textsuperscript{30}

With respect to cognitive skills, Raven’s-like progressive matrices (Raven, 1936), which measures fluid intelligence, was considered. The psychology literature usually makes a distinction between crystallized and fluid intelligence (Almlund et al., 2011). The former accounts for the ability to use past appropriation of the collective intelligence of the culture, giving one a distinct advantage in solving the problems involved, while the latter is “the ability to perceive complex relations, educe complex correlates, form concepts, develop aids, reason, abstract, and maintain span of immediate apprehension in solving novel problems in which advanced elements of the collective intelligence of the culture were not required for solution” (Almlund et al., 2011). Since the program should not directly affect fluid intelligence, this test is taken as a placebo.\textsuperscript{31}

Measurements of Non-cognitive Skills

There are different approximations to measure non-cognitive skills. A recent survey features the use of behaviors —drugs use, registered behavior at school, etc.— and peer’s and self-reported measures (Heckman and Kautz, 2012).

\textsuperscript{28} For a discussion see Runco (2010).

\textsuperscript{29} See Chapter 3 for details and examples

\textsuperscript{30} The last dimension comprises of an aggregate and richer assessment of creativity. Unusual visualization, internal visualization, extending or breaking boundaries, humor, richness of imagery, and colorfulness of imagery are some of the dimensions considered. See Chapter 3 on this dissertation for details.

\textsuperscript{31} See the Appendix for details.
For instance, Heckman et al. (2006) use risky behavior—i.e. marijuana consumption during adolescence—to predict later outcomes in the labor market. Hirschi and Gottfredson (1993) argue that objective behavioral measures might be preferred to self-reports, as filling out a survey requires some level of self-control. In addition, answering survey questions is another task that relies on skills beyond those targeted by the survey. Nevertheless, the measurement of risky and irresponsible behaviors in adolescents, such as marijuana consumption or theft, may be related to other context dependent factors rather than socio-emotional skills.

Using similar measures from administrative data, Jackson (2014) investigates the effect of teachers on students’ cognitive and non-cognitive skills. In particular, the author proxies cognitive skills using achievement test scores and non-cognitive skills using absences, suspensions, grades, and grade progression. These measures of non-cognitive skills predict adult outcomes with a strength similar to measures of cognitive ability. His measures of character are commonly available from schools’ administrative records. Some scholars criticize this approach, arguing that it is tautological to use measures of behavior to predict other behavior, even though the measures are taken early in life to predict later life behaviors (Heckman and Kautz, 2012). Moreover, it is not always feasible to have access to administrative—or third party—data in a field experiment context, which frequently is nominated and, thus, has restricted access.

Another option is to use others’ ratings. However, behaviors reported by peers, teachers or parents could be biased, especially if they know these reported behaviors are used to make decisions that could affect themselves. For instance, a school might like to assess the degree of grit among their students by asking teachers to rate them. Teachers can infer that the results of that assessment could influence their

32 For example, a meta-analysis by Pratt and Cullen (2000) finds that behavioral measures are at least as good at predicting crime as are measures based on self-reported taxonomies.
own evaluations, incentivizing teachers to over-rate particular skills in their students.

Moreover, self-reported tests’ performance can be influenced by the test setting (ONeil et al., 2014), reference bias (West et al., 2014; Kautz et al., 2014), and the examinee’s emotional and motivational state (Plucker and Makel, 2010; Querengsser and Schindler, 2014; Egana-delSol, 2016b). In the context of well designed assessment, it is plausible to argue that the test setting is, on average, similar to everyone. Reference bias is one of the most common caveat for self-reported measures of non-cognitive skills. For example, in the Grit Scale statement “I am a hard worker,” the assessment ranges from ”Very much like me” (1) to ”Not like me at all” (5). Thus, a subject’s reference group will determine their self-perception—a student from a high achieving school might think herself lazy whereas a similar student in a low achieving school may consider herself rigorous. Finally, as discussed in Chapter 3, emotional/motivational state also affects self-reported measures of non-cognitive skills.

In summary, self-reported tests are problematic instrument to measure non-cognitive skills. In response to these caveats, I suggest an alternative method to assess non-cognitive skills, in particular, emotional regulation. In the following section I will explain the relevance of emotional regulation as a non-cognitive skills as well as methods to proxy it eluding self-reporting or past behaviors.

33 See Duckworth et al. (2007) for details.

34 For a recent discussion about how reference bias can affect self-reported on non-cognitive skills measures see West et al. (2015).

35 As mentioned, in addition, I explore strategies to correct self-reported psychometric tests from emotional bias using methods from affective neuroscience in Chapter 3.
2.3 The Role of Emotions

The relation between emotion and cognition has been debated for centuries (Descartes, 1649[1989]). Prominent neuroscientists argue about the critical role that emotions play in cognition, perception, attention and memory (Damasio, 1994; Lakoff, 2008; Salzman and Fusi, 2010; Fuster, 2013). Furthermore, scientists who combine neuroscience, behavioral science, and economics also support the notion that emotions have a strong influence on economic behavior and decision-making, as well as labor market performance (e.g. occupational choice, salaries, entrepreneurship, etc.). For example, emotions experienced while making a decision —i.e. choice-option–elicited emotions— immediate emotions are at the base of traditional economic interpretations of utility as emotional carriers of value. Positive emotions increase value and elicit approach, whereas negative emotions decrease value and result in avoidance (Weber and Johnson, 2009). Moreover, emotions unrelated to the judgment or decision at hand, referred as incidental emotions, have also been shown to influence choice (Weber and Johnson, 2009).

Emotions have interesting behavioral implications. For instance, Durlak et al. (2011) conducted a meta-analysis on programs to develop emotional intelligence in educational settings. They found positive effects on targeted socio-emotional skills and attitudes about self, others, and school. They also argue that those types of programs foster students’ behavioral adjustment in the form of increased prosocial behaviors, reduced conduct and internalizing problems, and improved academic performance on achievement tests and grades. Similarly, a preschool and early primary school program to foster emotional intelligence improves classroom behavior as well as executive function, defined as higher level cognitive skills including inhibitory con-

36 See for example Weber and Johnson (2009); Lerner et al. (2012); Lempert and Phelps (2014); Loewenstein (2000).
Figure 2.1: Arousal-Valence Model of Emotions

In fact, educational or labor market programs that target socio-emotional skills are likely affecting emotional regulation. Complementarily, Haushofer and Fehr (2014) argue that poverty may have particular psychological consequences that can lead to economic behaviors making it difficult to escape poverty. The evidence indicates that poverty causes stress and negative emotional states, which in turn may lead to short-sighted and risk-averse decision-making, possibly by limiting attention and favoring habitual behaviors at the expense of goal-directed ones. Together, these relationships may constitute a feedback loop that contributes to the perpetuation of poverty, which they called a “psychological poverty trap.”
A variety of emotion models are used in affective neuroscience (Petrantonakis and Hadjileontiadis, 2010). The most frequent model used in psychology and affective neuroscience is continuous in nature and thus expresses emotions in a n-dimensional space, usually the two dimension arousal and valence model. In general, the literature uses James-Lange’s arousal-valence —also known as Circumplex— model of emotions, which will be followed here. As mentioned, valence accounts for judging whether a situation is positive or negative, while arousal expresses the degree of one’s excitation, spanning from calmness to excitement. In particular, the Ramírez and Vamvakousis (2012)’s neurophysiological version of the model will be considered as the framework for emotions. As an illustrative example, emotions could be classified into four categories in the valence-arousal locus, namely joy, anger, relaxation, and sadness. As Figure 2.1 reveals, joy is characterized by high arousal and high valence, anger by high arousal and low valence, relaxation by low arousal and high valence, and sadness by low arousal and low valence. Another approach, which is used relatively less frequently by affective scientists but is relevant to neuroeconomics, classifies different emotions, in particular valence, according to motivation (Davidson et al., 1990; Harmon-Jones et al., 2010; Lempert and Phelps, 2014). In this model, different emotional states lead to different goals for action. Positive (valence) emotional states in term of valence, such as happiness, evoke a motive or goal to approach a situation. On the contrary, negative (valence) emotional states such as sadness or disgust, are withdrawal emotions, evoke a motive or goal to withdraw from situations/stimuli linked to these emotions.

Regardless of whether approach/withdrawal motivations correlate with valence, researchers have argued that they represent a distinct dimension (Kassam et al., 2013). For instance, angered individuals are typically motivated to approach the source of their anger —i.e. to fight— despite the stimulus’ negative valence. Nevertheless, both arousal/valence and approach/withdrawal models have been supported
by research. They may capture distinct features of human function and may be better thought of as complementary or as simply different (e.g., Coombes et al., 2007; Wacker et al., 2003). The psychological phenomena that these models aim to explain allow for that ambiguity. In particular, the arousal/valence model focuses on emotional stimuli and information processing, rather than behavioral insights, whereas the approach/withdrawal model emphasizes neural activity associated with goal-related emotion, and thus actions, such as approaching or withdrawing from a situation (Spielberg et al., 2008). In summary, I will consider indistinctly positive valence/approach as well as negative valence/withdrawal emotions/motivations, and I will discuss the results considering the scope of both models.

The complex task of measuring emotions can be done using self-reported measures. These methods generally follow the Positive and Negative Affect Schedule (PANAS, Watson et al., 1988). Since the PANAS is self-reported, it experiences similar issues as those mentioned above for self-reported psychometric tests. In particular, performance can be influenced by the examinee’s emotional and motivational state (Plucker and Makel, 2010; Querengsser and Schindler, 2014; Egana-delSol, 2016b), the test setting (ONeil et al., 2014), and reference bias (West et al., 2014; Kautz et al., 2014). In fact, emotions may play a role in the measurement error of self-reported psychometric tests (Querengsser and Schindler, 2014; Egana-delSol, 2016b). Therefore, an alternative can be to use neurophysiological methods from affective neuroscience.\footnote{As mentioned, I explore strategies to correct self-reported psychometric tests using neurophysiological methods in Chapter 3 (Egana-delSol, 2016b). I argue that valence has a positive correlation: Positive/approach emotions (pleased, happy) imply self-reports are overestimated, while negative/withdrawal emotions (unhappy, sad) would underestimated the self-perception of certain socio-emotional skill. However, for arousal the effect is ambiguous (Weber and Johnson, 2009).}

Emotional regulation —pre-test emotional state and emotional responsiveness to positive and negative stimuli in this study— is the relevant feature respect to the impact evaluation of the mentioned educational program. The neurophysiological methods from affective neuroscience are discussed in next section.
Emotion-Detection Theory

Over the past decade, emotion-detection research has employed an array of different physiological measurements and methods, including pupil dilation, heart rate, and skin conductance for arousal, and voice and facial manifestations for valence (Takahashi et al., 2004; Partala et al., 2000; Brown et al., 2011; Verma and Tiwary, 2014; Bos, 2006; Petrantonakis and Hadjileontiadis, 2010; Ramirez and Vamvakousis, 2012; Yoon and Chung, 2013). But many of these measurements can be consciously modified, and thus the signals they produce are not purely objective (Partala et al., 2000; Ramirez and Vamvakousis, 2012), in particular those related to valence which is the main feature of interest in this study. Moreover, arousal effects on behavior, self-reporting and decision-making is ambiguous (Weber and Johnson, 2009; Egana-delSol, 2016b). Electroencephalogram (EEG) recordings allow us to measure brain activity and predict emotional state and physiological responsiveness, improving both accuracy and objectiveness with respect to the aforementioned physiological measurement.

The firing of neurons in the brain triggers voltage changes. The electrodes in an EEG headset capture the electrical activity corresponding to field potentials resulting from the combined activity of many individual neuronal cells in the brain cortex. However, cortical activity measures are distorted by the tissue and skull between the

---

38 Ekman et al. (1987) suggested the universality of six facial expressions based on Darwinian theory: happiness, surprise, anger, disgust, sadness, and fear (Darwin, 1872[2002]). However, Partala et al. (2000) argue that facial expressions can be easily simulated by subjects.

39 See for example Partala et al. (2000); Brown et al. (2011); Verma and Tiwary (2014); Bos (2006); Petrantonakis and Hadjileontiadis (2010); Ramirez and Vamvakousis (2012); Yoon and Chung (2013).

40 Electroencephalogram devices measure the voltage change that happens when a neuron fires. When a positive change in the voltage crosses a certain threshold, an action potential is triggered. Indeed, the voltage goes from the resting potential of about -60mV to +20mV. This electrical activity from group of neurons is measured in the cortex by the EEG’s electrodes. That is to say, the EEG measures the brain’s activity through voltage change in groups of neurons that had been fired due an action potential.
electrodes and the neurons. This introduces noise and reduces the intensity of the recorded signals. Regardless, EEG measurements offer important insights into the electrical activity of the cortex (Verma and Tiwary, 2014; Brown et al., 2011).

A number of authors have considered diverse methods to recognize emotions from EEG recordings, improving both accuracy and objectiveness relative to the aforementioned physiological measurement (Choppin, 2000; Takahashi et al., 2004; Lin et al., 2010; Ramirez and Vamvakousis, 2012; Kim et al., 2013). Accuracy accounts for the true and consistent results, that is to say, the likelihood that the model can correctly predict the elicited emotion, which is known by the characteristic of the stimuli or a self-reported emotional state after a given experiment. Numerous studies have attempted to compare those diverse methods of emotion recognition arguing that the most suitable methodology and strategy —i.e. the one with the highest predictive accuracy— is the use of EEG recordings (Choppin, 2000; Takahashi et al., 2004; Lin et al., 2010; Ramirez and Vamvakousis, 2012; Verma and Tiwary, 2014; Kim et al., 2013). For instance, Brown et al. (2011) estimate an 82% accuracy for arousal and valence. Yoon and Chung (2013) found a 70% accuracy for arousal and valence. Verma and Tiwary (2014) found 85% accuracy for arousal, valence and dominance.42

Furthermore, the "proof-of-concept" experiment that I conducted on graduate students at Columbia University exhibited 79% accuracy for valence (See Chapter 3 for details). The recent development of low cost portable EEG devices offers an unprecedented opportunity to incorporate methodologies from affective neuroscience in social programs evaluation and experiments in the field.43 In a recent study,

41 Verma and Tiwary (2014) propose a fusion model that encompass other physiological measures in addition to arousal and valence indices. Regardless, the authors estimations indicate that the Fusion and Circumplex —arousal-valence— models have similar accuracy.

42 The dominance scale ranges from submissive (or without control) to dominant (or in control, empowered).

43 In a lab setting, Ramirez and Vamvakousis (2012) use the same low cost portable EEG device
Martinez-Leon et al. (2016) compare the quality of data captured by a professional Biosemi Active II\textsuperscript{44} and a low cost Emotiv EPOC\textsuperscript{45} headset. The latter is identical to the device used in this dissertation. Their results are based on the comparison of the success rate of a Brain-Computer Interface (BCI) system. Higher precision and less variance are found on low cost Emotiv EPOC headset datasets. Moreover, the authors conclude that the Emotiv EPOC low-cost headset can be used on motor imagery BCI systems.

In fact, a suitable strategy to proxy emotional regulation —i.e. emotional state and responsiveness— for dimensions of arousal and positive (approach) and negative (withdrawal) valence (motivation) in the field can be the use of low cost EEG recordings (Friedman et al., 2015). It is important to note that these indices are not intended to describe and/or predict personality traits or character themselves.\textsuperscript{46}

Matlab 2014b software (Waltham, MA, USA), and EEGLAB open source toolbox (Delorme and Makeig, 2004), were used to do the neurophysiological offline data analysis. The EEG signal was passed through a low-pass filter with a 40-Hz cutoff frequency in order to remove noise coming from the power line and artifacts. The frequency of EEG measurements ranges from 1 to 80Hz, with amplitudes of 10 to 100 microvolts (Ramirez and Vamvakousis, 2012). Multi-taper Fourier transform for continuous data sets is used. By using a set of tapers, rather than a unique data taper or spectral window, the algorithm reduces the variance of spectral estimate. Moreover, it is particularly effective for short data segments (Castellanos and Makarov, 2006).\textsuperscript{47}

\textsuperscript{44} Biosemi. For details visit http://www.biosemi.com/.

\textsuperscript{45} Emotiv EPOC EEG. For details visit http://www.emotiv.com/.

\textsuperscript{46} In a recent study, Korjus et al. (2015) show that there is no correlation between resting state EEG waves and any of the five personality dimensions of the self-reported Big Five Inventory. They conclude that their results indicate that the extraction of personality traits from the power spectra of resting state EEG is extremely noisy, if it is even possible. That study does not tackle the mentioned issues around self-reported psychometric tests.

\textsuperscript{47} A MatLab toolbox developed by Mitra and Bokil (2008) to implement those methods is freely

used in the present study.
Many authors suggest that a maximum of approximately a 10 second signal would be required if the dependent measure was an EEG, since a longer period can include factors that are distinct to the elicited emotion (Davidson et al., 1990). Finally, in order to smooth the exponential nature of EEG signals we apply a logarithmic power transformation to the data, which is the standard in this literature (Verma and Tiwary, 2014; Ramirez and Vamvakousis, 2012).48

It was decided that the most suitable method to measure emotions from EEG signals is the one implemented by Ramirez and Vamvakousis (2012). Evidence has shown that the prefrontal cortex — in addition to the amygdala and the insula — coordinates consciousness and regulates emotions.49 Measuring emotions therefore requires consideration of the EEG signals using electrodes AF3, AF4, F3, and F4 in the 10-20 standard classification, which are located on the prefrontal lobe.50 The frequencies of interest in measuring emotions are both alpha (8-12Hz) and beta (12-30Hz) waves.51 Alpha waves are predominant in relaxed states and brain inactivation, while beta frequencies are associated with alertness and an excited state of mind.52 Consequently, the beta-alpha ratio is an effective indicator of the level of arousal. Therefore, arousal index can be estimated by the following equation:


48 The exact transformation follows Aspiras and Asari (2011). In particular, the EEG signals were transformed as follow: LogSignal$_j = 10 \times \log_{10}(Signal_j)$; where $j$ = positive, negative, resting state conditions.

49 Bear and Connors (2007). For a recent review see Salzman and Fusi (2010)

50 For the exact location of each electrode in the brain, see Figure 10 20 SYSTEM in the Appendix.

51 There is ample data linking changes in these frequency bands to various mental processes such as changes in mental state (Moretti et al., 2004), changes in attention allocated to a task (Klimesch, 1999), memory processes (Klimesch, 1996), motivation and emotional processes (Knyazev, 2007), and different sleep stages (Keenan, 1999), among others. Here I focus on arousal and positive valence/approach behavior and negative valence/withdrawal behavior (Davidson et al., 1990; Harmon-Jones et al., 2010; Verma and Tiwary, 2014).

52 See for example, Bos (2006); Ramirez and Vamvakousis (2012); Verma and Tiwary (2014); Yoon and Chung (2013); Brown et al. (2011); Friedman et al. (2015); Choppin (2000).
where $F$ indicates the simple average of electrodes located on the frontal brain areas AF3, F3, AF4 and F4 in the 10-20 universal system,\(^{53}\) while $i$ index an individual in the sample. On the other hand, previous neurophysiological evidence had shown that emotional valence has different representation in the right and left brain’s cortical hemisphere.\(^{54}\) Activity decrease over the frontal right region correlates with positive emotion, while activity decrease over the left frontal region is related to negative emotion. Since activity decrease —i.e. inactivation— could be measured as the inverse of arousal level, it is possible to estimate valence level by the following relation:

$$valence_i = \frac{\alpha_{F4,i}}{\beta_{F4,i}} - \frac{\alpha_{F3,i}}{\beta_{F3,i}}$$  \hspace{1cm} (2.10)

where $F3$ and $F4$ indicate electrodes located on the left and right frontal brain areas, respectively. Individuals are indexed by $i$.

As mentioned before, the literature interprets valence positive and negative indices as approach and withdrawal motivation to stimuli, respectively (Davidson et al., 1990; Harmon-Jones et al., 2010). Evidence from psychology and neurophysiology literature points out that frontal EEG asymmetry is associated with different emotional and psychological states rather than valence. In a seminal work, Davidson et al. (1983) \(^{54}\) For a discussion on the validity of estimate valence by comparing hemispherical activation, see, for example, Ramirez and Vamvakousis (2012); Kim et al. (2013); Verma and Tiwary (2014); Bos (2006); Harmon-Jones et al. (2010); Yoon and Chung (2013); Friedman et al. (2015); Davidson et al. (1990)
suggested a model called approach/withdrawal theory to investigate frontal EEG asymmetry during emotional states. He claimed that the left pre frontal cortex (PFC) activity is involved in a system facilitating approach behavior to appetitive stimuli, while the right PFC activity participates in a system facilitating withdrawal behavior from aversive stimuli. This model claims that processing related to emotional valence itself is not lateralized in PFC. Rather, emotion-related lateralization is observed because emotions contain approach and/or withdrawal components.

That is to say, frontal EEG asymmetry is perceived because emotions contain approach and/or withdrawal components. Thus, emotions will coexist with either right or left asymmetry, whether they are accompanied by approach or withdrawal behavior (Davidson et al., 1983, 1990; Harmon-Jones et al., 2010; Harmon-Jones and Gable, 2008). Approach/withdrawal motivational states have frequently been linked to asymmetries in left/right frontal cortical activation, especially using EEG, though meta-analyses of fMRI data have failed to find consistent localizations (Kassam et al., 2013).

Finally, there are a number of papers that consider the left, relative to right, frontal cortical activity (LFA) to build behavioral indices of approach or motivation. For instance, Hughes et al. (2014) examines the relation between LFA and effort expenditure for reward, a behavioral index of approach motivation. They found that subjects with greater resting LFA were more willing to expend greater effort in the pursuit of larger rewards, particularly when reward delivery was less likely.

### 2.4 Experimental Paradigm

As mentioned, there exists a common puzzle in the labor economics literature around social programs —i.e. educational or labor market programs— that exhibit positive
impacts on labor market outcomes, yet, contrary to expectations, do not seem to affect measures of non-cognitive skills (Card et al., 2010; Calero et al., 2014).

This paper studies the production function of non-cognitive skills —i.e. socio-emotional skills, creativity, etc.— in formal schools. In particular, I will explore how participants in the “Mining’s Rockstars” program change in terms of creativity, grit, locus control, and emotion regulation. I argue that the program affects participants’ emotional regulation —i.e. emotional state and responsiveness. In Chapter 3, I claim that the impact on emotional regulation modulates their self-reporting on tests about their own non-cognitive skills, and thus might be an explanation of the aforementioned puzzle.

As Figure 2.2 shows, the experiment was done in the context of the impact evaluation of the “Mining’s Rockstars” program. It targeted 1,380 students aged 17-18 years old at baseline in 8 public technical high schools in semi-urban towns in the north of Chile. Four schools were randomly selected to be in the treatment group, while the remaining four schools were assigned to the control group. In the treatment schools, all students in their 4th grade —i.e last year— of high-school were treated.

The typical self-reported method to evaluate emotions is the Positive and Negative Affect Schedule (PANAS). Besides its discussed caveats because of self-reporting, using PANAS makes salient the relevance of emotions in the study, potentially affecting both reaction to stimuli and self-reports on emotional state and other dimensions. Therefore, I do not use PANAS in this study.

As Figure 2.3 shows, I collect three streams of data, namely: pre-test resting emotional state from EEG recordings, a battery of psychometric tests (see previous section for details), and emotional responsiveness to both positive and negative stimuli. The pre-test emotional state was constructed using EEG scalp recordings while

---

55See section 1.2 for details.
Figure 2.2: Geographic Location of Schools in the Field Experiment
Figure 2.3: Field Experiment Timeline

- **Randomization of Mining’s Rockstars program**
  - (at school level)
  - [4 schools selected; 4 schools as control group]

- **Baseline Field Experiment**
  - (March 2015)
  - [Neurophysiological measures of emotional regulation; non-cognitive skills tests]

- **Intervention Begin**
  - (Mid. March 2015)

- **Follow-up Field Experiment**
  - (August 2015)
  - [Neurophysiological measures of emotional regulation; non-cognitive skills tests]

- February, 2015
- March, 2015
- Mid. March, 2015
- August, 2015
students watched a black cross in the center of a gray screen for a period of 30 seconds before taking the battery of psychometric tests (see Figure 2.4). Emotional arousal and valence indices at a resting state are estimated using those recordings. Hereafter, these measures will be called “baseline.” Then, I applied a battery of psychometric tests that includes grit scale, locus control scale, raven-like progressive matrixes, and torrance’s test of creative thinking. The emotional responsiveness measures were obtained right after the students finished the battery of tests. As Figure 2.5 exposes, the experiment consisted of showing an alternating series of positive and negative images in order to elicit emotional responses. Arousal and valence indices were also obtained from those emotional responses.\(^{56}\) Consequently, there are three estimates of both arousal and valence according to the nature—or absence—of stimuli, including:

\(^{56}\) Technical details are explained in Chapter 3 of this dissertation (Egana-delSol, 2016b).
pre-test resting state, positive and negative.

The final sample —after filtering the EEG data and accounting for attrition— is an unbalanced panel with a baseline in March and a follow up in August 2015, with a total of around 300 valid EEG recordings of students. This is an extraordinary number of measures, considering the time and complexity of collecting this data, especially in an out-of-the-lab setting.\footnote{Indeed, I collected almost 500 EEG recordings in students. Filtering for artifacts and data quality gives us the final sample. See Table 2.2 for details.} Figure 3.6 exhibit a real examples of how was the context of the out-of-the-lab setting in schools.
Emotional regulation—both emotional state and responsiveness—are innovative outcomes that try to capture the substantial qualitative evidence that many educa-
tional programs or ALMP designed to improve education and labor market outcomes through fostering socio-emotional skills do in fact affect them.

2.5 Results

This section has four subsections. First, descriptive statistics are shown. Second, the degree of attrition is presented. Third, the program’s impact on educational outcomes —i.e. SAT-like test registration and school dropout rates— as well as non-cognitive skills is estimated. The latter considers creativity and socio-emotional skills measures, which are the main expected mechanism accounting for program’s impact. Finally, I estimate the program’s effects on emotional regulation, in particular, on pre-test resting state emotional state and emotional responsiveness to both positive and negative stimuli.

Descriptive Statistics

This subsection presents the descriptive statistics and attrition in the field experiment. Table 2.1 shows that the sample is balanced at baseline —i.e. March 2014— in all variables but arousal index in resting state and Grit scale. As Deaton (2010) noted, randomized controlled trials are frequently unbalanced in randomized field experiments, even in large experiments such as the RAND Health Insurance Experiment. Moreover, in the context of this study, treated subjects were aware at the moment that the baseline experiments were taken that they belong to the treatment group of an incoming intervention. This situation can bias self-reported measures.

58 As mentioned, the data was collected in August instead of June because of national teacher strike at the end of the first semester. Then, many schools prefer to begin the follow up collection of data in August instead of the very beginning of the second semester —i.e. middle of July— because they already had many activities scheduled.

59 See for example Banerjee and Duflo (2008).
Table 2.1: Experimental Balance at Baseline

<table>
<thead>
<tr>
<th>Locus Control Test</th>
<th>Control</th>
<th>Treated</th>
<th>(1) vs. (2)</th>
<th>p-value diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.058</td>
<td>0.172</td>
<td>-0.230</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>(0.099)</td>
<td>(0.130)</td>
<td>(0.161)</td>
<td></td>
</tr>
<tr>
<td>Grit Test</td>
<td>-0.018</td>
<td>0.294</td>
<td>-0.312</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>(0.098)</td>
<td>(0.112)</td>
<td>(0.150)</td>
<td></td>
</tr>
<tr>
<td>Creativity Index</td>
<td>0.006</td>
<td>-0.032</td>
<td>0.039</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>(0.101)</td>
<td>(0.133)</td>
<td>(0.164)</td>
<td></td>
</tr>
<tr>
<td>13 Forces Creativity</td>
<td>-0.083</td>
<td>-0.137</td>
<td>0.054</td>
<td>0.729</td>
</tr>
<tr>
<td></td>
<td>(0.100)</td>
<td>(0.123)</td>
<td>(0.157)</td>
<td></td>
</tr>
<tr>
<td>Valence Resting State</td>
<td>0.047</td>
<td>0.078</td>
<td>-0.032</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>(0.092)</td>
<td>(0.078)</td>
<td>(0.127)</td>
<td></td>
</tr>
<tr>
<td>ValenceNegative/Withdrawal</td>
<td>-0.039</td>
<td>-0.065</td>
<td>0.026</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>(0.094)</td>
<td>(0.111)</td>
<td>(0.145)</td>
<td></td>
</tr>
<tr>
<td>ValencePositive/Approach</td>
<td>0.050</td>
<td>0.098</td>
<td>-0.047</td>
<td>0.764</td>
</tr>
<tr>
<td></td>
<td>(0.111)</td>
<td>(0.103)</td>
<td>(0.157)</td>
<td></td>
</tr>
<tr>
<td>Arousal Resting State</td>
<td>-0.021</td>
<td>0.260</td>
<td>-0.280</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>(0.101)</td>
<td>(0.113)</td>
<td>(0.153)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>104</td>
<td>76</td>
<td>180</td>
<td></td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.
Notes: Standard Errors are clustered at school level.

The differences between groups in the arousal index at resting state and the Grit scale becomes insignificant when controlling by baseline measures in the estimation. I will discuss these differences later in the difference-in-difference results. These simple results indicate that the puzzling result that typically arises in impact evaluations related to educational or labor market programs designed to foster non-cognitive skills is likely to happen in the present study as well. Moreover, as I will show later in this section, the program has significant impacts on at least two relevant educational outcomes —i.e SAT-like test registration and school dropout rates.

**Attrition**

It is important to analyze data attrition —i.e. missing data— in experimental studies. The main goal is to understand its sources and test whether the attrition was uneven across experimental groups. There are different sources of data attrition in this study.
First, experimental attrition generated in the baseline due the lack of precision in the collected metadata that was later used to merge different sources of information. Field experiments were implemented using several working stations —i.e. a laptop with a Emotiv EPOC headset. There were six and eight stations in the baseline and follow up, respectively. When conducting the experiment, we made several mistakes linking metadata on the computers, written records to match individuals with their data files, specially when an error happened and it was necessary to restart the whole experiment. In particular, it was necessary to record the initial and final time of each individual. With that information it was possible to match the EEG-data files with test scores.

Second, there was attrition as a result of the quality of the recording. Due to different issues, EEGLAB —the Matlab toolbox used to analyze EEG data— sometimes failed to read the EEG recordings properly. Highly dense amounts of hair, long hair, or both, and/or computers freezing during experiments were the usual suspects. Moreover, more subjects were rejected after signal processing —high and low band-pass filtering, Fourier transformation into frequencies, etc. In short, many subjects were lost.

### Table 2.2: Attrition

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th></th>
<th>Follow up</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Treated</td>
<td>Control</td>
<td>Total</td>
</tr>
<tr>
<td>Participants</td>
<td>287</td>
<td>123</td>
<td>164</td>
<td>205</td>
</tr>
<tr>
<td>Valid EEG Records</td>
<td>195</td>
<td>85</td>
<td>110</td>
<td>136</td>
</tr>
<tr>
<td>Valid EEG w/relevant info.</td>
<td>180</td>
<td>76</td>
<td>104</td>
<td>116</td>
</tr>
</tbody>
</table>

Notes: “Valid EEG records” indicates subjects who have EEG data that is readable. “Valid EEG w/relevant info.” accounts for subjects with readable EEG data and that was also possible to match it with test scores.

---

60 For example, Laptop number 5 recorded with a high proportion of error for two consecutive days, which generate 13 missing subjects.
Third, individual missed between experiments rounds —i.e. between baseline and follow up field experiments— can be an important source of selection bias due to attrition. There was a decrease in the total sample size of around 36%. Moreover, it is important to highlight that the attrition in the control group, which rose to 43%, is considerably larger than that in the treatment group. Nevertheless, the attrition in the control group was balanced with respect to the measurements considered in this study. In fact, there is no statistically significant difference at baseline between treatment groups that participated in the experiment only at the baseline and those whom participated at both baseline and follow up.\footnote{See Table 2.8 in the appendixes.}

Fourth, it is also possible to rule out some potential unobservables characteristics that could be present in a major proportion of the group of students who participated in the follow up. For example, the degree of intrinsic motivation could be different, which is allegedly a proxy in the locus control test, and did not show differences in the baseline (see Table 2.1). Anecdotally, references from teachers at the control schools explained some of the reasons behind the lack of participation in the follow up. For one thing, many students have a lack of motivation in general, especially to attend school. In fact, many of the control schools’ students who participated in the baseline experiment were absent from school that day. There were also a couple of cases of students who dropped out of school altogether.

Fifth, it is plausible that boredom contributed to the lack of incentive to participate in the study, especially in the control group which has no relation to the program. Students in the control group were completely uninformed about what and why we were doing this study on them. Participating in the baseline experiment could be boring. The experiment takes 30-40 minutes on average, requiring students to be seated in the classroom conditioned as a laboratory to set up the EEG headset, check and habituate the EEG headset to have a correct reading of all 14 electrodes,
and then take the actual test, which is 20 minutes on average. Indeed, the relative large number of takers—not considering whether they have a valid EEG datafile, see Table 2.2—in the control group in baseline (164 subjects) was unexpectedly high compared with those in the treatment group (123).

Potential selection bias due to attrition is unlikely because it happened apparently randomly in both treatment and control groups. See the Table 2.8 in the appendixes for the mean test of the difference of observable characteristics at baseline between participant and control group’s students missed between the baseline and the follow up, and those who had the change to do follow up as well.

Impact on Educational Outcomes

The Mining’s Rockstar program is expected to affect some educational outcomes. Here I consider two outcomes—i.e. SAT-like\textsuperscript{62} registration and dropouts rates—in order to argue that the program has observable impacts on the treatment group, aside from its impact on the emotional regulation indicators.\textsuperscript{63} This is particularly important since it shows that this program is experiencing a similar puzzle to the one found on the labor economics literature. That is to say, on the one hand, the program exhibits a positive impact on educational outcomes—i.e. SAT-like registration and dropouts rates. On the other hand, as shown in Table 2.6, the expected factors behind the impact—i.e. non-cognitive skills—are not affected.

\textsuperscript{62}The University Admission Test (Spanish acronym PSU) aims is to measure the mastery of secondary education topics. It is similar to the SAT in the US. This test is voluntary for students who have finished the secondary education, but required for applying to university. The PSU is the principal selection mechanism along with average secondary education grades, which are combined. The PSU, has a relative weight in the final university application score of around 30%. The PSU score is normalized to an average 500 points and a standard deviation of 100 points, with a fixed status ranging from 150 to 850 points. Thus the PSU and secondary school grades are the principal instruments to access the most prestigious careers and universities in the country.

\textsuperscript{63}See next subsection for details about the program’s impact on emotional regulation.
The administrative data on educational outcomes is only accessible at the end of the academic year. Thus, I estimate a simple difference model. In addition, since the data comes from administrative records, it was possible to consider additional control groups as a robustness check of the results. In particular, it was possible to have the same sample of individuals used socio-emotional skills and emotional regulation estimations for the SAT-like registration, but not for the dropouts rates. In fact, this sample considers those individuals who did the follow up field experiments, which is impossible for those who dropped out because these experiments were done at schools. The model is the following:

\[
Y_{j,i,t} = \alpha + \beta \cdot treat_i + \theta \cdot X_{i,t} + \varepsilon_{i,t},
\]

where \(Y_{j,i,t}\) indicates the \(j\) educational outcome of individual \(i\) in time \(t\), with \(j = SAT\text{registration}, \text{dropoutrates}\), \(i = \text{students}\), and \(t = \text{follow-up}\). Furthermore, \(treat_{j,i}\) indicate who had been treated by the program. Finally, \(X_{i,t}\) indicates school level dummy variables.

The program had a significant impact on educational outcomes —i.e. dropout rates and SAT-like registration; see Table 2.3 and 2.4, respectively.— yet had no impact on socio-emotional skills or creativity measures, which is consistent with the findings in the labor economics literature referenced above.

As Table 2.3 shows, there is a significant difference in registration for all models of SAT-like (PSU in Chile) registration. Registration for the SAT is voluntary, but is a requirement to enroll in any university. In order to attend a technical college —i.e. something similar to community college in the US— it is not necessary to take the SAT-like test. In fact, the positive impact found on the treated compared to both control groups in the SAT-like test registration rate can reflect an increase in their expectations about their futures.
Table 2.3: Registration to take PSU (SAT-like) test

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Registered</th>
<th>(2) Registered</th>
<th>(3) Registered</th>
<th>(4) Registered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (RCT’s Schools)</td>
<td>0.130***</td>
<td>0.221***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0276)</td>
<td>(0.0297)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment (RCT’s EEG Subsample)</td>
<td></td>
<td></td>
<td>0.309***</td>
<td>0.375***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0556)</td>
<td>(0.0585)</td>
</tr>
<tr>
<td>Non-Municipal School</td>
<td>0.295***</td>
<td></td>
<td>0.272***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0353)</td>
<td></td>
<td>(0.0865)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1,296</td>
<td>1,296</td>
<td>294</td>
<td>294</td>
</tr>
<tr>
<td>School FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Notes: School dummies and other controls are only available for registered students. Due to privacy restrictions of the PSU dataset, it was not possible to identify the schools in the sample. However, it was only possible to control for one school that is run by an external non-for-profit institution—instead of the DAEM—where students tend to enroll in the tests at a larger proportion.

Table 2.4: Students Dropout Rates

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Dropout</th>
<th>(2) Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (RCT’s Schools)</td>
<td>-0.0683*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0366)</td>
<td></td>
</tr>
<tr>
<td>Treatment (Vocational Sch. Same Region)</td>
<td>-0.0428***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00678)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>-0.00228</td>
<td>0.000849</td>
</tr>
<tr>
<td></td>
<td>(0.00968)</td>
<td>(0.00417)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,888</td>
<td>8,017</td>
</tr>
<tr>
<td>School FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Notes: “Dropout” is a dichotomous variable that equals to 1 if the student dropped out, 0 otherwise. Without loss of generality, only students in their last year of secondary school were considered controls in the analysis. Marginal effects of Probit model reported.

Table 2.4 shows a statistically significant decrease in dropout rates of about 6 percentage points among participants. This decrease in dropout is a relevant outcome. Dropout rates are high in Chile, especially among students who come from vulnerable socio-economic backgrounds, and who are over-represented in technical
schools (Valenzuela et al., 2013). Moreover, descriptive analyses of similar programs implemented previously by the same NGO that runs “Mining’s Rockstars” program, Emprende Joven, had shown a positive impact of the program decreasing dropout rates among participants with respect to historical school dropout rates (Albarran and Gonzalez, 2015). In addition to the social desirability of school retention for many reasons —e.g. in Chile it is mandatory to finish high school— the negative impact found is consistent with the positive impact on expectations. In summary, it is important to support the effectiveness of the program to affect observable educational outcomes. However, it is beyond the scope of this study to discuss further the impacts on educational outcomes because the focus is on the mechanisms, such as the program’s effects on emotional regulation or non-cognitive skills.

Impact on Self-reported measures of Non-cognitive Skills

It is expected that the Mining’s Rockstar program affects non-cognitive skills. In particular, using conventional self-reported tests, grit—perseverance— and locus control were measured. The program is also planned to impact on creative and innovative skills, which is a mixture between cognitive and non-cognitive skills. In addition, measuring creative skills is scarce in the literature, in particular in the context of an educational program evaluation. The model to be estimated will be the following:

\[
Y_{j,i,t} = \alpha + \beta \ast treat_i + \gamma \ast Post_t + \delta \ast treat_i \ast Post_t + \theta \ast X_{i,t} + \varepsilon_{i,t}, \quad (2.12)
\]

---

64 See previous section for details on these tests.
65 See Chapter 3 and 4 for details on creativity definitions and measures.
66 See Chapter 4 for a review of the literature on creativity and educational programs.
Table 2.5: Difference Model on Non-cognitive Skills: Creativity, Locus Control and Grit

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Creativity Index</th>
<th>(2) 13 Forces Creativity</th>
<th>(3) Locus Control Test</th>
<th>(4) Grit Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>-0.00600</td>
<td>-0.173</td>
<td>-0.115</td>
<td>-0.311</td>
</tr>
<tr>
<td></td>
<td>(0.989)</td>
<td>(0.677)</td>
<td>(0.813)</td>
<td>(0.473)</td>
</tr>
<tr>
<td>12th Grade (=1)</td>
<td>0.176</td>
<td>0.378</td>
<td>0.00129</td>
<td>1.050***</td>
</tr>
<tr>
<td></td>
<td>(0.544)</td>
<td>(0.218)</td>
<td>(0.997)</td>
<td>(0.00337)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.250</td>
<td>-0.144</td>
<td>0.191</td>
<td>-0.150</td>
</tr>
<tr>
<td></td>
<td>(0.233)</td>
<td>(0.565)</td>
<td>(0.446)</td>
<td>(0.541)</td>
</tr>
<tr>
<td>Observations</td>
<td>116</td>
<td>116</td>
<td>116</td>
<td>116</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.181</td>
<td>0.181</td>
<td>0.022</td>
<td>0.147</td>
</tr>
<tr>
<td>School Dummies</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student Level Cluster</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4th Grade</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Robust Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Notes: Standard Errors are clustered at individual level. “4th Grade” is dichotomic variable that index if student was in 4th grade or not. This variable was statistically non-significant in all specifications.

where $Y_{j,i,t}$ indicates the $j$ test’s score of individual $i$ in time $t$, with $j = \text{creativity, grit, locuscontrol}$, $i = \text{students} \in \text{FieldExperiment}$, and $t = \text{baseline, follow-up}$. Furthermore, $treat_{j,i}$ and $Post_{j,t}$ indicate who had been treated by the program and survey time —i.e. baseline or follow up— respectively. Finally, $X_{i,t}$ indicates school level dummy variables and 4th grade dummy variable. In order to have the highest possible statistical power in the estimations, it was considered both non-treated 4th grade ($N=61$) and 3rd grade ($N=95$) students as control in all model specifications, controlling by a dichotomic variable to differentiate them.

As Table 2.1 shows, there is a significant difference in pre-test resting state arousal and Grit scale in baseline measures in favor of the program’s participants. The estimation of balance on Table 2.1 considered standard errors that are clustered at school level. This allows to control for intra-cluster correlation within schools, which is particularly important here since the randomization was done among eight schools, a relatively small number for clusters according the literature (Cameron and Trivedi, 2005).
Furthermore Table 2.5 estimates the program’s impact on the treatment group, considering only the follow up measures. This implies $t = \text{august}$, and thus $\gamma = \delta = 0$. Since the experiment was randomized, the results contained in Table 2.5 are consistent and unbiased estimations of the average treatment effect on the treated. In particular, there is apparently no significant impact of the program in these dimensions. The only significant difference is in the Grit scale between those at 4th grade relative to 3rd grade.

However, in order to rule out differences that could arise in the baseline at individual level, a differences-in-differences model (DiD) is estimated. This kind of model incorporates the differences between the treated and control groups at baseline, before the program started. Moreover, it also contemplates a common time trend for both groups in the outcomes variables of interest.

Table 2.6 exhibits the estimates from equation 2.12 of the DiD model. The main parameter of interest is $\delta$ that isolates the average treatment effect taking into account the natural evolution of the outcome across time as well as the selection into the treatment group. In fact, Table 2.6 shows that program participation apparently does not positively impact grit or locus control scales, which is contrary to the expected impact by program design.

Consistent with the results on Table 2.6, the impact on Grit Scale is likely driven by a difference in the treatment related to those in 4th grade rather than the program itself. The Locus Control Scale exhibits no significant change, however the coefficient is negative, which indicates a trend towards internal locus control that is positively related to educational and labor outcomes (Puentes and Urzua, 2010). Furthermore, the Torrance Test of Creative Thinking comprises of three dimensions of creativity, plus an additional indicator of “overall” creative skill, namely: fluidity, flexibility, originality and “13 creative forces.” In fact, the estimated program’s impact on the
Table 2.6: Difference in Difference Model on Non-Cognitive Skills: Creativity, Locus Control and Grit

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Creativity Index</th>
<th>(2) 13 Forces Creativity</th>
<th>(3) Locus Control Test</th>
<th>(4) Grit Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment &amp; Post.</td>
<td>0.223</td>
<td>0.159</td>
<td>-0.118</td>
<td>0.0395</td>
</tr>
<tr>
<td></td>
<td>(0.381)</td>
<td>(0.502)</td>
<td>(0.631)</td>
<td>(0.842)</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.196</td>
<td>-0.105</td>
<td>0.188</td>
<td>0.0448</td>
</tr>
<tr>
<td></td>
<td>(0.479)</td>
<td>(0.697)</td>
<td>(0.563)</td>
<td>(0.872)</td>
</tr>
<tr>
<td>Post.</td>
<td>0.0751</td>
<td>0.331*</td>
<td>0.179</td>
<td>0.0423</td>
</tr>
<tr>
<td></td>
<td>(0.652)</td>
<td>(0.0537)</td>
<td>(0.270)</td>
<td>(0.760)</td>
</tr>
<tr>
<td>4th Grade (=1)</td>
<td>0.233</td>
<td>0.153</td>
<td>0.143</td>
<td>0.683***</td>
</tr>
<tr>
<td></td>
<td>(0.226)</td>
<td>(0.417)</td>
<td>(0.471)</td>
<td>(0.000169)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.0880</td>
<td>-0.183</td>
<td>0.00868</td>
<td>-0.203</td>
</tr>
<tr>
<td></td>
<td>(0.693)</td>
<td>(0.414)</td>
<td>(0.964)</td>
<td>(0.221)</td>
</tr>
</tbody>
</table>

Observations     | 296                   | 296                      | 296                    | 296          |
R-squared        | 0.078                 | 0.088                    | 0.031                  | 0.152        |
Student level Clustering | Yes | Yes | Yes | Yes |
4th Grade Control | Yes | Yes | Yes | Yes |
School Dummies   | Yes | Yes | Yes | Yes |
Exact pvalue (WildB.) | 0.382 | 0.510 | 0.618 | 0.852 |

Robust Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Note 1: Standard Errors are clustered at individual level.
Note 2: “4th Grade” is dichotomic variable that index if student was in 4th grade or not.
Note 3: Exact p-value is the results of 1,000 replications of Wild bootstrap, following the algorithm of Webb for few clusters. Results are similar if clusterization is at school level.

Creativity index as well as the “13 creative forces” index are not significant.67

Impact on Emotional Regulation

Table 2.7 describes the program’s impact following a DiD model that considers two novel outcomes attempting to capture the unobservable impact on emotional regulation. First, I use the indices of pre-test resting state valence and arousal following the methodology explained above. Second, I proxy emotional responsiveness to both positive and negative stimuli also in the arousal and valence dimensions. In simple

67 Analogous to the DiD model, estimations using either a simple average of the three dimensions of creativity —flexibility, fluidity and originality— or a principal component analysis index (pca) of creativity exhibited similar results in terms of magnitude, sign and significance. The only dimension on which the program had a positive and significant impact was originality, which is also similar to the difference-in-difference model.
words, emotional responsiveness captures the neurophysiological reactions of students when they view an emotionally laden set of images.68

In particular, the following DiD model is estimated:

\[
E_{j,i,t} = \alpha + \beta * treat_i + \gamma * Post_t + \delta * treat_i * Post_t + \theta * X_{i,t} + \varepsilon_{i,t}, \tag{2.13}
\]

where \(E_{j,i,t}\) indicates the \(j\) emotional state or emotional responsiveness associated with an emotionally laden stimuli of individual \(i\) in time \(t\), with \(j = arousal_{base}, valence_{base}, valence_{positive}, valence_{neg}, \) \(dif-valence_{pos}, dif-valence_{neg},\)

\(i = \text{students} \in \text{FieldExperiment}, \) and \(t = \text{baseline, follow - up}.\) Indicators \(dif-valence_{pos}\) and \(dif-valence_{neg}\) identified the difference for each individual \(i\) at time \(t\) between her \(valence_{base}\) and \(valence_{positive}\) and \(valence_{neg},\) respectively. That is to say, the indicators capture the emotional reaction considering the resting state level of valence. As before, \(treat_{j,i}\) and \(Post_{j,t}\) indices who had been treated by the program and survey time —i.e. baseline or follow up— respectively. Finally, \(X_{i,t}\) indicates school level dummy variables. It is expected that \(\delta < 0\) and \(\beta = \gamma = 0.\)

Tables 2.7 and 2.6 show Mining’s Rockstars program’s impacts on socio-emotional skills, creativity and emotional resting state and responsiveness. Both tables consider the same sample and specification of the models. Specifically, Tables 2.7 and 2.6 show the results from all (measured) potential mechanism of the effects —i.e. socio-emotional and creative skills, and emotional regulation using EEG recordings— using the sample that comprises all subjects for whom it was possible to get valid EEG recordings in both baseline and follow up surveys and experiments (296 observations). The model contemplates a difference-in-difference estimation (e.g equations 2.12 and

---

68 See Chapter 3 for examples images from the Geneva Affective Picture Database (GAPED).
### Table 2.7: Impact on Emotional Regulation: Emotional State and Responsiveness

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) ArouRest</th>
<th>(2) ValRest</th>
<th>(3) ValNeg</th>
<th>(4) ValPos</th>
<th>(5) ValNegDif</th>
<th>(6) ValPosDif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat &amp; Post.</td>
<td>-0.538**</td>
<td>-1.430*</td>
<td>0.740</td>
<td>-0.0646</td>
<td>-2.170**</td>
<td>-1.366</td>
</tr>
<tr>
<td></td>
<td>(0.0240)</td>
<td>(0.0911)</td>
<td>(0.264)</td>
<td>(0.939)</td>
<td>(0.0335)</td>
<td>(0.255)</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.000336</td>
<td>-0.144</td>
<td>0.180</td>
<td>0.186</td>
<td>-0.324</td>
<td>-0.330</td>
</tr>
<tr>
<td></td>
<td>(0.999)</td>
<td>(0.871)</td>
<td>(0.839)</td>
<td>(0.842)</td>
<td>(0.794)</td>
<td>(0.785)</td>
</tr>
<tr>
<td>Post.</td>
<td>-0.00431</td>
<td>0.296</td>
<td>-1.147**</td>
<td>0.513</td>
<td>1.443*</td>
<td>-0.217</td>
</tr>
<tr>
<td></td>
<td>(0.977)</td>
<td>(0.653)</td>
<td>(0.0201)</td>
<td>(0.399)</td>
<td>(0.0687)</td>
<td>(0.798)</td>
</tr>
<tr>
<td>12th Grade (=1)</td>
<td>0.0604</td>
<td>0.259</td>
<td>-0.183</td>
<td>0.423</td>
<td>0.442</td>
<td>-0.164</td>
</tr>
<tr>
<td></td>
<td>(0.737)</td>
<td>(0.738)</td>
<td>(0.752)</td>
<td>(0.570)</td>
<td>(0.669)</td>
<td>(0.865)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.0866</td>
<td>0.165</td>
<td>-0.258</td>
<td>-1.431*</td>
<td>0.423</td>
<td>1.596*</td>
</tr>
<tr>
<td></td>
<td>(0.658)</td>
<td>(0.785)</td>
<td>(0.757)</td>
<td>(0.6983)</td>
<td>(0.644)</td>
<td>(0.0993)</td>
</tr>
</tbody>
</table>

Observations: 296
R-squared: 0.051 0.024 0.037 0.018 0.034 0.027
School Dummies: Yes Yes Yes Yes Yes Yes
Student Level Cluster: Yes Yes Yes Yes Yes Yes
4th Grade Control: Yes Yes Yes Yes Yes Yes
Exact p-value (WildB.): 0.0300 0.0810 0.241 0.934 0.0300 0.262

Note 1: Standard Errors are clustered at individual level.
Note 2: “4th Grade” is dichotomic variable that index if student was in 4th grade or not.
Note 3: Exact p-value is the results of 1,000 replications of Wild bootstrap, following the algorithm of Webb for few clusters. Results are similar if clusterization is at school level.

2.13). The specification includes school dummies, 4th grade dummy, which control for being in 4th grade respect to 3rd grade in high-school, and robust standard errors clusterized at individual level. Moreover, exact p-value of our main coefficient of interest —i.e. $\delta$ in equation 2.12— is estimated due the econometric issues that arises because of the relatively small numbers of cluster (less than 20) in the randomized controlled trial design. The exact p-value is the results of 1,000 replications of Wild bootstrap, following the algorithm of Webb for few clusters (Cameron and Trivedi, 2005). Finally, Table 2.3 exhibits the program’s impact on registration to the PSU, which is the test to apply to college (SAT-like ), using the sample sample as Tables 2.7 and 2.6 in models (3) and (4). The approach and specification here is slightly different because many reasons. First, the registration to the PSU is a dichotomic variable, thus I follow a probit model instead of a linear regression as before. Second, school
dummies and other controls are only available for registered students, because the information comes from administrative data provided by the University of Chile’s DEMRE, and thus there is privacy restrictions. Nevertheless, it was possible to control for one school that is run by an external non-for-profit institution—instead of the municipal educational office—where students tend to enroll in the PSU test at a larger proportion, which is clearly shown in Table 2.3. In summary, the use of the same sample and as similar as possible specification gives consistency to the results.

As mentioned, the DiD model is a suitable choice considering baseline disparities between treatment and control groups, in particular the difference shown in the pre-test resting state arousal index. Table 2.7 shows that both participants’ pre-test resting state arousal and valence indices experience a statistically significant decrease compared to their control group. Impacts are $0.13\sigma$ and $0.44\sigma$ on pre-test resting state arousal and valence, respectively. Moreover, as Table 2.7 shows, there is a significant impact on responsiveness to negative emotionally laden stimulus—around $0.47\sigma$, which is near the upper bound on similar interventions (Murnane and Ganimian, 2014; Duflo et al., 2012). In other words, subjects’ neurophysiological responsiveness when faced with negative stimuli—i.e. pictures of mistreated animals or human rights violations—change significantly after treatment compared with their own baseline reaction and the control group emotional reaction. This might be interpreted as an increase resilience trait among participants.\footnote{See for example Troy and Mauss (2011)}

Similarly, I consider the model in differences on the dependent variable following the specification of the equation (2.14) as a robustness check exercise.\footnote{See details about the specification of this model in the Appendix.} By doing this, all individual time-invariant unobservable factors that are relevant for the model are cancelled out. It is important to note that here we restrict our analysis to the balanced panel data, which consists of only 68 observations. The significant decrease
on the emotional responsiveness to negative stimulus —Model (8) on Table 2.17 in the Appendix— is consistent with previous results.

In summary, the impacts found on emotional regulation are statistically relevant compared to those that similar programs have on other outcomes of interest (Murnane and Ganimian, 2014; Duflo et al., 2012). As mentioned, there is extensive literature about decision making and emotions, and also about behavioral change and emotions. However, the results here are not trivially translated into a behaviorally meaningful interpretation.

For instance, similarly, Ibarraran et al. (2014) argue that impacts from ALMP using self-reported measures of non-cognitive skills, yet statistically significant, have unclear practical and behavioral interpretation respect to outcomes in the labor market.

A discussion of the results and some examples can be found in the next subsection.

Discussion

Nowadays, evidence supports the argument that cognitive skills explain only a small fraction of the labor market outcome variance (Almlund et al., 2011; OECD, 2015). The literature has been focused on improving the robustness of the estimations of non-cognitive skills’ models using factor models (e.g. Cunha and Heckman, 2008; Attanasio et al., 2015a), with the idea of recovering a given underlying skill using at least two—or three, depending on the model— different proxy measures of the same skill. This strategy increases the reliability of a skill’s proxies, but does not solve the problem that the program potentially affects other unobservable dimensions such as

---

71 For recent literature reviews see for example Weber and Johnson (2009); Loewenstein et al. (2001); Lempert and Phelps (2014); Lerner et al. (2012)
emotional regulation, and thus the way in which subjects answer self-reported tests. This is particularly important for both surveys and single trial measurements, when it is not possible to have revealed outcomes —i.e. salaries or employment conditions— or when the aim is to understand the channels driving observed outcomes.

I argue that programs meant to enhance life skills likely impact dimensions that are too complex to measure through traditional self-reported psychometric tests. This complexity arises because programs that potentially impact emotional regulation can affect behaviors in many ways, including the actual self-perception of subjects at the moment in which they are answering self-reported tests. Furthermore, I explored the relevance of the development of emotional regulation, which was identified as a pretest arousal and valence indices, as well as emotional responsiveness to both positive and negative stimuli.

In short, I find no significant impact of participation in the program on different measures of non-cognitive skills that the program is supposed to foster, which is consistent with the literature. However, I found a significant program impact on pre-test emotional state, measured by a decrease in arousal and valence indices. In fact, the decrease on the indices of emotional state —i.e. arousal and valence— potentially accounts for the aforementioned puzzle in the literature. Moreover, I estimate a decrease in the emotional responsiveness to negative stimuli that is also significant. Considering the statistical significance of these results, and the novelty of the methodology and indices for a social program evaluation, I will offer a discussion with plausible explanations derived from psychology, behavioral economics, and affective neuroscience literature. The discussion will be structured to consider (i) the impact on resting-state emotional state, and (ii) emotional responsiveness to external stimuli.

See for example Calero et al. (2014); Heckman and Kautz (2012); Kautz et al. (2014).
Resting-state Emotional State

Consistent with previous studies, I found no short-run impact among “Mining’s Rockstars” participants in the expected outcomes, such as grit or locus control. This apparent lack of impact is consistent with the decrease on both arousal and valence indices generated by program’s participation. Indeed, since there is in general a positive correlation between emotional state and self-reporting on non-cognitive skills tests, nonsignificant impacts are expected. At first glance, this decreasing effect on emotional dispositions seems to have the opposite effect to that of a program aiming to improve socio-emotional skills following a methodology that relies on the principle of learning by failure. The program’s methodology also includes the analysis of failure — which is endogenously generated in every weekly activity — awareness about context, and potential opportunities to improve. The decrease of the emotional state due participation in the program might be theoretically driving the estimated lack of impact on the skills of interest. In the following sections I will review a number of potential interpretations of these counterintuitive results.

Emotions and Awareness

In terms of the empirical findings of the present study, a decrease in pre-test arousal can be completely positive. In particular, the level of arousal was higher at baseline for participants who can be thought to have an altered emotional state due external adverse factors, such as being surrounded by extreme poverty or violence. In fact, a decrease in the pre-test arousal indices can imply a more relaxed state of the individual.

Emotions can be interpreted in many ways. Darwinian emotions, modulated by the reptilian brain, are the most basic. Among these are emotions such as anger, fear

---

73 See for example Attanasio et al. 2015a; Calero et al. 2014; Card et al. 2011
74 See Chapter 3 for details in the relation between emotions and self-reporting.
of snakes and spiders, etc. The James-Lange theory argues that there is a correlation between physiological experience —i.e. feelings— and emotions. This theory is the basis for the arousal-valence model that I have used here. Damasio (1994) elaborated a similar theory, called the somatic-marker hypothesis. He argues that somatic-sensory feelings are coded —i.e. related— with past experience, which has associated emotions.

In general terms, valence can be understood as attitude and disposition to act. In other words, it is the departure state of body and mind before taking any action. Indeed, emotions modulate the course of an individual’s actions. They can also be seen as the framework for actions. For instance, if an individual is experiencing a low emotional state —i.e. she is low in arousal and valence— it is harder for her to engage in a given task. For example, Barraza et al. (2015) study the relation between emotional state and memory. The authors argue that positive emotions create an orientation towards the outside world. By contrast, a negative emotional state is associate with an orientation towards yourself, to the internal body/mind.

Moreover, low levels of arousal and valence could be thought of as depressive, and thus a more realistic and pessimistic state (Barraza et al., 2015). This claim is in opposition to those experiencing positive emotions, who are likely optimistic about their own situation and the environment. Moreover, subjects who are emotionally depressed focus their attention relatively more to the exterior, compared to those experiencing positive emotions who tend to focus more on themselves (Barraza et al., 2015). Therefore, the lower level of arousal and valence can be interpreted as an increase in external awareness. The program likely affects the sense of awareness among participants through its work on failure analysis. For instance, due to the high level of vulnerability of the context of this study, the withdrawal behavior can be thought of as a lower level of frustration because adverse outcomes are expected. Moreover,

---

75See for a recent review Harmon-Jones et al. (2010) or Salzman and Fusi (2010).
the results asymmetry respect to emotional responsiveness —i.e. there is significant impact on decreasing reaction only for negative stimulus— are consistent with those found in research about individuals who meditate performing mindfulness meditation. Meditation practitioners are less affected by stimuli with an adverse emotional load, while their processing of positive stimuli remains unaltered (Sobolewski et al., 2011).

In short, the results obtained in the present study can be interpreted as a change in attitude and disposition to act and the level of awareness among participants. Therefore, we should observe a dynamic complementarity between those changes and other skills acquisition by participants. For instance, changing attitudes towards failure would also change educational and labor market decisions, such as effort given to tasks, occupational choice, entrepreneurship, and the pursuit of creative and original work, among many others. The exploration of those dynamic complementarities is an interesting topic for further research.

**Approach/withdrawal Theory of Emotion**

As mentioned, Davidson et al. (1983) suggested, in a seminal work, a model called approach/withdrawal theory to investigate frontal EEG asymmetry during emotional states. Since then, psychology and neurophysiology literature points out that frontal EEG asymmetry is associated with different emotional and psychological states in addition to valence (Davidson et al., 1990; Harmon-Jones and Gable, 2008; Harmon-Jones et al., 2010). In fact, the approach/withdrawal model invites us to think of an additional interpretation of our results. The decrease in the frontal EEG asymmetry —i.e. valence index— are now consistent with a relative increase of withdrawal behavior. Due to the high level of vulnerability in the context of this study, that withdrawal behavior can be thought of as a lower level of frustration of participants relative to the control group because adverse outcomes are expected.
Since the Mining’s Rockstars program aims to master that attitude towards failure, this is an appropriate lens through which to interpret our results.

**Efficiency in the use of Resources**

Finally, an alternative explanation to the puzzle in the literature that motivates this study regards efficiency in the use of resources. I argue that emotional regulation, which modulates the attitude and disposition to act in the external world, is affected by the program. In particular, it is possible to frame the change under the concept of mental state. Mental state is a disposition to action —i.e., every aspect of the individual inner state that can contribute to its behavior or other responses— that is present at a given moment (Salzman and Fusi, 2010). Since many non-cognitive skills are thought to be non-malleable in the short term, it is possible that the program of study does not directly affect a certain personality trait or skill, but instead affects the degree of efficiency given the subject’s set of cognitive and non-cognitive skills or mental state. That is to say, participants somehow learn to use their personal resources more efficiently, take opportunities, modulate negative shocks, and so forth. These claims fit well into emotion theories that associate positive affects with approach motivation and negative affects with withdrawal motivation (Harmon-Jones et al., 2010). It is also plausible that the emotional state has a nonlinear relationship resembling an inverted U with respect to performance in cognitive or other behaviors. This would be consistent with the famous Yerkes-Dodson law that states that stress and performance may exhibit a nonlinear relationship resembling an inverted U (Haushofer and Fehr, 2014). According to the Yerkes-Dodson law, moderate increases in arousal lead to improvements in performance, whereas extreme levels of arousal lead to performance decrements. Since we observe a decrease in arousal and allegedly positive behavioral impacts, participants should be in the right side of the peak in the inverted U. This can be an interesting topic for further research.
Emotional Responsiveness

The impact on emotional responsiveness is of particular interest. We observe a decrease in the neurophysiological reaction to negative stimuli in participants, while observing no significant impact on the responsiveness to positive stimuli. This result could be phrased as a relevant outcome of the “Mining’s Rockstar” program. This result is consistent with recent evidence of late stage investments showing a positive impact on behaviors instead of on underlying non-cognitive skills (Blattman et al., 2015).

Reappraisal Strategy

A core aspect of the program’s methodology includes analyzing and rethinking failure, which is intentionally generated in each of the weekly activities contained in the program’s guidebook. Indeed, it is possible to frame that part of the methodology as a reappraisal of the emotions involved, especially in the case of failure. Lerner et al. (2015) review the recent findings for “Solutions that Seek to Minimize the Emotional Response.” They point out four dimensions: Time delay, Suppression, Reappraisal, and The “dual-emotion solution” (inducing a counteracting emotional state). In particular, reappraisal consists of reframing the meaning of stimuli that led to an emotional response. Reappraisal has consistently emerged as a superior strategy for dissipating the emotional response (Lerner et al., 2015; Gross et al., 2003). Specifically, reappraisal includes such behaviors as reminding oneself “it’s just a test” after receiving a poor exam grade, adopting the mind-set of a nurse or medical professional to minimize the emotional impact of viewing someone’s injury, or viewing a job layoff as an opportunity to pursue long-forgotten dreams (Gross et al., 2003). In contrast to suppression, reappraisal not only reduces self-reported negative feelings in response to negative events but also mitigates physiological and neural responses to those events (Lerner et al., 2012; Ochsner et al., 2002; Jamieson
et al., 2012). Additionally, regulating emotion with reappraisal-focused strategies that encourage taking a different perspective has been shown to reduce loss aversion in decision making (Sokol-Hessner et al., 2012). That result has been observed both in choices and in the relative arousal responses to actual loss and gain outcomes.

Finally, there is evidence that reappraisal-focused strategies increase resilience trait among participants, which is consistent with the hypothesis that the program is affecting that skill.\(^7\)

**Behavioral Corollaries**

In behavioral economics literature, even minor mood manipulations have a substantial impact on emotions and behavior (DellaVigna, 2009). For instance, individuals have been found to tip more at restaurants on sunnier days (DellaVigna, 2009). International soccer matches impact the daily stock returns for the losing country by 0.21 percent (Edmans et al., 2007), however there is no significant impact when a country wins. World Cup elimination games have larger effects. Whether or not the loss was expected does not affect the results. On a different topic, Dahl and DellaVigna (2009) estimate the short-run impact of exposure to media violence on violent crime, that can be phrased as a negative emotional stimulus. They find that violent crime is lower on days in which exposure to media violence is higher, which seems counterintuitive with respect to the laboratory evidence. Exposure to movie violence can lower violent behavior relative to the foregone alternative activity — i.e. the field findings— even if it increases violent behavior relative to exposure to nonviolent movies - i.e. the laboratory findings.

In a related work, Au et al. (2003) estimate that the performance of financial market traders is affected by music-elicited emotional states. The author found that a good mood resulted in inferior performance and overconfidence, while a bad mood

\(^7\)See for example Troy and Mauss (2011)
resulted in more accurate decisions and more conservative trading. Complimentarily, according to Lerner et al. (2012), sadness increases impatience and creates a myopic focus on obtaining money immediately instead of later. In fact, relative to median neutral-state participants, median sad-state participants accepted 13% to 34% less money immediately to avoid waiting three months for payment.

In another example, Coricelli et al. (2010) finds that the risk of public exposure of deception deters evasion, while the amount of fines encourages evasion. The authors conclude that those results imply that an audit policy that strengthens the emotional dimension of cheating favors compliance (Coricelli et al., 2010). Kushlev et al. (2015), using data from a diverse cross section of the U.S. population (N=12,291), shows that higher incomes are associated with experiencing less daily sadness, but have no effect on daily happiness.

A explanation for the asymmetric impact on only the negative reaction is consistent with Querengsser and Schindler (2014)’s findings. In particular, the authors refers to Nesse (1990), who argue that “Emotional states not only motivate action, they are also goals that we seek to achieve. Most human thought, plans, and actions are intended to induce positive emotions or to avoid negative emotions” (Nesse, 1990, p. 262). From this evolutionary point of view, a successful induction of negative emotion would be more relevant for participants’ behavior because negative emotions suggest a situation that should be altered, while positive emotions indicate situations that should be maintained (Nesse, 1990; Querengsser and Schindler, 2014).

Such evidence, along with related lines of work, have contributed to the conclusion that emotion is not epiphenomenal and can influence cognition and behavior in powerful ways.77

---

77 For reviews, see Loewenstein et al. (1992, 2001); Damasio (1994); Weber and Johnson (2009); Lempert and Phelps (2014); Lerner et al. (2015).
Finally, the literature has been emphatic arguing that it is crucial to be careful about directly comparing the results of laboratory and field studies (DellaVigna, 2009; Levitt and List, 2007).

### 2.6 Conclusions

This study contributes to the understanding of channels that could be observed—or qualitatively claimed—relative to social programs that foster socio-emotional skills among youth. Consistent with previous studies (e.g. Attanasio et al., 2015a; Calero et al., 2014; Card et al., 2015; West et al., 2015), the Mining’s Rockstars participants experience significant impacts on educational outcomes—i.e. dropouts and SAT-like registration rates—, yet no short-run impact among in the expected mechanism, such as perseverance (grit) or internal locus of control. Though evidence of interventions targeted at shaping non-cognitive skills in youth—and thus relieve poverty or improve anti-social behavior—is weak and scarce (Heckman and Kautz, 2012; Hill et al., 2011; Blattman et al., 2015), this dissertation shows that social programs designed to foster non-cognitive skills are likely to affect emotional regulation. In other words, emotional regulation can be the mechanism explaining the educational and/or labor market outcomes that are typically observed on assessment of social programs.

I argue that programs to enhance non-cognitive skills likely impact dimensions that are too complex to measure through traditional self-reported psychometric tests. In Chapter 2, I claim that emotions are correlated with psychometric test scores typically used to proxy non-cognitive skills. This chapter explores the relevance of developing emotional regulation represented by pre-test arousal and valence, as well as emotional response to both positive and negative stimuli. It is worth noting that is hard to disentangle the non-existent impact on non-cognitive skills with the impact
on emotions. The program is likely affecting emotion and self-reported tests scores through changes in emotion. In fact, besides the positive qualitative evidence in favor of change of non-cognitive skills, it is not possible to correctly separates these effects. However, there are interesting results.

The neurophysiological measures in the field experiment contribute to two dimensions. First, they make it plausible to hypothesize about the lack of impact of social programs on life skills. This lack of impact is consistent with the decrease in the emotional state among participants which is a direct corollary to the positive correlation between emotional state and test scores on non-cognitive tests (Querengssser and Schindler, 2014; Egana-delSol, 2016b). In other words, the increase in a withdrawal behavior —i.e. less positive valence— biases downward self-reported measures of non-cognitive skills, as showed in Chapter 2. In fact, the program might affect the self-perception of subjects at the moment they are answering self-reported tests. There is a behavioral response by students themselves. For instance, a program that encourages grit and locus control in individuals might not change their inherent skills, but may affect their emotional responsiveness to their own environment. It is possible that the treated students become more exigent with themselves than the non treated individuals, in addition to the positive effect that treatment has on their skills.

Second, I find a significant decrease in the neurophysiological reaction to a negative stimulus on the participant, which is a proxy of emotional/behavioral response. There is no significant impact with respect to positive stimuli. These behavioral findings are consistent with the asymmetric impact on emotional responsiveness (DellaVigna, 2009). For instance, higher income is associated with reduced daily sadness but not increased daily happiness (Kushlev et al., 2015). I highlight the relevance of emotional disposition and modulation as a key mechanism to overcome “Psychological Poverty Traps” (Haushofer and Fehr, 2014). According to Haushofer and Fehr (2014) poverty causes stress and negative affective states, which may lead to short-sighted
and risk-averse decision-making; possibly by, limiting attention and favoring habitual behaviors at the expense of goal-directed ones. This notion of poverty is consistent with recent results in reviews from OECD (2015) and WorldBank (2015).\footnote{In particular in the reports “Skills for Social Progress: The Power of Social and Emotional Skills” and “The World Development Report: Mind, Society, and Behavior”, respectively.}

Moreover, the impacts on emotional responses that I find are likely due to the reappraisal-focused strategies in the program methodologies. Indeed, these strategies have been identified as the most effective way to avoid emotional bias on decision making (Lerner et al., 2015).

The methodology proposed in this study is an alternative to the most frequently used approach in the labor economics literature: latent factor models and the use of revealed behaviors as proxies. The use of latent factor models increases the reliability of skills proxies, but does not solve the problem that the program is potentially affecting other unobservable dimensions such as emotional regulation, and thus self-reported tests. Furthermore, the use of revealed behaviors as proxies, in conjunction with information about past behaviors, is tautological, since it explains behaviors using them as proxies for non-cognitive skills (Heckman and Kautz, 2012).

Moving away from behavioral-based or self-reported measures of non-cognitive skills is a challenging topic for further applications and research. The methodology proposed in this paper has many benefits. First, it offers a way to incorporate emotion into the labor economics field. The importance of emotional regulation — i.e. emotional stability — in labor markets and overall life satisfaction has recently been highlighted for both developed and developing countries (Deming, 2015; OECD, 2015). This study shows that there are neurophysiological approaches to proxy emotional disposition and responsiveness with a high level of accuracy, and at a relatively low cost.

The results also may aid evaluation of similar programs attempting to foster non-
cognitive skills. In fact, the effects on emotional regulation —i.e. emotional state and responsiveness— also have implications for the experimental evaluation of educational interventions. Social programs for education and Active Labor Market Policies (ALMP) usually aim to impact socio-emotional or life-skills, such as perseverance, self-control, goal-oriented effort, and so forth. However, these factors suffer from measurement error —i.e. reference or emotional bias— due to self-reporting. The methodology proposed here allows us to measure emotional disposition and responsiveness from EEG recordings, which is a non-invasive and low-cost method. Therefore, further research could incorporate physiological measures of emotional state and modulation to study the human capital production function, educational interventions, and the ALMP effectiveness.
2.7 Appendix 1: Educational Outcomes and EEG features

Educational Outcomes and EEG features

This section provides empirical evidence showing that educational outcomes correlate with EEG features—i.e. valence and arousal indices. This provides support to the claim that emotional regulation, which is measured throughout EEG recordings, is a likely mechanism through which the program is affecting educational outcomes.

My analysis is focused on the registration to an SAT-like test as a main outcome, because it was not possible to get adequate administrative data from the Ministry of Education about dropout rates due legal privacy constraints.\(^79\) In order to support that EEG features and registration to SAT-like test are empirically related, I offer the following results:

1. There is no statistically significant linear correlation between EEG features and registration to SAT-like test. Pair-wise correlation and ANOVA tests fail to reject the null hypothesis of no significant correlation. This result is driven by a non-linear relationship between EEG features and outcomes. Finally, the presence of outliers, the presence of which are also a crucial caveat on ANOVA testing, is insinuated from the box plot presented in Figure 2.7 below.

2. The visual inspection of Figure 2.8 that contains the distribution of different EEG features by registration to SAT-like test suggests that there is a change in

\(^79\)It is not possible to use dropout rates because there are not enough observations to compare treatments and time: there are no observations for individuals who dropped out in follow up measures because I collected data directly in the schools, which would not have included students who were not present because they had dropped out.
the distribution, which may include second—i.e. variance— and higher moments, such as skewness and kurtosis.80

3. The null hypothesis of equal variance –i.e. the variance ratio test– is rejected for different EEG indices; in particular, valence of negative and positive variables (p-value<0.05) among participants, while valence positive and arousal at baseline (no stimuli) when considering the pooling sample with participants and controls. See the tables below.

4. Figure 2.9 below shows that the program is impacting not only the mean on some EEG features (See Results section), but also the variance of the distribution of EEG measures. Furthermore, as Tables 2.9 - 2.14 show, variance ratio tests are also rejected (null hypothesis equal variance) for different EEG features, which indicates that the program has an effect on the EEG features’ variance. These results provide support to the claim that the program is impacting outcomes, and that one of the mechanisms is affecting EEG features of emotional skills.

5. Probit models on the probability to be registered based on EEG features indicate that the EEG measures add information to the model.81 As Table 2.8 shows, EEG features are not linearly related to the outcome. This result can be due the non-linearity in the relationship between EEG features and educational outcomes. In fact, the Models (2) and (3) on Table 2.8 show significant marginal effects for the quadratic expression of EEG features; in particular, valence at baseline (no stimuli) and when facing negative stimulus. Moreover both Akaike Information Criteria and Bayesian Information Criteria support the argument that EEG features are adding information to the model, as Model (3), which

80 See Cameron and Trivedi (2005) for details.
81 See Cameron and Trivedi (2005) for details on Probit Models.
includes only significant variables in order to minimize the penalization for over-fitting, informs.\textsuperscript{82} These results are consistent with the non-linear relation between EEG features and outcomes that I have been suggesting exist.

In summary, the data suggests that there is an empirical relation between EEG features based on emotion-detection theories and educational outcomes. For the purpose of this study, the non-linear relation as well as in second moments that has been shown are sufficient to argue that emotional regulation and responsiveness, i.e. emotional resilience, is a possible mechanism for explaining the impact of programs that aim to affect socio-emotional skills. Moreover, together with the evidence presented in Chapter 3, the results support the necessity to incorporate measures of socio-emotional skills that are not self-reported, especially in the context of an intervention.

Future work can advance the understanding of the non-linear relation between EEG features and the outcomes.

\textsuperscript{82}See Cameron and Trivedi (2005) for details on model selection based on information criteria.
Figure 2.7: Box Plots: EEG features and Educational Outcome
Figure 2.8: Distribution of EEG Features by Registration to SAT-like test (PSU) status

- Arousal at Baseline (no stimuli) Distribution by SAT-ile Registration Status
- Valence at Baseline (no stimuli) Distribution by SAT-ile Registration Status
- Valence under Positive Stimulus Distribution by SAT-ile Registration Status
- Valence under Negative Stimulus Distribution by SAT-ile Registration Status
- Valence Difference of Positive and Baseline Measures by SAT-ile Registration Status
- Valence Difference of Negative and Baseline Measures by SAT-ile Registration Status
Figure 2.9: Distribution of EEG Features by Treatment status

Arousal at Baseline (no stimuli) Distribution by Treatment Status

Valence at Baseline (no stimuli) Distribution by Treatment Status

Valence under Positive Stimulus Distribution by Treatment Status

Valence under Negative Stimulus Distribution by Treatment Status

Valence Difference of Positive and Baseline Measures by Treatment Status

Valence Difference of Negative and Baseline Measures by Treatment Status
Table 2.8: Probit Model: Registration to SAT-like test (PSU) and EEG Features

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Registered for PSU (=1)</th>
<th>(2) Registered for PSU (=1)</th>
<th>(3) Registered for PSU (=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'La Providencia' School</td>
<td>0.298*** (0.0588)</td>
<td>0.312*** (0.0453)</td>
<td>0.306*** (0.0502)</td>
</tr>
<tr>
<td>Arousal RestState</td>
<td>0.0123 (0.0322)</td>
<td>0.0183 (0.0347)</td>
<td></td>
</tr>
<tr>
<td>Valence RestState</td>
<td>0.0289 (0.0582)</td>
<td>0.0199 (0.0653)</td>
<td></td>
</tr>
<tr>
<td>difvalenPOS</td>
<td>-0.00451 (0.0300)</td>
<td>-0.0198 (0.0343)</td>
<td></td>
</tr>
<tr>
<td>difvalenNEG</td>
<td>0.00457 (0.0380)</td>
<td>0.0230 (0.0383)</td>
<td></td>
</tr>
<tr>
<td>zarousal_base2</td>
<td></td>
<td>0.0132 (0.0120)</td>
<td></td>
</tr>
<tr>
<td>zvalen_base2</td>
<td></td>
<td>0.122*** (0.0458)</td>
<td>0.0898*** (0.0345)</td>
</tr>
<tr>
<td>zvalen_pos2</td>
<td></td>
<td>0.0324* (0.0195)</td>
<td></td>
</tr>
<tr>
<td>zvalen_neg2</td>
<td></td>
<td>0.0820** (0.0337)</td>
<td>0.0696** (0.0290)</td>
</tr>
<tr>
<td>difvalenPOS2</td>
<td></td>
<td>-0.0134 (0.0165)</td>
<td></td>
</tr>
<tr>
<td>difvalenNEG2</td>
<td>-0.0941*** (0.0322)</td>
<td>-0.0770*** (0.0276)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>201</td>
<td>201</td>
<td>201</td>
</tr>
<tr>
<td>School FE</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>AIC</td>
<td>253</td>
<td>255</td>
<td>246</td>
</tr>
<tr>
<td>BIC</td>
<td>272.9</td>
<td>294.7</td>
<td>262.5</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Legal restrictions on privacy issues from the DEMRE-University of Chile in order to merge the data set of this study with their administrative data made it possible to only keep identified the non-municipal school in our sample.

Program’s Impact on EEG variance

As was shown in the Results section, the program impacted EEG features of emotions; in particular, valence and arousal at the baseline, and the valence from the negative stimulus net of baseline measure.

In order to be consistent with respect to the previous evidence that the variance of the EEG measures correlates with educational outcomes, I provide evidence that
the program is also impacting the variance of the EEG measures. Therefore, emotional regulation is a possible mechanism behind the program’s estimated impact on educational outcomes.

In particular, as Tables 2.9 to 2.14 show, the variance ratio test between treated and control group is rejected at 90% of significance (p-value<0.1) for valence positive and negative indices. All in all, this empirical evidence supports the relationship between emotional regulation and educational outcomes.

Table 2.9: Variance Ratio Test: Arousal Baseline (no stimuli) by Treatment Status

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>-0.0427197</td>
<td>0.2963482</td>
<td>1.257299</td>
<td>-0.6679596 0.5825203</td>
</tr>
<tr>
<td>Treated</td>
<td>64</td>
<td>-0.2502685</td>
<td>0.1309358</td>
<td>1.047486</td>
<td>-0.5119227 0.0113858</td>
</tr>
<tr>
<td>combined</td>
<td>82</td>
<td>-0.204709</td>
<td>0.1206003</td>
<td>1.092082</td>
<td>-0.4446657 0.0352478</td>
</tr>
</tbody>
</table>

\[
\text{ratio} = \frac{\text{sd(Control)}}{\text{sd(Treated)}}
\]
\[
f = 1.4407
\]
Ho: ratio = 1
degrees of freedom = 17, 63
Ha: ratio < 1
Ha: ratio != 1
Ha: ratio > 1
\[
\begin{align*}
\text{Pr}(F < f) &= 0.8517 \\
2*\text{Pr}(F > f) &= 0.2965 \\
\text{Pr}(F > f) &= 0.1483
\end{align*}
\]

Note: Since it is intended to show that the program has an effects on the variance of the treated, we only consider subjects who were tested in the follow up survey.

Table 2.10: Variance Ratio Test: Valence Baseline (no stimuli) by Treatment Status

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>-0.0044791</td>
<td>0.270398</td>
<td>1.147202</td>
<td>-0.5749691 0.5660109</td>
</tr>
<tr>
<td>Treated</td>
<td>64</td>
<td>-0.2191908</td>
<td>0.1449804</td>
<td>1.159844</td>
<td>-0.5089111 0.0705296</td>
</tr>
<tr>
<td>combined</td>
<td>82</td>
<td>-0.1720589</td>
<td>0.12738</td>
<td>1.153475</td>
<td>-0.4255051 0.0813873</td>
</tr>
</tbody>
</table>

\[
\text{ratio} = \frac{\text{sd(Control)}}{\text{sd(Treated)}}
\]
\[
f = 0.9783
\]
Ho: ratio = 1
degrees of freedom = 17, 63
Ha: ratio < 1
Ha: ratio != 1
Ha: ratio > 1
\[
\begin{align*}
\text{Pr}(F < f) &= 0.5073 \\
2*\text{Pr}(F > f) &= 0.9854 \\
\text{Pr}(F > f) &= 0.4927
\end{align*}
\]

Note: Since it is intended to show that the program has an effects on the variance of the treated, we only consider subjects who were tested in the follow up survey.
Table 2.11: Variance Ratio Test: Valence Positive stimulus by Treatment Status

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>0.1948422</td>
<td>0.3095247</td>
<td>1.313202</td>
<td>-0.4581979 0.8478823</td>
</tr>
<tr>
<td>Treated</td>
<td>64</td>
<td>0.0815696</td>
<td>0.1282043</td>
<td>1.025634</td>
<td>-0.1746262 0.3377654</td>
</tr>
<tr>
<td>combined</td>
<td>82</td>
<td>0.1064343</td>
<td>0.1200773</td>
<td>1.087346</td>
<td>-0.1324818 0.3453505</td>
</tr>
</tbody>
</table>

\[ \text{ratio} = \frac{\text{sd(Control)}}{\text{sd(Treated)}} \]
\[ f = 1.6394 \]

Ho: ratio = 1  
Ha: ratio < 1  
Ha: ratio != 1  
Ha: ratio > 1

\[ \text{Pr}(F < f) = 0.9195 \]
\[ 2*\text{Pr}(F > f) = 0.1611 \]
\[ \text{Pr}(F > f) = 0.0805 \]

Note: Since it is intended to show that the program has an effects on the variance of the treated, we only consider subjects who were tested in the follow up survey.

Table 2.12: Variance Ratio Test: Valence Negative stimulus by Treatment Status

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>-0.1005964</td>
<td>0.0723581</td>
<td>0.3069896</td>
<td>-0.2532587 0.0520659</td>
</tr>
<tr>
<td>Treated</td>
<td>64</td>
<td>0.0456182</td>
<td>0.1176076</td>
<td>0.9408608</td>
<td>-0.1894018 0.2806382</td>
</tr>
<tr>
<td>combined</td>
<td>82</td>
<td>0.0135223</td>
<td>0.0931816</td>
<td>0.8437956</td>
<td>-0.1718799 0.1989245</td>
</tr>
</tbody>
</table>

\[ \text{ratio} = \frac{\text{sd(Control)}}{\text{sd(Treated)}} \]
\[ f = 0.1065 \]

Ho: ratio = 1  
Ha: ratio < 1  
Ha: ratio != 1  
Ha: ratio > 1

\[ \text{Pr}(F < f) = 0.0000 \]
\[ 2*\text{Pr}(F > f) = 0.0000 \]
\[ \text{Pr}(F > f) = 1.0000 \]

Note: Since it is intended to show that the program has an effects on the variance of the treated, we only consider subjects who were tested in the follow up survey.
Table 2.13: Variance Ratio Test: Difference Valence Positive and Baseline by Treatment Status

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>-0.1993213</td>
<td>0.3198542</td>
<td>1.357026</td>
<td>-0.8741546 0.475512</td>
</tr>
<tr>
<td>Treated</td>
<td>64</td>
<td>-0.3007603</td>
<td>0.1896044</td>
<td>1.516835</td>
<td>-0.6796545 0.0781338</td>
</tr>
<tr>
<td>combined</td>
<td>82</td>
<td>-0.2784932</td>
<td>0.1629671</td>
<td>1.47573</td>
<td>-0.6027467 0.0457602</td>
</tr>
</tbody>
</table>

\[ \text{ratio} = \frac{\text{sd(Control)}}{\text{sd(Treated)}} \]
\[ f = 0.8004 \]
\[ \text{Ho: ratio} = 1 \]
\[ \text{degrees of freedom} = 17, 63 \]

<table>
<thead>
<tr>
<th>Ha: ratio &lt; 1</th>
<th>Ha: ratio != 1</th>
<th>Ha: ratio &gt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(F &lt; f) = 0.3138</td>
<td>2*Pr(F &gt; f) = 0.6277</td>
<td>Pr(F &gt; f) = 0.6862</td>
</tr>
</tbody>
</table>

Note: Since it is intended to show that the program has an effects on the variance of the treated, we only consider subjects who were tested in the follow up survey.

Table 2.14: Variance Ratio Test: Difference Valence Negative and Baseline by Treatment Status

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>18</td>
<td>0.0961173</td>
<td>0.2905135</td>
<td>1.232544</td>
<td>-0.5168126 0.7090472</td>
</tr>
<tr>
<td>Treated</td>
<td>64</td>
<td>-0.264809</td>
<td>0.185047</td>
<td>1.480376</td>
<td>-0.6345959 0.1049779</td>
</tr>
<tr>
<td>combined</td>
<td>82</td>
<td>-0.1855813</td>
<td>0.1579573</td>
<td>1.430364</td>
<td>-0.4998667 0.1287042</td>
</tr>
</tbody>
</table>

\[ \text{ratio} = \frac{\text{sd(Control)}}{\text{sd(Treated)}} \]
\[ f = 0.6932 \]
\[ \text{Ho: ratio} = 1 \]
\[ \text{degrees of freedom} = 17, 63 \]

<table>
<thead>
<tr>
<th>Ha: ratio &lt; 1</th>
<th>Ha: ratio != 1</th>
<th>Ha: ratio &gt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(F &lt; f) = 0.2024</td>
<td>2*Pr(F &gt; f) = 0.4048</td>
<td>Pr(F &gt; f) = 0.7976</td>
</tr>
</tbody>
</table>

Note: Since it is intended to show that the program has an effects on the variance of the treated, we only consider subjects who were tested in the follow up survey.
## 2.8 Appendix 2: Balance of Attrition

Table 2.15: Balance of Attrition (missed subjects compared at baseline)

<table>
<thead>
<tr>
<th>Orthogonality Table</th>
<th>Control</th>
<th>Treated</th>
<th>(1) vs. (2)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus Control Test</td>
<td>-0.016</td>
<td>0.270</td>
<td>-0.286</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>(0.113)</td>
<td>(0.117)</td>
<td>(0.167)</td>
<td></td>
</tr>
<tr>
<td>Grit Test</td>
<td>0.053</td>
<td>0.312</td>
<td>-0.260</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>(0.115)</td>
<td>(0.110)</td>
<td>(0.165)</td>
<td></td>
</tr>
<tr>
<td>Creativity Index</td>
<td>0.111</td>
<td>-0.002</td>
<td>0.113</td>
<td>0.594</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
<td>(0.162)</td>
<td>(0.211)</td>
<td></td>
</tr>
<tr>
<td>13 Forces Creativity</td>
<td>-0.001</td>
<td>-0.095</td>
<td>0.094</td>
<td>0.641</td>
</tr>
<tr>
<td></td>
<td>(0.131)</td>
<td>(0.152)</td>
<td>(0.202)</td>
<td></td>
</tr>
<tr>
<td>Standardized values of (valen_neg)</td>
<td>-0.016</td>
<td>0.132</td>
<td>-0.148</td>
<td>0.506</td>
</tr>
<tr>
<td></td>
<td>(0.160)</td>
<td>(0.131)</td>
<td>(0.221)</td>
<td></td>
</tr>
<tr>
<td>Standardized values of (valen_pos)</td>
<td>-0.031</td>
<td>0.053</td>
<td>-0.084</td>
<td>0.657</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
<td>(0.114)</td>
<td>(0.188)</td>
<td></td>
</tr>
<tr>
<td>Standardized values of (valen_base)</td>
<td>0.189</td>
<td>0.061</td>
<td>0.128</td>
<td>0.401</td>
</tr>
<tr>
<td></td>
<td>(0.199)</td>
<td>(0.091)</td>
<td>(0.152)</td>
<td></td>
</tr>
<tr>
<td>Standardized values of (arousal_base)</td>
<td>-0.088</td>
<td>0.361</td>
<td>-0.449</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>(0.127)</td>
<td>(0.177)</td>
<td>(0.212)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>66</td>
<td>46</td>
<td>112</td>
<td></td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.05.
2.9 Appendix 3: Robustness Checks

Robustness Check: Impact on Cognitive Skills (Raven-like Test)

Table 2.16 presents the program’s impact on cognitive skills based on the Raven’s progressive matrices test, which capture logical and problem-solving skills through guessing the missing matrix between many Raven’s-like progressive matrices. The Mining’s Rockstars program is not designed to affect cognitive skills, thus the test was incorporated only in the follow-up survey as a placebo test. Regardless, since the program is affecting emotional state and responsiveness, we could expect an indirect effect, but as Table 2.16 shows, this is not the case here either.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raven Test</td>
<td>Grade Avg.</td>
</tr>
<tr>
<td>Treat &amp; Post.</td>
<td>-0.0530 (0.0525)</td>
<td>0.193 (0.153)</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.127 (0.154)</td>
<td></td>
</tr>
<tr>
<td>Post.</td>
<td>-0.108 (0.111)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.554*** (0.0705)</td>
<td>5.805*** (0.102)</td>
</tr>
</tbody>
</table>

Observations 116 166
R-squared 0.102 0.196
School Dummies Yes Yes
Schools’s Class Cluster Yes Yes

Robust p-values in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Notes: “Grade Avg.” is a variable that accounts for the average grade obtained the year before the baseline was conducted.
Robustness Check: In Difference Model: Impact on Emotional Regulation

Table 2.17 exposes the program’s impact following a Model in differences (Cameron and Trivedi, 2005). The model considers two novel outcomes attempting to capture the unobservable impact on emotional regulation on a subsample that considers only those individuals that did both baseline and follow up field experiments—i.e. the balanced panel. This model allows to clean all time-invariant non-observable characteristics on both participant and their control group. In particular, the following model is estimated:

\[ \Delta E_{j,i} = \delta_1 * treat_{i} * Post_{t} + \theta_1 * X_{i} + \varepsilon_{1i,t} \]  

(2.14)

where \( E_{j,i,t} \) indicates the \( j \) emotional state or emotional responsiveness associated with an emotionally laden stimuli of individual \( i \) in time \( t \), with \( j = arousal_{base}, valence_{base}, valence_{positive}, valence_{neg}, dif - valence_{pos}, dif - valence_{neg}, \)
\( i = students \in FieldExperiment, \) and \( t = baseline, follow - up. \) As before, \( treat_{j,i} \) and \( Post_{j,t} \) indices who had been treated by the program and survey time—i.e. baseline or follow up—respectively. Finally, \( X_{i,t} \) indicates school level dummy variables. It is expected that \( \delta, \delta_1 < 0 \) and \( \beta = \gamma = 0. \)
Table 2.17: Impact on Emotional Regulation: Emotional State and Responsiveness

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ArouRest</td>
<td>ValRest</td>
<td>ValNeg</td>
<td>ValPos</td>
<td>ValNegDif.</td>
<td>ValPosDif.</td>
<td>Delta_VaPosDif.</td>
<td>Delta_VaNegDi</td>
</tr>
<tr>
<td>Treat &amp; Post.</td>
<td>-0.591**</td>
<td>-1.350*</td>
<td>0.649</td>
<td>-0.0524</td>
<td>-1.999**</td>
<td>-1.298</td>
<td>-2.932</td>
<td>-3.383*</td>
</tr>
<tr>
<td></td>
<td>(0.233)</td>
<td>(0.802)</td>
<td>(0.636)</td>
<td>(0.802)</td>
<td>(0.954)</td>
<td>(1.135)</td>
<td>(2.679)</td>
<td>(2.011)</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.0669</td>
<td>0.129</td>
<td>0.0130</td>
<td>0.597</td>
<td>0.116</td>
<td>-0.468</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.196)</td>
<td>(0.436)</td>
<td>(0.592)</td>
<td>(0.571)</td>
<td>(0.669)</td>
<td>(0.695)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post.</td>
<td>0.00761</td>
<td>0.255</td>
<td>-1.079**</td>
<td>0.468</td>
<td>1.335*</td>
<td>-0.213</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.155)</td>
<td>(0.626)</td>
<td>(0.470)</td>
<td>(0.568)</td>
<td>(0.736)</td>
<td>(0.804)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.0639</td>
<td>0.191</td>
<td>-0.219</td>
<td>-1.319</td>
<td>0.410</td>
<td>1.510</td>
<td>-0.831</td>
<td>3.493</td>
</tr>
<tr>
<td></td>
<td>(0.190)</td>
<td>(0.588)</td>
<td>(0.774)</td>
<td>(0.814)</td>
<td>(0.857)</td>
<td>(0.924)</td>
<td>(2.282)</td>
<td>(2.798)</td>
</tr>
<tr>
<td>Observations</td>
<td>296</td>
<td>296</td>
<td>296</td>
<td>296</td>
<td>296</td>
<td>296</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.059</td>
<td>0.024</td>
<td>0.038</td>
<td>0.016</td>
<td>0.034</td>
<td>0.027</td>
<td>0.102</td>
<td>0.166</td>
</tr>
<tr>
<td>Student Level Clust.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>School FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1.
This page intentionally left blank
Chapter 3

How Much Should We Trust

Self-reported Measures of

Non-cognitive Skills?

3.1 Introduction

The labor force stock of skills is usually referred to as human capital. There is a long tradition of studying the relationship between human capital and economic outcomes, such as economic growth, salaries, or school attendance (Bowles and Gintis, 1976; Heckman, 2007; Bowles et al., 2001; WorldBank, 2015).

At the individual level, the economic literature argues that both cognitive skills — understood as the use of language, memory, and logical-mathematical reasoning — and non-cognitive skills — also called socio-emotional skills or personality traits, which include self-concept, self-control, perseverance, and motivation, among others — are similarly relevant in determining personal success in the workplace.¹

¹For example, measures of human capital at ages 6-8, can explain 20 percent of the variation in adult wages (Currie and Thomas, 1999).
Measures of cognitive skills—e.g., GPA, IQ—account for a small fraction of the variance in salaries and other economic outcomes (Bowles et al., 2001; Heckman et al., 2006). Therefore, there is an increasing interest in elucidating other factors that might explain that variance; in particular, the role played by non-cognitive skills. For instance, Deming (2015) shows that the labor market increasingly rewards social skills. In fact, employment and wage growth has been strongest in jobs that require high levels of both cognitive and social skills. In the economics literature, empirical attempts to measure the aforementioned dimensions—cognition, personality traits, and creativity—have been insightful but are noisy due to the low reliability of the proxies used in their measurement (Calero et al., 2014; Cunha and Heckman, 2008; Cunha et al., 2010; Attanasio et al., 2015a; Almlund et al., 2011). For example, Cunha and Heckman (2008) point out that the measurement error is high in all the proxies used to measure cognitive and non-cognitive skills, and also that slight changes to those measures substantially affect the direction and significance of the findings. In fact, a test’s setting (ONeil et al., 2014), reference bias (Kautz et al., 2014), and emotional or motivational bias (Plucker and Makel, 2010) are among the main caveats of self-reported measures of non-cognitive skills. For social-program evaluation, we are generally interested in single-trial, individual-level assessment of skills. Therefore, a test’s setting can be easily kept constant when comparing a group of individuals in similar circumstances. Moreover, due to its social nature, it is not trivial to control for reference bias, especially on individual-level assessments (West et al. (2015); Almlund et al. (2011); Heckman and Kautz (2012)).

The relationship between emotional and motivational states and behavior and socio-economic outcomes has been widely studied. Scientists who combine neuro-

---

2 see Chapter 1.

3 There is recent progress on strategies that attempt to palliate reference bias on self-reporting, such as the use of anchoring vignettes. See, for example, King and Hopkins (2010).

4 For recent reviews, see Weber and Johnson (2009); Lerner et al. (2015).
science, behavioral science, and economics also support the notion that emotions have a strong influence on economic behavior and decision-making, as well as on labor-market performance—i.e. occupational choice, salaries, entrepreneurship skills, etc.5 Furthermore, evidence regarding how emotions modulate self-reporting on non-cognitive tests, which are broadly used in the literature, is scarce. A recent study done by Querengsser and Schindler (2014) explores that relationship. In particular, the authors relate self-reported measures of emotional state to self-reporting on a well known non-cognitive test, and conclude that only negative affects are associated with lower self-reports in certain test dimensions.6 Consequently, this chapter seeks to explore the relationship between emotional and motivational states and self-reported measures of non-cognitive skills using objective measures.

Emotion-detection research has employed an array of physiological measurements and methods for the past decade, including pupil dilation, heart rate, skin conductance, and voice and facial manifestations (Takahashi et al., 2004; Partala et al., 2000). However, some of those measurements can be consciously modified, rendering the signal subjective. Low-cost, portable electroencephalogram (EEG) devices, which measure the scalp’s brain activity, yield improved accuracy and objectiveness as compared to the aforementioned physiological measurements. Recent studies offer a variety of EEG-based emotion-detection methods (Choppin, 2000; Takahashi et al., 2004; Lin et al., 2010; Ramirez and Vamvakousis, 2012; Kim et al., 2013). The Ramirez and Vamvakousis (2012) version of James-Lange’s arousal-valence model is followed as the framework for emotions.

I argue that there is a positive correlation between emotional state and test scores that rely on self-ratings. In particular, I claim that self-reported tests used to mea-

5See, for example Weber and Johnson (2009); Lerner et al. (2015); DellaVigna (2009); Loewenstein (2000).

6See next section for details.
sure both cognitive and non-cognitive skills are usually biased because of transient emotions that arise during testing. For instance, if two individuals have the same ex-post score on a given psychometric test, but one felt pleased during the test, and the other felt unhappy, the former’s score will overestimate skills, while the latter’s will underestimate them.

The development of this chapter is intrinsically related to the impact evaluation done in Chapter 1, and was designed as a deeper proof of concept of the relationship between emotions and self-reporting on measures of non-cognitive skills. The purpose is to show a plausible correlation between self-reported psychometric tests and transient emotional states, with the latter estimated from EEG recordings. To do so, a couple of relationships must be established: namely, (i) showing the behavioral correlation between self-reported psychometric tests and transient emotional states, (ii) detecting emotional state and responsiveness from scalp EEG recordings, and (iii) showing the correlation between self-reported psychometric tests and transient emotional states using features from the EEG recordings.

In order to test my hypothesis, I designed and implemented two experiments. The first studied the behavioral relationship between elicited emotions—positive and negative—and self-reported measures of non-cognitive skills, by inducing emotional states using emotionally-laden stimuli and psychometric tests. The resulting behavioral relationship was not mediated by EEG recordings or self-reporting. The study was done on a subsample of the population that participated in the program evaluation conducted in Chapter 1 of this dissertation.

The second experiment studied the possibility of detecting emotions from low-cost EEG devices and evaluated the correlation between transient emotions and self-

---

7 See details in next section.
8 Thereby differentiating this study from the experiment of Querengsser and Schindler (2014), which included self-reporting of emotions.
reporting on psychometric tests. This experiment induced emotional states through images from the GAPED database\(^9\) while recording brain activity. Studies have shown that emotions can be proxied from lab-quality EEG recordings (Choppin, 2000; Takahashi et al., 2004; Lin et al., 2010; Ramirez and Vamvakousis, 2012; Kim et al., 2013). In a study comparing the success rates of Brain-Computer Interface (BCI) systems, Martinez-Leon et al. (2016) also found higher precision and less variance in data captured with the low-cost Emotiv EPOC headset used in this study, as compared to that of a professional system (Biosemi Active II). In the experiment presented here, I found an emotion-detection accuracy of 79%, which is within the literature range of 70-85% accuracy. In addition, using the same experiment, I estimated the relationship between EEG-derived and self-reported emotional states. In particular, I estimated the correlation between emotional-state indices—arousal and valence—from EEG recordings and those from the self-rating in common psychometric tests. I found a positive correlation, which is significant at an 80% significance level. Combining these results—the behavioral correlation, the reliability of low-cost EEGs for emotion-detection, and the correlation between emotions proxied by the EEG and self-reporting—I support the empirical methodology used in Chapter 1. In fact, I was able to estimate emotional state and responsiveness in the field—i.e. in public schools in Chile—and to reveal the successful impact of participation in social programs aimed at fostering non-cognitive skills.\(^{10}\)

The main contribution of this study is the use of methods from affective neuroscience and neurophysiological recordings to proxy emotional state, allowing us to estimate an empirical relationship with self-reporting on psychometric tests, as well as to have quantitative measures by which to evaluate social programs.\(^{11}\)

\(^{9}\)See next section for details.
\(^{10}\) See Chapter 1 for details.
\(^{11}\) Please note that the literature has been emphatic in arguing that it is not possible to estimate or predict personality from resting-state EEG recordings directly (Korjus et al., 2015).
The paper is organized as follows: Section 2 describes theories of emotions, their implications on behavior, and their indices to be measured; Section 3 defines the experimental methods; Section 4 explains the experimental paradigm; Section 5 presents the results; and Section 6 presents the conclusions.

3.2 Conceptual Frameworks

Skill Measurements

The related economic literature has followed a quantitative approach to both cognitive and non-cognitive assessments (Almlund et al., 2011). Without loss of generality, I focus here on the measurements and methodologies regarding non-cognitive skills. The typical strategy for assessing non-cognitive skills is to apply a battery of tests —i.e. Grit Scale, Locus of Control Scale, Self-Esteem Scale, and so forth—that are likely to proxy a lower—and common—dimension of a non-cognitive skill. In particular, latent-factor models allow the measurement of latent —i.e. underlying—factors representing non-cognitive skills or their distributions (Cunha et al., 2010). Below, I present a schematic model of skill measurement.\footnote{For skills definition and dimensions see Chapter 1.}

Students’ skills are represented by $\theta_{i,t}^k$, with $i$ identifying individuals, $t$ indexing time, and $k = C, S, I$ indexing cognitive, socio-emotional, and innovative skills, respectively. Here, I follow Cunha et al. (2010) in using latent-factor models and the joint distribution of error-ridden latent factors measuring students’ skills at baseline ($\theta_{i,t}^C$, $\theta_{i,t}^S$ and $\theta_{i,t}^I$), students’ skills at follow-up ($\theta_{i,t+1}^C$, $\theta_{i,t+1}^S$ and $\theta_{i,t+1}^I$), parents’ skills ($P_{i,t}^C$, $P_{i,t}^S$) and parental investments ($I_{i,\tau}^\tau$ with $\tau \in \{\text{Material, Time}\}$).

Let $M_{i,k,t}^1$ represent measures of student $i$’s ($i = 1, \ldots, I$) skills of type $k$ ($k \in \{C, S, I\}$) at time $t$. Further, $M_{i,k}^2$ represents measures of parental skill of type $k$.\footnote{For skills definition and dimensions see Chapter 1.}
(assumed time invariant in this study), while \( M_{i,\tau}^3 \) represents parental investment of type \( \tau (\tau \in \{\text{Material, Time}\}) \) at time \( t \).

Therefore, we can represent \( M_{i,k,t}^n \) as:

\[
M_{i,k,t}^n = h_k(\theta_{ik}, \varepsilon_{ik}^n) = \mu_k^n + \lambda_k^n \log(\theta_{ik}) + \varepsilon_{ik}^n
\]

where terms \( \mu_k^n \) are intercepts, \( \lambda_k^n \), are factor loadings, and the terms \( \varepsilon_{ik}^n \) are measurement error for \( n = 1, 2, 3; k \in \{C,S,I\}; i = 1, ..., I \).

Then, let \( \theta_{i,k,t}^k \), be the vector of \( K \) different skills —i.e. factors or traits— for person \( i \) at time \( t \). In particular, \( \theta_{i,k,t}^k = (\theta_{ik,\theta_{ik},\theta_{i1}}) = (\theta_{ik}\theta_{i1}) \), where \( !k \) accounts for all but the \( k^{th} \) trait. Furthermore, \( m_{k,t,j}^1 \) indexes the \( j^{th} \) measure of a student’s skill of type \( k \) at time \( t \), \( m_{k,t,j}^2 \) indexes the \( j^{th} \) measure of a parent’s skill of type \( k \), and \( m_{k,t,j}^3 \) indices the \( j^{th} \) measure of parental investment of type \( \tau \).

Now, assuming separability of each measure in the latent factor it proxies, it is possible to express \( M \) as a linear (log) factor model\(^{13}\):

\[
m_{i,k,j,t}^1 = \mu_{i,k,j,t}^1 + \lambda_{i,k,j,t}^1 \log(\theta_{i,t}^k) + \varepsilon_{i,k,j,t}^1
\]

\[
m_{i,k,j}^2 = \mu_{i,k,j}^2 + \lambda_{i,k,j}^2 \log(P^k) + \varepsilon_{i,k,j}^2
\]

\[
m_{\tau,i,t,j}^3 = \mu_{\tau,i,t,j}^3 + \lambda_{\tau,i,t,j}^3 \log(I_{i,t}^\tau) + \varepsilon_{\tau,i,t,j}^3
\]

where independence is assumed between \( \theta_{i,k} \) and \( \varepsilon_{ik}^n \), and \( m_{i,k,t,j}^n \) is not affected by \( !k \neq k \). Note that the latent factors can be freely correlated with one other.

An important specificity of our application of latent-factors models is that we consider an intervention and aim to capture its effect on the entire distribution of

\[^{13}\text{The specification of the measurement equation implies that measures proxy the log of a latent factor, such that latent factors only take positive values.}\]
latent factors. To do so, we allow the joint distribution of the latent factors to differ in the two treatment states —i.e. in the case of this study, positive and negative emotional state \((d = 0, 1)\).

A quantitative assessment of personality traits includes an estimation of the measures’ internal reliability, stability, discriminant and convergent validity, and predictive power to explain behaviors and outcomes. In particular, a component of vector \(\theta_i\) has discriminant validity for trait \(k\) if \(\lambda^n_k\) is the only non-zero component of \(\lambda^n_{1,...,K}\). Likewise, a component of the vector \(\theta_i\) has convergent validity if measures within the construct are highly correlated. Following psychometric literature, it is assumed that each measure only proxies one latent factor (Gorsuch, 1983; Gorsuch, 2003). Nevertheless, this assumption can be relaxed to clarify the interpretation of latent factors (Attanasio et al., 2015a).

For the purpose of this study, performance measurement for trait \(k\) in situation \(n\) depends on the reward for manifesting the situation \((R^n_{i,k,t})\), on the transient emotional state during measurement \((E^n_{i,k,t})\), and on other factors operating in the situation \((W^n_{i,k,t})\):

\[
M^n_{i,k,t} = h_k(\theta_{ik}, \theta_{ik}, R^n_{i,k,t}, E^n_{i,k,t}, W^n_{i,k,t}, \varepsilon^n_{ik})
\]  
(3.5)

Finally, performance on task/test \(T_{ij}\) \((j = 1, ..., J\) tasks) by student \(i\) can be represented as a function of that student’s traits as:

\[
T^n_{i,k,t} = f_k(\theta_{ik}, \theta_{ik}, R^n_{i,k,t}, E^n_{i,k,t}, W^n_{i,k,t}, \varepsilon^n_{ik})
\]  
(3.6)

While the relevance of emotional processes to measurements is increasing (e.g. Almlund et al. (2011); Heckman and Kautz (2012)), previous studies have not con-

\[\quad\]

\[\quad\]

\[\quad\]
The Role of Emotions on Self-Reporting

As mentioned in Chapter 1\textsuperscript{15}, the relationship between emotion and cognition has been debated for centuries ([Descartes, 1649(1989)]). There is also an ongoing debate over the critical role played by emotions in cognition, perception, attention, and memory ([Damasio, 1994]; [Lakoff, 2008]; [Salzman and Fusi, 2010]; [Fuster, 2013]). Researchers combining neuroscience, behavioral science, and economics further posit that emotions have a strong influence on economic behavior and decision-making as well as on labor-market performance ([Weber and Johnson, 2009]). Emotions have a number of meaningful effects on behavior.\textsuperscript{16} With respect to decision-making, for example, emotions experienced while making a decision —i.e. choice option-elicited emotions— are at the base of traditional economic interpretations of utility as emotional carriers of value. Positive emotions increase value and elicit approach, whereas negative values decrease value and result in avoidance ([Weber and Johnson, 2009]). Moreover, emotions unrelated to the judgment or decision at hand, referred as incidental emotions, have also been shown to influence choice ([Weber and Johnson, 2009]).

Given their effect on perception (e.g. [Fuster, 2013]), emotions may also affect self-perception, and therefore play a role in biasing self-reported psychometric tests. A recent study by ([Querengsser and Schindler, 2014]) experimentally induced emotional states and compared baseline Big Five Inventory scores to those after the induction of

\textsuperscript{15}See the analogous section about the role of emotions on behavior and economic outcomes.

\textsuperscript{16}See Chapter 1 for a review.
happiness or sadness. The expected differential effect could be detected for neuroticism and extraversion, with sadness leading to increased neuroticism and decreased extraversion scores, and happiness leading to elevated extraversion scores. They conclude that sadness may trigger different self-concepts in susceptible people, biasing perceived personality, and that this bias could be minimized by tracking participants’ emotional states prior to personality measurement.

As mentioned earlier, I argue that there is a positive correlation between emotional state and scores on self-reported tests. In particular, I claim that self-reported tests used to measure both cognitive and non-cognitive skills are usually biased because of transient emotions that arise during test-taking. For instance, if two individuals have the same score, ex-post, on a given psychometric test, but one felt pleased during the test, and the other felt unhappy, the former’s score will overestimate skills, while the latter’s will underestimate them. I argue that valence would unambiguously affect self-reported tests in a corresponding manner. For instance, an individual who experiences positive valence will perceive himself as having a greater manifestation of a desirable trait, such as grit or conscientiousness, while one experiencing negative valence will have underestimated psychometric test scores. The degree to which the subject will over/under-estimate each statement about himself is related to his own reference-group perception. On the other hand, the level of arousal is theoretically ambiguously correlated with self-reported tests of non-cognitive skills (Weber and Johnson, 2009). That is to say, whether you are attentive, excited, distracted, or calm should not affect the direction of the self-reported score of a given test.

As another example, an individual experiencing high arousal and positive valence will be classified as having joy; therefore, his scores will be considered as overestimated when correcting for their aforementioned measurement error. Moreover, the psychometric test from an individual experiencing negative valence and low arousal will be underestimated. Low levels of arousal and valence could be thought as depres-
sive and thus a more realistic state (Barraza et al., 2015). Moreover, subjects who are emotionally depressed tend to focus their attention outward, as compared to those experiencing positive emotions, who tend to focus inward. Therefore, lower levels of arousal and valence can be interpreted as an increase in the level of awareness. This would also be consistent with underestimated self reporting of skills if the perception of reference groups is positive.

Figure 2.1 in Chapter 1 shows how emotional states are classified. Indeed, emotions can be classified in four different categories in the plane, namely: joy, anger, relaxation, and sadness. As Figure 2.1 exposes, joy is characterized by high arousal and high valence, anger by high arousal and low valence, relaxation by low arousal and high valence, and sadness by low arousal and low valence. Finally, the relationship between emotions and arousal is ambiguous (Weber and Johnson, 2009); thus, I will model emotions using the valence index instead of both arousal and valence indices. Both are kept throughout this conceptual framework, and in the emotion detection theory to preserve clarity.

3.3 Emotion-Detection Theory

The complex task of measuring emotions can be done using self-reported measures, which generally follow the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988). Since the PANAS is self-reported, it experiences similar issues to those mentioned for self-reported psychometric tests. Performance can be influenced by

---

17 There is an analogous section about emotion-detection theories in Chapter 1.

18 PANAS asks subjects to rate on a scale from 1 to 10 how much of each of 7 positive and 9 negative affects they feel. The positive affects are amusement, arousal, contentment, happiness, interest, relief, and surprise; the negative affects are anger, confusion, contempt, disgust, embarrassment, fear, pain, sadness, and tension. In addition, subjects were asked whether the film clip made them happier, sadder, or neither; and whether the film clip put them in a better mood, worse mood, or neither.

19 See Chapter 1.
the examinee’s emotional and motivational state (Plucker and Makel, 2010), test setting (ONeil et al., 2014), and reference bias (West et al., 2014; Kautz et al., 2014).

Another solution can be to use methods from affective neuroscience. That literature has studied emotion-detection from EEG recordings—and other sources—extensively, and a variety of emotional models have been developed (e.g. Petrantoniakis and Hadjileontiadis, 2010). Ekman et al. (1987) suggested the universality of six facial expressions based on Darwinian theory: happiness, surprise, anger, disgust, sadness, and fear (Darwin, 1872). However, the most frequent model used in psychology and affective neuroscience is continuous in nature and thus express emotions in an n-dimensional space; usually the two-dimensional arousal and valence model. Valence accounts for the judgement of a situation as positive or negative, while arousal expresses the degree of one’s excitation, spanning from calmness to excitement. The literature generally uses James-Lange’s arousal-valence model of emotions (Lang, 1995). In particular, the Ramirez and Vamvakousis (2012) version of James-Lange’s arousal-valence model will be considered as the framework for emotions.

Over the past decade, emotion-detection research has employed a variety of physiological measurements and methods, including pupil dilation, heart rate, and skin conductance for arousal, and voice and facial manifestations for valence (Takahashi et al., 2004; Partala et al., 2000). However, many of these measurements can be consciously modified, and thus, the signals they produce are not purely objective (Partala et al., 2000). In particular, those related to valence—e.g. facial expressions—are easily modifiable, and also unambiguously related to self-reporting. Thus, the use of electroencephalogram (EEG) becomes particularly appropriate. In a recent study, Korjus et al. (2015) showed that there is no correlation between resting-state EEG

---

20 See Kim et al. (2013) for a extensive review.
waves and any of the five personality dimensions of the self-reported Big Five Inventory (BFI, John and Srivastava, 1999b), and concluded that the extraction of personality traits from resting-state EEG power spectra is extremely noisy, if not impossible. In summary, EEG can relate self-reported tests for non-cognitive skills—e.g. BFI—to transient emotional state during testing, but does not allow us to predict psychometric test scores from a normal, resting-state EEG recording without stimuli.

The electrodes in an EEG headset capture the electrical activity corresponding to field potentials resulting from the combined firing of many individual neurons in the brain. However, cortical activity measures are distorted by the tissue and skull between the electrodes and the neurons. This introduces noise and reduces the intensity of the recorded signals. Regardless, EEG measurements offer important insights into the electrical activity of the cortex. Indeed, EEG recordings allow us to measure brain activity and, thus, predict emotional state and estimate physiological responsiveness.

A number of authors have considered methods for detecting emotions from EEG recordings, improving both accuracy and objectiveness as compared to the aforementioned physiological measurements (Choppin, 2000; Takahashi et al., 2004; Lin et al., 2010; Ramirez and Vamvakousis, 2012; Kim et al., 2013). Accuracy accounts for true and consistent results, that is to say, the likelihood that the model can correctly predict the elicited emotion, which is known by the characteristics of the stimuli or by a self-reported emotional state after a given experiment. Numerous studies have attempted to compare emotion-detection methods, arguing that the most suitable methodology and strategy—i.e. the one with the highest predictive accuracy—is the use of EEG recordings (Choppin, 2000; Takahashi et al., 2004; Lin et al., 2010; Ramirez and Vamvakousis, 2012; Verma and Tiwary, 2014; Kim et al., 2013). For instance, Brown et al. (2011) estimate an 82% accuracy for arousal and valence; Yoon
and Chung (2013) found a 70% accuracy for arousal and valence; and Verma and Tiwary (2014) found 85% accuracy for arousal, valence and dominance.\textsuperscript{21}

In a recent study, Martinez-Leon et al. (2016) compared the quality of data captured by a professional Biosemi Active II\textsuperscript{22} to that by the low-cost Emotiv EPOC\textsuperscript{23} headset used in this study, shown in Figure 3.1.\textsuperscript{24} Higher precision and lower variance are found in the EPOC datasets. Thus, a suitable strategy to proxy emotional state and responsiveness for dimensions of arousal and valence both in the lab and the field can be with low-cost EEG recordings.\textsuperscript{25}

\textsuperscript{21}The dominance scale ranges from submissive (or without control) to dominant (or in control, empowered). I do not consider that dimension in this study due to the requirement on self-reporting its degree.

\textsuperscript{22}For details visit http://www.biosemi.com/.

\textsuperscript{23}For details visit http://www.emotiv.com/.

\textsuperscript{24}Their results are based on the comparison of the success rate of a Brain-Computer Interface (BCI) system.

\textsuperscript{25}It is important to note that these indices are not intended to describe full personality traits or
Studies have shown that the prefrontal cortex—in addition to the amygdala and the insula—coordinates consciousness and regulates emotions.\textsuperscript{26} Measuring emotions therefore requires consideration of EEG signals from prefrontal-cortex electrodes AF3, AF4, F3, and F4 in the 10-20 standard classification, as referenced by Figure 3.2. Alpha (8-12Hz) and beta (12-30Hz) waves are of interest in measuring emotions.\textsuperscript{27} Alpha waves are predominant in relaxed states and brain inactivation, while beta frequencies are associated with alertness and an excited state of mind.\textsuperscript{28} Consequently, the beta-alpha ratio shown below is an effective indicator of the level of character.

\textsuperscript{26}Bear and Connors (2007). For a recent review see Salzman and Fusi (2010).
\textsuperscript{27}(Bos, 2006; Ramirez and Vamvakousis, 2012; Verma and Tiwary, 2014; Yoon and Chung, 2013; Brown et al., 2011; Choppin, 2000).
\textsuperscript{28}There is ample data linking changes in these frequency bands to various mental processes such as changes in mental state (Moretti et al., 2004), changes in attention allocated to a task (Klimesch, 1999), memory processes (Klimesch, 1996), motivation and emotional processes (Knyazev, 2007), different sleep stages (Keenan, 1999) and consciousness levels (John, 2002), among others.

101
arousal (Ramirez and Vamvakousis, 2012; Verma and Tiwary, 2014): 

\[ arousal_i = \frac{\beta_{F,i}}{\alpha_{F,i}} \]  

(3.7)

where \( F \) indicates the simple average of prefrontal-cortex electrodes AF3, F3, AF4 and F4, while \( i \) index an individual in the sample. On the other hand, previous neurophysiological studies have shown that emotional valence has different representations in the right and left hemispheres.\(^{29}\) Activity decrease over the right frontal lobe correlates with positive emotion, while activity decrease over the left frontal lobe is related to negative emotion. Since activity decrease, or inactivation, could be measured as the inverse of arousal, it is possible to estimate valence level by the following relation:

\[ valence_i = \frac{\alpha_{F4,i}}{\beta_{F4,i}} - \frac{\alpha_{F3,i}}{\beta_{F3,i}} \]  

(3.8)

where \( F3 \) and \( F4 \) indicate electrodes located on the left and right frontal lobes, respectively. Individuals are indexed by \( i \).

### 3.4 Methods

**Participants**

The first experiment, the Behavioral experiment, considered 60 high-school students from Northern Chile. In particular, the students were a subsample of those who

---

\(^{29}\)For a discussion on the validity of estimating valence by comparing hemispherical activation, see, for example, Ramirez and Vamvakousis (2012); Kim et al. (2013); Verma and Tiwary (2014); Yoon and Chung (2013); Bos (2006); Brown et al. (2011).
participated in the impact evaluation of the program “Rockstars para la Minería”. This is relevant as the results of the present study served as a pilot and proof of concept, experimental study of that research project whose results and analysis are presented in the Chapter 1 of this dissertation.

The second experiment, the Neurophysiological experiment, considered a total sample of 23 graduate students at Columbia University (16 males, aged 22-34). The final sample included 18 students due to the necessary rejection of some noisy datasets (e.g. due to physical movement, eye movement, poor signal quality, etc.).

**Treatments**

Both experiments induced emotional states using a series of pictures from the Geneva Affective Picture Database (GAPED). This database contains more than 700 images that were selected due to their power to elicit valence in humans. Scenes that induce emotions related to the violation of moral and legal norms (human-rights violations or animal mistreatment) were used to provoke negative valence, while positive pictures were those representing human and animal babies, and scenery. Finally, a series of neutral pictures, mainly depicting inanimate objects, were used.\(^{30}\) See Figures 3.3, 3.4, 3.5 for negative, positive, and neutral examples, respectively. As the pictures were originally rated based on valence and arousal in the Swiss population, it was important to ensure that they could elicit the same emotions in other populations, namely the Columbia University and Chilean students. In summary, subjects were induced to positive, negative, or neutral emotional states. Experimental paradigms of both experiments are explained below.

\(^{30}\)GAPED also contains a series of pictures of spiders and snakes designed to elicit basic “darwinian” emotions. These were not included in this experimental paradigm.
Figure 3.3: GAPED Negative Stimulus Examples

Figure 3.4: GAPED Positive Stimulus Examples
Measures

Battery of Tests

Both experiments considered a battery of self-reported psychometric tests designed to measure non-cognitive skills. Among these were the Big Five Inventory of Personality Traits (BFI, John and Srivastava, 1999a), the Grit Scale, (Duckworth et al., 2007), and the Rotter Locus of Control Scale (Rotter, 1966). The BFI is a widely accepted taxonomy of personality traits that describes personality as consisting of the following traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism (John and Srivastava, 1999a). The Grit Scale is one referring to the BFI’s conscientiousness measure, defined as “perseverance and passion for long-term goals.” It is measured on a likert scale, from 1 to 5 between “Strongly Agree” and “Strongly
Finally, the Rotter Locus of Control Scale (Rotter, 1966) indicates to what extent the subject attributes their fate to internal/personal factors as opposed to external factors, where some authors attribute the lack of locus of control to those that are highly fatalistic, and believe that their actions have little impact on the outcomes they experience (Bowles et al., 2001).

The Neurophysiological experiment considered the BFI and the Grit Scale. As mentioned, these have both become popular objects of study for non-cognitive skills and emotions. With regard to the BFI, there is evidence that some dimensions of the test are correlated with different kinds of emotional dispositions using both self-reported and peer-reported indices for 7 emotional dispositions. In particular, evidence suggests that there is a positive correlation between the BFI and positive emotional state, where the latter is often measured using the aforementioned, self-reported PANAS. For instance, Querengsser and Schindler (2014) estimated the impact of emotions on the self-reported BFI. They considered approximately 100 subjects who were induced to either a positive or a negative emotional state and were asked to perform the BFI. Their results show that induced sadness increased self-reported neuroticism while decreasing extraversion, while happiness was not associated with any significant correlation. However, since there is no agreement in the psychology literature about the degree of malleability of personality, the BFI is not a well suited index for a medium-term impact program evaluation. For this reason, and due to other practical and theoretical reasons explained in Chapter 1, the BFI was no considered in the initial field experiment, and was therefore excluded in this Behavioral experiment. Instead, the Behavioral experiment used the Rotter Locus

---

31 Statement’s examples of this test are: “I have overcome setbacks to conquer an important challenge”, “I become interested in new pursuits every few months”, or “Setbacks do not discourage me”. For details see, for example, Almlund et al. (2011)

32 See, for example, Almlund et al. (2011), and West et al. (2015), who recently administered more than 1,300 Grit Scale tests to 8th-grade students in Boston, USA.

33 See Chapter 1, Section 1.4 for an explanation of the low likelihood of affecting personality traits
of Control Scale (Rotter, 1966) and the Grit Scale (Duckworth et al., 2007).

**Electroencephalogram (EEG) measurements**

EEG recordings were conducted only in the Neurophysiological experiment. EEG devices measure the voltage change that occur when a neuron fires. When a positive change in the voltage crosses a certain threshold, an action potential is triggered, causing the voltage to increase from the resting potential of about -60mV to +20mV. This electrical activity from a group of neurons is measured over the cortex with EEG electrodes. While participants took the aforementioned tests in each experiment, their brain activity was measured using the low-cost, portable Emotiv EPOC EEG headset.\(^{34}\)

EEG recordings were transmitted wirelessly to a laptop computer. Online data was processed using OpenViBE (Renard et al., 2010). Moreover, MATLAB 2014b software (Waltham, MA, USA), and the EEGLAB open-source toolbox (Delorme and Makeig, 2004), were used to perform offline data analysis.

**EEG Data Filtering**

Using Fourier frequency analysis, the original signal was converted to the frequency domain to select data in frequency bands of interest, and was then transformed back to the time domain. Many authors suggest that a maximum of a 10-second signal is necessary for EEG measurements in order to include factors distinct to the elicited in the short or medium term. Moreover, the BFI measures some of its dimensions using only a couple of statements, which do not leave room for positive and negative transient emotional states to be determined for each of its dimensions. Therefore, when considering all five dimensions the correlation — and its sign — between emotions and the BFI was found to be asymmetrical — and heterogeneous — in the positive/negative locus with respect to each of the five dimensions (Querengsser and Schindler, 2014). However, as Querengsser and Schindler (2014) showed, there is a correlation between BFI and emotions, and that result is consistent with the results exhibit in the present study.

\(^{34}\)For details, visit http://www.emotiv.com/.
emotion (Davidson et al., 1990). Here, we consider an epoch window of 5 seconds for the analysis.

In conventional spectral analysis, the spectrum of a time-series $x(t)$ observed in a time-window of duration $T$ is estimated as the square of the absolute value of its tapered fourier transform. Namely:

$$S_{\text{conv}}(f) = |X(f)|^2 = |\int_0^T h(t)x(t)e^{2\pi ift}dt|$$  \hspace{1cm} (3.9)

where $h(t)$ is a taper or windowing function. Some standard choices for the windowing function are the constant, the Hanning window, the Hamming window, and the Parzen window (Bokila et al., 2010).

The standard method has three main caveats (Bokila et al., 2010). First, the choice of the type of windowing function is arbitrary. Second, the estimate is biased —i.e. the expectation of the spectral estimate differs from the population spectrum—for any type of windowing function. Third, the spectral estimate has a higher variance than that found using the multi-taper method.\footnote{The latter problem can be ameliorated by breaking up the observation window into shorter segments and averaging spectral estimates computed over those segments. However, this is typically impossible for neural time-series which are stationary only over relatively short durations of a few hundred milliseconds (Bokila et al., 2010).} Thus, in order to address the limitations of the standard method, the multi-taper Fourier transform is used for continuous data sets.

Instead of choosing an arbitrary windowing function, the multi-taper method finds $W$ for windowing functions that are maximally concentrated within $[-W,W]$, which are known as Slepian sequences (Bokila et al., 2010). For duration $T$, around $K = 2TW-1$ of these functions can be found (Bokila et al., 2010). Therefore, for $W > 1/T$ there are multiple such functions that are well-concentrated in frequency and therefore have bias-reducing characteristics. The multi-taper estimate of the spectrum is given by
using each of these tapers as windowing function and averaging the resulting spectral estimates, as the following equation illustrates:

\[
S_{MT}(f) = \frac{1}{K} \sum_{k=1}^{K} |X_k(f)|^2 = \frac{1}{K} \sum_{k=1}^{K} \left| \int_0^T u_k(t) x(t) e^{-2\pi if t} dt \right|^2
\]  

(3.10)

where \( u_k \) are the Slepian sequences or tapers. By using a set of tapers, rather than a unique data taper or spectral window, the algorithm reduces the variance of spectral estimate, due the orthogonality of the Slepian sequences. Moreover, it is particularly effective for short data segments (Castellanos and Makarov, 2006).  

Finally, in order to smooth the exponential nature of EEG signals we apply a logarithmic power transformation to the data, which is the standard in this literature (Verma and Tiwary, 2014; Ramirez and Vamvakousis, 2012).  

EEG activity is typically studied in the 1-80Hz range, with amplitudes of 10 to 100 microvolts (Ramirez and Vamvakousis, 2012). Here, the EEG signal was passed through a low-pass filter with a 40Hz cutoff frequency in order to remove noise from artifacts. The alpha (8–12Hz) and beta (12–30Hz) bands were described as particular areas of interest for emotion recognition for both valence and arousal. The influence of EOG artifacts (eye movement/BLINKING) is most dominant below 4Hz, ECG (heart) artifacts around 1.2Hz, and EMG (muscle) artifacts above 30Hz (Bos, 2006). Therefore, by extracting only the alpha and beta frequencies, the influence of much noise is already significantly reduced.  

Finally, arousal and valence indices were constructed following the equations 3.7 and 3.8, respectively.

---

\(^{36}\) A MATLAB toolbox for implementing those methods is freely available at http://chronux.org/chronux (Mitra and Bokil, 2008).

\(^{37}\) The exact transformation follows Aspiras and Asari (2011). In particular, the EEG signals were transformed as follow: \( \text{LogSignal}_j = 10 \log_{10}(S_j) \); where \( j = \text{positive, negative, restingstate} \) conditions.
3.5 Experimental Design

The Behavioral experiment examined the behavioral correlation between emotionally-laden image stimuli and self-reporting of emotions, without the use of EEG. The Neurophysiological experiment was designed to validate the correlation between emotional valence, which was measured from subjects’ EEG recordings, and self-reported psychometric test scores.

Behavioral Experiment

The non-EEG Behavioral experiment evaluated elicited and self-reported emotions in students in Northern Chile. The objective of the experiment was to determine whether there is a positive relationship between the emotionally-laden stimuli and the self-reported test scores, without considering the EEG recordings. This was particularly important since the context of the field experiment, where I later conducted an impact evaluation of a program designed to foster non-cognitive skills, is socioeconomically and demographically different than the one in which GAPED validated its database.\(^{38}\) The sample consisted of 60 student subjects, all of whom participated in the impact evaluation—i.e. field experiment—in Chilean public high schools. Schools and students were randomly selected to participate in this brief experiment. Each subject answered nine randomized selections from psychometric tests—five sets of Locus-of-Control questions, and four sets from the Grit Scale—with a series of 20 emotionally-laden positive, neutral, and negative images inserted between each part of the test.\(^{39}\) The order of the selected test questions and the valence of the initial stimuli were randomized in this experiment. The experiment was conducted using 7

\(^{38}\) The results of that evaluation is presented in Chapter 1.

\(^{39}\)There is a potential of 60*5*4=540 observations in the stimuli-test scores locus. However, this was lower due to the fact that some students did not answer the whole battery of tests.
laptop-based work-stations; MATLAB scripting boxes were used to present images and testing GUIs and to send time-locked stimulations.

**Neurophysiological Experiment**

This EEG experiment measured brain signals while asking portions of selected psychometric tests and while showing subjects emotionally-laden stimuli. The experiment consisted of subjects being shown images of a particular valence followed by a short non-cognitive test from either the BFI or the Grit Scale. In general, for each of the 18 students there were 7 measures in the valence-test score locus. In this particular case, however, I only considered the Grit Scale, as it had the potential for scores with both positive and negative stimuli, whereas the BFI’s dimensions are incomparable. The experiment aimed to identify the correlation between EEG-derived emotional valence and test scores, with the hypothesis that under a more positive emotional state, subjects tend to over-rate themselves in non-cognitive skills tests. The opposite should hold true under negative emotional states.

In this experiment subjects were instructed to watch set of 20 images from the GAPED database, and then to take short, untimed, non-cognitive tests. In particular, 10 image/test sets were presented with positive and negative image sets being randomized across subjects, while tests were administered in order: 2 parts Grit Scale, 5 parts BFI, and 3 parts Cognitive Reflection Test. Figure 3.6 demonstrates the utilized platform. Neuro-feedback was not provided during testing.

The experiment was conducted using 7 work-stations, each consisting of a laptop and an Emotiv EPOC headset. The latter was interfaced with the OpenViBE plat-

---

40For example, the self-reported measure of BFI-extraversion with a positive stimulus is not comparable to that of BFI-openness with a negative stimulus.
form. As shown in Figure 3.6, MATLAB-scripting boxes were used to present images and testing GUls and to send time-locked stimulations.

3.6 Results

The purpose of this chapter is to serve as a proof-of-concept of the correlation between self-reported psychometric tests and transient emotional states. In order to do so, a couple of relationships must be establish, namely: (i) showing the correlation between
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Loc of Ctrl. (z-Score)</th>
<th>(2) Grit (z-Score)</th>
<th>(3) Non-cog Skill Index (z-Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Stimulus (=1)</td>
<td>0.311**</td>
<td>0.0321</td>
<td>0.159*</td>
</tr>
<tr>
<td></td>
<td>(0.130)</td>
<td>(0.150)</td>
<td>(0.0917)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.136**</td>
<td>-0.0372</td>
<td>-0.0781*</td>
</tr>
<tr>
<td></td>
<td>(0.0651)</td>
<td>(0.0750)</td>
<td>(0.0458)</td>
</tr>
<tr>
<td>Observations</td>
<td>266</td>
<td>214</td>
<td>480</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.030</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>Number of Subjects</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Subject FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Subject-level Clustering</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Clustered at subject-level standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

Notes (1): The experiment considered 60 subjects. Each subject answered nine randomized portions of psychometric tests, with a series of 20 emotionally-laden—positive, neutral and negative—images intercalated between them.

Notes (2): Non-cognitive skills’ z-score outcome comprises both locus of control and grit scale test scores. This additional data in the stimuli-test score locus provides us with greater statistical power.

The first step is to establish a positive correlation between self-reported psychometric measures of a subject’s non-cognitive skills and transient emotional state during test taking, based on results of the Behavioral experiment.

The experiment estimates the degree to which the emotional state induced by the GAPED images affects self-reported skills in psychometric tests. Table 3.1 shows the correlation of psychometric tests with the stimuli, demonstrating that the stimuli generate altered self-perception, and consequently, psychometric test answers. All tests are exhibited as z-scores (mean of zero, standard deviation of one). Therefore, a
subject experiencing a positive emotional state, relative to a negative one, would self-report to 0.31σ in the Locus-of-Control Scale, 0.032σ—not statistically significant—in the Grit Scale, and 0.16σ when considering both. These impacts are economically relevant considering the evidence from educational policies and programs.\textsuperscript{41}

Given this positive correlation between emotional state and self-reported scores, we conclude that subjects experiencing a positive emotional state tend to overestimate their skills, while those experiencing a negative one underestimate their skills. These results differ from those of Querengsser and Schindler (2014) who established that only negative transient emotions affect self-reporting. In addition, the stimulus is arguably strong enough to trigger a change in behavior, which will become relevant for the estimation of emotional responsiveness as an outcome of the social intervention on Chapter 1.

**Emotion Detection from EEG Recordings**

As mentioned, affective neuroscience literature has studied emotion detection from EEG recordings—and other sources—extensively.\textsuperscript{42} The results shown here come from the Neurophysiological experiment. The aim of this subsection is to verify that the low-cost, portable EEG headset that would later be used in the field has comparable accuracy in emotion detection to that of the traditional, stationary, lab-based systems.\textsuperscript{43} As with most literature in this space, this study follows Lange’s arousal-valence model of emotions. Emotional state can be detected from EEG recordings with a reasonable level of accuracy, as measured by rate at which EEG features—e.g.

\textsuperscript{41} For example, relative to programs that seeks to improve academic achievement—i.e. cognitive skills—Duflo et al. (2012) set the impact range between 0-0.8σ, considering around 30 evaluations, while Murnane and Ganimian (2014) resent more than 100 impact evaluations in the range between 0-0.9σ.

\textsuperscript{42} For a review see Kim et al. (2013).

\textsuperscript{43} Results here are complementary to the recently published work of Martinez-Leon et al. (2016).
valence, arousal, and other robust indices—predict a known emotion. The emotion can be expected, or known, by the nature—positive, neutral, or negative—of the presented stimulus. Typical accuracy for arousal and valence is currently in the 70-85% range. This experiment considered support vector machines (SVM), Fisher linear discrimination analysis (LDA), expectation-maximization (EM), decision trees (DTs), Naive Bayes, and bagging classification methods. An average accuracy of 79% was found across subjects and classifiers, as shown in Figure 3.7.

Figure 3.7: Average Accuracy of Emotion Detection (Multiple Classifiers)

---

44 See, for example, Yoon and Chung (2013); Ramirez and Vamvakousis (2012); Verma and Tiwary (2014); Choppin (2000); Brown et al. (2011); Bos (2006).

45 Various MATLAB packages were used for classification. SVM was performed using the libsvm library (Chung and Lin, 2004); EM was performed using the emgm library (Chen, 2012). Fisher LDA, decision tree, naive bayes, and bagging methods were implemented using canned MATLAB functions, namely, classify(), treefit(), NaiveBayes.fit(), and voting, respectively. For further details on the classifiers used see, for example, Verma and Tiwary (2014).
Consistent with these findings and my hypothesis, Table 3.2 and Figure 3.8 indicate that the stimuli positively affect the level of valence. Table 3.2 displays the results of a linear, fixed-effects regression model, which controls for non-randomness in the explanatory variable —i.e. emotional valence—, considered, in this case, at an individual level. The p-value is 0.20, indicating that the correlation has a confidence level of 80%, which is considered reasonable for exploratory or pilot studies (Moore and Nietert, 2011; Iverson et al., 2003). These results indicate that measures of emotional state can be proxied by analysis of recordings from low-cost, portable EEG devices at an 80% confidence level and with a level of accuracy comparable to those found in the literature.  

\footnote{See previous section for details on the levels of accuracy found in the literature.}
Table 3.2: Emotionally-laden Stimuli and Valence from EEG Recordings

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Stimulus (=1)</td>
<td>0.0930</td>
</tr>
<tr>
<td>p-value (0.0700)</td>
<td>[0.202]</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.377***</td>
</tr>
<tr>
<td></td>
<td>(0.0317)</td>
</tr>
<tr>
<td>Observations</td>
<td>117</td>
</tr>
<tr>
<td>Number of Subjects</td>
<td>18</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.010</td>
</tr>
<tr>
<td>Subject FE</td>
<td>Yes</td>
</tr>
<tr>
<td>Subject-level Clustering</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

**Emotions from EEG and Self-reported Tests**

The objective in this subsection is to establish a correlation between emotional states derived from EEG recordings and from self-reported psychometric tests, based on the Neurophysiological experiment. It should be noted that the expected correlation in the Neurophysiological experiment is higher than that of the Behavioral experiment, given that it considers individual emotional responsiveness to the stimuli. Whereas the Behavioral experiment ignores individual subject reactions, and relies on a binary outcome in response to stimuli (i.e. 1 or 0 for positive or negative stimuli, respectively), it is expected that, in the Neurophysiological experiment, subjects more strongly affected by a particular stimulus will exhibit heightened emotional responses to it, evidenced by self-reporting on psychometric tests.
Table 3.3: Emotions (Valence) from EEG Recordings and Self-Reported Test Scores

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Grit (zscore)</th>
<th>(2) Grit (zscore)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>0.221</td>
<td>0.363**</td>
</tr>
<tr>
<td></td>
<td>[0.472]</td>
<td>[0.152]</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0827</td>
<td>0.136</td>
</tr>
<tr>
<td></td>
<td>[0.231]</td>
<td>[0.173]</td>
</tr>
<tr>
<td>Observations</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>Number of Subjects</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Model</td>
<td>FE</td>
<td>RE</td>
</tr>
<tr>
<td>Subjects</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Hausman Test</td>
<td>0.747</td>
<td></td>
</tr>
</tbody>
</table>

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Figure 3.9: Proof of Concept: Grit Test and Emotional Valence

Table 3.3 shows a significant correlation between the valence index of emotional state derived from EEG recordings and self-reported psychometric test scores, in particular, the Grit Scale (see, also, Figure 3.9). In this case, the Hausman test cannot reject the null hypothesis (p-value of 0.75) that the random effect (RE) model is the true model, thereby favoring the RE relative to the fixed effects (FE) model. The RE model considers an error form that is the weighed average of a pure between (BE) model and an FE model, allowing for randomness in the idiosyncratic component of the error term. In other words, the transient emotional valence—an exogenous covariate modulated by interchanging emotionally-laden stimuli—does not charac-
terize the individual fixed effect. In addition, the fixed part of the RE captures the relationship between each subject and the degree to which they react to the stimuli, a correlation not captured by a pure between model.\textsuperscript{47} Moreover, the RE has higher efficiency than the FE, which can explain the significance observed under this model.\textsuperscript{48}

In summary, a significant positive relationship between emotional state—proxied by the valence index from EEG recordings—and self-reported psychometric test scores was found in this experiment. These results support the behavioral correlation established earlier based on stimulus-based emotional state and self-reported test scores (see Table 3.1). Again, it is worth mentioning that this conclusion differs from that of Querengsser and Schindler (2014), who did not find any correlation for positive emotions when estimating the impact of emotions on the self-reported BFI.\textsuperscript{49}

It is interesting to note that it was not possible to compare an aggregate effect on a composite index of the dimensions of the BFI. As mentioned, this was to be expected given that many other factors relate to educational and labor market outcomes (e.g. economic sector and job characteristics) so a subject’s extroversion may not be related to such systematic differences (Deming, 2015). That said, the locus of control and some BFI dimension—e.g. openness to experience—have been shown to strongly correlate with labor market outcomes, including salary and higher educational attainment.

\textsuperscript{47}For details, see Cameron and Trivedi (2005).

\textsuperscript{48}In fact, when the RE is the correct specification, the RE is more efficient than the FE (Cameron and Trivedi, 2005). This increase in efficiency can explain why the relation between the emotions and test scores turn out significant under the RE model. It is also worth mentioning that the sample size is relatively small and demographically different to those in the Behavioral experiment found in Table 3.1.

\textsuperscript{49}For details see section 2.2 above.
3.7 Conclusion

In this study, I propose a methodology that overcomes the measurement error inherent in many psychometric tests designed to measure cognitive, socio-emotional and creative skills. In particular, I argue that emotion —i.e. transient emotional state— affects self-reporting on measures that proxy non-cognitive skills. This study also contributes to the understanding of the relationship —specifically, the positive correlation— between these skills. This study’s results differ from the asymmetric ones of Querengssser and Schindler (2014), who solely observed effects for negative emotions.

The present study contributes to the understanding of the relationship between emotional state and self-reported psychometric test scores. In particular, I conducted two experiments. The first one, referred as the Behavioral experiment, estimates the capacity of the GAPED database to elicit emotions by measuring the degree to which its images affect non-cognitive tests scores. The results demonstrated that a positive stimulus, increases test results, relative to a negative one, by $0.31\sigma$ and $0.16\sigma$ on the Grit Scale and for a non-cognitive skill index, respectively. These results are economically relevant considering the evidence from educational policies and programs (Duflo et al., 2012; Murnane and Ganimian, 2014).

The second, Neurophysiological experiment, enriches this study in two ways. First, it validates the capacity of a low-cost EEG headset —i.e. the Emotiv EPOC— to detect emotions through traditional methods of affective neuroscience, with a 79% average accuracy level across subjects and classifiers. Further, it establishes a positive correlation between EEG-derived emotional state and self-reported psychometric test scores. Specifically, a statistically significant positive correlation was found for a random effects model.

Emotions modulate both non-cognitive and cognitive processes (Damasio, 1994;
Salzman and Fusi, 2010; Weber and Johnson, 2009). I argue that transient emotional state affects self-perception and, consequently, the way individuals respond on self-reported psychometric tests. It is possible that EEG-based emotional state characterizations—i.e. valence indices—can be used to correct for such emotional bias in self-reported tests. It may also be possible to apply the proxies of emotions to cognitive measures. These are interesting topics for further research.\textsuperscript{50}

The experiments contributing to this study were performed in similar settings and with the same instruments and demographics—for the Behavioral experiment—to that of the impact evaluation examined in Chapter 1. This is particularly relevant given the often mismatched results between the lab and field experiments in the literature (DellaVigna, 2009).

In summary, this study supports a significant positive relationship between emotional state—proxied by a valence index—and self-reported psychometric test scores, evidenced both by induced stimuli and by scalp EEG recordings.

\textsuperscript{50}See the Appendix for an exploratory exercise on the correction of emotional bias in self-reported tests, based on data collected for the study examined in Chapter 1.
3.8 Appendix 1: A Simple Model of Emotionally-corrected Measures of Non-Cognitive Skills

Here, I explore the correction of emotional bias in non-cognitive skills tests, including the locus control and grit scale for data collected in the “Rockstars para la Minería” evaluation in Chile (see Chapter 1). As previously discussed, for two individuals with the same ex-post score on a self-reported psychometric test, the student who felt happy during the test will have overestimated their score, while the one who felt unhappy will have underestimated it. Here, I use EEG-based emotional indices to correct for such emotionally-driven measurement errors. First, I estimate the emotion-related variation in the scores across subjects, and then correct —i.e. subtract— that portion from the score. Empirically, we have:

\[ Y_{j,i,t} = \theta_0 + \theta_1 * \text{valence}_{i,t} + \theta_2 * \text{arousal}_{i,t} + \xi_{j,i,t} \] \hspace{1cm} (3.11)

where \( Y_{j,i,t} \) indicates the score of the \( j^{th} \) test for individual \( i \) in time \( t \), with \( j = \text{grit, locuscontrol}, \ i = \text{students in Field Experiment}, \) and \( t = \text{march, august} \). The covariates are the EEG-based arousal and valence indices as indicated previously. Then, using the coefficient estimates by an FE model:

\[ \hat{Y}_{j,i,t} = \hat{\theta}_0 + \hat{\theta}_1 * \text{valence}_{i,t} + \hat{\theta}_2 * \text{arousal}_{i,t} \] \hspace{1cm} (3.12)

Thus, the emotionally-corrected scores become:

\[ \bar{Y}_{j,i,t} = Y_{j,i,t} - \hat{Y}_{j,i,t} \] \hspace{1cm} (3.13)
According to the hypothesis, this should yield results closer to the true skill the psychometric test is meant to capture. Therefore, in an impact-evaluation context, where the program affects emotional regulation, I would expect the average treatment effects to be more consistent with the qualitative evidence. The following model represents the impact evaluation for the “Rockstars para la Mineria program”:

\[
\bar{Y}_{j,i,t} = \alpha + \beta \cdot treat_{j,i} + \gamma \cdot time + \delta \cdot treat_{j,i} \cdot time + \theta \cdot X_{i,t} + \varepsilon_{j,i,t} \quad (3.14)
\]

where \( \bar{Y}_{j,i,t} \) indicates the emotionally-corrected score on the \( j^{th} \) test for individual \( i \) in time \( t \), with \( treat_{j,i} \) and \( time_{j,i} \) indexing the individual treated by the program and their survey time —i.e. baseline or follow up—, respectively.

It is important to note that this bias correction is not without cost. In fact, more noise is added into the model given that the EEG-based emotional state is estimated. However, since we are correcting an independent variable —i.e. not a covariate— there is an improvement in efficiency, and thus potentially in significance. In addition, in models of educational or labor outcomes where life skills are inputs —e.g. academic achievement, dropouts, salaries, employment rate, etc.—, an analog methodology would potentially increase consistency when corrected psychometric tests are used as inputs in the model. The application of the methodology developed in this study is an interesting area for further research.

As an exploratory exercise, Table 3.4 presents the emotionally-corrected results. The procedure follows equations 3.12 through 3.13. Point estimates increased by around 0.02\( \sigma \) and 0.08\( \sigma \) on locus of control and grit tests, respectively, relative to the uncorrected estimations shown in Table 2.6.
Table 3.4: Emotionally-corrected Impact on Life Skills

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>LoC-emo-cor</th>
<th>Grit-emo-cor</th>
<th>Creativity-emo-cor</th>
<th>13Creative-emo-cor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat. &amp; Post.</td>
<td>-0.0836</td>
<td>0.0124</td>
<td>0.206</td>
<td>0.136</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.329*</td>
<td>0.723***</td>
<td>0.165</td>
<td>0.168</td>
</tr>
<tr>
<td>Post.</td>
<td>0.169</td>
<td>0.0212</td>
<td>0.0715</td>
<td>0.338**</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.428</td>
<td>-0.618**</td>
<td>0.181</td>
<td>0.0182</td>
</tr>
</tbody>
</table>

Observations 296 296 296 296
R-squared 0.031 0.104 0.075 0.085
Student FE Yes Yes Yes Yes
Student Clustering Yes Yes Yes Yes
School FE Yes Yes Yes Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In summary, I have presented an original methodological application designed to correct for latent, emotionally-driven measurement error on non-cognitive psychometric tests. While this exercise did not reveal a significant improvement, the change found was a trend in the expected direction. Further research can be conducted in this area.

3.9 Appendix 2: Complementary Figures on Emotion Classification

Event-related potentials (ERP) and penalized logistic regression (PLR) are demonstrated here for positive and negative valence.\textsuperscript{51} In these figures, the data is epoched at the onset of a given stimulus — i.e. positive or negative — and the corresponding valence index is estimated. Figure 3.10 reveals that there is a significant difference at the individual level between positive and negative stimuli. There is a positive

\textsuperscript{51}For details on these method see, for example, Bokila et al. (2010); Castellanos and Makarov (2006); Mitra and Bokil (2008); Delorme and Makeig (2004).
Figure 3.10: Valence: Positive-Negative

The spatial pattern in Figure 3.11 indicates that positive stimuli produce inactivity near the right side of the frontal cortex, while negative stimuli exhibit greater inactivity near the left side of the frontal cortex. Further, the grey shading indicates at which point the signal differences are significant at 5%. It is clear that the positive and negative stimuli are statistically identifiable between 250 and 500 ms.
post-stimulus onset.

Figure 3.11: Component via PLR
Chapter 4

Can Art-based Programs Nurture Human Capital? Evidence From Public Schools in Chile

4.1 Introduction

Since the publication of the seminal studies of Bowles and Gintis (1976), the literature has been emphatic in holding that both cognitive skills, understood as the use of language, memory and logic/mathematics, as well as non-cognitive or socio-emotional skills, which include self-concept, self-control, perseverance, and motivation, among others, are relevant in determining an individual’s likelihood of success in the workplace (Bowles and Gintis, 1976; Bowles et al., 2001; Heckman et al., 2006).

Moreover, the literature exploring the ability of innovation to foster economic development has argued the need for an educational process that fosters creativity, that is to say, that promotes active exploration, innovation, adaptation, and creation by students, in order to cultivate the tools and skills required to meet the challenges
of the modern economy.\(^1\) Furthermore, the discovery of effective policies that improve students’ academic results is rare in developing countries, both because of design and implementation factors.

In this context, creativity, conceived as a complex process that includes both cognitive and socio-emotional abilities, should play a key role in modern educational processes. Intending to change educational systems designed for knowledge and content memorization and repetition, recently several countries have undertaken reforms that seek to foster creativity in students.

This paper explores the role that participation in art-based programs has on the development of creativity as well as cognitive and non-cognitive skills. An understanding of this link would allow for the design of educational policies that pursue a more fruitful outcome than more traditional educational systems.

There is evidence that participation in artistic initiatives is correlated with higher academic achievement, creativity levels, and richer relationships with peers (e.g. Burton et al., 1999; Catterall et al., 2012). In general, those studies show correlations between individual or class-level participation in arts and various outcomes of interest, without taking into account the potential selection bias in their estimations (Fiske, 1999; Kaufman and Sternberg, 2010; Catterall et al., 2012). That is to say, there is scarce empirical evidence that shows causality between the participation on art-based programs and the increase of cognitive, non-cognitive and creativity skills on the part of students.

The hypothesis that underlies this study infers that involvement by students in intensive artistic and cultural activities has a positive impact on creativity, as well as on the development of cognitive and non-cognitive skills. Furthermore, the analysis is based on subjects engaged in the intensive practice of such activities during a

medium-time horizon —e.g. a semester— in different artistic disciplines. In addition, it is pertinent to only include initiatives that meet certain minimum quality standards in this analysis, as otherwise the probability of finding significant effects would be difficult (Steinberg, 1998; Durlak et al., 2011; Durlak and Weissberg, 2007). Indeed, to consider the regular art or music classes offered in schools makes it hard to identify impacts, given the high level of heterogeneity in quality and design of these forms of teaching (Catterall et al., 2012).

To satisfy the above-mentioned criteria, I selected the Chilean program for the promotion of creativity entitled Acciona Program (AP). The program takes place during normal school hours. The AP offers workshops in various artistic disciplines led by skilled artists who are paid competitively and have been selected through a competitive process organized through a public-private partnership. It was possible to survey four different high schools in which the program was implemented with a participant allocation rule that did not bias the results. Moreover, since AP is implemented at the high school level, impact estimations in the present document will be underestimated compared to the impact that students could have had in, for example, elementary school.

To test the aforementioned hypothesis, the impact of AP participation on different dimensions of creativity is estimated using two specially designed psychometric tests to measure the degree of creativity of students. Then, the impact of AP on both cognitive and non-cognitive skills is also estimated.

To the best of my knowledge, this study is the first to link psychometric measures of creativity with outcomes on both cognitive and socio-emotional skills. Similarly, with respect to the previous literature on the subject, our approach contributes improvements in methodology through correcting for the potential selection bias that compromises correlation analyses conducted previously.
The paper is organized as follows: the second section briefly reviews the related literature. The third section describes the nature of the program of interest and the means of data collection. The fourth section describes the methodology used. The fifth section presents the results, and finally, the sixth section concludes.

### 4.2 Literature Review

Several studies seek to elucidate the impact of individual participation in artistic activities. In particular, such studies examine how intensive participation in art-based initiatives may influence various dimensions of an individual’s human development, including creativity, self-concept, and academic performance. Indeed, the flexible use of language among students in artistic disciplines fosters socio-emotional skills, while promoting awareness, experimentation, creativity and encouraging communication and expression, which affect relationships. That argument is shared by various specialized researchers who argue that the aim of arts programs with educational purposes is to provide tools, boost creativity, and develop skills that enable individuals to understand their environment and engage in ‘better’ relationships with peers (Burton et al., 1999; Catterall et al., 2012). In addition, Seidel (1999) claims that the goal of such programs should not be to create artistic figures, but artistic communities that encourage group learning and integration.

Furthermore, there is little evidence on the impact of participation in artistic workshops on creativity and cognitive and non-cognitive skills as a whole. In general, studies cover only one of the dimensions mentioned above. For example, studies compiled at Fiske (1999) attest to the importance of art on creativity (defined using criteria such as fluency, flexibility and originality). They also argue that participation in artistic activities strengthens tolerance of frustration and stimulates the ability to create and express ideas. Similarly, several authors argue in favor of the link between
participation in arts initiatives and the development of non-cognitive skills, such as motivation, perseverance, frustration, tolerance, participation outside of school hours in other civil initiatives, self-criticism, and leadership among others (Bryce et al., 2002; Burton et al., 1999; Fiske, 1999). Further research has focused on the impact of participation in art programs on school performance (Bryce et al., 2002; Catterall et al., 2012; Egana-delSol et al., 2013). Those studies found greater effects on vulnerable populations, which are expected given the higher marginal contribution of art interventions in settings with low substitutability. Finally, there are studies that estimate correlations between programs for the development of creativity and other outcomes of interest, such as academic results. Scott et al. (2004) developed a meta-analysis of the literature. The authors conclude that effective programs focus on the development of cognitive skills and heuristics involved in the application of skills, using exercises typically relevant to the pursuit of these types of skills.

Meanwhile, Kaufman and Sternberg (2010) assemble recent state-of-the-art research on creativity in multiple dimensions. One of the most interesting arguments made by the authors is that there is a link between level of creativity and degree of openness to experiencing the world. The authors also refer to recent studies showing a positive link between creativity and motivation, which they then extrapolate to other areas of development. Therefore, the theoretical argument relies on the existence of a desire to improve the capacity in creative activities, which then affects other areas of intellectual and personal development (Kaufman and Sternberg, 2010, p. 358).

On the other hand, there is evidence that high school teachers, even among those who value creative students, prefer less creative students, or students performing tasks oriented towards instructor guidance and those who strictly follow the teacher’s instructions (Kim et al., 2010). In fact, teachers argue that students should be honest, disciplined, unselfish, humble, and diligent (Kim et al., 2010). This suggests that a program run by artists, many of whom have no previous experience working with
students and are unlikely to be biased in favor of “desirable” forms of behavior, can create a space for the creativity-prone to flourish. This fact is important to support the study of programs like AP instead of regular art classes in schools.

Nevertheless, the investigations referred to above for the most part represent case studies or comparative studies, which do not include an adequate strategy and robust identification of the effects found regarding similar populations who have not participated in the program. Many times they do not control for the observable characteristics of the students, comparing aggregate averages between students with high and low socioeconomic status compared to a high or low participation in artistic endeavors. Even studies that use longitudinal data do not control for time invariant unobservable variables, such as art skills or self-selection into art classes or programs (Catterall et al., 2012). Indeed, it is impossible to separate how much of the observed effect is due to a single act of selection bias, either self-selection of students, the parents or the establishment itself, with respect to the participation in art-based programs or classroom activity.

As the exception to the aforementioned common approaches to analysis, Garaigordobil and Perez (2002) performed an experimental design to study the effects of art on creativity, a robust design which substantially improves the researchers’ ability to identify impacts regardless the treated group comprises 89 individuals. The researchers randomly assigned students from six to seven years of age to a program to encourage creativity (Ikertze Art Program), and found a positive and significant impact on all aspects of creativity —i.e. flexibility, fluency and originality— using the Torrance Test of Creative Thinking (TTCT). In particular, the instrument of written creativity was positive in all three dimensions outlined above, while the graphic creativity tool was positive and significant only in the fluency dimension. From self-reported surveys, they also found a significant impact on social behaviors such as leadership, enthusiasm, respect and self-control.
In sum, I conclude that empirical evidence tends to suggest positive effects of artistic programs on the development of creativity and cognitive and non-cognitive skills in the subjects involved that favor personal, social and professional accomplishment. The literature also implies that such activities have a greater impact on vulnerable sectors. However, in general, the evidence presents serious deficiencies in identifying causal impact between participation in artistic and creative, academic or social-emotional skills. Finally, to the best of our knowledge, this study is the first to explore the causality between intensive participation in artistic activities and the development of creativity, cognitive and social-emotional skills using psychometric instruments in the context of high social vulnerability.

### 4.3 The Program and Data Collection

Acciona Program, formally called the Program to Enhance Creativity in the Full School Day Time (FSDT), was implemented in 2007 by the National Council for Culture and the Arts in collaboration with the Ministry of Education of Chile and the Balmaceda Youth Art Foundation. The program aims to contribute by improving the quality of arts and cultural education in the hours of free programming in the FSDT as well as helping to improve the quality of education in the development dimensions of creativity and cultural and artistic training of students. The program uses workshops in various artistic disciplines developed by skilled artists in the area, who are paid competitively and have been previously selected by open competition through a public-private institution, the Balmaceda Youth Art Foundation. The selected artist, with the support of a school teacher, works during the entire semester with a group of students.

The data collection was conducted by research assistants and the author of this study. In order to establish which schools would be considered in this study, we
performed an initial screening of the educational institutions that met the following exclusionary criteria: (i) AP was implemented in 2009 or prior; (ii) Schools were geographically accessible; and (iii) Schools were not directly affected by the tsunami and earthquake of February 27, 2010 (i.e. Lota, Coronel, among others). Subsequently, we contacted every pre-selected school in order to identify: (1) years of implementation of AP in each class in order to get a sense of the intensiveness of treatment; (2) criteria under which it was decided to implement the program, to avoid selection bias of more or less skillful students in the dimensions of this study, and (3) whether the institutions have another class of the same grade level that has not participated in AP, in order to generate suitable control groups.

As discussed above, the characteristics of the selected schools would allow unbiased and suitable analysis to estimate the impact of the program. Moreover, we tried to choose schools in which the criteria that determine whether a student participates were the mere fact of belonging to a particular class, as opposed to the possession of certain skills or characteristic behaviors. Consequently, we identified four educational institutions in the whole country that met the above conditions, in particular, three in the metropolitan area of Santiago and one in the Los Rios Region. However, at the time of application of instruments to students in the selected institutions, it was found that students in the treated group (those who were exposed to AP) did not always meet the condition of having participated in the workshops for at least two years as had been originally gleaned from preliminary research. Thus the final sample included students who had participated zero, one and two years in the workshops. Therefore, it was possible to estimate the impact for those participants who had one workshop and for those who had two or more workshops, considering that in one year a student may participate in two semiannual workshops. Finally, the sample of both treatment and control group participants comprises a total of 297 students in their freshman and sophomore years in four different secondary schools.
In addition, we conducted a survey in order to obtain valuable information about AP workshops, similar-substitute workshops outside of and within the schools of interest, socio-economic characteristics and metrics of non-cognitive skills, specifically self-concept, and creativity of students. In particular, through a written version of the Torrance’s test of creative thinking (TTCT) we obtained a measure of fluidity, flexibility, and originality, while through the graphical form of the TTCT we obtained a measure of eighteen dimensions of creativity that will be explained in the next section. Finally, grades were standardized in order to allow for comparison between classes and schools.

4.4 Methodology

The methodology section of this study consists of four subsections. The first explains how both the graphical and written form of creativity of students are measured. The second describes how the self-concept scale was computed. The third explains the statistical approach to identifying impacts, while the fourth draws on the empirical strategy to estimate the impacts.

Approach to Creativity

In general terms, creativity is characterized by the ability to establish new and different relationships between elements, creating innovative productions. This ability is a greater or lesser extent in any subject, involving the emotional and cognitive dimensions, and whose development depends on the received stimulation from contextual variables. Theoretically there are several approaches to creativity. In a recent article, Kozbelt et al. (2010) assemble the mainstream approaches, namely: Developmental, Psychometric, Economic, Stage and Componential Process, Cognitive, Problem Solv-
The present study considers the psychometric approach to creativity, which advocates being independent of any theoretical model of creativity, and focusing exclusively on the measurement of creativity, having indeed the purpose of informing alternative theories of creativity, as the aforementioned approaches. In particular, it follows the psychometric approach as it assumes that creativity can be measured by psychometric tests or proxy variables following the foundational models of Guilford (1950) and Torrance (1988, 1992, 2000). The TTCT is the most widely recognized and employed test of creativity, or more specifically, divergent thinking (Runco, 2010; Plucker and Makel, 2010).

Accordingly, TTCT was used both in graphic and written form. Unlike other tests, such as the Piers-Harris Test of self-concept, which will be explained below, the TTCT instruments require a particular scoring because it is necessary to apply certain criteria to objectify subjective information. To ensure the correct application of the criteria throughout the correction process, the corrections made by a team of three artists were monitored daily by a psychologist expert in order to assess the consistency between the remedial corrections. In addition, a second correction was conducted for each test. If the difference between scores was greater than the standard deviation of the test scores, the scoring of the whole day was cancelled.

**Torrance Test of Creative Thinking: Graphical Form**

The graphical form of the TCTTG is a psychometric instrument developed by Torrance (1966) and subsequently revised in 1988, 1992 and 2000 (Torrance, 1988, 1992, 2000). It was translated and validated in Chile in 1992 for students in the seventh and eighth grades of middle school by Parra (1993), which means that the rules of the test apply only for students of those cohorts. However, the test remains valid for
other ages whenever conclusions are drawn from the comparison of results between similar groups that answered the test, as is done in this study.

The TCTTG is an instrument for measuring creative skills through graphic exercises, evaluating a total of eighteen skills. Five of them are considered fundamental by the author, and the remaining thirteen offer insights to consider the creative potential of the subject and possible areas of strength. The test consists of three sections, each of which evaluates some or all of the eighteen creative thinking skills. For the present study, we applied activities number 2 and 3, which include the measurement of eighteen skills that the test attempts to assess.

The five basic skills, also referred to as norm-referenced measures (NRM), are: fluency, originality, elaboration, abstractness of title, and resistance to premature
closing. Examples of fluency and elaboration are exhibit in Table 4.1 and Table 4.2, respectively. These five core fundamental skills are behind the score obtained. However, an important part of the results are also provided through the thirteen remaining creative forces — i.e. skills, dimensions — that the test evaluates, called criterion-referenced measures (CRM). The thirteen creative forces are: emotional expressiveness, storytelling articulateness, movement or action, expressiveness of titles,
synthesis of incomplete figures, synthesis of lines, unusual visualization, internal visualization, extending or breaking boundaries, humor, richness of imagery, colorfulness of imagery, and fantasy. An example of high scored relative to CRM is shown in Table 4.3.

The thirteen CRM evaluate creative forces that the author considers as important as the NRM. However, a response may not contain any of the forces or more of them. Some of the forces tend to occur simultaneously, although they may materialize in an isolating way. Many of these measures occur frequently while others are unusual. The score obtained from the creative forces of the CRM is a bonus that adds to the score obtained on the basis of the five NRM measures, resulting in the final score of the TCTTG.

The TCTTG was evaluated with a 20 percent rate of double correction. Therefore, at the end of each day of correction, two evaluators corrected the same 20 percent of the surveys. Then, if a difference greater that 1 standard deviation in the dimension’s score was found between the evaluators, all surveys that were scored that day were declared invalid, and therefore, the 100 percent of surveys collected on that specific day had to be corrected again.

**Torrance Test of Creative Thinking: Written Form**

The written form of the TCTTW is an assessment tool of creativity that was developed in Chile by the firm Consultancy for Development. Following Torrance (1966), the authors shaped an instrument to measure three dimensions of creativity, namely: fluency, flexibility and originality. The test consists of two open questions that assess these three dimensions. For the present study Test Form 2 was randomly selected. Test form 2 includes the following questions:

---

1. What do you think might be the disadvantages of having a cell phone?

2. What do you think could be the similarities between a carpet and a washing machine?

The answers were firstly classified into categories and then into thematic fields. The number of categories alluded determines the fluency score, while the number of thematic fields referred defines the flexibility score. Finally, the originality score is assigned to the answer’s category with a frequency less than or equal to 1 percent of respondents, that is to say, the categories that were mentioned by three or fewer subjects in our sample. Moreover, when scoring the fluency dimension of the test, answers which did not follow directions (e.g. presented advantages as opposed to disadvantages), as well as the answers “do not know”, “no” or leaving the question blank were considered invalid.

Each of the three dimensions evaluated yields a raw score. Afterward, all scores were standardized so that the values are expressed in equivalent scales and thus it is possible to add them and get a total score. It is crucial to note that this procedure assumes that the three dimensions are equally weighted. The TCTTW was evaluated with a 100 percent rate of double correction.³

**Approach to Self-Concept: Piers-Harris Scale**

The Piers-Harris Self-Concept Scale (PHSCS) was created in 1967 and subsequently revised and amended in 1976 and 1984 (Gorostegui, 1992). The original version consisted of eighty statements with the opportunity to answer “Yes” or “No.” The instrument was adapted and validated in Chile in 1992 for students in the third to sixth year of primary education by Gorostegui (1992), who created a version of

³Same methods were applied in Chapter 1.
seventy claims with possible answers “Yes” and “No.” One point is awarded for each item that has been answered in the positive direction of the self-concept and 0 points if the item has been answered in the negative direction of the self-concept.

The PHSCS is based on self-perception among individuals and not on observations from parents or teachers. The test provides an overall score, and scores for six sub-scales, which correspond to specific dimensions of self-concept, namely: behavioral adjustment, intellectual and school status, physical appearance and attributes, freedom from anxiety, popularity, happiness and satisfaction.

As in the case of the TTCT, there are no national standards for the cohort considered in our sample, however, this does not interfere with the objectives of the investigation since the analysis is performed based on the comparison of results between groups belonging to the same cohort.

**Propensity Score Matching**

Propensity Score Matching (PSM) was implemented to estimate the quantifiable impacts of the program, which correspond to a semi-parametric estimator of the differences in the averages of the relevant outcome (i.e. TCTTG, TCTTW, PHSCS, GPA) between the participant in AP and their control group. Specifically, the PSM was constructed doing matching based on the probability of participating in the program, using a binary Probit model. In other words, I created a statistical clone of each participant based on the probability of participating in the program given his or her socio-economic and demographic characteristics. Nevertheless, I did not perform a matching method conditional on the vector of characteristic given the inaccuracy in small samples and the dimensionality problem posed by using vectors. Indeed, the

---

4See Cameron and Trivedi (2005).
propensity score transforms the characteristic vector of each individual into a scalar number in the range zero to one.

In the estimation of the propensity score the gender, parent’s level of education, family per capita income, class and school, number of books at home, and whether the student participated in artistic activities before implementation of the program, both within and outside of the school were considered. All of this information was derived from the survey, though it was checked with administrative data from the school. In the definition of the PS, $Y$ index the relevant outcome, and in particular:

$Y_{0,i} = \text{If the student did not participate in AP.}$

$D_i(0,1) = \text{Treatment dummy. It takes the value 1 if student } i \text{ participates, and 0 otherwise.} $

$X = \text{Student’s characteristic vector}$

$\text{Prob}(X) = \text{Prob}(D = 1/ X)$

Additionally, following Rosenbaum and Rubin (1983), it must be hold:

$$0 < \text{Prob}(X) < f$$ \hfill (4.1)

$$\left( Y_{i0}, Y_{i1} \right) \perp D_i / \text{Prob}(X)$$ \hfill (4.2)

Where the constraint (1) ensures that the probability is well defined, while the assumption (2) is known as “unconfoundedness” or selection on observables, conditional on student’s characteristics, the participation probability is the same for both groups, and therefore the actual allocation is random. Consequently, if both (1) and (2) hold you can properly obtain the average impact of the AP as:

$$\Delta(X) = E(Y_1 - Y_{0}/\text{Prob}(X), D = 1) = E(Y_1/\text{Prob}(X), D = 1) - E(Y_{0}/\text{Prob}(X), D = 1)$$ \hfill (4.3)
Where the term \( E(Y_0/\text{Prob}(X), D = 1) \) identifies the interest outcome of the Program participant did not participate in it, which it is impossible to observe in reality. Therefore, it is necessary to estimate it using the control group, using 3 different methods, namely: Propensity Score Matching Radius 1 Closest Neighbor. Propensity Score Matching Radius 5 Closest Neighbors. Propensity Score Matching Kernel-Epanechnikov (bw 0.06.). For simplicity, the only the results from last methods is showed.

These three methods differ essentially in the metric and procedure performed to look for the clone in the control group given the existence of a trade-off between bias and variance of the estimator of the average treatment impact. Then, bias reduction quality after the matching analysis was performed.

**Matching Quality**

Rosenbaum and Rubin (1983) suggested measuring the quality of different matching used by estimating the percentage reduction in bias between the average estimated standardized variables for participants and controls. Basically, for each of the variables considered in estimating the propensity score a standardized bias is calculated for both before and after matching, which is defined as:

\[
SE_{\text{BeforeMatch}} = 100 \times \frac{X_1 - X_0}{\sqrt{0.5[V_1(X) - V_0(X)]}} \tag{4.4}
\]

\[
SE_{\text{AfterMatch}} = 100 \times \frac{X_{1\text{match}} - X_{0\text{match}}}{\sqrt{0.5[V_{1\text{match}}(X) - V_{0\text{match}}(X)]}} \tag{4.5}
\]

where \( X \) (\( V \)) represents the mean (variance) of the sample and the subscripts 1 and 0 identify treated and control participants, respectively.

\[5\text{See, for example, Todd (1999); Leuven and Sianesi (2015); Imbens (2004); Abadie and Imbens (2006); Augurzky and Kluve (2004).}\]
4.5 Empirical Strategy

In the literature there is no agreement on the number of variables used to estimate the propensity score, i.e. the model should be parsimonious or over-parameterized (Caliendo and Kopeinig, 2005). According to Bryson et al. (2002), over-parameterized model does not generate inconsistent estimates, however, the variance increases, causing inefficiency in the estimators. Additionally, Heckman et al. (1997) show that omitting relevant variables biases the estimates. Also, Zhao (2005) indicates that it is necessary to omit variables that affect treatment, since they can be affected by or affect participation, which prevents the researcher from properly identifying the impact of a given treatment. Finally, Caliendo and Kopeinig (2005) state that economic theory, previous research in similar areas and knowledge of incumbent institutions are an effective guide for model selection.

The selection of variables to define the propensity score of the students was based both on the results of the literature regarding the main variables have a direct effect on creativity, school performance and socio-emotional skills, as well as some degree of knowledge of the class tracking in the Chilean public education system. It was also considered whether students had participated in a similar workshop, both inside and outside of the school, which could be affecting the results on both psychometric tests and academic outcomes. In sum, I considered student’s gender, mother’s and father’s education level, family per capita income, class, number of books at home, and if the student had participated in artistic activities before program implementation both inside and outside the school.

Therefore, the Probit Model is defined by the following equation:

\[ y^* = X\beta + \varepsilon \]  \hspace{1cm} (4.6)
\[ y = [y^* > 0] \] 
(4.7)

with \( \varepsilon \sim \eta(0, 1) \)

where \( Y \) is the participation dummy (\( =1 \) if the student participates) and \( X \) is the matrix of the individual characteristic aforementioned.

Furthermore, Bryson et al. (2002) argue that use common support could exhibit problems if the sample is small or the proportion of underlying treated individuals who rely outside the common support is high. Besides, when the results were compared between the case with and without conditioning on common support the impacts did not change substantially; nonetheless, the results presented in this paper are those who consider the common support.

Also, the standard deviations of the estimates of impact are considered overestimated, because the propensity score is considered as true, being that is estimated in the previous stage. However, what matters most is that the propensity score is well defined and meet the quality conditions that were outlined previously.

Furthermore, standard errors of the estimators are not bootstrapping considering the unfavorable results found by Imbens and Abadie (2008). Consequently, as a measure of robustness’ checks it was estimated impacts of AP for at least one workshop to at least two workshops using the standard errors correction and asymptotic distribution proposed by Abadie and Imbens (2006), which is generalized for various functions recently in Abadie and Imbens (2015). This last procedure checked that the statistical significance of the findings is robust to the correction of standard errors.

4.6 Results

This section presents the impacts of the participation on the AP program considering three dimensions, namely: creativity, socio-emotional and academic outcomes.
Firstly, it presents results for those participants that experienced only one AP workshop with respect to those who had no participation. It then turns to those participants who had at least two AP workshops, comparing again with those who never participated in AP. By doing this I can capture the hypothesis that the size of the impact could be in function of time of exposure. In fact, I observed positive and significant results mostly for the group that was treated more intensively (those who have at least two workshops). Furthermore, this increasing function of the impact is coherent with a no-sample bias, since if I would have sample bias, for example both groups differ in some unobservable skills, I found a positive impact even in those who participated only in one workshop. This holds if I assume that those who had only one workshop are similar in unobservable skills to those who had at least two.

Table 4.6 shows the difference in the considering characteristic variables between participants and their controls in the “one workshop” sample, using the Kernel (band width 0.06) just to illustrate. Complementarily, Table 4.6 also shows the results obtained for the Matching Quality for differences between the means of the participants and their counterfactual, after matching, all methods are not significant for most of the variables that are used as a control and therefore it can be argued that the sample is well balanced after conditioning on the propensity score.

Table 4.6 shows that one semester of AP affects positively and significantly language GPA, which considers all grades in the language class during a year, and graphical creativity. For all the remaining outcomes impacts are not statistically different from zero.

As Table 4.6 and 4.6 show, the variables for both treated and control groups. As before, after PSM the groups are balancing, accepting all null hypotheses of no difference between groups. Furthermore, Table 4.6 shows positive and significant impact of at least two semester of AP on Avg. General Grade (GPA) and on Math, Lan-
Table 4.1: Descriptive Statistics and Matching Quality of One AP Workshop

<table>
<thead>
<tr>
<th>Variable</th>
<th>Condition</th>
<th>Treated</th>
<th>Control</th>
<th>%bias</th>
<th>Bias</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre (0 if art)</td>
<td>Unmatched</td>
<td>0.58</td>
<td>0.35</td>
<td>47.30</td>
<td>2.98</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.57</td>
<td>0.53</td>
<td>6.60</td>
<td>86.00</td>
<td>0.37</td>
<td>0.71</td>
</tr>
<tr>
<td>Mother’s Scholing (years)</td>
<td>Unmatched</td>
<td>11.75</td>
<td>12.01</td>
<td>-11.40</td>
<td>-0.71</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>11.75</td>
<td>11.98</td>
<td>-10.20</td>
<td>10.70</td>
<td>-0.61</td>
<td>0.54</td>
</tr>
<tr>
<td>Father’s Scholing (years)</td>
<td>Unmatched</td>
<td>11.88</td>
<td>12.00</td>
<td>-5.20</td>
<td>-0.33</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>11.84</td>
<td>12.20</td>
<td>-15.10</td>
<td>-188.50</td>
<td>0.37</td>
<td>0.71</td>
</tr>
<tr>
<td>Income1</td>
<td>Unmatched</td>
<td>0.13</td>
<td>0.10</td>
<td>7.10</td>
<td>0.45</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.13</td>
<td>0.09</td>
<td>12.40</td>
<td>-0.33</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>Income2</td>
<td>Unmatched</td>
<td>0.44</td>
<td>0.53</td>
<td>-17.90</td>
<td>-1.13</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.51</td>
<td>0.51</td>
<td>-10.00</td>
<td>44.00</td>
<td>-0.58</td>
<td>0.56</td>
</tr>
<tr>
<td>Income3</td>
<td>Unmatched</td>
<td>0.31</td>
<td>0.27</td>
<td>7.20</td>
<td>0.45</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.33</td>
<td>0.26</td>
<td>14.20</td>
<td>-96.80</td>
<td>0.82</td>
<td>0.42</td>
</tr>
<tr>
<td>Income4</td>
<td>Unmatched</td>
<td>0.11</td>
<td>0.08</td>
<td>10.70</td>
<td>0.68</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.06</td>
<td>0.10</td>
<td>-14.80</td>
<td>-38.70</td>
<td>-0.92</td>
<td>0.36</td>
</tr>
<tr>
<td>Other Similar Activity (1 if yes)</td>
<td>Unmatched</td>
<td>0.33</td>
<td>0.45</td>
<td>-24.80</td>
<td>-1.56</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.34</td>
<td>0.32</td>
<td>5.50</td>
<td>77.80</td>
<td>0.33</td>
<td>0.74</td>
</tr>
<tr>
<td>Books (1 if books&gt;30)</td>
<td>Unmatched</td>
<td>0.56</td>
<td>0.57</td>
<td>-2.50</td>
<td>-0.16</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.54</td>
<td>0.56</td>
<td>-4.40</td>
<td>-75.60</td>
<td>-0.26</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Total Obs: 72 Treated and 88 Controls.

Note: The total number of observations is reduced because it is imposed common support between both groups. Variables Income1-Income4 index dummies of income’s intervals, with 1 the lowest and 4 the highest.

These results are particularly interesting, since usually art-based programs are not planned to impact cognitive skills. Still, the results are consistent with previous evidence that shows that relationship (e.g. Kaufman and Sternberg, 2010; Catteral et al., 2012; Egana et al., 2013). Specifically, the impact on Avg. General Grade was 0.55 standard deviations (s.d), while it was 0.61 s.d. on Avg. Language Grade, 0.36 on Avg. Math Grade, and 0.33 s.d. on Avg. Art Grade. These results are relatively high compared with other more structural and more expensive interventions done in the same cultural and geographic context. Indeed, Chile implemented an extended school day policy on 1997 which recent eval-
Table 4.2: Impacts of One AP’s Workshop (Kernel, bw 0.06)

<table>
<thead>
<tr>
<th>One AC Workshop</th>
<th>Mean Treated</th>
<th>Mean Control</th>
<th>Difference</th>
<th>S.E.</th>
<th>T-stat</th>
<th>Impact (Dif/S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. General Grade</td>
<td>ATT 101.40</td>
<td>96.02</td>
<td>5.38</td>
<td>4.41</td>
<td>1.22</td>
<td>0.27</td>
</tr>
<tr>
<td>Avg. Art Grade</td>
<td>ATT 102.10</td>
<td>100.04</td>
<td>2.06</td>
<td>4.59</td>
<td>0.45</td>
<td>0.10</td>
</tr>
<tr>
<td>Avg. Language Grade</td>
<td>ATT 103.83</td>
<td>95.69</td>
<td>8.14</td>
<td>4.17</td>
<td>1.95</td>
<td>0.41</td>
</tr>
<tr>
<td>Avg. Math Grade</td>
<td>ATT 101.85</td>
<td>97.46</td>
<td>4.39</td>
<td>4.40</td>
<td>1.00</td>
<td>0.22</td>
</tr>
<tr>
<td>Personality Test Score</td>
<td>ATT 52.04</td>
<td>52.65</td>
<td>0.61</td>
<td>1.90</td>
<td>-0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>Total TTCTW</td>
<td>ATT 99.42</td>
<td>96.51</td>
<td>2.91</td>
<td>4.04</td>
<td>0.72</td>
<td>0.15</td>
</tr>
<tr>
<td>Total TTCTG</td>
<td>ATT 108.89</td>
<td>106.17</td>
<td>2.72</td>
<td>3.26</td>
<td>0.84</td>
<td>0.14</td>
</tr>
<tr>
<td>13 Forces Total TTCTG</td>
<td>ATT 104.66</td>
<td>96.23</td>
<td>8.43</td>
<td>4.63</td>
<td>1.82</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Total Obs: 50 Treated and 60 Controls on Common Support.
Note: The total number of observations is reduced because it is imposed common support between both groups. Variables in the Propensity Score are the same that those used as control in the double robust regression analysis (See Appendix).

valuation showed minimum effects on the improvement of school performance, ranging from 0.05 to 0.07 s.d. in Language and from 0.00 to 0.12 s.d. in Math (Valenzuela, 2005; Bellei, 2009). Moreover, Table 4.6 also shows positive and significant impact on flexibility of the written form of TTCT as well as on Abstractness of Title of the graphical form of TTCT.

An explanation of these findings could be that since the workshops develops creativity, and in particular the flexibility dimension –between thematic fields- as well as the abstractness of title capacity (a conceptual skill), it would impact academic achievement by producing a more flexible and sharp understanding of novel knowledge or theoretical frameworks. In fact, according to Torrance (1966, 1988), abstractness of title measures the ability of the individual to highlight the important or essential to a given situation and requires the ability to synthesize, discarding erroneous or irrelevant information. Specifically, when applying the test, the subject is invited to think of titles for drawing objects or situations, capturing the essential of the figures.
Table 4.3: Bias Reduction: Two Workshops

<table>
<thead>
<tr>
<th>Variable</th>
<th>Condition</th>
<th>Mean</th>
<th>Mean</th>
<th>%reduction</th>
<th>Bias</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Treated</td>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genre (0 if ale)</td>
<td>Unmatched</td>
<td>0.48</td>
<td>0.35</td>
<td>25.60</td>
<td>1.42</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.48</td>
<td>0.43</td>
<td>10.60</td>
<td>58.50</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>Ed. Mother</td>
<td>Unmatched</td>
<td>11.46</td>
<td>12.01</td>
<td>-22.60</td>
<td>-1.24</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>11.46</td>
<td>11.50</td>
<td>-1.70</td>
<td>92.50</td>
<td>0.50</td>
<td>0.93</td>
</tr>
<tr>
<td>Ed. Father</td>
<td>Unmatched</td>
<td>11.33</td>
<td>12.00</td>
<td>-29.00</td>
<td>-1.56</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>11.33</td>
<td>11.44</td>
<td>-4.70</td>
<td>83.70</td>
<td>0.02</td>
<td>0.82</td>
</tr>
<tr>
<td>Income1</td>
<td>Unmatched</td>
<td>0.13</td>
<td>0.10</td>
<td>8.70</td>
<td>0.49</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.13</td>
<td>0.11</td>
<td>7.10</td>
<td>18.30</td>
<td>0.34</td>
<td>0.74</td>
</tr>
<tr>
<td>Income2</td>
<td>Unmatched</td>
<td>0.59</td>
<td>0.53</td>
<td>10.60</td>
<td>0.58</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.59</td>
<td>0.59</td>
<td>0.30</td>
<td>96.80</td>
<td>0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>Income3</td>
<td>Unmatched</td>
<td>0.15</td>
<td>0.27</td>
<td>-29.60</td>
<td>-1.57</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.15</td>
<td>0.17</td>
<td>-5.60</td>
<td>81.20</td>
<td>0.15</td>
<td>0.77</td>
</tr>
<tr>
<td>Income4</td>
<td>Unmatched</td>
<td>0.13</td>
<td>0.08</td>
<td>16.50</td>
<td>0.94</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.13</td>
<td>0.12</td>
<td>-3.30</td>
<td>79.90</td>
<td>0.15</td>
<td>0.88</td>
</tr>
<tr>
<td>Other Similar activity (1 if yes)</td>
<td>Unmatched</td>
<td>0.33</td>
<td>0.45</td>
<td>-26.30</td>
<td>-1.44</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.33</td>
<td>0.29</td>
<td>6.50</td>
<td>75.30</td>
<td>0.33</td>
<td>0.75</td>
</tr>
<tr>
<td>Books (= 1 if &gt; 30)</td>
<td>Unmatched</td>
<td>0.57</td>
<td>0.57</td>
<td>-0.60</td>
<td>-0.03</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>0.57</td>
<td>0.53</td>
<td>6.80</td>
<td>-1050.00</td>
<td>0.33</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Total Obs: 46 Treated and 88 Controls.
Note: The total number of observations is reduced because it is imposed common support between both groups. Variables Income1-Income4 index dummies of income’s intervals, with 1 the lowest and 4 the highest.

Thus, the subject transform figural in verbal information.

Furthermore, there is also a positive impact on the CRM, also called 13 creative forces, which are as important as the Norm Referenced Measures (NRM). This is one of the main findings of the present study considering the impact of AP on creativity itself. Indeed, the empirical findings implies that most likely this kind of programs does not affect NRM broadly, which could be associated to other or previous experiences among students from a given cohort, but it does affects CRM, which could be interpreted as a more sophisticated or elaborated expression of creativity. In fact, as it was mentioned, CRM is related to unusual visualization, internal visualization, extending or breaking boundaries, humor, richness of imagery, colorfulness of imagery,
Table 4.4: Impacts of At least two AP’s workshops (Kernel, bw 0.06)

<table>
<thead>
<tr>
<th>Two AC Workshops</th>
<th>Mean Treated</th>
<th>Mean Control</th>
<th>Difference</th>
<th>S.E.</th>
<th>T-stat</th>
<th>Impact (Dif/S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. General Grade</td>
<td>ATT 109.49</td>
<td>98.55</td>
<td>10.94</td>
<td>6.08</td>
<td>1.80</td>
<td>0.55</td>
</tr>
<tr>
<td>Avg. Art Grade</td>
<td>ATT 105.03</td>
<td>98.38</td>
<td>6.66</td>
<td>5.89</td>
<td>1.13</td>
<td>0.33</td>
</tr>
<tr>
<td>Avg. Language Grade</td>
<td>ATT 109.92</td>
<td>97.73</td>
<td>12.20</td>
<td>5.85</td>
<td>2.09</td>
<td>0.61</td>
</tr>
<tr>
<td>Avg. Math Grade</td>
<td>ATT 107.11</td>
<td>99.92</td>
<td>7.19</td>
<td>6.00</td>
<td>1.20</td>
<td>0.36</td>
</tr>
<tr>
<td>Personality Test Score</td>
<td>ATT 50.58</td>
<td>55.27</td>
<td>-4.69</td>
<td>3.15</td>
<td>-1.49</td>
<td>-0.23</td>
</tr>
<tr>
<td>Total TTCTW</td>
<td>ATT 104.12</td>
<td>99.31</td>
<td>4.81</td>
<td>5.44</td>
<td>0.88</td>
<td>0.24</td>
</tr>
<tr>
<td>Total TTCTG</td>
<td>ATT 113.33</td>
<td>109.57</td>
<td>3.75</td>
<td>4.30</td>
<td>0.87</td>
<td>0.19</td>
</tr>
<tr>
<td>13 Forces Total</td>
<td>TTCTG ATT 112.26</td>
<td>101.25</td>
<td>11.01</td>
<td>5.87</td>
<td>1.87</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Total Obs: 24 Treated and 48 Controls on Common Support
Note: The total number of observations is reduced because it is imposed common support between both groups. Variables in the Propensity Score are the same that those used as control in the double robust regression analysis (See Appendix).

Among others, which are fairly more complex than NRM.

It is important to note that the negative impact found on the “personality” dimension of the PHSCS becomes insignificant as students participate in more workshops. This result is surprising, because one of the channels to explain the link between art-intensive initiatives and academic achievement would be through the development of non-cognitive skills. Nevertheless, it is possible to have two explanations. Firstly, as we can subtract from the difference between the impact of one and at least two workshops, the PHSCS was low in average for the participant respect to the control group. Secondly, the PHSCS scale is not a good measure of the socio-emotional skills that a program like AP promotes, and moreover, are the key element to have a subsequent impact on academic achievement, such as perseverance, motivation, self-control or responsibility. As robustness check, there is implemented a double-robust re-weighted regression model (DRRW) (Hill and Reiter, 2005; Ho et al., 2007). The main difference from a standard linear regression model is that each observation is
weighted by the inverse frequency of the propensity score. In fact, for each outcome and intensity of treatment, it was estimated a DRRW controlling for the same observable characteristics as the PSM analysis. The main conclusions remain the same. See Tables in the Annexes.

4.7 Conclusions

In this paper, the impact on the development of creativity, cognitive and socio-emotional skills by intensive participation in an art program is studied. One of the main contributions of this paper is to improve the identification strategy with respect to previous studies in the literature. In fact, the identification strategy is more robust in terms of self-selection bias and also it is controlled for many relevant variables that could be driving the impacts, such as parent’s education, art-based activities beyond the program, among others.
Firstly, the intensity of treatment is crucial. In fact, the impacts of an art based program on the aforementioned dimensions become statistically and economically significant when the students had taken at least two workshops.

Secondly, it is shown that AP would impact academic achievement by producing a more flexible and sharp understanding of novel knowledge or theoretical frameworks. Specifically, the impact on Avg. General Grade was 0.55 standard deviations (s.d), while it was 0.61 s.d. on Avg. Language Grade, 0.36 on Avg. Math Grade, and 0.33 s.d. on Avg. Art Grade.

Thirdly, in terms of CRM the impact of AP on creativity is positive and significant when students had at least two AP workshops. However, the NRM did not demonstrate a significant impact. Particularly, it is found a positive and significant
impact on flexibility of the written form of TTCT as well as on Abstractness of Title of the graphical form of TTCT. In sum, the impact on creativity is positive in the thirteen creative forces (CRM), in the Flexibility dimension measured in the written form of TTCT as well as the Abstractness of Title dimension of the NRM of the graphical form of the TTCT.
4.8 Appendix: Doubly Robust Re-weighted Regressions
Table 4.6: Doubly Robust Re-weighted Regression: One Workshop

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>GPA</th>
<th>GPA Math</th>
<th>GPA Leng</th>
<th>GPA Art</th>
<th>PHSC</th>
<th>TTCTW</th>
<th>TTCTG</th>
<th>TCTG 13F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (1=F)</td>
<td>-7.166</td>
<td>-8.430*</td>
<td>1.698</td>
<td>-3.035</td>
<td>-2.019</td>
<td>-5.369</td>
<td>-3.708</td>
<td>-7.517*</td>
</tr>
<tr>
<td>Ed. Mother</td>
<td>0.223</td>
<td>1.058</td>
<td>-0.824</td>
<td>1.079</td>
<td>-0.273</td>
<td>-0.264</td>
<td>0.0451</td>
<td>0.327</td>
</tr>
<tr>
<td>Ed. Father</td>
<td>1.168</td>
<td>0.712</td>
<td>0.691</td>
<td>1.708</td>
<td>-0.231</td>
<td>-0.103</td>
<td>0.0593</td>
<td>-0.493</td>
</tr>
<tr>
<td>Family has car</td>
<td>1.577</td>
<td>3.761</td>
<td>-0.684</td>
<td>3.959</td>
<td>0.128</td>
<td>0.218</td>
<td>-0.285</td>
<td>7.444*</td>
</tr>
<tr>
<td>Computer at home</td>
<td>2.342</td>
<td>-5.437</td>
<td>2.469</td>
<td>5.913</td>
<td>-3.764</td>
<td>-9.653**</td>
<td>-3.193</td>
<td>-12.08*</td>
</tr>
<tr>
<td>Art outside</td>
<td>1.399</td>
<td>0.502</td>
<td>-0.684</td>
<td>4.398</td>
<td>-1.599</td>
<td>5.924</td>
<td>8.027***</td>
<td>13.45***</td>
</tr>
<tr>
<td>Books (=1 less than 30)</td>
<td>-0.998</td>
<td>3.958</td>
<td>-2.881</td>
<td>4.784</td>
<td>1.436</td>
<td>-3.058</td>
<td>-2.339</td>
<td>-5.479</td>
</tr>
<tr>
<td>Const.</td>
<td>83.19***</td>
<td>84.40***</td>
<td>98.32***</td>
<td>59.17***</td>
<td>60.01***</td>
<td>108.4***</td>
<td>105.4***</td>
<td>108.8***</td>
</tr>
<tr>
<td>Obs.</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>113</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.104</td>
<td>0.145</td>
<td>0.088</td>
<td>0.097</td>
<td>0.1</td>
<td>0.108</td>
<td>0.146</td>
<td>0.313</td>
</tr>
</tbody>
</table>

Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1. Doubly Robust Re-weighted Regression estimation combines inverse probability weighting, in which each individual observation is given a weight equal to the inverse of the probability of the treatment he/she received conditional on baseline covariates— i.e. the estimated propensity score, as equation (4.6) shows— with standard regression modeling.
### Table 4.7: Doubly Robust Re-weighted Regression: Two Workshops

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>11.93 (4.280)</td>
<td>9.855 (4.133)</td>
<td>14.57 (4.369)</td>
<td>3.833 (3.699)</td>
<td>-0.822 (2.019)</td>
<td>2.463 (4.899)</td>
<td>2.409 (3.683)</td>
<td>9.902 (4.513)</td>
</tr>
<tr>
<td>Gender (1=F)</td>
<td>-1.399 (4.729)</td>
<td>-6.872 (4.541)</td>
<td>0.559 (5.260)</td>
<td>1.966 (4.767)</td>
<td>-2.615 (2.265)</td>
<td>-9.849 (5.300)</td>
<td>-5.296 (3.932)</td>
<td>-7.651 (4.989)</td>
</tr>
<tr>
<td>Ed. Mother</td>
<td>-0.701 (0.907)</td>
<td>0.213 (0.785)</td>
<td>-1.864 (1.152)</td>
<td>0.079 (0.807)</td>
<td>8 -0.344 (0.427)</td>
<td>0.063 (0.807)</td>
<td>1.360 (0.869)</td>
<td>0.716 (0.898)</td>
</tr>
<tr>
<td>Ed. Father</td>
<td>1.312 (1.024)</td>
<td>1.374 (0.965)</td>
<td>1.940 (1.151)</td>
<td>0.867 (0.838)</td>
<td>0.067 (0.564)</td>
<td>-1.165 (1.165)</td>
<td>-1.038 (0.765)</td>
<td>-1.744 (0.997)</td>
</tr>
<tr>
<td>Family has car</td>
<td>3.405 (4.370)</td>
<td>3.327 (4.209)</td>
<td>-4.218 (4.406)</td>
<td>8.402 (4.317)</td>
<td>0.698 (2.229)</td>
<td>-3.083 (3.714)</td>
<td>2.572 (3.758)</td>
<td>5.903 (4.137)</td>
</tr>
<tr>
<td>Computer at home</td>
<td>22.44 (5.901)</td>
<td>26.00 (6.110)</td>
<td>15.23 (6.113)</td>
<td>7.259 (6.391)</td>
<td>-1.4 (3.401)</td>
<td>-2.071 (6.858)</td>
<td>-5.839 (5.010)</td>
<td>-2.474 (7.069)</td>
</tr>
<tr>
<td>Art outside</td>
<td>0.854 (5.156)</td>
<td>0.739 (5.054)</td>
<td>-7.751 (5.179)</td>
<td>7.331 (5.552)</td>
<td>-3.489 (2.092)</td>
<td>5.677 (5.877)</td>
<td>10.58 (3.724)</td>
<td>10.85 (5.094)</td>
</tr>
<tr>
<td>Internet at home</td>
<td>-9.778 (5.261)</td>
<td>-14.05 (5.331)</td>
<td>-4.101 (5.399)</td>
<td>-2.329 (4.924)</td>
<td>4.21 (3.199)</td>
<td>-10.73 (6.071)</td>
<td>-0.745 (4.762)</td>
<td>-8.384 (6.745)</td>
</tr>
<tr>
<td>Books (=1 less than 30)</td>
<td>-11.76 (4.731)</td>
<td>-10.65 (4.777)</td>
<td>-12.59 (4.390)</td>
<td>0.941 (4.806)</td>
<td>-2.877 (1.908)</td>
<td>-14.59 (7.255)</td>
<td>-1.095 (4.289)</td>
<td>-8.195 (5.675)</td>
</tr>
<tr>
<td>Const.</td>
<td>85.36 (15.10)</td>
<td>76.61 (15.75)</td>
<td>95.02 (13.32)</td>
<td>80.75 (16.84)</td>
<td>56.92 (7.93)</td>
<td>129.8 (26.07)</td>
<td>107.2 (14.92)</td>
<td>125.2 (19.93)</td>
</tr>
</tbody>
</table>

Obs. 898
R-squared 0.30
R-squared School Dummies

Robust standard errors in brackets. *** p < 0.01, ** p < 0.05, * p < 0.1. Doubly Robust Re-weighted Regression estimation combines inverse probability weighting, in which each individual observation is given a weight equal to the inverse of the probability of the treatment he/she received conditional on baseline covariates —i.e. the estimated propensity score, as equation (4.6) shows—with standard regression.

Regression estimation combines inverse probability weighting, in which each individual observation is given a weight equal to the inverse of the probability of the treatment he/she received conditional on baseline covariates —i.e. the estimated propensity score, as equation (4.6) shows—with standard regression.
Bibliography


162


Kluve, J., 2010. The effectiveness of European active labor market programs. Labour Econ 17 (6), 904.


Leuven, E., Sianesi, B., 2015. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components.


