CALL FOR TENDERS
100001311

LONGITUDINAL STUDY OF SOCIAL AND EMOTIONAL SKILLS IN CITIES

The deadline date for the receipt of Tenders is Monday November 23, 2015 at 10.00 AM (Paris time).

INTRODUCTION

The OECD brings together the governments of countries committed to democracy and the market economy from around the world to:

- Support sustainable economic growth
- Boost employment
- Raise living standards
- Maintain financial stability
- Assist other countries' economic development
- Contribute to growth in world trade

The OECD also shares expertise and exchanges views with more than **100 other countries and economies**, from Brazil, China, and India to the least developed countries in Africa.

Fast facts

Established: 1961
Location: Paris, France
Membership: 34 countries
Budget: EUR 357 million (2014)
Secretariat staff: 2 500
Secretary-General: Angel Gurría
Publications: 250 new titles/year
Official languages: English/French

Monitoring, analysing and forecasting

For over 50 years, the OECD has provided statistical, economic and social data comparable with the most important and most reliable in the world. In addition to its collection of data, the OECD monitors trends, analysis, and forecasts economic developments. The Organisation studies changes and developments in trade, environment, agriculture, technology, taxation and more.

The Organisation provides a setting where governments can compare their experiences in developing public policies, seek answers to common problems, identify good practices and coordinate both domestic and international policies.

Enlargement and Key Partners

OECD member countries agreed to open accession discussions with Colombia and Latvia in 2013, and with Costa Rica and Lithuania in 2015.
The Organisation is also reinforcing its engagement with its Key Partners – South Africa; Brazil, China, India and Indonesia.

Publishing

The OECD is one of the world's largest publishers in the fields of economics and public policy. OECD publications are a prime vehicle for disseminating the Organisation's intellectual output, both on paper and online.

Publications are available through the Online Information System (OLIS) for government officials, through OECD iLibrary for researchers and students in institutions, corporate, subscribed to our online library and through the Online Bookshop for individuals who wish to browse titles free-of-charge and to buy publications.

BETTER POLICIES FOR BETTER LIVES

INSTRUCTIONS TO TENDERERS

ARTICLE 1 - PURPOSE AND OBJECT OF THE CALL FOR TENDERS

The OECD is issuing this Call for Tenders with a view to identifying credible partners to work with the OECD on a Longitudinal Study of Social and Emotional Skills in Cities.

The **Longitudinal Study of Social and Emotional Skills in Cities** is organised into four Cores:

- Core A: Social and emotional skills instrument development (i.e. the feasibility study),
- Core B: Background questionnaire development,
- Core C: Survey design, sampling and quality assurance of the field trial and the main study,
- Core D: Management and implementation of the field trial and the main study

Tenderers are encouraged to respond to all the four Cores in English. Tenderers may establish partnerships or a consortium with a diverse range of in-depth competencies and track records to execute the required tasks. Those tenderers responding to all Cores will detail the added benefits and price reductions that the OECD will obtain in case of contracting them all with the same supplier. Tenderers submitting proposals for one or more Cores (but not for all the four Cores) will be asked to demonstrate how their proposal will facilitate the performance of the outcomes of the Cores for which they didn't submit a proposal.

ARTICLE 2 - TERMS AND CONDITIONS OF THE CALL FOR TENDERS

2.1 Composition of the Call for Tenders

The documentation relating to the Call for Tenders includes the following parts:

a) Instructions to Tenderers and its Annex;
b) Terms of Reference and their Annexes;
c) Minimum General Conditions for OECD Contracts.

2.2 Tenders

All Tenders will be treated as contractually binding for the Tenderer and the Tenderer shall consequently issue in response to this Call for Tenders a Letter of Application dated and signed including all the provisions set out in clause 3.2 below.

2.3 Duration of Tender validity

Tenders shall remain valid for two hundred ten days (210) calendar days, as from the deadline for receipt of Tenders.

2.4 Additional information

Should any problems of interpretation arise in the course of drawing up the Tender documents, Tenderers may submit their questions to federica.darida@oecd.org and denis.elices-rejon@oecd.org , no

later than seven (7) calendar days before the deadline for the receipt of Tenders. All Tenderers will be advised of the answers given to such questions.

2.5 Acceptance and rejection of Tenders

There is no commitment on the part of the Organisation to accept any Tender or part thereof that is received in response to the Call for Tenders.

The OECD reserves the right:

- To accept Tenders with non-substantial defects
- To reject Tenders received after the deadline for receipt of Tenders, without indemnity or justification.

2.6 Modification or cancellation of Call for Tenders

The Organisation reserves the right to modify or cancel all or part of the Call for Tenders, should the need arise, without having to justify its actions and without such action conferring any right to compensation on Tenderers.

2.7 Partnerships

Partnerships must jointly meet the administrative requirements set out in the Call for Tenders. Each partner must also meet full requirements individually.

2.8 Extension of the deadline for receipt of Tenders

The OECD reserves the right to extend the deadline for receipt of the Tenders. In that case, all the Tenderer's and Organisation's rights and duties and in particular Article 2.3 above will be subject to this new deadline.

2.9 Expenses

Tenders are not paid. No reimbursement of expenses related to the preparation of any Tender will be made by the OECD.

2.10 Confidentiality

Any information communicated to the Tenderer or which come to his/her knowledge in the course of the Call for Tenders and/or the performance of the work are confidential and are strictly dedicated to the purpose of the Call for Tenders. The OECD reserves the right to request that all material be returned at the end of the Call for Tenders process.

ARTICLE 3 - PRESENTATION, SUBMISSION AND CONTENTS OF TENDERS

3.1 Tender presentation and conditions for submission

Tenders shall be entirely drafted **in English** and shall be **received** by the Organisation:

Before the deadline date of **Monday November 23, 2015 at 10.00AM (Paris time).**

- In three paper copies and one electronic version (e.g. USB Key):
- In an envelope bearing the words:

« NE PAS OUVRIR par le service courrier
Appel d'Offres n°100001311 »

To the following address:

OECD
EXD/PBF/CPG
To the attention of Federica Darida and Denis Elices-Rejon / Central Purchasing Group
2 rue André Pascal
75775 Paris Cedex 16
FRANCE

3.2 Contents of the Tender

- The Tender in **three copies and one electronic version** (e.g. USB Key);
- A Letter of Application, signed by the Tenderer, confirming the following:
  - That all the elements of the offer are contractually binding;
  - That the person signing the offer has the authority to commit the Tenderer to a legally binding offer;
  - That the Tenderer accepts all of the Minimum General Terms and Conditions without any modification. If there is an exception, please state the exception and the rationale for that exception.
  - That the Tenderer, and each of the partners in the case of a partnership, have fulfilled all its legal obligations with regards to tax declarations and payments in its home country and must supply all the requisite certificates to that effect;
- Moreover, the Tenderer shall provide, to the extent possible in accordance with the national regulations of the Tenderer, certificate(s) identifying the Tenderer, including its name, legal form, address, registration number or equivalent, date founded, areas of activity and number of employees;
- The signed Declaration detailed in Annex to these Instructions to Tenderers.

Please note that the Tenderer, ***should it be shortlisted***, will be asked to provide the following:

- Any relevant existing agreements with intermediaries or third parties;
- Financial information for the last three (3) years;
- Proof of completed legal obligations with regards to tax declarations and payments in its home country and all the requisite certificates to that effect.

3.2.1 Financial Conditions

Prices quoted must include everything necessary for the complete execution of an eventual contract (insurance, transport, guarantees). Charges for items essential to the execution of the contract and not identified in the Tender will be borne by the Tenderer.

ARTICLE 4 - INTERVIEWS

The Organisation reserves the right to organise interviews and request the Tenderers to explain in more details the content of their Tenders.

ARTICLE 5 - SELECTION CRITERIA

Main criteria for Tenderer evaluation are detailed within the Terms of Reference.

ARTICLE 6 - INFORMATION TO TENDERERS

All Tenderers will be informed, whenever possible, of the decision taken on their Tenders.

I declare having read the terms of the present instructions and agree to comply with said terms should (please insert here the name of your entity)…….be selected to carry out the Contract.

Done at:

Date: Signature:

Stamp:

Annex

Declaration for Call for Tenders n° 100001311

As part of the offer in response to the OECD call for Tenders n° 100001311, the Tenderer (company or individual) declares on oath the following:

- That it is not bankrupt or being wound up, is not having its affairs administered by the courts, has not entered into an arrangement with creditors, has not suspended business activities, is not the subject of proceedings concerning those matters, and is not in any analogous situation arising from a similar procedure provided for under national legislation or regulations;

- That it has not been convicted of an offence concerning its professional conduct by a judgment which has the force of *res judicata*;

- That it has not been the subject of a judgment which has the force of *res judicata* for fraud, corruption, involvement in a criminal organisation or any other illegal activity which may be detrimental to the financial interests of the OECD, its members or its donors;

- That it is not guilty of misrepresentation in supplying the information required as a condition of participation in this Call for Tenders or has failed to supply any relevant information;

- That it is not subject to a conflict of interest;

- That its employees and any person involved in the execution of the work to be performed under the present Call for Tenders are regularly employed according to national laws to which it is subject and that it fully complies with laws and regulations in force in terms of social security and labor law;

-That it has not offered and will not offer, has not granted and will not grant, has not sought and will not seek to obtain, and has not accepted and will not accept any advantage, financial or in kind, to or from any party whatsoever, constituting an illegal practice or involving corruption, either directly or indirectly, as an incentive or reward relating to the award or the performance of the contract that would result from the OECD call for Tenders n 100001311.

I, the undersigned, …….. on behalf of the company ……., understand and acknowledge that the Organisation may decide not to award the contract to a Tenderer who is one of the situations indicated above. I further recognise that the Organisation may terminate for default any contract awarded to a Tenderer who has been found guilty of misrepresentation in supplying, or has failed to supply, the information required as a condition of participation in this Call for Tenders. Finally I understand and acknowledge that the Organisation may inform any third party, including its members and donors in case a Tenderer is in one of the above mentioned situations or when should it be found guilty of making false declarations, committing fraud, or to be in serious breach of its contractual obligations.

The .. / .. / ..

Signature

TERMS OF REFERENCE

TABLE OF CONTENTS

TERMS OF REFERENCE

1. This document presents the Terms of Reference for an International Contractor to support the development and implementation of the OECD's Longitudinal Study of Social and Emotional Skills in Cities. The OECD Secretariat is seeking a competent contractor with track records to drive the preparation of survey instruments to measure social and emotional skills over time among school-aged children (Grades 1-12). The successful contractor will also develop and validate measures of learning contexts and outcomes that are closely related to social and emotional skills. Once the instruments have been validated and other key survey parameters defined, the Contractor will engage in the preparation and implementation of the main longitudinal study which is scheduled to take place from 2020 onwards.

2. This study will be the first international longitudinal study designed to track the development and outcomes of social and emotional skills over a long period of time. As the evidence-base and measurement technologies are relativelty limited in this area, the successful contractor must be able to: (a) demonstrate the capacity to employ creative yet practical solutions to measurement challenges, (b) undertake and manoeuvere risks and uncertainties and (c) possess strong networks of diverse technical expertise from different disciplines and countries. The Terms of Reference (including Annex A and B) includes some descriptions of initial ideas related to survey design and assessment methodologies. Bidders with alternative ideas are encouraged to present them with supporting arguments.

3. When developing the proposal, bidders may assume 5 major cities, one from East Asia, two from Europe, one from North-America and one from South America, as core participants of this study. The OECD Secretariat estimates up to seven additional cities to engage in this study.

4. This study is designed to follow two cohorts of children, i.e. those starting from Grades 1 and 7, until early adulthood (around age 25). Hence, the Contractor will be asked to develop and implement robust strategies to reduce survey attrition over a long period of time. For the purpose of this Call for Tenders, bidders are asked to submit proposals and budgets that cover the first 3 years of the main study, i.e. until 2022.

5. Given the diversity and high-levels of expertise required to successfully execute the tasks outlined in this Call for Tenders, bidders may consider establishing a consortium of contractors, with a lead contractor assuming overall managerial responsibility for all the tasks involved. In this case, participating contractors will be required to establish a clear and effective co-ordination strategy. The Contractor(s) will be required to seek national expertise from the participating cities; manage the flow of information and decision making process; work with National Project Managers (NPMs) on project implementation; and, build capacity of the participating countries, particularly the National Centres.

TERMS OF REFERENCE

SECTION 1: OVERVIEW

Introduction

6. The OECD is scheduled to launch the Longitudinal Study of Social and Emotional Skills in Cities from 2019/20 onwards. There will be approximately 4 years of developmental work, including the feasibility study which is designed to develop and validate social and emotional skills instruments. The main longitudinal study will follow the lives of two child cohorts (Grades 1 and 7) by measuring social and emotional skills, learning contexts and socioeconomic outcomes over time. The data collected will be used, in the short-term, to assess the distribution of social and emotional skills and to identify learning contexts that are associated with the development of these skills. In the long-term, the data will be used to identify the sensitive period of socio-emotional skills development and the types of these skills that help improve children's economic and social prospects. Box 1 summarises the OECD's initial considerations of the main characteristics of the study. The rest of this section presents an overview of the proposed study.

Box 1. OECD's Longitudinal Study of Social and Emotional Skills in Cities (initial considerations)

- **Objectives**	To identify the drivers and consequences of social and emotional skills
- **Survey delivery**	Schools (students and teachers) and home (parents)
- **Target cohorts**	Children in Grades 1 and 7 (approximately ages 6 and 12, respectively)
- **Survey coverage**	Cities (with an option of state or nation-wide coverage)
- **Sampling method**	Random selection of schools. Full sampling of grade 1 and 7 cohorts within schools
- **Cycle**	Annual data collection of target cohorts
- **Duration**	Minimum of 3 years. Ideally until both cohorts reach adulthood (or age 25)
- **Measures of skills**	Social and emotional skills
- **Measures of contexts**	School, family and community learning contexts
- **Measures of outcomes**	Education, labour market, health, civic engagement, violence, life satisfaction, etc.
- **Timeline**	Developmental work: 2015-19; Main study: 2019/20 onwards

Objectives

7. The OECD's Longitudinal Study of Social and Emotional Skills in Cities aims at:

- Identifying the **social and emotional skills** that drive children's future outcomes, including educational attainment, labour market, health status, relationships and civic engagement;
- Better understanding how **investments made by families, schools and communities** influence the development of social and emotional skills; and,
- Developing **recommendations and measurement tools** for policy-makers and practitioners to better monitor and enhance social and emotional skills.

Rationale

8. ***Why do we focus on social and emotional skills?*** It is difficult to imagine any policy-maker, teacher, let alone parent, who are not convinced about the importance of children's character, personality

and socio-emotional abilities for their future success. However, there are those who still hold that children's social and emotional skills are largely set upon before early childhood, and that these skills are difficult to conceptualise and measure meaningfully. While assessing socio-emotional skills is indeed highly challenging, our conceptual understanding and measurement technologies around socio-emotional skills have improved over the past decade. There is also an increasing enthusiasm across diverse education stakeholders regarding the potential role socio-emotional skills can play in improving our education system and societal well-being. Now is a good moment to build on these interests and methodological progress by launching an international study on social and emotional skills.

9. ***Why do we need a longitudinal study?*** Such a question naturally arises given that a longitudinal study often demands a significant amount of time and financial resources to prepare and run. The reason is due to our main study objective, which is to identify the specific learning contexts that drive socio-emotional skills formation and improve children's life outcomes. To this end, it is important to accurately and repeatedly measure children's socio-emotional skills, learning contexts and diverse measures of socioeconomic prospects and well-being over time. One may argue that a cross-sectional survey would suffice to address our study objectives. Such surveys can play an important role in providing policy-makers with detailed information on the distribution of socio-emotional skills within/across countries, and how they are associated with demographic background, socio-economic circumstances, learning contexts and policies. These surveys are, however, not ideal for examining the process of skills formation; the role learning contexts play in this dynamic process; and, the medium/long-term outcomes of socio-emotional skills. Longitudinal surveys of skills already exist in a few OECD countries. However there is no international survey that is specifically designed to examine the developmental process of socio-emotional skills along with diverse measures of learning contexts and outcomes.

10. ***Why do we propose following children from Grades 1 and 7 until they finish schooling?*** Mid-childhood (6-12 years) and adolescence (13-18 years) are considered appropriate time periods to study social and emotional skills development as these are the periods during which meaningful changes in skills are experienced while exposed to policy-relevant learning contexts (e.g., formal schooling). The choice of Grades 1 and 7 is based on the fact that they correspond to the beginning of primary and lower-secondary schools in many OECD countries. Simultaneously starting an assessment of Grade 7 cohort would also allow countries to benefit from understanding the impact of socio-emotional skills on adult outcomes within a shorter time-span.

11. ***Why do we need an international study?*** Participating cities would benefit from understanding how their children fare in terms of the development of socio-emotional skills *via-a-vis* children from other major cities. Cities would also benefit from learning what other high-performing cities are doing to improve certain socio-emotional skills of high policy priority. For instance, this study will, shed light on how certain learning/policy contexts help enhance socioeconomic prospects of disadvantaged children through improved levels of children's perseverance, self-esteem and social skills. Lastly, cities and countries that do not have a tradition of running longitudinal studies would also benefit from developing or improving their expertise on longitudinal survey design and administration.

Background

12. The OECD Education and Social Progress (ESP) project dedicated the past 3 years developing a conceptual framework to better understand the dynamics of skills formation and their socioeconomic outcomes. This involved conducting extensive literature reviews; launching empirical analyses of existing

longitudinal data in 9 OECD countries; evaluating available measurement instruments; and, synthesising existing policies and practices. The outcome of this work is summarised an international report titled "Skills for Social Progress: the Power of Social and Emotional Skills", published in March 2015.

13. While so much have been learned from the past efforts made on this topic by researchers and educators around the world, one of the main conclusions from the OECD's conceptual work is that we still know very little about:

- The nature of education policies and practices that work in enhancing social and emotional skills.
- The types (and the combination) of social and emotional skills that drive children's lifetime success.
- Robust instruments to measure social and emotional skills.

14. The OECD's Longitudinal Study of Social and Emotional Skills in Cities will be launched to shed light on these issues and to provide better knowledge-base to policy-makers, practitioners and researchers.

15. Designing and launching an international longitudinal study of skills is a highly complex enterprise. This is in part due to difficulties in ensuring that the resulting multi-year panel data render rich empirical analyses with maximum information power. The OECD therefore proposed an extensive preparatory stage including: (a) the development of a conceptual framework of social and emotional skill (see Annex A); (b) a Feasibility Study specifically designed to develop and validate social and emotional skills assessment instruments; (c) a pilot study to test the background questionnaire (i.e., learning contexts and outcomes); and, (d) a field trial to test the whole survey procedures.

Concepts

Social and Emotional Skills

16. Social and emotional skills are the kind of skills involved in working with others, achieving goals and managing emotions. As such, they manifest themselves in countless everyday life situations. Children and adults live in a highly interconnected world in which, 'who you know' and 'how you interact' matter critically. Children start pursuing goals from a very early age (e.g., playing games, solving puzzles) and this becomes ever-more important during adulthood (e.g., pursuing academic degrees and jobs). Capacity to regulate positive and negative emotions and managing stress and frustration play an indispensable role when dealing with life changes such as unemployment, divorce and long-term disabilities.

17. The OECD proposes a framework of social and emotional skills that takes into account the most recent conceptual and empirical literature on these type of skills (see Annex A). Social and emotional skills are defined as "individual capacities that (a) are manifested in consistent patterns of thoughts, feelings and behaviours, (b) can be developed through formal and informal learning experiences, and (c) can influence important socioeconomic outcomes throughout the individual's life. The scope of skills is limited to those that are malleable (i.e. they can be learnt) and relevant for diverse future outcomes (e.g. education, employment, healthy behaviours, civic participation, etc.).

18. At the highest level of abstraction, socio-emotional skills constructs can be classified into five broad domains (see Annex A):

- Emotional regulation (emotional stability)
- Engaging with others (extraversion)
- Collaboration (agreeableness)
- Task performance (conscientiousness)
- Open-mindedness (openness)

19. These five domains are subdivided into several more narrowly defined constructs labelled facets. Annex A provides a definition of constructs and facets, as well as a rationale for the dimensions selected, and the coherence with other frameworks and educational goals.

Outcomes

20. One of the most important goals of education is to help children achieve the highest level of well-being possible. The framework for individual well-being and social progress envisioned here is in line with the OECD Framework for Measuring Well-Being and Progress, which emphasises the broad spectrum of outcomes relevant in the modern world. They include education, labour market outcomes, health, life satisfaction, family life, relationships, civic engagement, safety and environmental outcomes. This study will measure indicators that are age-appropriate and that can be reliably measured and analysed. A preliminary list of possible outcomes is presented below.

- ***Education:*** educational attainment, academic grades, grade repetition and truancy.
- ***Labour market:*** work status (e.g. employment, unemployment, looking for job), type of occupations and earnings.
- ***Health:*** behaviours (e.g. exercising, visiting the doctor regularly and risky behaviours) and outcomes (e.g. body mass index, self-reported health status and depression).
- ***Civic engagement:*** volunteering, voting and interpersonal trust.
- ***Violence:*** bullying, violent acts, and criminal activities (e.g. personal theft, vandalism and assault).
- ***Family and social connections:*** single parenthood; family breakdown; teenage pregnancy; contact with, and support from family and friends.
- ***Subjective well-being:*** life satisfaction, experiences of stress and other measures of subjective happiness.
- ***Environment:*** individual's pro-environmental behaviours, such as recycling, use of public transportation.

- ***Material conditions:*** income, assets, consumption and housing.

Learning contexts

21. Learning takes place in a variety of social settings such as: school, family, the community and the workplace. Within each type of context, we can distinguish a number of specific elements, with examples presented in Figure 1. Each context are expected to contribute to the development of social and emotional skills, however their relative importance will change depending on the individual's stage in life. For instance, parents are clearly crucial during infancy and early childhood, but school and the community can become increasingly important as a child enters formal education and interacts with diverse social networks. The workplace, in turn, can be a key learning context particularly during late adolescence and (early) adulthood.

Figure 1. Framework of learning contexts

Source: OECD (2015), Skills for Social Progress: The Power of Social and Emotional Skills, OECD Skills Studies, OECD Publishing, Paris.

22. The impact of learning contexts on skills can be divided into direct inputs, environmental factors and policy levers (Table 1). They represent different ways in which schools, parents, workplaces and communities can shape skills. Direct inputs intentionally and explicitly affect skill development; for example, parental involvement in child-rearing activities. Environmental factors, on the other hand, influence skill development indirectly by providing the context in which skills can develop; for instance, the civic and cultural activities available to a child growing up in a particular community. Policy levers, on the other hand, are the elements of a learning context which are directly malleable by policy making and

can be used to foster skill development; for instance, teacher training, which informs teachers' approaches to teaching social and emotional skills. These learning contexts do not function in isolation from each other; rather they constantly interact and mutually influence each other. In fact, the patterns of interactions between contexts can themselves be related to the development of skills.

23. This study will look into diverse learning inputs: not only ones that are formal (e.g., classroom), but also non-formal (e.g., extra-curricular activities) and informal (e.g., peer interactions and media) in nature. It is also important to address diverse learning contexts since key learning inputs are likely to vary by type of social and emotional skills. For instance, the ability to 'work with others' is likely to be more strongly enhanced through frequent interactions with peers, rather than teacher's classroom instructions.

Table 1. Examples of direct inputs, environmental factors and policy levers

	Family	School	Workplace	Community
Direct inputs	Parenting styles (e.g. democratic parenting, allowing autonomy, sensitive parenting); time use (e.g. explicit teaching time with a child)	Teaching styles (stimulating teamwork, disciplining) Classroom climate	Work-based training targeting skills development; management styles (e.g. supervisors nurturing skill development)	Activities offered in the community (e.g. art classes in cultural centres, sports associations)
Environmental factors	Availability of learning aids, technology in the household Trauma (negligence, malnutrition, abuse)	School resources, school safety	Workplace resources	Neighbourhood safety
Policy levers	Parents' employment, parental leave systems, out-of-school-hours care services	Teacher training, curriculum and recruitment	Apprenticeship systems	Training of social workers, cultural agents

Source: OECD (2015), Skills for Social Progress: The Power of Social and Emotional Skills, OECD Skills Studies, OECD Publishing, Paris.

Cycle

24. The OECD Secretariat proposes an annual data collection of skills, learning contexts and outcomes in order to precisely analyse the dynamics of social and emotional skills development. One could also argue for biennial/ triennial data collection, if these skills do not change much over time. To date, we are not aware of any evidence as such. Moreover, there are several reasons why annual data collection can be considered appropriate. First, even if the change in skills is small, collecting annual measures of skills has the advantage of providing means to minimise measurement errors when treating the data. Second, it is likely that annual assessments would increase the likelihood of maintaining contact with survey respondents over time. Third, annual data collection will help collect an accurate information on learning contexts (particularly school and home learning contexts) that children experience during the year. The feasibility study will be able to inform on whether this option should be the way forward for the main study.

TERMS OF REFERENCE

Sample

25. The study will be conducted at a city level with an option of a state or country wide coverage. A city is considered an optimal geographical boundary for addressing the study objectives, providing several advantages over a nationally representative sample. First, it is likely to be easier and less costly to conduct a longitudinal study at a city/state/province level than at a national level, especially in countries with a federal system. Second, some local jurisdictions may be more capable of engaging in a long-term project that requires sustained political and financial commitments. Countries are also welcome to opt for conducting the study at the national or federal level. The criterion for cities to be eligible to participate is that within their country they have to be a major city in terms of population and/or size of the economy. Selecting a big city has several advantages compared with selecting smaller locations, including making easier comparisons between different participating cities.

Platform

26. Computer-based survey delivery platform is increasingly mobilised in skills assessments. In PISA 2015, all the new test items and background questionnaires, except the optional parent questionnaire, were delivered on computer. In light of the progress made in assessment technologies and the longitudinal nature of this study, there is a considerable benefit from employing a computer-based assessment. For parental, teacher and school administrator questionnaires, computer-based or paper-and-pencil based delivery may be considered.

Deliverables

27. The following provides a list of key outputs from the main longitudinal study:

- *Social and emotional skills indicators* that describe the types, levels, distributions and malleability of relevant social and emotional skills in their country.
- *Learning context indicators* that provide insight into how schools, families and communities drive the development of such skills.
- *Recommendations and measurement tools* for policy-makers, school administrators, practitioners and parents to better monitor and promote social and emotional skill development.
- The above mentioned deliverables will be presented in *international reports* to be produced after each survey cycle. Assuming annual survey cycle, the OECD Secretariat anticipates producing short international reports after the first and second survey cycle. This will mainly describe levels and distributions of socio-emotional skills and their associations with learning contexts and outcomes. The Secretariat plans to prepare an in-depth international report after three survey cycles by fully exploiting the longitudinal aspect of the data and drawing policy implications.

Duration

28. The OECD proposes the main study to be continued until both cohorts reach early adulthood (or age 25) in order to maximise information power and usefulness for policies and practices. Nevertheless, political and financial circumstances in participating cities/countries may hinder long-term sustainability of

this study. In light of these risks and uncertainties, the study will provide policy-relevant outputs in the short-term, i.e., in 1-3 years. This includes the analyses of a) the levels and distribution of social and emotional skills, b) the relationship between learning contexts/education policies and skills formation and c) the relationships between skills and early social outcomes – e.g., health-related or antisocial behaviours. In the long-term, the study will provide analyses of the impact of education policies and skills on adult educational, economic and social outcomes. Policy recommendations and measurement tools will be developed progressively during the course of the longitudinal study.

The Feasibility Study

29. The initial stage of the preparatory work will be dedicated to developing and validating measures of social and emotional skills for school-age children (Grades 1 to 12). The goal is to ensure that the proposed measures not only demonstrate relevant construct coverage, construct validity, measurement reliability and robustness across ages, but also cross-cultural and cross-linguistic validity. The feasibility study is also expected to improve our understanding on the malleability of social and emotional skills. This will give us ideas of the optimal frequencies of assessing each of the social and emotional skill measures.

Target population

30. The Secretariat proposes a feasibility study that covers the whole school-cycle (all grades: from grade 1 to grade 12) given the objective of the longitudinal study to study the developmental trajectory of social and emotional skills during children's school years. This approach would help identify upfront which relevant skills can be reliably measured between Grades 1 and 12.

Deliverables

31. The proposed deliverables of the feasibility study are as follows:

- A *validated conceptual framework* of social and emotional skills of school-aged children;
- *Instruments* to assess social and emotional skills of school-aged children;
- *Cross-sectional datasets* with information on social and emotional skills and basic demographics;
- An *instrument validation report*;
- The above mentioned deliverables will be presented in an *international report* which includes recommendations on the type of instruments that can be used in an international survey to study social and emotional skills of school-age children.

Governance, Management and Implementation Entities

32. This study involves a number of governance, management and implementation entities as described in this section. They include the CERI Governing Board and the group of representatives of the Longitudinal Study of Social and Emotional Skills in Cities which will be responsible for approving and driving the development of the proposed study which will be centrally managed by the Contractor under close guidance of the OECD Secretariat and the Technical Advisory Group. The Contractor will be

responsible for co-ordinating the activities of the National Centre which will administer the proposed survey in each city.

CERI Governing Board

33. The OECD Centre for Educational Research and Innovation (CERI) Governing Board is the governing body for all CERI activities, including the Longitudinal Study of Social and Emotional Skills in Cities. The CERI Governing Board is composed of representatives of OECD and partner countries. The OECD Secretariat will regularly report to the CERI Governing Board on the progress made in the preparation and implementation of this study. This includes updates on the survey design, survey instruments, data collection and preliminary analysis. The CERI Governing Board will (a) endorse the development and implementation of this study, (b) monitor the quality and timeliness of activities and outputs, and (c) enable representatives of member countries to be fully informed of all aspects of this study.

OECD Secretariat

34. The OECD serves as the Secretariat of the CERI Governing Board. The OECD Secretariat will assume overall managerial responsibility for this activity, and closely monitor the preparation and implementation of the proposed study. The Secretariat's responsibilities entail: (a) preparing the terms of reference for each study cycle under the guidance of the CERI Governing Board; (b) engaging and monitoring Contractors for quality assurance purposes; (c) building consensus at the policy level among participating cities, (d) updating the CERI Governing Board and the group of representatives of this study with progress of project, financial and contractual management, (d) ensuring decisions of the CERI Governing Board and the group of representatives of this study are implemented, (e) ensuring that risks are regularly monitored and appropriately mitigated, (f) monitoring budgets and milestones of the Contractor and resolving budgetary or contractual issues, (g) ensuring that the Contractor is kept fully informed of any decisions which impact on project structure or timelines, and (h) providing support to National Centres and National Project Managers (see below).

35. The OECD Secretariat produces indicators and analyses based on verified data sets and analytical outputs provided by the Contractor. Based on this information, the OECD Secretariat will prepare an international report that summarises the results of the indicators and analyses.

Group of representatives of the Longitudinal Study of Social and Emotional Skills in Cities

36. The group of representatives of the Longitudinal Study of Social and Emotional Skills in Cities will be responsible for reporting to the CERI Governing Board for strategic advice and decisions. The role of this body is to guide the development and implementation process of the longitudinal study. This body will be composed of representatives of participating countries and cities (or of other relevant jurisdictions). This body is scheduled to meet twice a year in order to facilitate the preparation of the study.

Participating cities

37. Each participating city will be asked to establish a **National Centre** which is a local co-ordination body. The National Centre includes representatives of the relevant local education board,

Ministry of Education and/or education agencies, and research institution(s) that will be in charge of local implementation of the longitudinal study. Participating cities shape and guide the project as follows:

- As core members of the implementation body, they help to guide the design parameters for the project within the context of the proposed framework and governance arrangements.
- Through the National Centres and in collaboration with the Contractor, they implement the project at the city level subject to agreed-upon administrative procedures.
- Through the National Centre and relevant authorities, they provide inputs for the design of the analytical outputs and the content of the report that reflect the policy priorities of the cities.

38. Each participating city will nominate a **National Project Manager (NPM)** who will be responsible for driving the whole survey process within the city. NPMs are the primary means of day-to-day contact between participating cities and the Contractor. They shall communicate with the Contractor on all issues related to the implementation of the assessments in their city. NPMs play a vital role in maintaining the quality of the project with results that can be verified and evaluated. They can also play an important role in the development and review of reports and publications, in consultation with their respective country and city representatives of this study.

39. The NPMs decide how best to facilitate the co-ordination needed at the city/national level during data collection, analyses and reporting. This includes interactions with the Contractor. The NPMs in each participating city is expected to develop a Project Implementation Plan (PIP), the principal tool for project management, in collaboration with the OECD. The PIP will be progressively revised within each participating city as activities develop and more input and information is available.

Technical Advisory Group

40. The role of the Technical Advisory Group (TAG) is to provide objective guidance to the OECD Secretariat on technical issues around survey design, instrument development and sampling procedures proposed by the Contractor. TAG will be appointed by the OECD Secretariat in consultation with the CERI Governing Board and the group of representatives of this study. Some members of the group may remain constant during and after the feasibility study, and new members may be appointed as required. TAG will be managed by the OECD Secretariat.

Contractor's relationship with different entities

41. The Contractor will act as the focal-point of this study to ensure that the project plans approved by the CERI Governing Board and the group of representatives of this study will be implemented successfully, in a timely manner and within the budget envelope. The Contractor will lead and guide the works of the National Centres by developing operational strategies, implementing quality control procedures, closely monitoring the progress of the local work and providing timely advice and guidance. The OECD Secretariat works very closely with the Contractor to ensure quality, timeliness and relevance of its work and that it fully respects the needs and constraints of participating cities. To this end, the Contractor will regularly discuss with the Secretariat on the detailed proposal, work progress, budgets and any survey related questions and concerns transmitted by the CERI Governing Board, the group of representatives of this study and the Technical Advisory Group (TAG). Should there be more than one

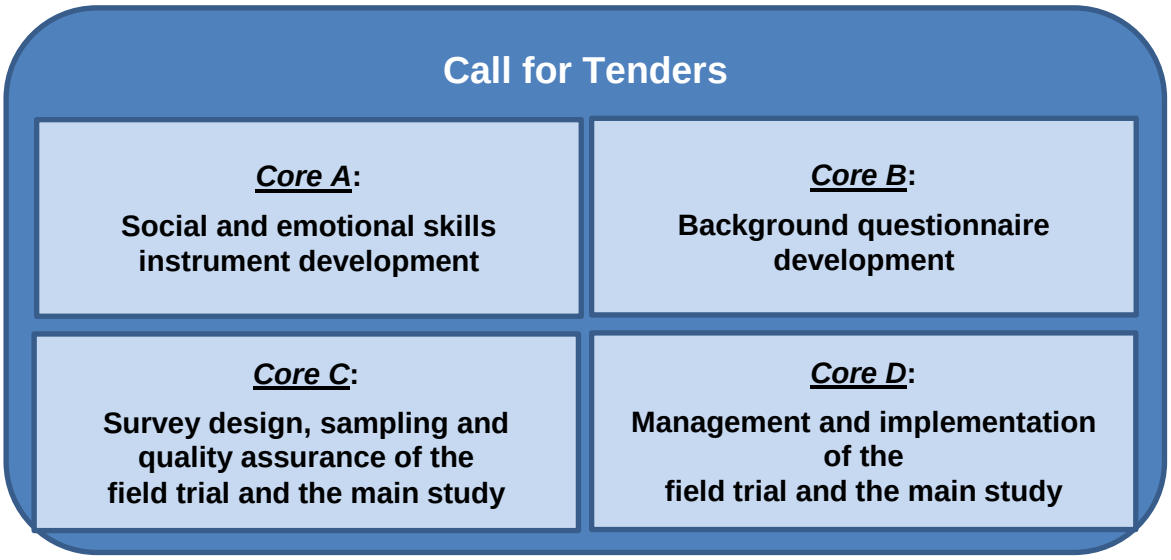
contractor (or consortium) working on Cores B, C and D, the contractors should closely liaise as all the work elements are interconnected. See Section 2 for more details on intra-Core co-ordination.

TERMS OF REFERENCE

SECTION 2: ORGANISATION OF THE CALL FOR TENDERS

42. The Organisation for Economic Co-operation and Development (OECD) invites proposals for the development and implementation of the Longitudinal Study of Social and Emotional Skills in Cities. The terms of reference covers *Core A, Core B, Core C* and *Core D* (see Figure 2).

Figure 2. Structure of the Call for Tenders



43. The Contractor's tasks for each of the Cores are outlined below:

Core A: Social and emotional skills instrument development (i.e. the feasibility study)

Task 1: Develop an assessment strategy and validated instruments

- Draft an assessment strategy.
- Draft the items.
- Analyse and validate the items.

Task 2: Ensure quality of translations of instruments

- Prepare a guideline for translations and adaptations to be used by the National Centre.
- Ensure the quality of translations of instruments into national languages.

Task 3: Prepare a computer-based survey platform

- Develop a computer-based platform.

- Prepare training materials to facilitate NPM's usage of the survey delivery platform.

Task 4: Undertake project management

- Appoint a project director.
- Facilitate the implementation of the agreed project management approach.
- Establish and manage an expert group.
- Support the work of National Project Managers (NPMs).
- Prepare a technical report of the feasibility study.
- Support the OECD Secretariat in preparing an international report.
- Document the data-base.

Core B: Background questionnaire development

Task 1: Develop a conceptual framework and validated instruments

- Develop a conceptual framework.
- Draft the instruments.
- Analyse and validate the instruments

Task 2: Ensure quality of translations of instruments

- Prepare a guideline for translations and adaptations to be used by the National Centre.
- Ensure the quality of translations into national languages.

Task 3: Develop a survey platform

- Develop a survey delivery platform if computer-based assessment will not be fully employed for the background questionnaire.

Task 4: Undertake project management

- Appoint a project director.
- Facilitate the implementation of the agreed project management approach.
- Establish and manage an experts group.
- Support the work of National Project Managers (NPMs).
- Prepare a technical report of the background questionnaire validation study.
- Document the data-base.

Core C: Survey design, sampling and quality assurance of the field trial and the main study

Task 1: Identify longitudinal design features and sampling strategy

- Define key longitudinal design features.

- Develop a sampling strategy.

Task 2: Develop and support quality control procedures

- Develop quality control procedures.
- Support National Centres and NPMs.

Task 3: Develop technical standards and sampling guidelines

- Develop technical standards for the main study.
- Develop sampling guidelines for the field trial and the main study.
- Prepare sampling weights for the field trial and the main study.

Core D: Management and implementation of the field trial and the main study

Task 1 Operate the field trial and the main study

- Undertake project management.
- Support the work of the NPMs.
- Establish plans for monitoring adherence to the technical standards.
- Develop methods to deal with ethical issues.

Task 2: Process micro-data, analyse, scale and report

- Clean all collected data and conduct analyses of the field trial and the main study.
- Provide a fully documented database which will allow the OECD Secretariat and participating cities to conduct their own analyses.
- Finalise instruments for the main study.
- Develop an analysis and reporting plan, which will guide the OECD Secretariat in preparing and designing international reports.
- Develop technical reports.

Core A: Social and emotional skills instrument development (i.e. the feasibility study)

44. The Contractor will be asked to develop and validate instruments based on the conceptual framework presented in Annex A. Bidders that would like to mobilise an alternative conceptual framework are requested to provide supporting arguments and detailed explanations.

45. This is a critical stage of the development work where various assessment innovations can be explored, tested and validated. The value-added of proposed innovations should however be weighed against their likely consequence on costs. Bidders interested in exploring multiple assessment methods are requested to provide separate cost estimates for each of the suggested methods to facilitate the comparisons of value-added from different bidders.

Task 1: Develop an assessment strategy and validated instruments

46. The Contractor will be asked to:

- *Develop an assessment strategy.* This will involve evaluation of existing instruments for measuring social and emotional skills and sampling plans. Bidders are invited to propose the types of instruments (e.g., self-reports) that could be explored in the feasibility study, and present an outline of the assessment strategy to rigorously validate the instruments while taking into account financial and survey time constraints.

- *Draft the items.* This stage will involve intensive interactions between the participating cities, the group of representatives of this activity, the OECD Secretariat and the experts group (see Task 4 below). The Contractor will be asked to prepare all instruments (including scoring guides) in English. Translation from English into other languages will be handled by the participating country (apart from the quality assurance described in Task 2 below). The Contractor will prepare Item Submission Guidelines for the experts to develop the Item Pool and all other associated materials including scoring guides and training materials for scoring. The Contractor will also develop measures to collect basic background information that can be used to validate the social and emotional skills instruments. The Contractor may be called on to develop and implement additional tests and questionnaires which the CERI Governing Board and the group of representatives of this study may decide (in 2016) to include in the feasibility study. Additional questions may relate to child outcomes such a bullying, engagement in risky behaviours and life satisfaction. Bidders are invited to outline the process of item development.

- *Analyse and validate the items*. The Contractor will be asked to clean all data collected. Bidders are invited to indicate the types of checks that they will carry out on the data, and the mechanisms which will be put in place to ensure that necessary checks are also carried out by National Centres. The Contractor shall subsequently conduct psychometric analyses to investigate the measurement properties of the instruments; and carry out other necessary analyses to inform the selection of items and methods that can be employed during the field trial and the

main study. Bidders are invited to outline the range of psychometric analyses to be performed. The Contractor will also be responsible for conducting the training and standardisation of the scoring of open-ended items, responding to queries from countries during the marking process, and checking the reliability of scoring at the international level. Bidders are also invited to outline how they plan on organising this part of the work.

Task 2: Ensure quality of translations of all the instruments

47. The Contractor will be asked to:

- *Prepare a guideline for translations and adaptations to be used by the National Centre.* Countries will be responsible for translating the English version into their own language. The Contractor will be asked to work with National Centres to ensure that the translations are of a quality which will ensure cross-national comparability of the assessments. The Contractor will be asked to prepare a guideline for translations and adaptations to be used by the National Centre.

- *Ensure the quality of translations of instruments into national languages.* The Contractor will also review the technical standards regarding the translation of assessment instruments, questionnaires and manuals of the feasibility study, to be considered and adopted by the group of representatives of this study.

Task 3: Prepare a computer-based survey platform

48. The Contractor will be asked to:

- *Develop a computer-based platform.* The Contractor will be requested to develop a computer-based survey delivery platform for assessing social and emotional skills. This platform will also house a limited set of instruments to collect learning contexts, outcomes and background information, to be subsequently used for validation purposes. The Contractor will be asked to prepare protocols and standards for instrument development to ensure that a single delivery platform can be used for collecting information on socio-emotional skills, learning contexts and outcomes. Bidders who would like to propose an alternative survey delivery platform are asked to outline the methodology and discuss its advantages over computer-based platform. The OECD will, in principle, keep all the intellectual property rights within the survey delivery system. Bidders with different views are invited to specify aspects of the intellectual property (e.g., certain instruments, delivery platform, data and analytical results) that will involve different arrangements. Bidders are asked to make clear their position regarding the intellectual property, describe implications of their proposed solutions and clarify where third party rights are being used and therefore cannot be assigned to the OECD.

- *Prepare training materials to facilitate NPM's usage of the survey delivery platform.* Bidders are invited to outline how they will provide support services to National Centres.

Task 4: Undertake project management

49. The Contractor will be asked to:

TERMS OF REFERENCE

- *Appoint a project director*. The project director will act as an overall leader of the work as well as providing leadership for National Project Managers (NPMs), and to this end should have strong management and team-building skills. The project director will work closely with the OECD Secretariat and attend the meetings of the group of representatives of this activity to present updates on project activities. These meetings are scheduled to take place twice each year and are generally hosted by the participating countries. The Contractor will be responsible for covering travel, accommodation and subsistence expenses for either their own personnel or members of expert groups who attend these meetings, and should describe the extent of such attendance they have assumed in their budget. The person in this role should also have sufficient track record to provide the intellectual leadership with experts, and to work with the Secretariat in identifying technical issues to be discussed by the Technical Advisory Group (TAG). Bidders are asked to specify the project director as well as the percentage of time he/she will spend on this part of the project.

- *Facilitate the implementation of the agreed project management approach.* This will involve (a) developing and maintaining a project plan and timeline that joins the works in a coherent and cost-efficient way, (b) negotiating and resolving timeline amendments, for example those which might arise from unanticipated requests from the CERI Governing Board and the group of representatives of this study, unavoidable operational delays or other unforeseen project changes, (c) keep the OECD Secretariat fully updated on amendments and any timeline issues which cannot be resolved or which may have implications for achievement of project milestones, (d) establish procedures for monitoring and managing risks, (e) establish mechanism for submission of all documents, materials and databases to the OECD archive, (f) discuss additional requests from participating cities, (g) negotiate and co-ordinate additional national requirements or requests with the Secretariat and with National Centres or the group of representatives of this study of this activity as appropriate, and (h) provide regular progress reports to the OECD Secretariat and the group of representatives of this study. The nature and frequency of such reports will be agreed between the Contractor and the Secretariat.

- *Implement survey operations.* The contractor will develop and implement survey operations and related aspects of quality control, including the development of test and questionnaire administration procedures, the development of scoring procedures, and the training of all necessary and relevant city representatives in these procedures (e.g., NPMs, test administrators). The contractor shall develop all related training materials and procedures in consultation with the TAG and the OECD Secretariat. All training materials shall be developed in English. The contractor will develop and monitor procedures to ensure technical standards are met. The Contractor will establish frequent communication with NPMs, and formalise survey operations in instruction manuals so that National Project Managers and/or other local institutions have the required guidance for survey implementation.

- *Establish and manage an expert group*. The Contractor will be asked to establish an experts group and lead the technical discussions during the feasibility study. The Contractor and the OECD Secretariat will jointly determine members of the experts group. Bidders are asked to describe the number of expert group meetings they have included in their proposed budget, and to explain how they would call on the expertise of group members outside the formal meetings. Bidders should include in their cost proposal all expenses associated with holding expert group meetings for which they are responsible, such as conference venue, travel, accommodation,

subsistence and honorariums to expert group members. A member of the OECD Secretariat team will generally attend meetings and the Secretariat will cover their own associated costs. The bidder is invited to submit a proposed list of experts.

- *Support the work of the National Project Managers (NPMs).* The Contractor will be asked to develop a description of the role and profile of the NPMs and specify the Contractor's working relationships with NPMs. The Contractor will be asked to support the NPMs with the implementation of the feasibility study. This includes providing and maintaining tools for the NPMs to track progress with the implementation of the tasks involved with the survey in each country and to keep track of any potential problems with cities' abilities to meet project timelines or technical standards. The Contractor will be asked to organise and host meetings of the NPMs. Provisions for meeting venues and facilities as well as for travel and compensation of experts, as required, should be included in bidders' proposals. No compensation of travel costs for the NPMs or representatives from the OECD Secretariat should be included in the cost proposal. Participating cities will bear the costs of their NPMs' participation in these meetings. Bidders are invited to outline procedures to support the NPMs and propose frequency of NPM meetings.

- *Prepare a technical report of the feasibility study.* This report will be designed to validate the conceptual framework and provide recommendations on the appropriate instruments and methods that can be used to assess social and emotional skills from grade 1 to12. The Contractor is expected to deliver the final drafts of the report no later than 31 December 2018, in line with an outline agreed with the OECD Secretariat. Upon feedback from participants and the Secretariat, the Contractor is expected to provide the necessary revisions that will allow the Secretariat to finalise the production of the report by 31 July 2019. Note that participating cities may wish to learn from preliminary analysis based on the feasibility study data. Bidders are invited to outline the main contents of this technical report and present preliminary ideas for the types of analysis that could be conducted and the policy insights that could potentially be drawn from such an analysis which can be presented in the technical report.

- *Support the OECD Secretariat in preparing an international report.* The OECD Secretariat will be responsible for the preparation of an international report of the feasibility study. The report will (i) present the validated conceptual framework of social and emotional skills, (ii) provide evidence on the feasibility of assessment delivery, and (iii) discuss evidence on whether reliable cross-cultural comparisons of social and emotional skills of school-age children can be made, and (iv) identify the best methods to reduce response-style biases. To support the development of the international report, the Contractor will be asked to (a) provide statistical and technical support to the Secretariat, (b) deliver descriptive tables following the OECD's standard format, and (c) review the draft report for technical consistency and coherence. Responsibility for the production of tables and analyses from the international database will be shared between the Contractor and the Secretariat.

- *Document the database.* The Contractor will be asked to prepare a fully documented database to be delivered to the OECD Secretariat in complete form no later than the end of December 2018. The OECD Secretariat will also request an initial dataset to be compiled in March 2018, containing such data that have been processed by that time, in order to allow the Secretariat to carry out initial data exploration. These datasets shall cover all data sources, from both tests and questionnaires. The overall quality control of the international database rests with the Contractor.

Bidders are invited to describe the workflow that will be necessary in order to allow these datasets to be produced according to this timeline.

Core B: Background questionnaire development

50. Background questionnaires will measure learning contexts, outcomes and background information of children in Grades 1-12. This information may be available at the individual, school, local community and the city level. One of the main challenges in background questionnaire development is that little is known about learning contexts that drive social and emotional skills development. Hence, the Contractor is likely to engage in this work based on limited evidence and existing instruments. For background information and some of the learning context measures, frameworks developed in PISA and TALIS can be a useful point of reference.[1] Note that this work will affect the work of Core D. Should the chosen Contractor for Core B be different from that of Core D, the Contractor for Core D will be asked to co-ordinate the work across the Contractors of Cores B and C, in close co-operation with the OECD Secretariat.

Task 1: Develop a conceptual framework and validated instruments.

51. The Contractor will be asked to:

- *Develop a conceptual framework.* The conceptual framework describes the domains of learning contexts, outcomes measures and other background information. The framework will justify the choice of domains, questionnaire types and respondents. The framework will also describe sampling strategies to validate the instruments. The contactor will be asked to work closely with the National Centres to take account the policy priorities of participating cities. The Contractor will also be asked to present the draft framework to the OECD Secretariat and the group of representatives of this study, and make adjustments as necessary. Bidders are invited to briefly outline key learning contexts to address, the type of background questionnaires to introduce, the best respondents to collect robust information on learning contexts, and the sampling strategy to validate the instruments. Bidders are encouraged to provide separate cost estimates by the type of questionnaires introduced.

- *Draft the instruments.* The Contractor will be asked to develop a questionnaire separately for Grade 1 and 7 cohorts. The questionnaire may also vary across grades. The Contractor may be asked to revise the questionnaires after discussing with the OECD Secretariat and the group of representatives of this study, and after testing them in the pilot phases. The Contractor will also be asked to provide guidelines for participating cities should they request adding national components to the background questionnaires.

- *Analyse and validate the instruments*. The Contractor will be asked to clean all data collected. Bidders are invited to indicate the types of checks that they will carry out on the data, and the mechanisms which will be put in place to ensure that necessary checks are also carried out by

[1] PISA 2015 questionnaire framework can be found in: http://www.oecd.org/pisa/pisaproducts/PISA-2015-draft-questionnaire-framework.pdf. TALIS 2013 conceptual framework can be found in: http://www.oecd.org/edu/school/TALIS%20Conceptual%20Framework_FINAL.pdf.

National Centres. Moreover, the Contractor shall subsequently conduct analyses to investigate the measurement properties of the instruments; and carry out other necessary analyses to inform the selection of background questions that can be employed during the field trial and the main study. Bidders are invited to outline the range of psychometric analyses to be performed. Bidders are invited to describe the process they will follow to sample the data and validate the questionnaires, and to separately budget the instrument development for each cohort.

Task 2: Ensure quality of translations of instruments

52. The Contractor will be asked to:

- *Prepare a guideline for translations and adaptations to be used by the National Centre*. The Contractor will provide the questionnaire and translation guidelines in English for National Centres. Countries are responsible for translating the questionnaire into their own national languages from the English source versions. This may be an adapted version of the guideline described in Core A.

- *Ensure the quality of the translations of instruments into national languages.* The Contractor will also review the technical standards regarding the translation of the questionnaires to be considered and adopted by the group of representatives of this study.

Task 3: Develop a survey platform

53. The Contractor will be asked to:

- *Develop a survey delivery platform should computer-based delivery not be fully employed for the background questionnaire.* If, for reasons specific to some of the participating cities, computer-based delivery is not an option, the Contractor will be asked to prepare an alternative survey platform such as paper-and-pencil. Bidders are invited to discuss if/when the collection of information on contexts and outcomes should be operationalized though paper-pencil questionnaires or other methods.

Task 4: Undertake project management

54. The Contractor will be asked to:

- *Appoint a project director*. The project director for Core A may continue managing this part of the study. The project director will be expected to attend the meetings of the group of representatives of this activity and present updates on project activities. These meetings are scheduled to take place twice each year and are generally hosted by participating countries. The Contractor will be responsible for covering travel, accommodation and subsistence expenses for either their own personnel or members of expert groups who attend these meetings, and should describe the extent of such attendance they have assumed in their budget. The person in this role should also have sufficient track record to provide the intellectual leadership with experts, and to work with the Secretariat in identifying technical issues to be discussed by the Technical Advisory Group (TAG). Bidders are asked to specify the project director as well as the percentage of time he/she will spend on this part of the project.

- *Facilitate the implementation of the agreed project management approach.* This will involve (a) developing and maintaining a project plan and timeline that joins the works in a coherent and cost-efficient way, (b) negotiating and resolving timeline amendments, for example those which might arise from unanticipated requests from the CERI Governing Board and the group of representatives of this study, unavoidable operational delays or other unforeseen project changes, (c) keep the OECD Secretariat fully updated on amendments any timeline issues which cannot be resolved or which may have implications for achievement of project milestones, (d) establish procedures for monitoring and managing risks, (e) establishing mechanism for submission of all documents, materials and databases to the OECD archive, (f) discussing additional requests from participating cities, (g) negotiate and co-ordinate additional national requirements or requests with the Secretariat and with National Centres or the group of representatives of this study as appropriate, and (h) providing regular progress reports to the OECD Secretariat and the group of representatives of this study.

- *Establish and manage an experts group*. The Contractor will be asked to establish the background questionnaire experts group and lead the technical discussions. The Contractor and the OECD Secretariat will jointly determine members of the experts group. Bidders are asked to describe the number of expert group meetings they have included in their proposed budget, and to explain how they would call on the expertise of group members outside the formal meetings. Bidders should include in their cost proposal all expenses associated with holding expert group meetings for which they are responsible, such as conference venue, travel, accommodation, subsistence and honorariums to expert group members. A member of the Secretariat team will generally attend meetings and the Secretariat will cover their own associated costs. Bidders are invited to submit a proposed list of background questionnaire experts.

- Support the work of National Project Managers (NPMs). The Contractor will be asked to support National Project Managers (NPMs) with the data collection and validation of the background questionnaire. This includes developing and maintaining tools for NPMs to track progress with the implementation of the tasks involved with the survey in each country and to keep track of any potential problems with cities' abilities to meet project timelines or technical standards. The Contractor will be asked to host meetings of NPMs as necessary.

- *Prepare a technical report of the background questionnaire validation study.* This report will be designed to validate the background questionnaire framework and provide recommendations on the appropriate learning contexts, outcomes and other background questions to be retained for the field trial and the main study. Bidders are invited to outline the contents of the technical report.

- *Document the data-base.* The Contractor will be asked to prepare a fully documented database to be delivered to the OECD Secretariat in complete form.

Core C: Survey design, sampling and quality assurance of the field trial and the main study

55. Identification of robust longitudinal survey structure and rigorous sampling strategy is indispensable for ensuring the integrity of the data collected. The technical requirement for this part of the developmental work will therefore be considerable with various mechanisms employed for quality

assurance purposes. Bidders may refer to sampling strategies employed in PISA as our baseline model that would help maximise the quality of the data collected in the proposed longitudinal study. Bidders are also encouraged to study some of the best sampling strategies adopted in existing longitudinal studies on education. Note that this work will closely relate to the works of Core D. Should Contractor for Core C be different from that of Core D, the Contractor for Core D will be asked to co-ordinate the work across the Contractors of Cores B and C, in close co-operation with the OECD Secretariat.

Task 1: Identify longitudinal design features and sampling strategy

56. The Contractor will be asked to:

- *Define key longitudinal design features.* The Contractor will be asked to define key elements of the longitudinal survey structure (e.g. survey frequency, duration, respondents), by building on the initial proposal by the OECD Secretariat (outlined in Section 1) and results from the feasibility study (Core A). This will be done in consultation with the Secretariat the group of representatives of this study. Bidders are invited to provide initial reflections on the optimal frequency of assessing social and emotional skills, learning contexts and outcomes that would help ensure success and long-term viability of this study. Bidders are invited to specify the pros and cons of various survey frequencies.

- *Develop a sampling strategy*. The Contractor will be asked to develop a sampling strategy to collect a representative sample of each of the starting cohorts in Grades 1 and 7. A sampling strategy includes sampling plans that describes procedures that each city will be asked to follow in drawing a robust sample from the sampling frame. A robust sample provides adequate demographic representation of students and schools in a city. A major challenge for this study is that the sampling strategy must be longitudinally viable. Section 1 of this document provides initial sampling considerations by the OECD Secretariat. Note that participating jurisdictions are typically major cities in terms of population, political significance or the size of the economy. The OECD proposes data from the two cohorts to be simultaneously collected at the city level with an option for state-wide or nation-wide coverage. The Contractor will also be responsible for defining exclusion criteria for students and/or schools (e.g., children with special education needs). Bidders are invited to (a) outline and briefly justify their proposed sampling strategy including sampling plans and exclusion criteria, (b) describe how they would work with cities to ensure that the sampling strategy suits the contexts of each city and the needs of this international study, (c) discuss the process of identifying the minimum sample size (i.e., number of schools, teachers and students) and sampling weights for each city, separately for the two cohorts, and (d) provide an approximate starting sample size for both the Grades 1 and 7 cohorts by making assumptions about attrition rates for two possible scenarios: (a) each cohort is followed for 6 years (i.e., Grade 1 cohort until grade 6 and Grade 7 cohort until grade 12) and b) each cohort is followed until they reach age 25. While bidder's proposals need not be identical to the Secretariat's initial considerations outlined in Section 1, they need to be well justified. Bidders may consult the Sampling Manual for the PISA 2012 main survey and the Sampling Manual for the PISA 2015

field trial. These documents are both available on the Call for Tender section of the OECD PISA website2.

Task 2: Develop and support quality control procedures

57. The Contractor will be asked to:

- *Develop quality control procedures.* The Contractor will be asked to develop procedures for identifying and dealing with samples that do not meet the predetermined sampling standards. Note that the group of representatives of this study will appoint a sampling referee whose task is to verify that the quality of data collected in a particular city is sufficient to construct city-based indicators. The sampling referee will: (a) identify problems with sampling or response rates that may jeopardise cities' compliance with the agreed-on sampling procedures, (b) provide an explanation for the problems or concerns and, when possible, (c) suggest remedies for them. The sampling referee will also make recommendations regarding the use of individual cities' data in the reporting process. If the sampling referee identifies issues which could threaten the integrity of the sample, the Contractor will be asked to provide an explanation to the participating city and, when possible, suggest remedies or work with the country if further investigation of sample quality is required. The OECD Secretariat shall arbitrate disagreements between participating cities, the Contractor and the sampling referee. Bidders are invited to outline the proposed quality control procedures including a description of the range of potential difficulties anticipated and ways to address them. Bidders are also invited to describe how they would propose to work with the sampling referee most effectively to ensure the integrity of city samples.

- *Support National Centres and NPMs.* The Contractor will be responsible for developing sampling guidelines (see Task 3) and training materials for National Centres. The Contractor will also be asked to attend NPM's meetings to conduct training sessions in sampling procedures and to carry out individual consultations with NPMs as necessary. Bidders are asked to describe how they would use alternative methods such as online training materials and webinars to support National Centres.

Task 3: Develop technical standards and sampling guidelines

- *Develop technical standards for the main study*, As one of the core background document for this study, the technical standards describes standards that need to be maintained for the main study to ensure data quality assurance, management integrity and national/local relevance. The Contractor will be responsible for developing the technical standards building on the Tasks 1 and 2 outlined above and working closely with the Technical Advisory Group (TAG), NPMs and the group of representatives of this activity. Bidders are invited to consult the PISA Technical Standards for reference. These are available on the Call for Tender section of the PISA website (see www.pisa.oecd.org).

- *Develop sampling guidelines for the field study and the main study.* As one of the core background document for National Centres to follow the sampling procedures, this report presents guidelines

[2] http://www.oecd.org/pisa/pisaproducts/PISA-2018-documents-for-bidders.htm

for the submission and approval of sampling and population information. The Contractor will be responsible for preparing the sampling guidelines building on Tasks 1 and 2 outlined above and working closely with the Technical Advisory Group and the NPMs. Bidders are invited to consult PISA's Sampling Guidelines for the Field Trial for reference3.

- *Prepare sampling weights for the field trial and the main study*. The Contractor will develop sampling weights for each participating city, to be used in the preparation of the international database. This activity will be carried out as part of the broader data analysis process, which will be managed by the contractor for Core D.

Core D: Management and implementation of the field trial and the main study

58. The management and implementation of this study is likely to be highly complex and intensive given the state of measurement technologies and paucity of expertise and evidence available to guide the technical work. There are considerable uncertainties in how the survey instruments will perform, whether the students will remain in the study, and the extent to which political support can be sustained during the course of this study. The OECD Secretariat will therefore require the Contractor to demonstrate the capability to overcome these challenges by demonstrating strong leadership, technical capacity and diverse networks and experience to run a large-scale international data collection enterprise. Note that the Contractor chosen for Core D may be different from that of Core B and C. In this case, the Contractor for Core D will be asked to co-ordinate the work across the Contractors of Cores B and C, in close co-operation with the OECD Secretariat.

Task 1: Operate the field trial and the main study

59. The Contractor will be asked to:

- *Undertake project management.* The Contractor, in collaboration with the OECD Secretariat, will assume a significant role in the oversight and management of this study. The Contractor will be responsible for: (a) developing and maintaining a project plan and timeline that joins the work in a coherent and cost-effective way, (b) negotiating and resolving timeline amendments, unavoidable operational delays or other unforeseen project changes, (c) developing procedures for monitoring risks; provide regular updates on risks, issues and deviations from timelines to the Secretariat, (d) establishing mechanisms for submission of all documents, materials and databases to the OECD archive (e) discussing additional requests from participating cities with the Secretariat, National Centres and/or the group of representatives of this study, and (f) providing regular progress reports to the OECD Secretariat and the group of representatives of this study. The nature and frequency of such reports will be agreed between the Contractor and the Secretariat. The Contractor will be required to appoint an **International Survey Director,** who may be the same individual as the project director described in Core A and B. This person will act as overall leader of the work as well as providing leadership for the NPMs. The person in this role would also have a track record of providing the intellectual leadership among experts. Bidders are invited to propose the name of

[3] http://www.oecd.org/pisa/pisaproducts/PISA2015FT-SamplingGuidelines.pdf

the International Survey Director and specify the percentage of time he/she is scheduled to spend on this part of the project.

- *Support the work of the NPMs*. The Contractor shall develop a description of the role and profile of the NPMs and specify the Contractor's working relationships with the NPMs. The Contractor will be asked to implement procedures that promote excellent communication with the NPMs. As part of its co-ordinating role, the Contractor will be asked to provide and maintain tools for the NPMs to track progress with the implementation of the tasks involved with the survey in each country and to keep track of any potential problems with countries' abilities to meet project timelines or technical standards. The Contractor shall call, organise, and host meetings of the NPMs. Provisions for meeting venues and facilities as well as for travel and compensation of experts, as required, should be included in bidders' proposals. No compensation of travel costs for the NPMs or representatives from the OECD Secretariat should be included in the cost proposal. Participating countries will bear the costs of their NPMs' participation in these meetings. Bidders are invited to outline procedures to support the NPMs and propose frequency of the NPM meetings.

- *Establish plans for monitoring adherence to the technical standards.* Under close consultation with the OECD Secretariat and the TAG, the Contractor will implement survey operations and related aspects of quality control, including the development of the test and questionnaire administration procedures, the development of scoring procedures, and the training of all necessary and relevant country representatives in these procedures (e.g., NPMs, test administrators). All training materials shall be developed in English. To ensure technical standards are met, the Contractor establishes frequent communication with NPMs, and formalise survey operations in instruction manuals so that NPMs and/or institutional coordinators have the required guidance for survey implementation. The Contractor shall establish plans for monitoring adherence to the technical standards during field operations in all cities. This will include National Centre procedures and survey operations in the centres that participate. These plans shall include a requirement that the Contractor shall appoint and pay quality monitors to visit a number of centres in each of the participating cities to assess their compliance with the project's guidelines for sampling and data collection. The Contractor will produce a quality monitoring report, outlining cities' compliance with quality standards throughout the project and which will be taken into consideration along the adjudication report in decisions about data inclusion for the final reporting. The Contractor will also be asked to develop a strategy for assisting countries in attaining acceptable response rates and reducing attrition and non-response bias. The Contractor shall therefore assign each participating city a set of follow-up procedures aimed at achieving the required response rates. These should be included in the sampling guidelines. Bidders are asked to identify the procedures to be undertaken to ensure high response rates and low attrition during the whole survey cycle so that there will be sufficient sample sizes when the cohorts reach early adulthood (age 25).

- *Develop methods to deal with ethical issues*. The Contractor will be asked to deal with all ethical issues related to this study. Ethical issues are an important aspect of planning any survey, particularly involving children and adolescents as respondents. Throughout the preparation of the survey outmost attention should be paid not to offend or upset the participants, or introduce any harmful content. One of the most important ethical concerns is to assure data confidentiality, by for instance appropriately storing the data and maintaining anonymity, in accordance to the countries' data protection acts. The alignment with the standards will be coordinated by the OECD together with participating cities. International ethics and guidelines should be used to develop these ethical

standards. The bidders are invited to outline their strategy to deal with ethical issues across participating cities.

Task 2: Process micro-data, analyse, scale and report

60. The Contractor will be asked to:

- *Clean all collected data and conduct analyses of the field trial and the main longitudinal study.* The Contractor will be asked to ensure that high-quality data verification and processing is undertaken during all stages of the survey. Data processing shall include the merging of national datasets from individual cities into an international dataset. It will also require the merging of the international dataset for each of the two cohorts. Descriptive statistics should be calculated from the merged datasets and appropriate investigations of statistical anomalies undertaken. Bidders are asked to indicate the types of checks that will be carried out on the data, and the mechanisms which will be put in place to ensure that checks are carried out by National Centres as required. This task is necessary for the field trial and the main study.

- *Provide a fully documented database which will allow the OECD Secretariat and participating cities to conduct their own analyses.* The Contractor shall ensure the data is cleaned and weights and variance estimations are computed. Datasets should be prepared containing the relevant sampling and variance estimation information. Bidders are invited to outline the quality control procedures that will ensure the delivery of an error-free, reliable and comparable dataset. The Contractor shall also provide all products accompanying the dataset. These include user friendly data files for each cohort, including a clear mechanism for merging and analysing the data files of each of the two cohorts, file descriptions, codebooks, and any indicators and indices formulae. These may be provided in the format of a user's guide for the international database. In addition, the Contractor shall test and compile the derived variables, scales and indices for inclusion in the international database.

- *Finalise instruments for the main study.* The Contractor shall work with the experts group to develop proposals for the main study instruments, based on the analyses of the field trial data.

- *Develop an analysis and reporting plan, which will guide the OECD Secretariat in preparing and designing international reports.* The Contractor shall submit a draft analysis and reporting plan to the participating cities and the Secretariat for review and approval. The plan shall discuss the kinds of analyses that will be possible with the data collected in this study. Most importantly, the plan should summarise and explain the types of analyses that can be conducted to address the key policy questions, and discuss how the data can best be presented and reported. Once approved by the participating cities, the plan will serve as the basis for the international report that the Secretariat will coordinate. To support the preparation of the report, the Contractor will be asked to: (a) develop an analysis and reporting plan, (b) provide statistical and technical support for the Secretariat during the development of the report, (c) design and provide basic descriptive tables following a standardised format specified by the Secretariat, (d) review the tables and drafts of the report for technical consistency and coherence, (e) establish and maintain an archive of all project resources, documents, materials and databases. Responsibility for the production of tables and analyses from the international database will be shared between the Contractor and the Secretariat. Given the level of coordination that will be necessary between the Secretariat and participating

cities, bidders are reminded of the need to discuss how such coordination will be facilitated and managed successfully. One issue that should be addressed in this discussion is the consistency of results in the international and national reports (should countries wish to undertake them). The Contractor cannot guarantee such consistency but should be available to assist those preparing national reports should questions arise about procedures for data analysis, scaling procedures, weighting, software, etc.

- *Develop technical reports.* The Contractor will be responsible for preparing technical reports which detail all the data and statistical analyses conducted for each of the two cohorts per survey cycle. Technical reports should serve the needs and address the likely questions of the most sophisticated users of the dataset. It should also provide guidance for future waves of the survey if particular issues and/or difficulties were encountered or identified. The Contractor will ensure that the technical report has been thoroughly edited and written according to the OECD Style Guide. All tables to be included in the technical report shall be provided in Excel format. The OECD Secretariat will be responsible for final formatting and copy-editing of the report for publication. The PISA 2009 Technical Report (OECD, 2012) is an example of the type of publication intended[4]. Bidders are invited to suggest mechanisms for ensuring that the technical reports serve the needs of the users.

[4] http://www.oecd.org/pisa/pisaproducts/pisa2009technicalreport.htm

TERMS OF REFERENCE

SECTION 4: SCHEDULE, DELIVERABLES AND BUDGET GUIDELINES

Indicative timeline

61. The following indicative timeline provides major milestones of the Longitudinal Study of Social and Emotional skills in Cities. Bidders are asked to provide detailed project plans for the scope of work. The project plans should include tasks, milestones and deliverables as well as an allocation of personnel to tasks. The following timeline contains only selected major milestones, whereas the project plans submitted by bidders should cover the totality of the activities. Note that the schedule presented here is tentative and subject to the proposals made by the successful bidder in response to this terms of reference.

Dates	**Responsibilities**
2016	
May	*Core A*: Provide the OECD Secretariat and TAG with initial project plans and proposals
June	*Core A: P*rovide the OECD Secretariat with a description of the role and responsibilities of National Project Managers (NPMs)
July	*Core A:* Submit initial consolidated timeline for the feasibility study as well as proposal for the assessment strategy; social and emotional skills instruments, and computer-based platform for review by the OECD Secretariat
September	Meeting of the group of body of this study to discuss work progress
October	*Core A:* Submit final proposal for the assessment strategy; social and emotional skills instruments; and computer-based platform. Submit translation guidelines and translator manual to national centres.
December	*Core A:* Provide manuals and training for assessment administrators
2017	
January	*Core A*: Completion of quality assurance for translating social and emotional skills instruments *Core B:* Submit 1st draft background questionnaire framework for review by the OECD Secretariat *Core D:* Submit description of the role and responsibilities of NPMs for the main study
March- June	*Core A*: Data collection of feasibility study
February	Meeting of the group of representatives of this study to discuss work progress
May	*Core C*: Provide initial sampling forms and guidance for National Centres *Core D:* Submit consolidated timeline for the main study
June	*Core D*: Submit field trial NPM manual
July	*Core C*: Submit draft Technical Standards
July - December	Core *A*: Data analysis and adaptation of instruments

September	*Core B:* Submit 2nd draft background questionnaire framework for review by the OECD Secretariat
October	*Core B:* Submit background questionnaire items for country review
November	Meeting of the group of representatives of this study to review background questionnaires

---

2018

January	Core *A*: Submit the proposed social and emotional skills instruments for field trial phase for review by the OECD Secretariat and preliminary results of the feasibility study
	Core B: Submit final draft background questionnaire framework. Submit translation guidelines and translator manuals.
February	Meeting of the group of representatives of this study to review final instrument proposal
March	*Core A*: Submit final proposed social and emotional skills instruments for field trial phase. Submit initial database
	Core C: Release field trial sampling forms and guidelines
March- December	*Core A:* Draft feasibility study report
March- May	*Core B*: Pilot data collection of background questionnaires
June- September	*Core B*: Data analysis and adaptation of background questionnaires
September	Core *B:* Submit the proposed background questionnaires for field trial phase for review by the OECD Secretariat
November	Meeting of the group of representatives of this study to review background questionnaires for field trial phase
	Core D: Release manuals for Test Administrators and School coordinators
December	*Core D*: Dispatch field trial instruments
	Core A: Submit final draft feasibility study report. Deliver complete database

---

2019

January	*Core C*: Release Main Study sampling forms and guidelines
	Core D: Field trial Coder Training
February	Meeting of the group of representatives of this study to review work progress
March – April	Core *D*: Conduct field trial data collection in the Northern Hemisphere
July	*Core C*: Finalise field trial sampling data in the Northern Hemisphere
October – November	Core *D*: Conduct field trial data collection in the Southern Hemisphere
November	Meeting of the group of representatives of this study to review work progress
December	*Core D*: Propose main study item selection

2020

January	*Core D*: Main Study Coder Training
	Core D: Dispatch instruments for the main study
February	*Core C*: Finalise field trial sampling data in Southern Hemisphere
	Meeting of the group of representatives of this study to review work progress

March – April	Core *D*: Collect 1st wave of data collection main study in Northern Hemisphere
July	*Core C*: Finalise main study sampling data in Northern Hemisphere
October - November	Core *D*: Collect 1st wave of data collection main study in Southern Hemisphere
November	Meeting of the group of representatives of this study to review work progress

---

2021

February	*Core C*: Finalise main study sampling data in Southern Hemisphere
	Meeting of the group of representatives of this study to review work progress
March – April	Core *D*: Collect 2nd wave of data collection in Northern Hemisphere
June	*Core C*: Finalise sampling weights of 1st wave of data collection
July	*Core C*: Submit final weighting summaries of 1st wave to National Centres
	Core D: Deliver first draft complete database of 1st wave
September	*Core D*: Deliver final complete database of 1st wave
October - November	Core *D*: Collect 2nd wave of data collection in Southern Hemisphere
November	Meeting of the group of representatives of this study to review work progress
December	*Core D*: Deliver all data products of 1st wave

2022

February	Meeting of the group of representatives of this study to review work progress
March	*Core D*: Submit technical report of 1st wave
March – April	Core *D*: Collect 3rd wave of data collection in Northern Hemisphere
	Core D: Deliver first draft complete database of 2nd wave
June	*Core C*: Finalise sampling weights of 2nd wave of data collection
July	*Core C*: Submit final weighting summaries of 2nd wave to National Centres
September	*Core D*: Deliver final complete database of 2nd wave
October - November	Core *D*: Collect 3rd wave of data collection in Southern Hemisphere
November	Meeting of the group of representatives of this study to review work progress
December	*Core D*: Deliver all data products of 2nd wave

2023

February	Meeting of the group of representatives of this study to review work progress
March	*Core D*: Submit technical report of 2nd wave
March – April	*Core D:* Deliver first draft complete database of 3rd wave
September	*Core D*: Deliver final complete database of 3rd wave
June	*Core C*: Finalise sampling weights of 3rd wave of data collection
July	*Core C*: Submit final weighting summaries of 3rd wave to National Centres
November	Meeting of the group of representatives of this study to review work progress
December	*Core D*: Deliver all data products of 3rd wave

Budget guidelines and assumptions

62. The OECD is in favour of bids that demonstrate good value for money. Bidders are asked to provide budgets until the end of the first 3 years (or, 3rd cycle) of the main longitudinal study, i.e. until 2023. Budgets should be presented in EUR and detailed according to the table below. For each type of expenditure, a total should be given as well as a breakdown of individual staff costs, roles and profiles. Costs should be given separately for each of the tasks in the Statement of Work of the relevant Core and when appropriate, also disaggregate cost by cohort (Cohort starting at Grade 1 and 7).

63. Bidders are invited to calculate the Core Budget which assumes participation of 5 cities, including one from East Asia, two from Europe, one from North-America and one from South America. Bidders are asked to also provide a cost estimate for the participation of each additional city in the feasibility and main studies.

Table 2. Budgetary worksheets

Core A: Budget for Social and emotional skills instrument development (i.e., the feasibility study)

TASKS	2016	2017	2018	TOTAL
Task 1: Develop an assessment strategy and validated instruments				
Task 2: Ensure quality of translations of instruments				
Task 3: Prepare a computer-based survey platform				
Task 4: Undertake project management				

Core B: Budget for Background questionnaire development

TASKS	2017	2018	TOTAL
Task 1: Develop a conceptual framework and validated instruments			
Task 2: Ensure quality of translations of instruments			
Task 3: Develop a survey platform			
Task 4: Undertake project management			

Core C: Budget for Survey design, sampling and quality assurance of the field trial and the main study

TASKS	2017	…	2023	TOTAL
Task 1: Identify longitudinal design features and sampling strategy				
Task 2: Develop and support quality control procedures				
Task 3: Develop technical standards and sampling guidelines				

TERMS OF REFERENCE

Core D: Budget for Management and implementation of the field trial and the main study

TASKS	2017	…	2023	TOTAL
Task 1: Operate the field trial and the main study				
Task 2: Process micro-data, analyse, scale and report				

TERMS OF REFERENCE

SECTION 5: EVALUATION CRITERIA

64. The evaluation criteria for each of the Cores A, B, C and D and their weights will be as follows.

Core A, Core B, Core C and Core D (maximum 100 points each)

Technical Evaluation Criteria (70 points)

Technical quality (30 points)

- Demonstrate the capacity to employ creative yet practical solutions to measurement challenges.
- Extent to which the proposal demonstrates an understanding of the project design and assessment domains.
- Clear, convincing and feasible proposals for each of the tasks in the Statement of Work.
- A survey design that is in-line with the study objectives.
- Should proposals of equal technical quality be submitted, the proposal offering more innovation and efficiency gains shall be rewarded.

Organisational and management capabilities (20 points)

- Proven capacity to develop a collaborative working relationship with the other actors and to promote consensus-building activities through effective communication and management.
- Proven ability to put effective management and financing structures in place.
- Demostrate the capacity to undertake and manoeuvere risks and uncertainties.
- Clear and convincing proposals on how the Contractor will work with the National Project Managers and the Expert Groups.
- A commitment to work within a fixed price envelope and to work flexibly and in partnership with the OECD Secretariat and the group of representatives of this activity.

Staff qualifications and previous experience (20 points)

- Possess strong networks of diverse techncial expertise from different disciplines and countries.
- Past experience and track record, preferably in an international context.
- Capacity to enlist the best expertise in providing the deliverables required under the terms of reference.
- The qualifications and experience of the proposed Project Director and the International Survey Director.
- Experience that demonstrates relevant and successful project management and coordination of large-scale assessments and/or projects involving multiple countries.

Financial Evaluation Criteria (30 points)

- Each bid should be evaluated based on the proposed pricing and on the justification it provides for the real costs associated with each component and activities (including alternative and optional activities) of the proposal. The evaluation should also consider the justification given for the stated marginal cost associated with the participation of each additional city for two

possible scenarios: (a) an additional city in a country with a city participating in the study (e.g. London and Manchester) and (b) an additional city in a country without another city participating in the study (e.g. London when England was not in the original list of participating cities).

MINIMUM GENERAL CONDITIONS FOR OECD CONTRACTS

The following articles constitute of the minimum general conditions of the contract to be signed between the OECD and the Contractor to whom the Call for Tenders would have been awarded (the "Contract"). These minimum general conditions are not exclusive and could, as the case may be, be modified and/or complemented with additional conditions in the Contract.

ARTICLE 1 – GOODS OR SERVICES

The goods and/or services provided under the Contract (hereinafter "The Work") shall strictly comply with the standards mentioned in the Terms of Reference. It is expressly agreed that the Contractor shall perform the Work in strict accordance with all Standards or, where no such standards have yet been formulated, the authoritative standards of the profession will be the applicable norms.

ARTICLE 2 - PRICES

Prices charged by the Contractor for the Work shall not vary from the prices quoted by the Contractor in its Tender, with the exception of any price adjustment authorised in the Contract.

ARTICLE 3 - PAYMENTS AND TAXES

Payment will be made in Euros.

In case the Contractor is located outside of France, the Organisation is exempt from taxation, including from sales tax and value added tax (V.A.T.). Therefore, the Contractor shall not charge any such tax to the Organisation. All other taxes of any nature whatsoever are the responsibility of the Contractor.

ARTICLE 4 - DELAY IN EXECUTION

The Contractor shall perform the Work in accordance with the time schedule and the terms specified in the Contract, this being an essential element of the Contract. Any delay will entitle the Organisation to claim the payment of penalties as negotiated between the Contractor and the Organisation.

ARTICLE 5 - ACCESS TO THE PREMISES

If the Work requires at any time the presence of the Contractor and/or of the Contractor's employees, agents or representatives ("Personnel") on the premises of the Organisation, they shall observe all applicable rules of the Organisation, in particular security rules, which the Organisation may enforce by taking any measures that it considers necessary.

ARTICLE 6 - IMPLEMENTATION OF THE WORK

The Contractor undertakes that the Work shall be performed by the individual(s) named in the Contract or otherwise agreed in writing by the Organisation. The Contractor may not replace said individual(s) by others, without the prior written consent of the Organisation.

ARTICLE 7 - AUTHORITY

The Contractor hereby declares having all rights and full authority to enter into the Contract and to be in possession of all licences, permits and property rights, in particular intellectual property rights, necessary for the performance of the Contract.

ARTICLE 8 - LIABILITY

The Contractor shall be solely liable for and shall indemnify, defend and hold the Organisation and its personnel harmless from and against any and all claims, losses, damages, costs or liabilities of any nature whatsoever, including those of third parties and Contractor's Personnel, arising directly or indirectly out of or in connection with Contractor's performance or breach of the Contract.

It is the responsibility of the Contractor to possess adequate insurances to cover such risks, including any risks related to the execution of the Contract.

ARTICLE 9 - REPRESENTATIVES

Neither the Contractor nor any of its Personnel:

- shall in any capacity be considered as members of the staff, employees or representatives of the Organisation;
- shall have any power to commit the Organisation in respect of any obligation or expenditure whatsoever;
- shall have any claim to any advantage, payment, reimbursement, exemption or service not stipulated in the Contract. In particular and without limitation, it is understood that neither the Contractor, nor any of the Contractor's Personnel may in any manner claim the benefit of the privileges and immunities enjoyed by the Organisation or by its personnel.

ARTICLE 10 - INTELLECTUAL PROPERTY

The copyright and any other intellectual property rights arising from the Work carried out in performance of this Contract, including the intermediate and final results thereof, shall, on an exclusive and worldwide basis, automatically vest in the Organisation as the Work is created, or be assigned to the Organisation, as the case may be under any applicable legal theory. The price agreed between the Contractor and the Organisation is deemed to include this transfer of rights.

The Contractor undertakes not to use the Work for any purpose whatsoever that is not directly necessary to the performance of the Contract, except with the prior written consent of the Organisation. The Contractor shall ensure that the Contractor's Personnel are expressly bound by and respect the provisions of the present clause.

ARTICLE 11 - TRANSFER OF RIGHTS OR OBLIGATIONS

The Contractor shall not transfer to any third party any rights or obligations under this Contract, in whole or in part, or sub-contract any part of the Work, except with the prior written consent of the Organisation.

ARTICLE 12 - TERMINATION

Without prejudice to any other remedy for breach of Contract the Organisation may claim, the Organisation reserves the right to terminate the Contract without any prior notice or indemnity:

i) in the event of failure by the Contractor to comply with any of its obligations under the Contract; and/or

ii) if the Contractor, in the judgment of the Organisation, has engaged in corrupt or fraudulent practices in competing for or in executing the Contract.

The Organisation may also, by written notice sent through registered mail with recorded delivery to the Contractor, terminate the Contract, in whole or in part, at any time for its convenience. The notice shall specify that termination is for the Organisation's convenience, the extent to which Work of the Contractor under the Contract has been completed, and the date upon which such termination becomes effective. The Work that is complete on receipt of notice by the Contractor shall be accepted by the Organisation, at the Contract terms and prices. For the remaining, the Organisation may elect:

i) To have any portion completed at the Contract terms and prices; and/or;

ii) To cancel the remainder and pay to the Contractor the amount corresponding to the completed work.

ARTICLE 13 – FINANCIAL INFORMATION

During the Contract and at least seven years after its termination, the Contractor shall :

i). keep financial accounting documents concerning the Contract and the Work ;

ii). make available to the Organisation or any other entity designated by the Organisation, upon request, all relevant financial information, including statements of accounts concerning the Contract and the Work, whether they are executed by the Contractor or by its any of its sub-Contractors.

The Organisation or any other entity designated by the Organisation may undertake, including on the spot, checks related to the Contract and/or the Work.

ARTICLE 14 - ARBITRATION CLAUSE

Given the status of the Organisation as an international organisation, the rights and obligations of the Contractor and the Organisation shall be governed exclusively by the terms and conditions of the Contract.

Any dispute arising out of the interpretation or implementation of the Contract, which cannot be settled by mutual agreement, shall be referred for decision to an arbitrator chosen by agreement between the Organisation and the Contractor or, failing such agreement on the choice of the arbitrator within three months of the request for arbitration, to an arbitrator appointed by the First President of the Court of Appeal of Paris at the request of either Party. The decision of the arbitrator shall be final and not subject to appeal. The arbitration shall take place in Paris, France. All proceedings and submissions shall be in the English language.

Nothing in the Contract shall be construed as a waiver of the Organisation's immunities and privileges as an international organisation.

ARTICLE 15 – CONFIDENTIALITY

Any information, on any medium whatsoever, sent to the Contractor to which the Contractor obtains access on account of the Contract, shall be held confidential. In consequence, the Contractor shall not disclose such information without the written prior consent of the Organisation. The Contractor shall ensure that the Contractor's Personnel is expressly bound by and respect the provisions of the present clause.

ARTICLE 16 - DURATION OF THE CONTRACT

Unless otherwise stated in the Call For Tenders, the duration of the Contract shall be for one year. It may be renewed twice by tacit agreement for periods of one year, but the total duration may not exceed three years.

ANNEX A.

FRAMEWORK FOR THE LONGITUDINAL STUDY OF SOCIAL AND EMOTIONAL SKILLS IN CITIES

Oliver P. John, University of California and Filip De Fruyt, University of Ghent

TABLE OF CONTENTS

1. SCOPE

The OECD's Longitudinal Study of Skill Development in Cities aims to track the development of children's and adolescents' socio-emotional skills, and to assess how they predict a broad set of outcomes, including educational attainment, employability and labour market position, job performance, health conditions, well-being, interpersonal connectedness, civic engagement and environmental awareness, and crime/safety (OECD, 2015a). These developmental processes will be studied in a variety of cities, each characterized by specific political, socioeconomic and cultural contexts and challenges, forming a natural experiment to examine processes of continuity, change, time course, impacts, and outcomes of socio-emotional skills. The investigated outcomes are directly consequential at the level of the individual, but also impact upon the societal level. Therefore, the findings emerging from such a project should provide much-needed information about the life paths of individuals as they experience particular home, school, and community environments as well as generalizable research findings that have fundamental implications for education policies and practices.

The Longitudinal Study of Social and Emotional Skills in Cities extends prominent OECD projects that include socio-emotional measures, including the Programme of International Student Assessment (PISA). This is done by expanding the coverage of social and emotional assessment domains and introducing an assessment design to identify the growth trajectory of these skills. These ambitious objectives will not only introduce challenges in terms of study design and subsequent research, but also impose extra requirements on the key constructs that will be examined. The proposed Longitudinal Study aims to trace the developmental pathways of the same group of individuals across a long period of time. Instead of assessing the level of educational performance within and across countries and its associations with contextual factors, the Longitudinal Study of Social and Emotional skills in Cities will examine growth trajectories within developing contexts and document how these affect each other in explaining a broad variety of outcomes that are valued by individuals and society in general.

This report presents a conceptual framework of social and emotional skills among school-aged children and adolescents, relying on a review of the most relevant and recent literature. It first defines what socio-emotional skills are and how they have been characterised and structured in different literatures. The next section provides a summary of the proposed framework, beginning with the three core issues identified by the OECD (2015): (a) *Working with Others,* (b) *Managing Emotions,* and *(c) Achieving Goals.* The next section explains the rationale and research background for the particular constructs included in the framework, including predictive power and comprehensiveness, malleability, and temporal stability. The next section examines the coherence of the proposed framework with other frameworks and educational goals. The report ends with two briefer sections, one outlining strategies for validating the proposed framework over the next several years and the other outlining policy questions.

The proposed framework has a strong empirical foundation. It is based on a large number of psychometric analyses of micro-data conducted by psychologists across countries, languages and cultures. This framework may not be consistent with similar social and emotional skills frameworks proposed by educators due to the differences in the methodologies employed. For instance, some of these frameworks have been derived by synthesizing other existing frameworks, complemented by qualitative interviews with diverse stakeholders including teachers, parents and employers.

The proposed framework has a strong empirical foundation. It is based on a large number of psychometric analyses of micro-data conducted by psychologists across countries, languages, and cultures. This framework may not be consistent with similar social and emotional skills frameworks proposed by educators due to the differences in the methodologies employed. For instance, some of these frameworks have been derived by synthesizing other existing frameworks, complemented by qualitative interviews with diverse stakeholders including teachers, parents and employers.

2. DEFINITION OF SOCIAL AND EMOTIONAL SKILLS

Interest in social and emotional skills (e.g., goal-setting, perseverance, optimism, emotional control, gratitude, social intelligence, curiosity, etc.) has a long history. Education researchers, school administrators, teachers, parents, and even the children themselves have long been aware that education involves interactions among people; people are not only inherently social creatures but also experience and express a wide range of emotions. In other words, most schools are places that are both intensely social and intensely emotional. Previous OECD publications have shown that the way in which students, parents and teachers navigate these social and emotional processes can have powerful consequences for a multitude of important life outcomes (e.g., Kautz, et al., 2014; OECD, 2015).

In recent years, socio-emotional characteristics have also been referred to as a key component of *21st century* or *employability skills* (e.g., Trilling and Fadel, 2009) because they include a set of competencies considered increasingly crucial for individuals' development, employment, and healthy functioning in current and future societies (National Academy of Sciences, 2012). As individuals and jobs become increasingly interconnected, complex, and collaborative, socio-emotional characteristics are expected to become ever more important. Many different 21st century skills have been proposed over the years (see Annex 1 for a list of more than 160 individual skills described in Trilling and Fadel, 2009 and Fadel, 2014). They include such concepts as abnegation and altruism, engagement and enthusiasm, innovation and inquisitiveness, self-discipline and self-control, stability and tranquillity, and many more.

Many of these characteristics have also interested psychologists, who have studied them under the broad rubric of "personality traits" (John, 1990). As described below, extensive research has shown that these personality concepts can be organized into five broad and relatively independent and distinct domains of individual differences in thinking, feeling, and behaving, often referred to as the *Big Five* or the *Five Factor Model*. Moreover, as will be reviewed below, modern research has shown that traits are not in-born and fixed; in contrast to popular views, personality traits develop through the interplay of personal and environmental factors (i.e., learning) and they show considerable plasticity, especially during childhood and adolescence.

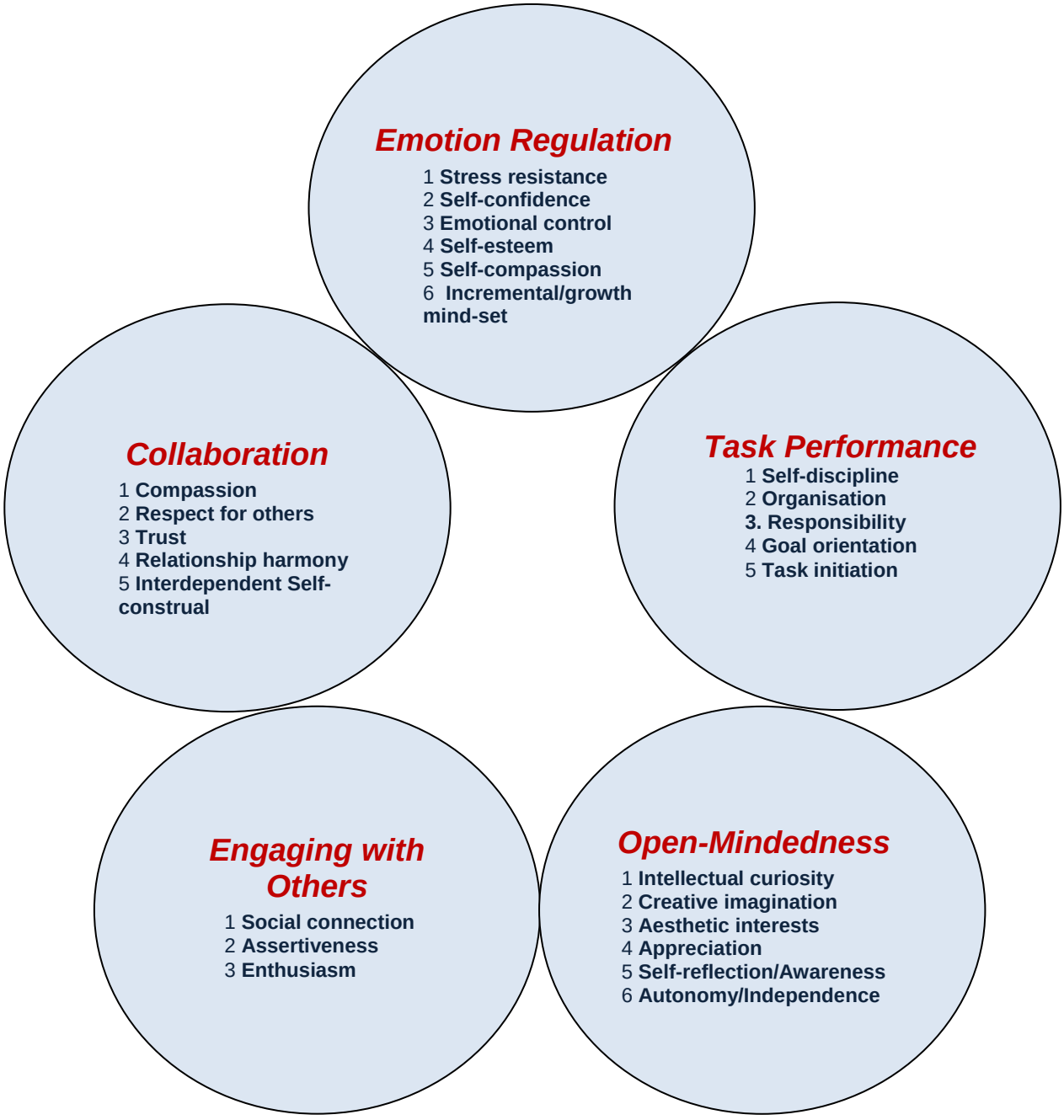
At the same time, developmental psychologists and educational researchers have studied how socio-emotional learning can be improved through school-based interventions. They have focused on specific interventions, and their particular contexts and unique outcomes. These researchers have been less interested in developing a single, generally accepted model that could organize the hundreds of individual social and emotional skills into one coherent taxonomy, like the table of elements in chemistry or the taxonomy of the animal kingdom in biology. Instead, they have developed multiple heuristic models. For example, Elias et al. (1997) proposed 6 (or 7) major domains of socio-emotional learning; Durlak et al. (2011) proposed 5 domains; and Saarni (2011) proposed 8. These three models differ in the number of major skill constructs they include, and they are not fully consistent with each other. Neither of them is comprehensive; they emphasize different aspects of the construct space. Section 5 of this report describes in more detail how these frameworks relate to, and differ, from each other.

The OECD (2015) defines social and emotional skills as: "*individual capacities that (a) are manifested in consistent patterns of thoughts, feelings and behaviours, (b) can be developed through formal and informal learning experiences, and (c) influence important socioeconomic outcomes throughout individual's life"*. This definition captures the essential features of social and emotional skills that are reflected in all the constructs proposed in the next sections as well as in the literature of 21st Century skills, personality psychology, developmental psychology, and social and emotional learning.

3. SUMMARY OF THE SOCIAL ND EMOTIONAL SKILLS FRAMEWORK

Figure 1 summarizes the proposed Social and Emotional Skills framework that builds on the vast conceptual and empirical literature that pertains to social and emotional skills, which will be described in the subsequent sections of this report. This section briefly describes this framework and the next section justifies the choice of the five broad domains and the corresponding lower-level facets.

Figure 3.1. Proposed Social and Emotional Skills Framework



The OECD's Longitudinal Study of Skill Development in Cities specified three domains of individual functioning that are of particular interest: Managing Emotions, Working with Others, and Achieving Goals. The conceptual framework should be able to account for the most important socio-emotional skills that are relevant to these three domains. However, these three domains are not themselves elements of the empirical framework because they are formative latent constructs: they are called formative because they are (like the Consumer Price Index) formed for practical or descriptive purposes and often identified through consensus opinion or traditions in a field. Formative latent variables have been described as a "stew"—a mixture of more basic elements that might or might not be related. For example, "Working with Others" mixes together more basic elements such as helping others, leading others, and following others that are psychologically distinct from each other and empirically rather different.

These more basic elements are called reflective latent variables because they reflect the essence or commonality of the various specific skills that can be measured. In psychometric models, these reflective latent variables are also called factors because they are often discovered through the statistical method of factor analysis. As will become apparent in Section 5 of this report, the elements of the framework proposed here can be linked systematically to formative frameworks that articulate and advocate particular goals for education. Indeed, it has been argued that "The various lists of 21st century skills that have been proposed to date are formative variables, identified by consensus opinion" (National Academy of Sciences, 2012, p. 27).

What are these more basic, reflective latent constructs? At the highest level of abstraction, socio-emotional skills constructs can be divided into five broad domains as shown in Figure 3.1. However, these five distinctions are at a very abstract level, and each of the five domains has therefore been subdivided into several more narrowly defined, specific constructs labelled facets. Table 3.1 lists a brief label for each of these facets, provides a short definition, and (if available) illustrates the concepts with an example item in parentheses.

Table 3.1. Proposed Framework: Five Broad Skill Domains and More Specific Socio-Emotional Characteristics within Each Domain Derived from the Literature Review

1. Engaging with Others
1 **Social approach and connection**: Able to approach others, both friends and strangers, initiating and maintaining social connections; skilled at teamwork, including communication and public speaking skills (Is outgoing, comfortable around people) 2 **Assertiveness (or courage)**: Able to voice opinions, needs, and feelings, and exert social influence; capacity to assert own will to accomplish goals in the face of opposition, such as speaking out, taking a stand, and confronting others if needed; courage (Takes on leadership roles) 3 **Enthusiasm**: Passion and zest for life; approaching daily life with energy, excitement, and spontaneity (Is full of energy, shows enthusiasm)
2. Collaboration: "Tending and Befriending" Others
1 **Compassion**: Kindness and caring for others stems from perspective taking and empathic concern for their well-being, and leads to valuing and investing in close relationships (Considerate and kind to everyone) 2 **Respect for others (politeness)**: Treating people with respect and politeness, the way oneself would like to be treated, according to notions of fairness, justice, and tolerance (Is respectful; treats others with respect vs. breaking rules) 3 **Trust**: assuming that others generally have good intentions and forgiving those that have done wrong; avoid being harsh and judgmental, giving people another chance (Assumes the best about people) 4 **Relationship harmony**: Living in harmony with others and valuing interconnectedness among all people; being inclusive of others who have different backgrounds, customs, and beliefs (It is important to me to respect decisions made by the group) 5 **Interdependent self-construal**: Experiencing self as part of a collective, interconnected and inseparable from important groups, such as family (I feel my fate is intertwined with the fate of those around me)

3. Emotion Regulation
1 **Stress resistance**: Effectiveness in modulating anxiety and response to stress; untroubled by excessive worry and able to calmly solve problems (Is relaxed, handles stress well) 2 **Self-confidence**: Positive and optimistic expectations for self and life; anticipates success in actions undertaken; a "can-do" mind-set (Feels secure, comfortable with self) 3 **Emotional control**: Effective strategies for regulating temper, anger, and irritation; able to maintain tranquillity and equanimity in the face of frustrations; not moody or volatile (Keeps their emotions and temper under control) 4 **Self-esteem**: Acceptance and positive evaluation of oneself (I am a person of worth.) 5 **Self-compassion**: Taking a mindful, kind, and accepting approach towards oneself, rather than being overly critical or self-blaming (When going through a hard time, I give myself the caring and tenderness I need) 6 **Incremental (or Growth) mind-set**: Believing that things are changeable; that humans can improve, learn, and grow; and that effort will improve one's personal future (When bad things happen, I think about ways to make things better, rather than "what's wrong with me") 7 **Fear of happiness**: Beliefs and worry that happiness will lead to bad outcomes (I prefer not to be too joyful, because usually joy is followed by sadness)
4. Task Performance
1 **Self-discipline**: Grit, perseverance, and effortful control are related concepts that involve concentration skills: the ability to focus attention on the current task and avoid distractions in order to achieve personal goals (Is efficient, gets things done) 2 **Organisation**: Organisational skills are critical for planning and executing plans to reach longer-term goals (Keeps things neat and tidy) 3 **Responsibility**: Time management, punctuality, and honouring commitments are critical to reliability and consistency, and engender trustworthiness (Is reliable, can always be counted on) 4 **Goal orientation**: Setting high standards for oneself and working hard to meet them, as illustrated by a strong "work ethic", consistent effort, and high levels of productivity (Wants to be excel at everything s/he does) 5 **Task initiation**: Ability to get started on a task or goal, rather than engaging in prolonged procrastination (Leaves difficult tasks for later vs. tackles them immediately)
5. Open-Mindedness: Interest and devotion to matters of the mind
1 **Intellectual curiosity**: Interest in ideas and love of learning, understanding, and intellectual exploration; an inquisitive mind-set (Likes to think, play with ideas) 2 **Creative imagination** : Generating novel ways to do or think about things through tinkering, learning from failure, insight, and vision (Is original, comes up with new ideas) 3 **Aesthetic interests**: Valuing art and beauty that may be experienced or expressed through music, writing, visual and performing arts, and other forms of self-actualization (Is fascinated by music, art, or literature) 4 **Appreciation**: Valuing and noticing the environment, living in harmony with nature, spirituality, awe, and reverence 5 **Self-reflection/Awareness of inner experiences**: Awareness of inner processes and subjective experiences, such as thoughts and feelings, and the ability to reflect about and articulate such experiences (meta-cognition) 6 **Autonomy/Independence of judgment and self-construal**: Thinking for yourself; grounding beliefs, attitudes, and values on a critical analysis through independent thought (I enjoy being unique and different from others in many respects)

3.1. Working with Others: Two Sets of Interpersonal Skills

We begin with the constructs most relevant to the OECD domain "Working with Others." Indeed, schools are intensely social settings, with students, teachers, parents, and school administrators all interacting with each other and forming relationships. Thus, schools provide rich environments for acquiring and practicing many kinds of social skills. These fall into two distinct sets within the framework. One set, labelled here *Engagement with Others*, captures the basic interpersonal direction of the individual, *towards* engagement and interaction with others, as contrasted with avoidance or withdrawal *away from* interpersonal contact. The other set, *Collaboration*, captures the quality of the interactions and relationships: is the child able to construe others as likely friends and sources of pleasure and support that can be trusted and loved (amity), or as adversaries (enmity)?

Within the superordinate domain *Engagement with Others*, the research literature agrees on three further distinctions. The first facet is *Social approach and connection*: learning to approach others and initiate and maintain connections with them is a critical learning task during childhood and adolescence. This involves practicing communication and public speaking skills (e.g., being able to tell a joke or story to a group of friends or give a short presentation in front of the class), which are critical to all kinds of team work. The second facet is Assertiveness (or courage), which involves children finding their "voice" and practicing it by speaking out, taking a stand, and confronting others if needed to achieve goals or meet their needs. The third facet is Enthusiasm, defined as passion and zest for life, approaching every day with energy, excitement, and spontaneity. These three facets are clearly interpersonal and thus most relevant to the OECD theme of "Working with Others"; however, these social and communication skills may also help children with the third OECD theme, namely Achieving Goals.

The second interpersonal domain in the framework, *Collaboration*, involves perceiving and treating others as friends, as expressed by the idea that others exist for our "Tending and Befriending". Again, three core facets have been identified. *Compassion* refers to kindness and caring for others that stems from perspective taking and empathic concern for their well-being; *Respect for others* involves treating people with respect and politeness, the way oneself would like to be treated; and *Trust* involves the belief that others generally have good intentions and forgiving those that have done wrong. In addition to these three core facets, cross-cultural research reviewed below suggested that Western notions of Collaboration constructs may need to be supplemented with other ways of experiencing the connection between self and other: one additional construct is *Living in harmony* with others and valuing interconnectedness among people; the other is *Interdependent Self-Construal* defined as experiencing the self as part of a collective, interconnected and inseparable from important groups, such as one's family.

3.2. Managing Emotions: Regulation Skills

The first set of constructs reviewed so far involved characteristics that are primarily social or *inter*-personal. However, learning environments in general, and schools in particular, are also rich with emotions, which are primarily *intra*-personal phenomena. Researchers generally agree that the way students learn to manage their emotions is of critical importance for their concurrent and subsequent adjustment and well-being. Psychologists have long recognized three basic and universal negative emotions that can potentially undermine students' well-being: fear, sadness, and anger (Ekman, 1972). Fear (and anxiety) arises when students are stressed by danger or uncertainty; anxiety can lead to a cascade of negative consequences through avoidance behaviour when students try to control their anxiety by avoiding the situations that make them anxious. Sadness (and depression) occurs when students experience disappointments, failures, and losses; sadness is sapping the individual of energy and can lead to social withdrawal (i.e., lower engagement). Anger typically arises from frustrations and the perception that one is not getting what one wants or deserves (i.e., when our actions or wishes are blocked by specific others, by rules, etc.). Thus, effectively regulating these negative emotions is of considerable importance, and emotion researchers (e.g., Saarni, 1999) have repeatedly called for teaching students emotion regulation skills in school. In the proposed framework, the broad domain of Emotion Regulation specifies three core facets. *Stress Resistance* reflects effectiveness in modulating anxiety and stress responses. *Self-confidence* includes positive expectations for self and life, anticipating success, and a "can-do" mind-set that helps buffer the child from sadness and depression. And *Emotional Control* refers to effective strategies for regulating temper, anger, and irritation that help maintain tranquillity and equanimity in the face of frustration. In addition to these three core facets, the framework includes a construct that has been widely studied as an emotion-protective factor, namely *Self-esteem*, defined as acceptance and positive evaluation of oneself. Three additional constructs were added on the basis of research on cognitive factors and cross-cultural variation. Beliefs about controllability have been widely studied, and the most promising recent intervention research points to the power of mind-sets; the *Incremental (or Growth) Mind-Set* involves

beliefs that things are changeable, that humans can improve, learn, and grow, and that effort will improve one's personal future. The other construct was suggested by recent research on mindfulness and emotion regulation: *Self-compassion* involves taking a mindful, kind, and accepting approach towards oneself when making mistakes or experiencing failure or set-backs, rather than being overly critical or self-blaming; this approach to self-care tends to improve emotional functioning. Finally, cross-cultural research has demonstrated substantial differences among countries and cultural groups in the ways emotions (especially happiness, pride, and love) are experienced and regulated; thus the framework includes a concept that has shown substantial cultural differences, namely *Fear of Happiness*, which involves beliefs and worries that happiness will lead to bad outcomes (e.g., "I prefer not to be too joyful, because usually joy is followed by sadness").

3.3. Achieving Goals: Task Performance and Open-Mindedness

Finally, even though some conceptions of socio-emotional functioning do not include skills related to goal achievement, the OECD did emphasize that achieving goals ought to be addressed in the longitudinal study. Indeed, considerable research has accumulated evidence that there are two distinct sets of skills that predict achievement behaviours (e.g., such as completing homework, studying regularly, and school attendance) as well as achievement outcomes (e.g., scores on standardized achievement tests and grades).

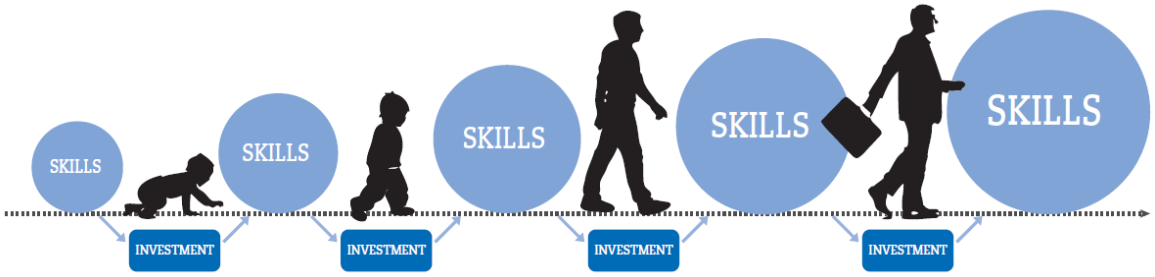
The first set involves *Task Performance* (labelled Conscientiousness in the Big Five personality work) and includes three commonly studied core facets. First, a number of constructs defined and studied separately in the research literature (e.g., grit, perseverance, effortful control), are in fact related concepts that all involve what here is called *Self-discipline*, namely concentration skills and the ability to focus attention on the current task and avoid distractions in order to achieve personal goals. The second facet is *Organization:* organizational skills are critical for planning and executing plans to reach longer-term goals. The third facet, *Responsibility*, also has some interpersonal implications because time management skills, punctuality, and honouring commitments are critical to being perceived as reliable and consistent and engender trustworthiness. In addition, we included two other facets to ensure adequate coverage of this important domain. *Goal orientation* involves setting high standards for oneself and working hard to meet those standards (e.g., work ethic), and *Task initiation* involves the ability to get started on a task or goal immediately, rather than engaging in prolonged procrastination.

The second set is here called *Open-Mindedness: Interest and devotion to matters of the mind.* The three most commonly studied facets are *Intellectual Curiosity* (defined as a passionate interest in ideas and the desire to learn and understand, and intellectual exploration); *Creative Imagination* (Generating novel ways to do or think about things through tinkering, learning from failure, insight, and vision); and Aesthetic Interests (Valuing art and beauty that may be experienced or expressed through music, writing, visual and performing arts, and other forms of self-expression and self-actualization). In addition, research on positive psychology has emphasized the importance of a spiritual connection to nature and its meaning-making potential, which led us to add the concept of *Appreciation* (Valuing and noticing the environment, living in harmony with nature, spirituality, awe, and reverence). Research on cognitive approaches and meta-cognition suggested that awareness of mental processes may be important, while research on emotional competence suggested that awareness of emotional experience may be important; thus, an additional facet was included to capture skills related to *Self-reflection, Introspection, and Awareness of inner experiences* (Awareness of inner processes and subjective experiences, such as thoughts and feelings, and the ability to reflect about and articulate such experiences; meta-cognition). The final facet shown in Table 3.1 was identified in research on cultural differences between Western and East Asian countries and involves the relative importance of *Autonomy and Independence of Judgment and Self-construal* (Thinking for yourself; grounding beliefs, attitudes, and values on a critical analysis through independent thought) versus following societal traditions and conventions.

3.4. Developmental pathways

Social and emotional skills are assumed to develop progressively over time, building on skills already accumulated as well as learning inputs (or, investments) from parents, teachers and the community (Figure 3.2).

Figure 3.2. Skills beget skills



Source: OECD (2015)

Past research conducted by developmental psychologists and economists are broadly consistent with this model (see OECD, 2015 for some evidence on this). However, little is known about the precise nature of the dynamic formation of social and emotional skills.

Policy-makers, parents and teachers would benefit from knowing the optimal learning inputs (e.g. parenting, specific curricular activities) that are conductive to social and emotional development during each period of the child's development. They would also be interested in learning about the degree of malleability of social and emotional skills and the optimal mix of social and emotional skills during each developmental period. This information would help education stakeholders prepare a sequence of learning environments that would allow children to accumulate sufficient levels of social and emotional skills before they enter adulthood. There is a need to develop better data and analyses to disentangle such a complex nature of dynamic skill formation. The proposed Longitudinal Study of Social and Emotional skills in Cities is expected to contribute to this much-needed evidence.

4. RATIONALE

The proposed social-emotional skills framework was developed according to the following set of **general principles**. More detailed information will be provided later in this section.

Strong empirical foundation

The proposed framework should be based on a strong empirical foundation. Recent advances in the field of personality psychology have identified the *Big Five personality taxonomy* (John, 1990), also referred to as the Five-Factor Model (FFM; McCrae & Costa, 1996) as the most empirically compelling model that can serve as the starting point of developing a more comprehensive framework that meets the needs of the Longitudinal Study of Social and Emotional skills in Cities. The core dimensions of this model are usually referred to as Extraversion (vs. Introversion), Agreeableness (vs. Antagonism), Conscientiousness (vs. Lack of Direction), Emotional Stability (vs. Neuroticism), and Openness to experience (vs. Closed-mindedness). As literally thousands of concepts have been proposed to describe and explain individual differences in personality functioning, the emergence of this integrative taxonomy and its general acceptance in the 1990s has brought order and coherence into the field and has led to a remarkable surge in research productivity and findings (John, Naumann, & Soto, 2008). Although the Big Five taxonomy was initially derived from research on adults, it has been well-documented that the Big Five are suitable to describe personality differences from childhood to old age (e.g., De Fruyt & De Clercq, 2014; Soto, John, Gosling, & Potter, 2008, 2011). The proposed framework should also be **cross-culturally relevant** given the diverse countries and cultures that the Longitudinal Study of Social and Emotional Skills in Cities is scheduled to cover.

High predictive power and comprehensiveness

Given the ambition of this study is to explain and predict a variety of life outcomes, a comprehensive model of skills will have to be assessed across the developmental trajectory. The Big Five dimensions broadly capture the underlying core qualities of the individual—typical patterns of thoughts, feelings, and behaviours—that drive their lifetime success in life, and thus provide a parsimonious and highly efficient summary. However, such a parsimonious model with few concepts is by necessity very broad and limited in predicting specific outcomes (e.g., Hampson, John, & Goldberg, 1986). Thus, the broad level of the Big Five domains is unlikely to represent the most appropriate level of assessment for all the goals of this project, including understanding growth trajectories, examining the impact of different sorts of environmental factors, or explaining consequential outcomes. The measurement model proposed here will hence need further specification at a more fine-grained level, ultimately with about 3 to 5 facets comprising each of the broad Big Five domains. The final set of non-cognitive constructs should thus enable researchers to follow individuals across both broad and more specific levels of the personality trait hierarchy. Moreover, traits measured at the more specific facet level have the potential to be combined into skill compounds, better reflecting the complex nature of socio-emotional skills. A construct like Grit (Duckworth, Peterson, Matthews, & Kelly, 2007), for example, may be conceptualized as the aggregate of particular facets of Conscientiousness and possibly Extraversion (i.e., the enthusiasm facet here). As explained below, although there is a widespread consensus on the Big Five, there has been less research and thus agreement on the structure of the lower-order level of facet traits (e.g., John et al., 2008, Table 4.3). A crucial task for future work will hence be to identify those facets that are most likely to predict key outcomes of interest.

Malleability

The socio-emotional skill battery will have to include measures that are sensitive to detecting reliable changes, including normative change pattern (Roberts, Walton, & Viechtbauer, 2006; Roberts, Wood, & Smith, 2005) that so far have been studied much more widely in adults than in younger age groups. For example, researchers have hypothesized increases in conscientiousness, agreeableness and emotional stability to begin in mid-to-late adolescence, but evidence has been mainly come from large studies that used cross-sectional (not longitudinal) designs studies (e.g., Soto, John, Gosling, & Potter, 2011). Moreover, theory also suggests that individual change patterns will be important, as the onset and timing of age-related changes will differ across individuals; again, supporting evidence is available from adult samples but there are only a few longitudinal studies (De Fruyt et al., 2006; Prinzie et al., 2005) that have followed the same children or adolescents over longer periods of time, and all have been done in Western societies. Thus, the present design will include a variety of cultures and multiple assessment points to be able to chart non-linear forms of growth, and to distinguish latent groups of persons following specific developmental trajectories, beyond or deviant from normative change (De Fruyt & Van Leeuwen, 2014).

Temporal stability

The socio-emotional skills presented in the framework tend to be "manifested in consistent patterns of thoughts, feelings and behaviours"; that is, they show sufficient temporal consistency over short time intervals (e.g., one month) to allow for reliable measurement. Indeed, if a characteristic was largely determined by temporary circumstances and showed no temporal stability, then that characteristic would not be considered a skill but a transient state or behaviour. Empirical evidence suggests that many of the lower-order facets presented in Table 3.1 are likely to satisfy the condition of temporal stability to permit reliable measurement (see Almlund, et al., 2011; and Kautz et al., 2014).

4.1. Strong Empirical Foundation: The Big Five Personality Domains

The Empirical Development of a Taxonomy for Personality Attributes

Until the late 1980s, the Big Five personality dimensions, now seemingly ubiquitous, were hardly known (see John et al., 2008). Researchers and practitioners were faced with a bewildering array of personality scales from which to choose, with little guidance and no organizing theory or framework at hand. What made matters worse was that scales with the same names might measure concepts that were quite different, and scales with different names might measure concepts that were quite similar. Systematic accumulation of research results and communication across researchers was impossible amidst this cacophony of competing concepts and scales.

In the 1970s and 1980s, researchers at the University of California, Berkeley, for example, measured personality with only two concepts (e.g., Block's two dimensions of Ego-resilience and Ego-control), with the four scales on the Myers-Briggs Type Indicator (MBTI) as well as the 20 scales on the California Psychological Inventory. Many personality researchers were hoping to be the one who would discover the "true" structure that all others would then adopt, thus transforming the fragmented field into a community speaking a common language. However, we now know that such integration was not to be achieved by any one researcher or any one theoretical perspective. As Allport once put it, "Each assessor has his own pet units and uses a pet battery of diagnostic devices" (1958, p. 258).

The field of personality psychology lacked a descriptive model (or taxonomy), of its subject matter. One of the central goals of scientific taxonomies is the definition of overarching domains within which large numbers of specific instances can be understood in a simplified way (John, 1990). Thus, in personality, a generally accepted taxonomy would specify domains of related personality characteristics

that researchers could study, rather than studying separately each of the thousands of particular attributes that make human beings individual and unique. Moreover, such a taxonomy can help accumulate, summarize, and communicate empirical findings by offering a standard nomenclature.

After decades of research, personality psychology has finally agreed upon a general taxonomy of personality traits, the so-called Big Five personality domains. Table 4.1 presents a brief definition and summary of each domain. Although there remain some critics and contrarians (e.g., Block, 2010), there is now considerable agreement about the Big Five in the personality literature (for a review, see John et al., 2008).

However, there is less agreement about the more specific components or *facets* that define each Big Five domain at a lower level of abstraction. As summarized in Table 4.2, Soto and John (2015) recently reviewed the major facet models, which range from a minimum of only 2 facets per Big Five domain (DeYoung et al., 2006) to 6 facets per domain (Costa and McCrae, 1992). Considering the communalities among the existing models, three major facets emerged as core themes that virtually all of the different models had in common. Soto and John thus concluded that these three facets are widely accepted as important in the field and thus the most worthy to focus on. For example, each of the three facets for Negative Affect reflects one of the three basic emotions of anxiety/fear, sadness, and anger that are commonly accepted as fundamental and universal by emotion researchers (Ekman, 1972).

These three lower-order facet traits are also listed in Table 4.1 with the definition for each Big Five domain. In other words, the Big Five model specifies more than 5 basic constructs because it is hierarchical; instead, each of these five big (i.e., broad) domains consists of several more narrowly defined constructs at a lower level of abstraction (see John et. al., 2008, Table 4.3).

Finally, the Big Five dimensions do not represent a particular theoretical perspective but were derived from analyses of the natural language terms people use to describe themselves and others, much like the 21st Century Skills concepts listed in Annex 1. Rather than replacing previous systems, the Big Five taxonomy can serve an integrative function: it represents the various and diverse earlier systems of personality description within a single, common framework (see John et al., 2008, Table 4.1). Could the same taxonomy be applied to bring some order to the myriad number of diverse socio-emotional skills constructs?

Table 4.1. The Big Five OCEAN of Personality Domains: First-Letter Abbreviations, Verbal Labels, Conceptual Definitions, and Three More Specific Facet Traits in Each Domain

O: Openness, Originality, Open-mindedness
The attributes in this domain describe the breadth, depth, originality, and complexity of an individual's mental and experiential life. Facet traits include *Intellectual curiosity*, *Imagination/creativity*, as well as *Aesthetic and spiritual awareness*.
C: Conscientiousness, Constraint, Control of Attention
The attributes in this domain describe socially prescribed, effortful self-control that facilitates task- and goal-directed behaviour, such as thinking before acting, delaying gratification, following rules and norms, and planning, organizing, and prioritizing complex and long-term tasks. Facet traits include *Self-discipline, Orderliness*, and *Reliability*.
E: Extraversion, Energy, Enthusiasm
The characteristics in this domain involve an energetic approach toward the social and material world and include more specific facet traits such as *sociability, assertiveness*, and *positive activity*.
A: Agreeableness, Altruism, Affection
These characteristics contrast a pro-social and communal orientation towards others with antagonistic or antisocial tendencies, and include facet traits like *compassion, respect /politeness*, and *trust*.
N: Negative Emotionality, Nervousness, Neuroticism
These characteristics contrast emotional stability, confidence, and even-temperedness with the tendency to experience negative emotions, such as feeling *Anxious/nervous, Sad/depressed,* or *Angry/frustrated*.

An Exploratory Pilot Study of Self-reported Socio-emotional Skills: Elaborating the Socio-Emotional Content of the Five Personality Factors

Consider the 21st century attributes listed in Annex 1. To identify links between these socio-emotional skills and the Big Five model in adulthood, John and Mauskopf (2015) conducted a pilot study of self-rated socio-emotional skills and personality characteristics. Specifically, 452 volunteers were presented with an on-line questionnaire that included both socio-emotional skill items and the items from the standard Big Five Inventory. Correlational and factor analyses of these self-ratings showed that the socio-emotional skill list contained much content that was related to the Big Five dimensions. Table 4.3 shows examples of the 21st Century Skills items that had the strongest correlations with the Big Five personality dimensions.

Table 4.2. Minimal Set of Three Facets for Each Big Five Domain Based on a Review of Previous Facet Models

Facets minimally necessary to represent each Big Five	NEO PI-R (Costa & McCrae, 2008)	AB5C (Goldberg, 1999; Hofstee et al., 1992)	Lexical subcomponents (Saucier & Ostendorf, 1999)	Big Five aspects (DeYoung et al., 2006)
Extraversion				
Social connection	Gregariousness	Gregariousness	Sociability	Enthusiasm
Assertiveness	Assertiveness	Assertiveness	Assertiveness	Assertiveness
Enthusiasm	Activity	—	Activity-Adventurousness	Enthusiasm
Agreeableness				
Compassion	Altruism	Understanding	Warmth-Affection	Compassion
Respect for others	Compliance	Cooperation	Gentleness	Politeness
Trust	Trust	Pleasantness	—	—
Conscientiousness				
Organization	Order	Orderliness	Orderliness	Orderliness
Self-discipline	Self-Discipline	Efficiency	Industriousness	Industriousness
Responsibility	Dutifulness	Dutifulness	Reliability	—
Negative Emotionality				
Anxiety	Anxiety	Toughness (R)	Emotionality	Withdrawal
Depression	Depression	Happiness (R)	Insecurity	Withdrawal
Volatility	Angry Hostility	Stability (R)	Irritability	Volatility
Openness to Experience				
Aesthetic interests	Aesthetics	Reflection	—	Openness
Intellectual curiosity	Ideas	Intellect	Intellect	Intellect
Creative imagination	—	Ingenuity	Imagination-Creativity	—

Note. Based on Soto and John (2015), who used the Big Five labels Negative Emotionality (rather than the older Neuroticism) and Open-mindedness (rather than Openness)

However, despite many clear similarities, there were some noteworthy differences when the socio-emotional item content was considered. First, corresponding to Trilling and Fadel's (2009) emphasis on positive characteristics and interpersonal strengths needed for interlinked *Collaboration*, the largest factor was interpersonal and related to the Big Five domain described earlier as "A" (i.e., Altruism, Affection, Agreeableness). This factor attracted by far the largest number of 21 century skills items. The emphasis here was more explicitly on genuine mutuality and reciprocal exchange. In addition to standard facets of compassion, respect/politeness, and trust, several additional items suggested a potentially new facet, captured well by the item *Living in harmony with others*, along with other interpersonal skills related to interconnectedness and inclusiveness.

As shown in Table 4.3, the second 21st Century Skill factor may be described as *Task Performance*. It was defined by a large number of attributes and was conceptually quite similar to the personality domain of Conscientiousness. In addition to items representing Self-discipline, Organisation, and Responsibility facets, another group of socio-emotional skill items suggested a facet best described as Goal orientation, which highlights positive motivational characteristics like effort, work ethic, and productivity. Again, the strength-based, positive-psychology origin of the 21st Century Skills items rounds out this version of the Big Five "C" factor in a more substantial way.

The third factor, *Emotion Regulation skills*, also highlights positive strengths, whereas the traditional personality literature had focused on the negative, distressing emotions defining the other pole of this dimension. In terms of the more specific facets, we see the opposite of *anxiety*, worry, and avoidance, namely: self-esteem and self-confident, decisive tackling of tough problems. Instead of *sadness* and depression, the focus is on happiness and cheerful optimism. Instead of *anger*, temper, and frustration, there is equanimity, tranquillity, and balance. An additional aspect of emotional strength here is self-compassion (and self-kindness), similar to Kristin Neff's (2007) constructs anchored in research on mindfulness that help the individual avoid self-blame and respond to failures and set-backs in a measured, self-accepting, and normalizing way.

The fourth socio-emotional factor emphasized skills that allow the individual to constructively and joyfully engage with others in their social world. Even though "only" 4th largest in number of items, this factor reminds us of the great importance of positive engagement with the school environment for children and adolescents. Interestingly, the cluster of items related to the Assertiveness facet is enriched by items highlighting proactive strengths, such as leadership and charisma, courage, and the willingness to take a stand. The Enthusiasm facet is enriched by positive-psychology concept like passion and zest for life, spunk and spontaneity, as well as playfulness and humour. We hope there is a place for these positive characteristics in our schools, and that our schools will encourage and nurture these positive emotions in our children.

Table 4.3. Socio-Emotional Elaboration of the Big Five: Examples of Self-reported 21st Century Skills

Factor I: Collaboration (Related to Big Five Agreeableness)
1 Compassion, care, cooperation, kindness
2 Respect for others, empathy, tolerance, fairness
3 Trust, forgiveness, gratitude, appreciation of others
4 Living in harmony with others, interconnectedness, inclusiveness
Factor II: Task Performance (Related to Big Five Conscientiousness)
1 Self-discipline, focus, perseverance, self-control at school, grit
2 Organization, diligence, precision
3 Dependability, reliability, consistency, trustworthiness
4 Goal orientation, motivation, work ethic, effort, productivity
Factor III: Emotion Regulation (Related to low levels of Negative Emotionality)
1 Self-confidence, self-esteem, decisiveness, tackling tough problems
2 Cheerfulness, happiness, optimism
3 Tranquillity, balance, stability, equanimity (composure and even temper in difficult situations)
4 Self-compassion, self-kindness (being positive and understanding towards yourself when you suffer, fail, or feel inadequate)
Factor IV: Engagement with Others (Related to Extraversion)
1 Social connection, teamwork, social awareness, public speaking
2 Assertiveness, leadership, courage, charisma, speaking out/taking a stand, bravery
3 Enthusiasm, passion, zest, inspiration, spunk, spontaneity, playfulness, humour
Factor V: Open-mindedness: The Inquiring Mind (Related to Openness)
1 Curiosity, inquisitiveness, willingness to try new ideas, receptivity
2 Innovation, vision, insight, tinkering (inventing), learning from mistakes and failures, excitement of creating something new
3 Appreciating beauty in the world, living in harmony with nature, spirituality, mindfulness, existentiality, awe, wonder, reverence
4 Self-reflection, self-awareness, consciousness, self-actualization, authenticity

Note. Based on John and Mauskopf (2015).

The fifth socio-emotional domain was, *Open-Mindedness*: *The Inquiring Mind*. Just as in the earlier research on personality, this factor was defined by the smallest number of socio-emotional skill items, even though theoretical writings on 21st Century skills greatly emphasize the importance of intellectual curiosity and exploration as well as innovation and creativity. Nonetheless, the items classified in the *Innovation* facet here in Table 4.3 included interesting and novel features, such as having vision and insight, tinkering and learning from mistakes, and the excitement of creating something new. Interestingly, there were few, if any, of the usual Openness items that relate to the standard Openness facet of aesthetic interests or sensitivity, with its emphasis on art, music, and literature. It is possible that these characteristics are underrepresented in more technology and employment oriented collections of 21st century skills; instead, this facet emphasized appreciating beauty, living in harmony with nature, and emotions relevant to spirituality, such as awe, wonder, and reverence. Also of interest is an additional facet that we have included in the present OECD framework, namely self-reflection and awareness of self and inner experiences (like the facet *Openness to feelings* that is included on the NEO PI-R). We selected that facet because it also appeared in our review of the literature on emotional competence (see below).

In conclusion, these broad domains defined by socio-emotional skill characteristics bear enough similarity to the familiar and well-studied Big Five personality to give us some confidence about their likely replication and generalizability. At the same time, the content of these five socio-emotional skill factors emphasizes their unique origin in 21st century skills and positive-psychology, with its approach based in strengths and virtues (Seligman, Steen, Park, and Peterson, 2005), and can thus advances our understanding beyond from the hierarchical personality taxonomy of the Big Five and the three core facets shown in Tables 4.1 and 4.2. The socio-emotional characteristics summarized in Table 4.3 provide a starting place for a new integrative and operational definition of socio-emotional characteristics that can be implemented in a longitudinal study on course and impact of social-emotional skills. More generally, socio-emotional skills are best defined as *individual characteristics that (a) originate in the reciprocal interaction between biological predispositions and environmental factors, (b) are manifested in consistent patterns of thoughts, feelings and behaviors, (c) continue to develop through formal and informal learning experiences, and (d) influence important socioeconomic outcomes throughout the individual's life* (OECD, 2015; De Fruyt, Wille, and John, 2015; Primi, Santos, John, and De Fruyt, submitted).

Reasons for Beginning the Conceptual Framework with the Big Five Facets as a Minimal Set

Previous OECD reports (e.g., Kautz et al., 2014) as well as economists working on socio-emotional skills have simply adopted the Big Five personality dimensions as the conceptual framework of choice because they found that most of the empirical research on the development and longer-term impact of socio-emotional characteristics has been conducted with Big Five measures. Specifically, in Kautz et al.'s (2014) report *Fostering Non-Cognitive Skills to Promote –Lifetime Success*, five economists from three countries concluded:

> "Although non-cognitive skills are overlooked in most contemporary policy discussions and in economic models of choice behavior, personality psychologists have studied these skills for the *past century*.
>
> Psychologists primarily measure non-cognitive skills by using self-reported surveys or observer reports.
>
> They have arrived at a relatively well-accepted taxonomy of non-cognitive skills called the Big Five, with the acronym **OCEAN**, which stands for: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism."

Similarly, the recent report by the National Academy of Sciences (2012) in the United States, entitled "*Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century"* observes: "For the past two decades, the big five model of personality has been widely accepted as a way to characterize competencies in the interpersonal and intrapersonal domains" (p. 28).

In contrast to the view, commonly held by laypersons, that personality traits are fixed, this report further cites extensive research demonstrating that personality traits are malleable and subject to multiple influences even during adulthood (e.g., Srivastava et al., 2003), and concludes that "these traits can be altered by experience, education, parental investments, and targeted interventions" (p. 24). Reviewing the available research evidence, the report then concludes:

> "The five major factors provided a *small number of research-based constructs* (emphasis added) onto which various terms for 21st century skills could be mapped. The facets helped to define the range of skills and behaviors encompassed within each major factor to serve as a point of comparison with the various 21st century skills."

Further, important evidence comes from the pilot study conducted jointly by the OECD and the Ayrton Senna Institute in Brazil, which sampled more than 21,000 students from 5th to 12the grade in public schools in the State of Rio de Janeiro (OECD, forthcoming). This research used both careful conceptual reviews and extensive analyses of the new data collected in Brazil. It reached a similar conclusion, namely that the Big Five, supplemented with an assessment of positive core self-evaluations (e.g., self-esteem), would provide a comprehensive framework for the socio-emotional skills of children aged 10 to 19 in Brazil. A related background study (Primi and Santos, 2015) carefully reviewed all existing child and adolescent instruments assessing socio-emotional constructs and then selected the most promising ones, which were then translated into Portuguese. Through a series of pilot studies, these researchers arrived at a final set of 65 socio-emotional skills items for younger children (ages 9-14) and 96 items for older children (14-19). Remarkably, in their highly diverse samples of Brazilian children in public schools, analyses of these items again produced the now familiar Big Five, plus a 6th dimension capturing extremely negative core self-evaluations (i.e., low self-esteem and external locus of control). For example, their careful empirical mapping studies (see also Primi et al., 2014, submitted) showed that the scales on such child assessment instruments as the *Strengths and Difficulties Questionnaire* (SDQ) and the three-dimensional *Self-Efficacy Scales for Children* (SES) could be represented well within the more comprehensive Big Five model. Social skills items like "I am able to tell a joke or story to a group of friends at school" loaded with other items from the Engagement with others (or extraversion) factor; The Prosocial strengths (respect) item "I try to be nice to other people. I care about their feelings" loaded on the Collaboration (or agreeableness) factor. The time management item "I am able to complete all my homework" loaded with other items from the Task performance (or conscientiousness) factor; and the item indicating a lack of emotion regulation skills "I get very angry and lose my temper" loaded on the Emotion regulation (or stability) factor.

In short, both conceptual and empirical evidence point to the promise of a framework that has an empirical foundation in the insights and 20-year research record accumulated for the Big Five. However, it is the view of the expert committee charged with developing the conceptual framework for the new OECD longitudinal study of skill development that this framework ought to *go beyond the Big Five*, in two ways. First, in contrast to the two previous OECD reports (Kautz et.al, 2014; OECD, 2015b), which examined only the superordinate-factor level of the Big Five, we propose to assess socio-emotional skills not only at the factor level but also at the lower level of more specific facets (as illustrated in Tables 4.2 and 4.3). Second, in the remainder of this report, we will conduct comprehensiveness checks of the proposed

framework by reviewing other frameworks, approaches, cultural perspectives, and research findings, supplementing the initial draft framework based on the Big Five with additional constructs as needed. Before we turn to these comprehensiveness checks, we first review the evidence that is already available for the consequential validity of some of the socio-emotional skill concepts in the proposed framework.

Cross-cultural relevance of the framework

The OECD Longitudinal Study of Social and Emotional Skills in Cities intends to launch longitudinal surveys in a variety of countries across the world. This multi-national approach poses specific challenges for the cross-cultural validity of the proposed framework and its constructs, and the psychometric requirements for its operationalization.

Is the proposed framework cross-culturally relevant?

Nowadays, there is evidence that the Big Five dimensions can be recovered in the cultural and language communities to be included in the OECD Longitudinal Study of Social and Emotional skills in Cities (e.g., Mexico, Brazil, South Korea, Japan, Norway, and Russia) and that in these countries the Big Five capture core qualities underlying a broad range of personality characteristics. Conceptually, the Big Five have to be understood as the basic colours of character and personality, expressed as more pure or blended manifestations of individual differences at the phenotypic level. Although the Big Five are a guiding framework to structure personality descriptions across the world, this does not necessarily implies that these are the only important constructs to consider.

Reviewing the literature, it is clear that, when transported to another culture, the Big Five factors can be recovered in self and observer ratings. Early studies compared the factor structure of personality ratings across individual countries, such as the USA, Spain, and Mexico (e.g., Benet-Martinez and John, 1998). More recently, large international teams of investigators have collaborated on larger-scale studies. Schmitt and colleagues (2007), for example, could retrieve the five-factor structure when the Big Five Inventory (BFI) was translated from English into 28 different languages administered to 17,837 individuals from 56 countries, including Argentina, Australia, Austria, Bangladesh, Belgium, Bolivia, Botswana, Brazil, Canada, Chile, Chinese Taipei, Croatia, Cyprus, Czech Republic, Democratic Republic of Congo, Estonia, Ethiopia, Fiji, Finland, France, Germany, Greece, Hong Kong, India, Indonesia, Israel, Italy, Japan, Jordan, Latvia, Lebanon, Lithuania, Malaysia, Malta, Mexico, Morocco, Netherlands, New Zealand, Peru, Philippines, Poland, Portugal, Romania, Serbia, Slovakia, South Africa, South Korea, Spain, Switzerland, , Tanzania, Turkey, Ukraine, United Kingdom, United States and Zimbabwe. Likewise, McCrae and colleagues (2005) found support for the FFM structure when analysing college-students' NEO-PI-R descriptions (N=11,985) of college-aged (18-21) and adult-aged (>40 years) individuals from 50 nations, including Argentina, Australia, Austria, Belgium, Botswana, Brazil, Burkina Faso, Canada, Chile, China, Croatia, Czech Republic, Denmark, Estonia, Ethiopia, France, Germany, Hong Kong, Iceland, India, Indonesia, Italy, Japan, Kuwait, Lebanon, Malaysia, Malta, Mexico, Morocco, New Zealand, Nigeria, Peru, Philippines, Poland, Portugal, Puerto Rico, Russia, Serbia, Slovakia, Slovenia, South Korea, Spain, Switzerland, Thailand, Turkey, Uganda, United Kingdom: England, United Kingdom: Northern Ireland, and the United States. De Fruyt et al. (2009) provided similar evidence analysing descriptions of adolescents (12-17) obtained in 24 cultures, including Argentina, Australia, Chile, People's Republic of China, Croatia, Czech Republic, Estonia, France, Hong Kong, Islamic Republic of Iran, Japan, Malaysia, Peru, Poland, Portugal, Puerto Rico, Russia, Serbia, Slovak Republic, South Korea, Thailand, Turkey, Uganda, and the USA. The explicit listing of countries illustrates that these are well-spread across North- and South America, Western, Eastern and Southern Europe, the Middle East and Africa, Oceania, and South/South-East Asia and East Asia.

Whereas the previous work concentrated on the replication of the structure of imported inventories constructed in the US (i.e. the BFI or the NEO-PI-R) and administered to adolescents (De Fruyt et al., 2009) or adults (Schmitt et al., 2007; McCrae et al., 2005) in a broad variety of cultures, the International Consortium for the Developmental Antecedents of the Five-Factor Model (ICDA-FFM; Kohnstamm, Halverson, Mervielde, and Havill, 1998) worked bottom-up in various cultures examining the content and structure of parental free descriptions to define the structure of personality for children. This group of developmental, temperament and personality researchers examined the active instead of the passive (like represented in dictionaries) personality descriptive vocabulary. At the same time, they aimed to assemble item sets that represented age-specific indices of individual differences that were more sensitive to describe developmental differences in childhood. Nearly all studies conducted before 1995 had used adjective lists or inventories initially developed for adults preventing the emergence of childhood specific personality dimensions or facets.

Their methodology was simply asking parents with children in the age-range of 6 to 12 (primary school) to describe what they found characteristic of their child, without any additional prompts to avoid influencing the content of their descriptions. All descriptors collected this way were subsequently sorted in a personality descriptive lexicon including 14 major categories, containing the Big Five supplemented with a number of temperament categories. This age-grouped descriptor database served as the starting point for developing age-specific item sets to enable bottom-up research on the structure of personality in youth. The ICDA-FFM group used this approach in various countries, including Belgium, China, Germany, Greece, the Netherlands, Poland, and the United States, to be in a position to study the impact of culture on the resulting structure. This basic and innovative approach provided a strong test of the validity of the comprehensiveness of the Five-Factor Model from a cross-cultural *and* a developmental perspective.

Mervielde and De Fruyt (1999, 2002) used this approach and assembled a pool of near to 10.000 parental free descriptions of Flemish children aged between 6 and 13 years. Lexicon categories were further split in about 100 homogeneously descriptive categories, and their content was represented by 2 to 3 personality items for each age-group (6 year-, 9-year, and 12-year olds). These item sets were subsequently administered to large samples of parents and teachers requested to rate children aged 6 to 12 years. An analysis of the factor structures within and across age groups and gender clearly pointed towards the same five factors, identified as extraversion, benevolence (agreeableness), conscientiousness, emotional stability or neuroticism, and imagination (openness). Benevolence referred to a broader set of traits than the adult agreeableness factor, referring to content associated with the concepts ‘easy-difficult’ child described in the temperament literature (Thomas, Chess, Birch, Herzig, and Korn, 1963) and ‘manageability’ from the perspective of the parent or teacher informant.

Despite the initial focus to work with trait indicators for specific age-groups, the resulting sets represented a very similar behavioural content and a highly similar higher-order structure. Mervielde and De Fruyt (1999, 2002) additionally examined the lower-level structure across age-groups and proposed a common set of 18 facets (with 8 items per facet) to describe children from 6 to 12 in the Hierarchical Personality Inventory for Children (HiPIC; Mervielde and De Fruyt, 1999; Mervielde, De Fruyt, and De Clercq, 2009). The emotional stability domain included two facets, i.e. anxiety and self-confidence, whereas extraversion grouped four facets: energy, expressiveness, optimism and shyness. Imagination included creativity, intellect and curiosity, whereas Benevolence distinguished among altruism, dominance, egocentrism, compliance and irritability. Finally, conscientiousness grouped the facets concentration, perseverance, orderliness and achievement striving. The American team of the ICDA-FFM, led by Halverson, constructed the Inventory for Child Individual Differences (ICID; Halverson et al., 2003), following a similar starting point. The resulting factor solution of the ICID was comparable to that of the HiPIC (Tackett, Kushner, De Fruyt, and Mervielde, 2013).

When researchers operationalized personality with indigenous concepts rooted in a particular culture. some studies have encountered difficulties recovering all of the Big Five factors, or have found factors that were slightly different from those in Western studies. Cheung and colleagues (2001), for example, jointly examined the structure of the NEO-PI-R and the Chinese Personality Assessment Inventory (CPAI; Cheung et al., 1996). Their results provided powerful evidence for the replication of four of the Western Big Five domain (namely, Neuroticism, Conscientiousness, Agreeableness and Extraversion). In addition their results suggested a separate 'interpersonal relatedness' factor (see the Relationship Harmony concept included in the present framework) and three specific Openness facets (O3: Feelings, O2: Aesthetics, and O1: Fantasy) chiefly loading an openness factor. The additional factor interpreted as "interpersonal relatedness" had loadings only from CPAI scales, including Harmony, Optimism (versus Pessimism), Ren Qing (relationship orientation), Flexibility, Defensiveness (Ah-Q mentality), Face, and Logical versus Affective Orientation. This additional factor could not be absorbed by the NEO-FFI factors across three different samples; the variances explained by the FFM ranged from as low as .08 (Ren Qing; sample of 372 Chinese managers) to .31 (Flexibility; same sample).

A project with impressive bottom-up work has been conducted in South Africa, making use of the lexical approach to identify the local vocabularies of personality description and then derive basic dimensions of personality from that emic material. Nel et al. (2012) developed the South African Personality Inventory (SAPI) by starting with the personality descriptive language obtained in semi-structured of more than 1,200 individuals representing the 11 major language groups in South-Africa. This vocabulary was grouped into 9 broad content-based clusters: conscientiousness, emotional stability, extraversion, facilitating, integrity, intellect, openness, relationship harmony, and soft-heartedness. Those clusters that were not represented by the Big Five were all related to social-relational functioning and tapped into ways of maintaining positive interpersonal relationships with others, and could thus be conceptually linked to the Agreeableness in the Big Five.

Valchev and colleagues (2014) conducted a series of follow-up studies, and reported that the social-relational scales of the SAPI generated two factors not presented in the Big Five, assessed by the BTI (Taylor and De Bruin, 2005) with items from the International Personality Item Pool (IPIP; Goldberg et al., 2006). The positive social-relational factor was defined by facilitating, integrity, relationship harmony, active support, empathy, facilitating, and integrity, and South-African Blacks scored higher on this factor than did Whites. Moreover, social-relational concepts explained substantial variance in pro-sociality beyond the Big Five. Finally, there was evidence that the Big-Five-Plus-Two factor-structure could be also recovered from a joint factor analysis of 50 IPIP items and SAPI social relational scales administered to a mixed sample of 452 mainstream Dutch, 427 Western, 225 Antillean, Surinamese, and Indonesian, and 179 non-Western participants. In sum, evidence for the cross-cultural generalizability of the basic Big Five taxonomy is substantial overall, but *additional dimensions may be necessary to provide a comprehensive structural representation of personality in particular cultures*. The studies reviewed so far suggest the most promising candidates for additional constructs are likely to be found in the relational domain.

More specifically, psychological studies of cultures suggest that the structure and meaning of social relationships may differ across cultures (e.g., Markus and Kitayama, 1991; see Benet-Martinez and Oishi, 2008 for a review). We therefore expect that additional constructs may be needed particularly in the interpersonal domain, supplementing constructs broadly situated in the Agreeableness domain. One important candidate construct involves the *construal of self as interdependent in relationships* (e.g., Markus and Kitayama, 1991; Singelis, 1994); research has shown that individuals in East Asian societies with more "collectivistic" values (e.g., China, Korea, Japan) score higher on interdependence than individuals in Western societies with more "individualistic" values (e.g., USA; Australia); the opposite pattern is obtained for interdependent self-construal. These variables are potentially important because they

can help explain, at the individual level, important cultural differences in life satisfaction and well-being, which are important life outcome indicators and considered central by the OECD.

Another promising candidate is *relationship harmony*, mentioned above in the South African studies. Kwan, Singelis, and Bond (1997) proposed a relationship-based measure for this concept and compared East Asian countries (e.g., Hong Kong) with the US. Their careful analyses showed that self-esteem influenced life satisfaction more than relationship harmony in the US whereas both factors were equally important in Hong Kong (Kwan et al., 1997). They also measured individual differences in independent and interdependent self-construal in both cultures; the effect of *independent* self-construal on life satisfaction was mediated through *self-esteem*, whereas the effect of *interdependent* self-construal was mediated through relationship harmony. In short, these variables capture cultural differences of "self-in-relationships" that can be measured at the level of the individual. Indeed, members of East Asian cultures, even when they live abroad in individualistic cultures (like Asian-Americans in the US) consistently report somewhat lower levels of life satisfaction, happiness, and self-esteem than do Western countries (like the US and Western Europe), and these kinds of differences need to be anticipated and conceptualized.

More recently, one additional concept has been suggested in the Journal of Cross-Cultural Psychology and there is reason to believe it will be relevant to the Emotion Regulation domain in the present framework. Specifically, it is *fear of happiness*, which can be measured with a short self-report scale (Joshanloo and Lepshokava, 2014). This research on individual differences is consistent with the work of Jeanne Tsai (e.g., Tsai, Knutson, and Fung, 2006), who suggested that cultures differ in the way they value particular emotions, especially emotions considered positive and exciting (or arousing) in the West, especially intense happiness states like joy, enthusiasm, and excitement (as well as love and pride). East Asian cultures have been shown to value these intense positive emotions less than Western cultures and thus express them less in publically observable behaviour. The Fear of Happiness scale explicitly measures various specific beliefs about happiness that individuals may learn, to varying degrees, from their culture and socialization experience. Some individuals tend to be more suspicious of feelings of great happiness, expecting something bad to happen when they allow themselves to be too happy, whereas others tend to embrace and indulge feelings of intense happiness. As expected, in an initial cultural comparison, East Asian countries (e.g., Japan) scored substantially higher on fear of happiness than Westerners (e.g., Western Europe), who in turn scored higher than Brazilians who had by far the lowest scores of all countries studied and seemed to embrace happiness without fear or suspicion (Joshanloo and Lepshokava, 2014). These cultural differences were substantial in size but need to be considered with caution until replicated. However, they hold the promise to understand, at the level of the individual, why East Asian consistently report lower levels of positive emotion and many other positively balanced attributes, such as self-esteem, life satisfaction, well-being, and even extraversion.

It is important to realize that this research on interdependent self-construal and relationship harmony did *not* suggest that these constructs could not be measured reliably in one of the two cultures involved. On the contrary, reliable measurement in both cultures made possible the demonstration of mean-level cultural differences and their explanations. In other words, individual differences in interdependence and relationship harmony exist and can be measured in the USA but they are less important and therefore less expressed there than in China. Indeed, the discussion on structural replicability (or invariance) has been complicated by arguments about the importance of particular factors across cultures. Structural replicability does not imply that factors have equal importance across cultures. For example, more collectivistic cultures may put higher value on Agreeableness facets related to politeness and compliance, whereas more individualistic cultures may value individual achievements and thus put higher value on traits associated with "standing out" from the group, such as assertiveness (i.e., Extraversion) and

achievement striving (i.e., Conscientiousness). These weighting differences reflect important cultural variations but the Big Five may be structurally replicable in both groups of cultures.

Measurement invariance and comparisons between- and within- cultures

Cultural differences not only play a role when comparing scores of individuals between cultures and countries, but societies within single countries have also become increasingly heterogeneous the past years in terms of the cultural backgrounds of their members. Half of the population at schools in capital cities, for example, may have various ethnic origins and pupils' cultural identity may be a mixture of characteristics from the host and the culture of origin. Moreover, people within a culture, may identify with multiple groups at the same time, such as ethnic origin, ethnic identity, gender, sexual orientation, age, and social-economic group. These different group attributes may interact and affect individuals' score patterns on psychological constructs (De Fruyt and Wille, 2013), introducing different forms of construct, method and item biases (Van de Vijver and Leung, 1997). The absence of bias is labelled as equivalence or measurement invariance.

Construct bias refers to the phenomenon that constructs only partly share meaning across groups. For example, the trait of Assertiveness is a relatively factor pure indicator of Extraversion in US samples, though taps into Extraversion and Emotional stability in Germanic languages like Dutch and German. Method biases refer to the differential impact on groups of the scale formats that are used or the way the assessment is conducted. For example, some groups may find it difficult to use a sorting procedure like the Q-sort, sorting different values, or choosing the item "most like you" and "least like you" from an ipsative item set. A final threat to measurement equivalence is differential item functioning (DIF): "DIF occurs when individuals with the same level or amount of a trait, but from different cultural groups, exhibit a different probability of answering the item in the keyed direction" (Church, 2010, p. 154).

Measurement invariance can further be demonstrated at different levels (Vandenberg and Lance, 2000), distinguishing among configural (same number of factors and pattern of loadings), metric (loading patterns constrained to be equal across groups), and scalar invariance (item intercepts are equal across groups). Church and colleagues (2011) recently examined DIF in the data that were collected in the USA, the Philippines, and Mexico, and showed that 40 to 50 percent of NEO-PI-R (Costa and McCrae, 1992) items exhibited some form of DIF, suggesting that one should be careful with making comparisons of mean scores across cultures. For example, Schmitt et al. (2007) compared self-reported BFI means observed in 10 world regions and found that the level of negative affect was very high in Japan, but also that East-Asians scored the lowest on conscientiousness, whereas the mean for Africans was the highest. Schmitt et al. (2007) raised that it is unlikely that Japanese would be perceived as low in self-discipline, order and achievement, and suggested that for some constructs and in some cultures, culturally endorsed response styles may be responsible for such effects. There are alternative, powerful methods to examine measurement invariance across cultural groups, such as the bilingual approach where the same samples are administered the same instrument in each of the two languages (e.g., see Benet-Martinez and John, 1998, Studies 2 and 3). When assessments are equivalent, the resulting mean scores should be alike (McCrae and Terracciano, 2008). More recently, Kyllonen and Bertling (2013) have suggested the use of vignettes to correct for potential biases in the use of response scales across cultures, and Primi et al. (2014, under review) have demonstrated that the vignette approach can be effective in studies of personality self-ratings made by children.

4.2. High Predictive Power and Comprehensiveness

Although the description of social-emotional skill development trajectories is a valuable research objective in itself, the OECD's Longitudinal Study of Social and Emotional Skills in Cities also aims to clarify the mechanisms of how these trajectories lead or contribute to a broad series of consequential outcomes. National contexts, in addition to individual circumstances, are considered in this respect as moderators and mediators of the outcomes.

Social-emotional skills may affect the listed outcomes below in multiple ways at the same time, but the driving mechanisms of outcomes can also differ across development. Social-emotional skills may affect outcomes directly or influence an outcome indirectly via another construct. Such effects can be independent effects, i.e. one or two social-emotional skills exerting an influence independent of each other, but there may be also interactive effects, substituting (resource substation model) or strengthening (Matthew) effects. Social-emotional skills may further act as a moderator of an association between a predictor and outcome. Traits and skills will not only affect the outcomes, but will also be shaped by the outcomes (i.e. reciprocal relationships). Finally, social-emotional skills can have short and long term effects.

Evidence on the predictive power of social and emotional skills

Educational attainment

School systems around the world focus on both knowledge acquisition and developing social-emotional skills in youth. In the impressive meta-analysis summarizing results of 213 school-based social-emotional learning programs conducted with pupils (N= 270,034) from kindergarten to high school already discussed in the previous section, Durlak and colleagues (2011) reported the effects of social-emotional learning ranging from .22 (conduct problems) to .57 (social-emotional skills). They also found, however, that training of social-emotional skills had direct effects on academic performance. Social-emotional skills are hence both targets and means in formal education programs.

There is a growing interest in educational psychology to examine more social-emotional variables, including interests, motivational factors and personality traits to explain the trends?/outcomes? in educational attainment. This research, focusing more in the typical behavioural indices affecting the study performance, complemented the well-established research line on more maximal predictors of learning outcomes such as cognitive abilities. Research by Strenze (2007), for example, showed that intelligence explained nearly a quarter of the academic attainment scores, with a corrected correlation coefficient of .56. A key question was to what extent does academic achievement also predicted by socio-emotional skills.

Although a series of specific traits have been examined with respect to educational outcomes, such as procrastination (Steel, 2007), grit (Duckworth and Seligman, 2006; Duckworth, Peterson, Matthews, and Kelly, 2007), and goal setting-engagement (Bipp and Van Dam, 2014), most work has been conducted using the broad Big Five domains rather than the specific facets, with a large number of replicated studies now dating back to more than 20 years (e.g., John et al., 1994). Poropat (2009) meta-analytically investigated the relationship between the dimensions of the FFM and academic performance. Some of the primary studies included in this meta-analytic summary also reported correlations between intelligence and academic performance, and were also quantitatively summarized. The sample-weighted correlation corrected for scale reliability between intelligence and academic performance was .25, hence substantially lower than the .56 reported by Strenze (2007). The larger effect size estimated by Strenze may be due to

sample variation or to the absence of range restriction, so the .25 reported by Poropat for intelligence provides a benchmark to interpret the relative weight of the FFM traits. The corrected correlations for the FFM scales and academic performance were .22 for Conscientiousness, .12 for Openness, .07 for Agreeableness, .02 for Emotional Stability, and -.01 for Extraversion; three of the Big Five, namely Conscientiousness, Openness, and Agreeableness were significant explanatory constructs. The correlation found for Conscientiousness alone almost equalled the one found for intelligence. Academic level was found to significantly moderate the FFM-academic achievement association, with the largest coefficients found in primary education for both intelligence and all FFM factors, and declines from primary to secondary and tertiary level for Intelligence, Agreeableness, Emotional stability, and Extraversion, and linear declines across the three levels for Openness. The correlation for Conscientiousness did not significantly alter across academic levels. The correlations in primary education were .58, .30, .28, .20, .18, .24 between academic achievement and respectively intelligence, Agreeableness, Conscientiousness, Emotional Stability, Extraversion and Openness. Poropat (2009) further found that Conscientiousness added little to the prediction of tertiary GPA when partialed out for secondary GPA, though still slightly performed better than intelligence. In a meta-analytic investigation of adult-rated child personality and academic performance in primary education (Poropat, 2014a), corrected correlations of .43, .18 and .50 were reported for Openness, Emotional Stability and Conscientiousness respectively, significantly outperforming the effects observed for self-ratings in the case of Openness and Conscientiousness. These relationships were not moderated by age or year of education (grades 1 to 7). The positive associations between Conscientiousness and Openness on the one hand and academic achievement on the other were also extended to other-rated personality (Poropat, 2014b).

A broad series of studies have directly related personality traits to education performance outcomes, with the FFM Conscientiousness, Openness and Neuroticism factors as key dimensions to describe achievement-relevant personality (Briley, Domiteaux, and Tucker-Drob, 2014). Spengler and colleagues (Spengler, Ludtke, Martin, and Brunner, 2013), following a large representative sample of 15-year-old students and another sample of students of the 9th and 10th grade, showed that conscientiousness better predicted grades (r= .15-.30), whereas openness was more strongly associated with performance (.15-.32) on math and reading items culled from the Programme for International Student Assessment (PISA; OECD, 2009).

Performance at work and employability

Industrial and Organizational psychologists have been interested for a long time in the predictors of job performance when assigning job applicants to vacancies. More recently, due to the volatile and changing employment market and requirements of life-long learning, this professional group got also strongly interested in the concept of 'employability', referring to individuals' labour market fitness and requirement to be in charge of their own careers. The interests of industrial and organizational psychologists hence overlapped considerably and increasingly with the concept of 21st century skills, given their increased attention for meaning of working (work or employment flows better?) for people, happiness and mental health at work, and work-life balance, going beyond mere indicators of job performance such as quantity and quality of task performance.

Parallel to the educational domain, also the Industrial and Organizational field do not have an agreed upon taxonomy to describe the basic qualities underlying individual difference predictors of job performance and employability. Selection psychologists and human resources professionals frequently introduce a new vocabulary to refer to the individual qualities they are looking for, labelling these the past 10 years as (behavioural) 'competencies', and more recently as 'talents' but also as '21st century skills'. Although the introduction of these new concepts underscores the dynamics of this field, it does not

necessarily facilitate the communication among various labour market stakeholders, and it imposes major challenges for assessing these qualities reliably and efficiently among job applicants or incumbents. There is usually not a one-to-one relationship between the listed competencies, socio-emotional skills, or talents on the one hand and the concepts and constructs for which differential psychologists have constructed particular models and assessment tools. Competencies like "Resiliency or low stress Vulnerability", for example, are very trait-like and easy to map into a personality descriptive model such as the FFM, and thus relatively straightforward to assess. However, competencies such as "having Impact", "having a Helicopter view" or ability to articulate "Vision" or "Deal with disputes" are more complex "hybrid" constructs that are best conceived as blends between more cognitive and more trait-liked characteristics. These latter examples nicely illustrate that the bifurcation between cognitive and non-cognitive factors is artificial, and often rather useless. Skills that help turn a conflict into a fruitful discussion require abilities to analyse, to communicate, and also to regulate emotions, hence, requiring a mixture of cognitive and non-cognitive resources in a single labelled skill.

Selection and assessment psychologists, with a background in individual differences, started to develop conceptual models to relate competency models to constructs from differential psychology. Hoekstra and Van Sluijs (2003) consider FFM personality traits and intelligence as building blocks of behavioural competencies, with formal and informal learning processes impacting upon the competency level during development. De Fruyt, Bockstaele, Taris and Van Hiel (2006) illustrated how this model can be used to link the FFM trait model and police interview competencies. The central position as building blocks of competencies given to traits and intelligence in Hoekstra and Van Sluijs' model (2003) makes these constructs important assets to examine an individual's employability. These approaches further illustrate that the applied field can borrow constructs and assessment methodology from the differential psychology fields.

Nowadays, there is convincing evidence in Industrial and Organizational Psychology that individual differences' constructs are key variables to assess when discussing performance at work and employability. There is considerable meta-analytic evidence that cognitive abilities are among the best predictors of job performance and training proficiency across a range of jobs and in different cultures (Salgado, Anderson, Moscoso, Bertua, and De Fruyt, 2003; Salgado et al., 2003), with also good validity early on to predict outcomes that are observed later in life such as income and occupational attainment (Judge, Higgins, Thoresen, and Barrick, 1999; Woods, Lievens, De Fruyt, and Wille, 2013). Validities for intelligence measures and job performance do not vary much across jobs (Salgado et al., 2003), showing that intelligence is predictive across jobs, though type of job moderates this relationship, with stronger associations for more complex occupations. In addition to explaining job performance directly, there is also evidence that intelligence is probably more important when individuals get a new employment, because cognitive capacity facilitates learning and helps the individual adapt to new challenges, whereas its validity seems to decline somewhat when individuals get more acquainted with their job, underscoring the importance of cognitive ability for labour market fitness or employability.

In a meta-analysis of FFM measures and job performance, Hurtz and Donovan (2000) reported estimated mean operational (true) validity coefficients (corrected for unreliability in the criterion measure and range restriction on the personality measures) for self-reported personality of .22 (Conscientiousness), .14 (Emotional Stability), .10 (Agreeableness), .09 (Extraversion), and .05 (Openness). However, using theory to align Big Five dimensions to specific job-performance aspects, Hogan and Holland (2003) reported true estimated validities of .43 (Emotional stability), .35 (Extraversion-Ambition), .34 (Agreeableness), .36 (Conscientiousness) and .34 (Intellect-Openness to experience). These findings unequivocally demonstrate that the underlying broad dimensions of personality description are related to various performance indicators valued by employers.

In the past years, there has been increasing efforts to raise the predictive validities of personality measures to understand performance at work. First, there was considerable debate about the level of the trait hierarchy at which predictions of job performance were best made. That is, at the level of broader domain factors such as the Big Five dimensions or at the level of more specific facets. A similar discussion was conducted at the level of the outcomes, further subdividing job performance into quantity and quality of task performance, in addition to contextual and adaptive performance. The latter approach is also in line with Hurtz and Donovan (2000) arguing to use theory to better align the predictors and outcomes.

A second improvement involve contextualized personality assessments; that is, rather than asking for a description of one's personality "overall" or "in general," the more specific and relevant *work context* is included as the frame-of-reference for the personality description (Lievens, De Corte, and Schollaert, 2008), for example by adding a tag "at work" to items. Not surprisingly, these more contextualized personality assessments are better aligned with the work criteria they are supposed to predict and thus show generally better predictive validities. De Fruyt and Rolland (2013) illustrated the combined effects of aligning predictors and criteria and using a work frame-of-reference, showing that self-rated Conscientiousness at work correlated .36 with colleague-rated task performance (relative to a correlation of .27 using a general personality Conscientiousness scale). Self-rated Neuroticism and Openness to experience correlated -.21 and .26 with adaptive performance rated by colleagues (relative to correlations of .16 and .12, respectively, for non-contextualized general measures). A third established improvement is to include additional observers beyond self-descriptions (Connelly and Ones, 2010). Oh, Wang and Mount (2010) meta-analytically examined the effect of adding 1 to 3 observer personality ratings to self-ratings, convincingly showing that validity increases adding observers, reporting coefficients of .41 (Conscientiousness), .24 (Emotional stability), .34 (Agreeableness), .29 (Extraversion) and .29 (Openness/Intellect) when 3 observer ratings were added to the earlier reported self-ratings by Hurtz and Donovan (2000). Oh and colleagues' meta-analysis hence suggest that all FFM traits explain job performance ratings, and should be incorporated in a comprehensive assessment of a job candidate to evaluate employability.

Whereas the previous innovations are getting progressively integrated into professional assessment practice, a few other routes are currently under investigation, and their merits to improve prediction have yet to be established. A primary group of innovations is situated at the assessment side, examining the predictive validity of implicit measures of personality (Vecchione, Dentale, Alessandri, and Barbaranelli, 2014) or using a situational judgment paradigm (Lievens and Sackett, 2012) to describe an individual's trait positions. These approaches try to find alternative indices of personality beyond the traditional questionnaire approach. A second valuable line of research is more state-oriented research expanding the traditional between-individual paradigm with approaches to look at within-individual variability (Debusscher, Hofmans, and De Fruyt, 2014; Minbashian and Luppino, 2014; Minbashian, Wood, and Beckmann, 2010). These research lines suggest that in addition to between-individual differences (i.e. some people are generally more conscientious than others), there is also a huge variability within the person (i.e. a person may have a certain variability in conscientiousness during the day).

De Fruyt, Wille and John (2015) defined employability in terms of five key characteristics, defined as whether the person (a) demonstrates task-engagement and goal-setting, (b) can get along with other people, (c) adapts to/fits in an organizational structure, or has the capacity to deploy such structure (for those pursuing in self-employment), (d) learns on the job and can prepare for future challenges, (e) can deal with short and long term perspectives. This minimal set of employability indices taps into all basic personality dimensions, with task-engagement and goal-setting related to Conscientiousness, interpersonal skills related to emotion regulation and the core dimensions of the interpersonal circumflex (Extraversion and Agreeableness), and 'fitting in', 'learning and adapting', and 'time perspective' related to Openness to

experience and Conscientiousness. Employability and the impact of FFM traits are further reflected in how individuals navigate on the employment market and develop their career paths (Wille, De Fruyt, and Feys, 2010, 2013).

Finally, two other prominent related though distinct areas indexing employability are the demonstration of entrepreneurship or self-employment. Again, the FFM dimensions turned out to be associated with these outcomes (Obschonka, Schmitt-Rodermund, Silbereisen, Gosling, and Potter, 2013; Obschonka, Schmitt-Rodermund, and Terracciano, 2014; Obschonka, Silbereisen, and Schmitt-Rodermund, 2012).

Mental health

There is strong evidence that social-emotional skills and their underlying trait building blocks show strong relations to a variety of mental health problems. Tackett (2006) and De Bolle, Beyers, De Clercq, and De Fruyt (2012) described five different ways how basic personality dimensions are connected to broad psychopathology dimensions, such as internalizing and externalizing behaviour, but also to specific disorders. The *vulnerability* model stipulates that particular traits make an individual more susceptible to develop a particular form of disorder. For example, there is convincing evidence that those higher on neuroticism have an increased likelihood to develop one or more depressive episodes later in life (Fanous, Neale, Aggen, and Kendler, 2007; Kotov, Gamez, Schmidt, and Watson, 2010). In the Tracking Adolescents Individual Lives' Survey (TRAILS), Laceulle, Ormel, Vollebergh, van Aken, and Nederhof (2014) found that personality at age 11 was predictive of internalizing and externalizing disorders at age 19. They also showed that changes in personality between age 11 and 16 were predictive of both internalizing and externalizing psychopathology between 16 and 19, controlling for basal personality scores, demonstrating that both cross-sectional and dynamic trait indicators are indicative of later forms of psychopathology.

The *pathoplasty* mechanism further explains how traits can affect the manifestation, course and prognosis of mental disorders, although both may have independent origins and developmental paths. A well-documented example is the co-occurrence of callousness-unemotional traits and the diagnosis of conduct disorder in adolescence, having a worse prognosis (Hawes, Price, and Dadds, 2014). The direction of the effect can also be the other way around, i.e. mental disorders affecting the underlying trait dimensions. This causal model is also called the *complication* or *scar* model, with the disorder leaving a 'scar' on the individual's personality. For example, multiple depressive episodes may have their influence on the persons' neuroticism score (Fanous, et al., 2007), or recurrent symptoms of paranoia may have affected a persons' trust (agreeableness) level. A fourth mechanism proposes that traits and broad psychopathology dimensions show systematic phenotypic co-variation and may form a single continuum, i.e. the so-called *continuity* hypothesis. Evidence for such relationships between dimensions assessed by the HiPIC and the CBCL internalizing and externalizing dimensions of psychopathology has been provided by De Bolle et al. (2012). Finally, an extension of this continuity hypothesis is the *spectrum* model, assuming that trait and disorder covariance has a common origin. This model has been recently supported in youth by Martell, Gremillion, Roberts, Zastrow, and Tackett (2014) describing longitudinal relations between personality traits and Attention-Deficit/Hyperactivity Disorder (ADHD) symptoms, and by Tackett and colleagues (2013) identifying common genetic influences on a general psychopathology factor and the negative emotionality trait in young twin pairs. This review of different mechanisms makes clear that the relationship between both sets of constructs is complex and De Bolle and colleagues (2012) convincingly demonstrated in youth that for many disorders, multiple of these mechanisms explain part of the association across time between traits and disorders.

Socio-emotional skills and their underlying traits also have many indirect effects, for example via the selection of specific situations that may independently or jointly contribute with skills/traits to mental health (De Fruyt and De Clercq, 2014). For example, an impulsive and aggressive adolescent may end up in a deviant peer-group, because s/he is rewarded and respected in this environment, further reinforcing dysregulation.

Physical fitness, health and longevity

In a recent review, Friedman and Kern (2014) distinguished among six core health outcomes covered in public health policy research: physical health, subjective well-being, social competence, productivity, cognitive function, and longevity. Currently there is substantial evidence that socio-emotional skills and their building blocks are associated with a broad range of health outcomes, such as smoking (Munafo, Zetteler, and Clark, 2007), obesity (Gerlach, Herpertz, and Loeber, in press; Sutin, Ferrucci, Zonderman, and Terracciano, 2011), alcohol craving and consumption (Papachristou et al., 2013; Stautz and Cooper, 2013), resilience to Alzheimer's disease (Terracciano et al., 2013) and health status in cardiovascular populations (Versteeg, Spek, Pedersen, and Denollet, 2012). There is abundant evidence that traits related to impulsiveness are associated with eating problems and obesity. In a meta-analysis of 50 individual studies, Fischer, Smith, and Cyders (2008) found that emotion-based impulsivity indices such as negative urgency were substantially correlated with bulimic symptoms. Likewise, Sutin and colleagues (2011) followed 1988 individuals over a time span of more than 50 years and reported that low Agreeableness and high impulsivity traits predicted a larger increase in body mass index across adulthood.

Socio-emotional skills have been further successfully related to constructs from the positive psychology area, such as happiness (Hills and Argyle, 2001), quality of life, and subjective well-being (Diener, 2013; Steel, Schmidt, and Shultz, 2008). Many of these health and well-being related outcomes have subsequent effects on other criteria. For example, in a series of experiments, Oswald, Proto and Sgroi (2014) showed that happiness affects productivity.

Friedman and Kern (2014), however, warn against an oversimplified view of the association between dispositions and indicators of well-being and health, because many studies report the association between concurrently assessed self-reported dispositions and health indicators using the same informants and often using items that are included at both the "predictor" and the "outcome" side. This cautionary note aside, there is nevertheless enough evidence that dispositional constructs do contribute substantially to objective health indices. Friedman and Kern (2014) urge the field to consider more comprehensive and complex causal models of relationships among personality variables and correlated outcomes, taking into account mediator and moderator variables. For example, genetic predispositions, the environment and personality affect lifestyle patterns across time, manifested into correlated subjective well-being and physical health scores at Time 1 in development and these three may influence subjective well-being and physical health parameters observed at Time 2. Lifestyle patterns may mediate well-being/physical health from Time 1 to Time 2, and different contextual variables, such as psychological or biomedical intervention programs rolled out at school may moderate these developmental trajectories (Friedman and Kern, 2014).

In an impressive meta-analysis of 20 independent samples summarizing findings obtained on over 9000 participants, Kern and Friedman (2008) provided evidence on the lifelong significance of conscientiousness for individuals' health and longevity. Results were straightforward and showed that higher levels of conscientiousness were significantly associated with longevity ($r = .11$, 95% confidence interval = .05-.17), with the strongest correlations observed for the goal achievement (persistent, industrious) and inhibitory (organized, disciplined) facets of conscientiousness. Kern and Friedman (2008) argue that the protective effects of conscientiousness may probably work via multiple ways across the life course ultimately contributing and combining into longevity. People high in conscientiousness for example

may engage in healthier behaviours, or select safer and healthier family, living and work environments. Conscientiousness is a predictor of employability, and as a result persons with higher conscientiousness scores may end up in better jobs, having higher incomes and building more successful careers. Conscientiousness may further buffer and moderate the relationship between neuroticism and negative outcomes (e.g. poor mental health; see previous section). Another potential mechanism is that conscientiousness and health are influenced by similar genes, and hence are associated at the phenotypic level. Finally, and probably the most difficult to investigate, conscientiousness may contribute to health via the accumulation of small positive actions and/or the reduction of very small risks across the lifetime.

Civic engagement and environmental awareness

Civic engagement, environmental awareness and sustainable behaviour are outcomes that have become increasingly important during the past decade. Omoto, Snyder and Hackett (2010) examined motivational and personality predictors of activism and civic engagement, showing that other-focused motivation predicted AIDS activism and civic engagement better than self-focused motivation, interpersonal orientation and traits. Schnittker and Behrman (2012) examined the effects of schooling on civic engagement (participation and volunteering) and social cohesion (density of social network and quality of social relations) tempering somewhat previous optimism on the effects of education on achieving these outcomes. The effects of schooling on volunteering and participation in civic organizations disappeared almost entirely when taking into account different confounders. They concluded that increased schooling may generate some tension between navigating on the employment market and non-market commitments, as well as between independence and interpersonal reliability, making those who invest in schooling also more apt to pursue career-oriented interests, with less time left to engage in volunteering activities or civic engagement.

Developmental psychologists have paid attention to a related construct with high social significance called *generativity* (Erikson, 1950). During middle adulthood, somewhere between the ages of 40 and 65, people strive to create or nurture things that will outlast them. This can be achieved by having children or by contributing to positive changes that benefit other people, society in general, but especially future generations (e.g., building the Golden Gate Bridge). The generativity stage of development in Erikson's model refers to "making your mark" on the world, through caring for others, creating things and undertaking things that make the world a better place. The lack of generativity, also described as *stagnation*, refers to failure of some individuals to find a way to contribute to these goals. These individuals may feel disconnected or disengaged with their community and even with the society as a whole (Van Hiel, Mervielde, and De Fruyt, 2006). Van Hiel and colleagues (2006) showed that the "making your mark" generativity construct was related to low Neuroticism (-.22), and high levels of Extraversion (.36); Openness (.21), and Conscientiousness (.26), but not to Agreeableness (.04).

Finally, raising environmental awareness and engagement has been promoted as one of the most recent major challenges to achieve in social-educational learning programs. Milfont and Sibley (2012) examined the relationships between Big Five traits and different indices of "green" (environmentally sound) behaviour at both the level of the individual and countries. At the level of the individual, they examined the association with valuing protecting the environment, whereas at the level of countries, they examined the association between aggregated personality traits (within countries) and country-level measures of sustainability, environmental attitudes, and values. At both levels of analyses, Agreeableness, Conscientiousness and Openness were significantly related to engagement in green behaviours.

Crime/safety

A final group of outcomes that are associated with social-emotional skills is the entire spectrum of externalizing disorders, including drug abuse, bullying, conduct problems, vandalism, youth and adult criminality, but also more socially camouflaged deviant behaviours such as unethical comportment, fraud, greed and corporate psychopathy (Furnham and Taylor, 2004). A key difficulty in this area of research has been the co-occurrence of symptoms and specific disorders making it very complex to study associations between deficiencies in specific skills and traits and particular disorders. For example, Attention-Deficit/Hyperactivity Disorder has shown high comorbidity with Oppositional-Defiant disorder, suggesting to structure childhood disruptive disorders along major dimensions (Martel, Gremillion, Roberts, von Eye, and Nigg, 2010). The common denominator across a broad range of studies (de Haan, Dekovic, van den Akker, Stoltz, and Prinzie, 2013; Decuyper et al., 2013; Decuyper, De Pauw, De Fruyt, De Bolle, and De Clercq, 2009; John et al., 1994; Klimstra, Luyckx, Hale, and Goossens, 2014; Nigg et al., 2002; van den Akker, Dekovic, and Prinzie, 2010; Van den Akker et al., 2013) is that across development, the externalizing spectrum is negatively related to Agreeableness and Conscientiousness and, depending on the type of disorder, positively with traits related to Neuroticism.

Distinctiveness from cognitive measures

The socio-emotional skill constructs should ideally exhibit incremental explanatory power over cognitive constructs. The Big Five dimensions can potentially fulfil this requirement, with the highest cross-correlation (between Openness and verbal IQ) being only about .30 (John et al., 1994; Loehlin, McCrae, Costa, & John, 1998). At the same time, the socio-emotional battery should include constructs that can be examined in interaction with cognitive measures to explain a variety of outcomes. For example, the interaction between basic cognitive ability and the socio-emotional differences in goal setting may help to explain school achievement, beyond the main effects of cognitive and socio-emotional measures (Poropat, 2009, 2014).

Learning from other conceptual frameworks

We review several other approaches to socio-emotional skills. Our goals in the following review will be to ensure that we identify any additional concepts with potentially high predictive validity at lower levels of abstraction, like those already presented in Figure 1.

Social-emotional Learning Approaches

We are now ready to return to the three models, mentioned earlier, that define social and emotional skills in terms of 5, 6, or even 8 domains. These three models are summarized in Table 4.4, each in one column. As shown there, Elias and colleagues (1997) described six major domains of social and emotional learning, defined as “core competencies to recognize and manage emotions, set and achieve positive goals, appreciate the perspectives of others, establish and maintain positive relationships, make responsible decisions, and handle interpersonal situations constructively.” More recently, in their review of learning effects in intervention programs, Durlak, Weisberg, Dymnicki, Taylor, and Schellinger (2011) explain that there is no single taxonomic or measurement model guiding past or present research on socio-emotional interventions; instead, interventions are typically driven by specific school or district contexts and needs, and thus tend to focus on a diverse set of outcomes that may vary as widely as obesity problems in elementary school, depressive symptoms in middle-school girls, or vandalism in high school classrooms.

Table 4.4. How Do the Major Domains of the Proposed Framework Relate to Models of Socio-Emotional Learning Domains and Emotional Competencies?

Socio-emotional framework	Socio-emotional learning models		Emotional competence models
	Elias et al. (1997)	Durlak et al. (2011)	Saarni (1999, 2011)
Engage-ment with Others(E)	4 - Establish and maintain positive relationships	4 - Relationship skills 3 - Social awareness	7 - Genuine emotion expression and reciprocity for relationships
Collabo-ration (A)	3 - Appreciate the perspectives of others 6 - Handle interpersonal situations constructively		2 - Discern and understand emotion of other 4 - Empathic involvement in other
Task Per-formance (C)	5 - Make responsible decisions 2 - Set and achieve positive goals	5 - Responsible decision-making	
Emotion Regulation (N)	1b - Manage emotions	2 - Self-management	6 - Coping with aversive or distressing emotions 8 - Accept own emotions (emotional self-efficacy)
Open-mindedness (O)	1a - Recognize emotions	1 - Self-awareness	1 - Aware of own emotional state 3 - Use local emotion vocabulary correctly 5 - Appreciate emotion experience differs from expression

Note. If the model in this column does not include a relevant construct, the space is left empty. Labels given in the first column are the terms in the socio-emotional framework proposed here. In parentheses, we provide the acronym abbreviations of the old labels for the Big Five (E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Negative emotionality; O = Openness).

Researchers have commented on this lack of standard constructs and instruments. For example, Furlong et al. (2007) commented that "Assessments have begun to be developed, but have not had time to fully mature" (p. 2). When they examined the structure of the Behavioral and Emotional Rating Scale (BERS-2; Epstein, 2004), which is now more commonly used to evaluate the outcome effectiveness of intervention studies, they noted, "The BERS was developed using a mixture of professional judgement and empirical procedures; however—it had no prespecified theoretical foundation or set of psychological constructs" (p. 2). The *Intrapersonal Strength* subscale measures a youth's outlook on his or her competence and accomplishments (e.g., "I believe in myself"). The *Affective Strength* subscale measures the ability of a child to accept affection from others and express feelings towards others (e.g., "It's okay when people hug me"). The *Interpersonal Strength* subscale measures a youth's ability to control his or her emotions or behaviours in social situations (e.g., "I can express my anger in the right way"). The *School Functioning* subscale measures competence in school and classroom tasks (e.g., "I complete tasks when asked"). The *Family Involvement* subscale measures a child's participation in and involvement with his or her family (e.g., "My family makes me feel wanted").

On the basis of their extensive review of social and emotional learning programs, Durlak et al. (2011) concluded that socio-emotional intervention researchers tend to focus on five broad competency sets, including self-awareness, self-management, social awareness, relationship skills, and responsible decision-making. As shown in Table 4.4, these attributes, as well as the six defined by Elias et al. (2007), map reasonably well onto the socio-emotional Big Five shown in the first column of Table 4.4. For example, Elias et al.'s domain 3 (Appreciating the perspective of others) should be well represented by the facets of our Collaboration (Agreeableness) domain, especially by our facet 2 "Respect for others, empathy, tolerance, fairness." Similarly, Durlak et al.'s first domain is Self-awareness (which involves the ability to introspect and take note of one's inner experiences, like thoughts and feelings) should be well-represented by our Interest & awareness (Openness) domain, especially by facet 4 "Self-reflection, self-awareness, consciousness, self-actualization, authenticity."

The Emotional Competence Approach and Its Major Measures

The *emotional competence* approach originated in developmental and clinical analyses of what a child needs to *learn to become an emotionally and socially competent adult*. One major theoretician and proponent of the competence approach is Saarni (e.g., 1999; 2011) and her conceptual work has been quite influential. She analysed emotional functioning from the perspective of how well it serves the adaptive and instrumental goals of the individual and then defined emotional competence as a set of affect-oriented behavioural, cognitive and regulatory skills. Simply put, the child needs to learn both what it means to *feel* something and to *do* something about those feelings.

As shown in Table 4.4, Saarni (e.g., 1999; 2011) postulated 8 basic skills she considered prerequisites for emotional competence: (1) Awareness of one's own emotional state; (2) Skills in discerning and understanding the emotions of others; (3) Skill in using the common vocabulary of emotion and expression; (4) Capacity for empathic and sympathetic involvement in others' emotional experiences; (5) Skill in realizing that inner emotional states need not correspond to outer expression; (6) Capacity for adaptive coping with aversive or distressing emotions by using self- regulatory strategies that ameliorate the intensity or temporal duration of such emotional states; (7) Awareness that relationships are defined by emotional genuineness of expressive display and reciprocity; and (8) Capacity for emotional self-efficacy (i.e., individuals can accept their own emotional experience and view themselves as generally feeling the way they want to feel). Although not necessarily accepted by all researchers in the field, these 8 competencies provide a formidable set of complex socio-emotional skills for researchers to define and measure.

Note that competency *(6), capacity for adaptive coping* with negative emotions like sadness, anxiety, or anger, would be most directly relevant for individual differences in *emotional outcomes* in the child. The other competencies specify important processes that may help (rather than hinder) the enactment of effective regulatory strategies in both emotional and social situations. For example, if an individual has no competency (1), *awareness of his or her current emotional state*, then the individual would hardly proceed with (6) attempts to cope with those emotions. Similarly, lack of emotional self-awareness is likely to interfere with (7) genuine emotional expressions and reciprocity in social contexts.

Self-report Questionnaires of Emotional Competencies

Although the emotional competence approach has not generated a commonly accepted taxonomy and measurement model, it has generated several individual difference measures, mostly for adults. These measures and their strengths and weakness have been reviewed twice in the last decade (see John and Gross, 2007; John and Eng, 2014), and we therefore provide an abbreviated version here. For more details, tables, and references, we refer the reader to these original sources.

The Generalized Expectancy for Negative Mood Regulation Scale (NMR), developed by Catanzaro and Mearns (1990), was one of the earliest measures. It focused on individuals' *beliefs* that some "behaviour or cognition will alleviate a negative state or induce a positive one" (p. 547), and asks participants to indicate the extent to which they believe that their attempts to alter their negative moods will work. Many of the items focus on ways to avoid negative emotions. Thus, the measure has been criticized for equating mood regulation with the avoidance of negative affect (e.g., Gratz and Roemer, 2004); simply avoiding negative emotion is assumed to be an indication of effective regulation, as shown by items such as "When I'm upset, I believe that I can forget about what's upsetting me pretty easily" versus "When I'm upset, I believe that I won't be able to put it out of my mind" (reverse scored).

Consistent with the emotional competence perspective, Salovey, Mayer, Goldman, Turvey, and Palfai (1995) aimed to understand the reflective or "meta" processes that accompany many mood states. These "meta-mood" processes capture how individuals reflect on their feelings, including how they monitor, evaluate, and regulate them (Mayer and Gaschke, 1988). Salovey et al. (1995) assumed that emotions serve as an important source of information for the individual, and that individuals differ in how skilled they are at processing this kind of information, particularly in "their understanding of and ability to articulate their affective states" and in their ability to "regulate such feelings and use them adaptively to motivate behavior" (p. 147). The *Trait Meta-Mood Scales (TMMS)* were designed to measure stable and general attitudes about moods and the degree to which individuals attempt to manage (or repair) mood experiences. The TMMS measures Saarni's construct *awareness of one's own emotional state* in terms of two scales: (a) the tendency to attend to one's moods and emotions (*attention*) and (b) to discriminate clearly among them (*clarity*). The third scale aims to assess efforts to *repair* one's emotional state if needed. These individual differences were considered "fundamental to the self-regulatory domain of emotional intelligence" (Salovey et al., 1995, p. 147).

The TMMS Attention scale refers to paying close attention to feelings, accepting feelings, valuing them positively, and letting oneself experience them fully and intensively, using items such as "I often think about my feelings" versus "I don't think it's worth paying attention to your emotions or moods" (reversed-scored). As expected, the Attention scale correlated with the Private Self-consciousness Scale, which measures awareness and attention to private aspects of the self (like thoughts and feelings), which is represented in the Big Five personality taxonomy by the Openness to feelings facet; this facet is also well-

represented in Table 4.3. The TMMS Clarity scale assesses feeling at ease and clear about one's emotions, as contrasted with a deep and troubling confusion about one's emotions and what they mean. This should interfere with effective mood repair and, indeed, low clarity was related to vulnerability to negative affect, distress, and depression. The TMMS Repair scale assesses attempts to improve negative mood by thinking positively and taking an optimistic (rather than pessimistic) attitude more generally. Item examples include "Although I am sometimes sad, I have a mostly optimistic outlook" and "I try to think good thoughts no matter how badly I feel," as contrasted with "Although I am sometimes happy, I have a mostly pessimistic outlook" (reverse-scored). This scale seems conceptually similar to coping measures and was found to correlate substantially with low vulnerability to distress and depression as well as with greater optimism. More generally, these findings are consistent with the structural model proposed in Table 4.4: competencies or skills in emotion regulation, like mood repair, should lead to better emotional outcomes when individuals face aversive or stressful situations. Again, these socio-emotional characteristics are well-represented in the model laid out in Table 4.3.

Gratz and Roemer (2004) followed Saarni's (1999) approach to emotional competencies, and were influenced by Salovey et al.'s TMMS scales, as well as by the older NMR (p. 44). However, they deviated from the "strengths" or "skill" based approach and instead devised a measure they called *Difficulties with Emotion Regulating Scale (DERS)* during times of distress. Specifically, their items used the format of the older NMR, and all items begin with the sentence stem "When I'm upset, I… ." This focus on global negative affect (upset) is a feature that both the NMR and the DERS share with measures of coping. Even though the DERS has six subscales, they are substantially inter-correlated and are often aggregated into a single overall dysregulation score, which correlates substantially with various indicators of negative affect, psychopathology, and low well-being (e.g., Weinberg and Klonsky, 2009).

Example of a Situational Judgment Test: Emotional Intelligence Test (MSCEIT)

It should be clear by now that the emotional competence approach conceptualizes emotion regulation in terms of a number of specific *abilities*. Yet, all the measures we have discussed so far have used self-report questionnaire methodology, asking about self-perceptions (including self-efficacy beliefs) and about typical experiences and behaviours. In fact, we have seen that some of these "competence" scales seem indistinguishable from coping styles. This is hardly a compelling way to assess constructs defined as abilities such as intelligence, which psychologists measure with maximum performance tests of the behaviour or process in question.

Mayer, Salovey, and Caruso (2002) acknowledged this methodological inconsistency and undertook the challenging task of constructing an "emotional intelligence test" (abbreviated MSCEIT) scored objectively in terms of correct and incorrect answers. They define emotional intelligence as a set of skills involved in the processing of emotion-relevant information. Here we address, as an example, only emotion management ability, which is the most relevant of the MSCEIT components and is defined as the capacity to reduce, increase, or maintain particular emotions in both oneself and other people. The tasks used to measure these abilities follow the format known as a *situational judgment test*; these tests aim to assess the ability to choose the most appropriate action from a pre-specified set of options and are typically used with adults in workplace and job selection contexts (but see a recent application to collaboration skills in adolescents by Richard Roberts and colleagues, 2012). The MSCEIT requires respondents to react to hypothetical scenarios and evaluate the effectiveness of various behaviours and subjective construals for emotion management purposes. For example, participants are asked to judge the effectiveness of strategies to help a friend enhance a joyful mood or reduce feelings of sadness.

The test itself is owned by a commercial "test publisher (who) does not authorize reproduction of actual test items" (e.g., Lopes et al., 2005, p. 114). This has been an impediment to research on the MSCEIT and its scientific evaluation has been hindered; thus, , we reprint below the only two abridged item examples available, both from Lopes et al. (2005, pp.114-115). Each item consists of a vignette paired with separate response options:

"*Debbie just came back from vacation. She was feeling peaceful and content. How well would each action preserve her mood? (1) She started to make a list of things at home that she needed to do. (2) She began thinking about where and when to go on her next vacation. (3) She called a friend to tell her about the vacation . . .*"

"*Ken and Andy have been good friends for over 10 years. Recently, however, Andy was promoted and became Ken's manager. Ken felt that the new promotion had changed Andy in that Andy had become very bossy to him. How effective would Ken be in maintaining a good relationship, if he chose to respond in each of the following ways? (1) Ken tried to understand Andy's new role and tried to adjust to the changes in their interactions. (2) Ken approached Andy and confronted him regarding the change in his behavior*."

These examples are from the fourth (Managing Emotions) "branch" of the MSCEIT, which consists of two distinct tasks. Five vignettes measure *ability in emotion management*, and each describes a person (like Debbie above) who is experiencing a mood or emotion. For each of the 5 vignettes, the respondent rates (on a 5-point scale) how effective four different actions would be for obtaining a specified effect on the person's experience (here, to preserve Debbie's good mood), yielding a total of 20 separate ratings. The second task measures *emotional relationship abilities* and consists of three vignettes describing relationships between persons (like Ken and Andy above). In each vignette, the respondent rates how effective three different actions would be *to maintain a good relationship* between the persons, for a total of 9 separate ratings. Each of the 29 individual ratings is scored according to a normative effectiveness rating provided by a panel of emotion experts or the group consensus.

Although abbreviated, these two examples are very instructive. First, the total emotion management ability score includes more than 30% of the ratings that do not involve emotional but relational skills, raising questions about content validity. Second, each vignette and action includes a lot of detailed contextual information specific to that rating, which adds error and lowers inter-item correlations and thus reliability; with 29 ratings aggregated into the total score, reliability in this study was a modest .63, and that is higher than in other studies (see Føllesdal and Hagtvet, 2009, for a thoughtful psychometric analysis and critique). Third, as the MSCEIT authors readily acknowledge, these vignette ratings do not actually measure individual differences in skilful or effective regulation scored or observed objectively in an emotional situation; instead they tap the individual's *knowledge, and capacity to reason,* about emotions and emotional situations (e.g., Lopes et al., 2005, p. 114). Fourth, the emphasis on knowledge and complex reasoning processes is likely to introduce correlations with measures of other abilities, creating discriminant validity problems. Fifth, there are interpersonal themes even in the emotion management vignettes (e.g., calling a friend to share one's mood nature and thus capitalize on the experience) and thus performance on these items may yield surprising correlations with personality variables, again introducing potential problems with discriminant validity. In response to these discriminant validity concerns, the test authors and their collaborators have argued that the MSCEIT scores predict social, emotional, and leadership outcomes even when intelligence and broad personality traits are controlled. So far, however, many researchers have remained unconvinced; the MSCEIT, and emotional intelligence research more generally, is viewed with scepticism among researchers (e.g., Landy, 2005; Joseph and Newman, 2010).

For example, Lopes et al. (2005) obtained 8 criterion measures (e.g., interpersonal sensitivity; socio-emotional competence; friendship nominations) with self-ratings or peer nominations. When the Big Five Inventory personality scales (John et al., 2008) were controlled, the MSCEIT emotion-management ability scale still significantly predicted two of these criteria. However, with 2 out of 8 correlations significant, it is hard to say whether the predictive-validity goblet is a quarter full or three quarters empty. Of greater interest to personality researchers is a finding not highlighted by the authors: by far the highest correlation was not found in predicting any of the 8 socio-emotional outcome measures but the Big Five dimension of Agreeableness; the *r* of .40 is a very substantial correlation once compared with the modest reliability of .63 of the MSCEIT scale in this study. Again, we have defined several facets relevant to reciprocal collaboration that should well capture any individual differences related to Agreeableness in the MSCEIT.

In conclusion, the MSCEIT, though an admirable and conceptually interesting undertaking, has not proven the decisive fix for the self-report measures of emotional "competencies" that have come before it. Even though as outsiders we do not know much about the inner workings of the MSCEIT, it seems unlikely that scores on its 5 emotion management vignettes and the 3 relationship vignettes can provide the conceptual building blocks needed to construct a comprehensive measure of socio-emotional abilities.

More generally, the conceptual richness, reach, and resulting complexity of the emotional competence approach (e.g., Saarni, 1999) may be a strength as well as a major limitation. It may include too many cognitive, behavioural, self-perception, and emotion perception processes under one broad rubric. Fewer constructs, more narrowly delineated distinctions, and tighter links between construct definitions and actual measures may prove a fruitful avenue for future research. Nonetheless, we were able to draw on this approach for a better understanding of our social-emotional constructs in Table 4.3 and to supplement the conceptual framework for the OECD longitudinal study.

We now turn to two other concepts that educational psychologists have developed over the past decades and that are often mentioned in the context of social-emotional learning and 21st century skills, namely metacognitive skills and learning styles. Conceptually speaking, however, these two are probably better described as social-cognitive instead of social-emotional skills. Similarly, an important social-cognitive belief construct that we plan to add to the model in Table 4.3 is Carol Dweck's implicit theories concept (e.g., Dweck et al., 1993); she contrasts entity (fixed and unchanging) beliefs about ability (and intelligence) with incremental (growth-oriented) beliefs that abilities can change and grow; growth beliefs have been shown to predict much better academic and well-being outcomes than beliefs that abilities are fixed and unchanging.

Meta-cognitive skills

Schraw and Dennison (1994, p. 460) define meta-cognition as: "The ability to reflect upon, understand, and control one's learning", or differently phrased 'thinking about your thinking' (Flavell, 1979) in the context of learning. The metacognition concept taps into higher-order mental processes referring to knowing what strategies work best for learning and how and when to activate these strategies (metacognitive knowledge), but also to the capacity to regulate these skills reflected in activities such as planning, information management strategies, comprehension monitoring, debugging strategies, and evaluation (subsumed under metacognitive regulation). Schraw and Dennison (1994) found that the knowledge and the regulation components are distinct components though are associated in the .40 to .50 range. There is support for the predictive validity of metacognition for academic performance (.21; Coutinho, 2007), and there is evidence that it mediates the relationship between mastery goal setting and academic success (Coutinho, 2007).

Schraw and Dennison (1994) developed the Metacognitive Awareness Inventory (MAI), a 52-item self-report instrument to assess metacognitive awareness and its components. The MAI has a two-factor structure, including of facets. A sample item is written between parentheses, to be in a better position to examine coverage by the proposed social-emotional framework. The first 'knowledge of cognition' dimension includes three facets: (a) Declarative knowledge, described as "knowledge about one's skills, intellectual resources, and abilities as a learner" (p. 474) ("I am good at organizing information"), (b) Procedural knowledge, i.e. knowing how to apply learning strategies ("I try to use strategies that have worked in the past"), and (c) Conditional knowledge, i.e. knowing when and why to use particular learning skills ("I use different learning strategies depending on the situation"). The second factor, 'regulation of cognition' has five facets: (a) Planning, i.e. the process of planning and preparation, goal-setting and allocation of investments ("I set specific goals before I begin a task"), (b) Information Management, i.e. the efficient and timely processing of available information ("I slow down when I encounter important information"), (c) Monitoring, i.e. controlling and assessing the learning process and use of strategies ("I ask myself periodically if I am meeting my goals"), (d) Debugging, i.e. correcting and redirecting learning strategies when necessary ("I ask others for help when I don't understand something"), and (e) Evaluation, i.e. analysing and reflecting on the learning performance and the obtained result after learning ("I summarize what I've learned after I finish"). A content analysis of the 52 MAI-items showed that the majority of items could be described as indicators of Conscientiousness and Openness to experience, though contextualized in a learning situation.

The construct of metacognition is represented in a number of 21st century frameworks, including other frameworks discussed below, such as those proposed by the Center of Curriculum Redesign (CCR) and by DeSeCo (Definition and Selection of Key Competencies), in which reflection is at the core concept.

Learning styles

Students show a huge variety in how they perceive and learn, and this variability has been represented in models and theories on learning styles or learning approaches. Learning style is usually narrowly conceived, i.e. as a combination of different learning activities or as a learning strategy, but when depth of information is emphasized, the concept 'approach' is mostly used. Different learning style instruments are around which are fairly similar to one another (Furnham, 1996). Honey and Mumford's (1982) Learning Style Questionnaire (LSQ) was based on Kolb's (1976, 1984) theory, and distinguished four different types of learning: activists, reflectors, theorists, and pragmatists. Activists are open-minded and get fully and enthusiastically engaged in new experiences, whereas 'reflectors' like to stand back and (re-)evaluate the different elements before deciding or acting. Theorists adapt, analyse and integrate distinct facts into coherent theories, but 'pragmatists' check out whether something new works in practice. Another measure is the Approaches to Studying Inventory (ASI) developed by Entwistle and Tait (1995) distinguishing among level of engagement and depth of processing when learning. This inventory assesses deep (intention to understand, relating ideas, use of evidence and active learning), surface (intention to reproduce, unrelated memorizing, passive learning and fear of failure), and strategic (study organization, time management, alertness to assessment demands and intention to excel) learning approaches.

The association between learning strategies and the Big Five dimensions were first described by Diseth (2003a) among samples of psychology and philosophy students. Across the samples, the deep learning approach was associated with Openness (.46 and .54), the surface approach positively with Neuroticism (.42 and .49) and negatively with Openness (-.25 and -.21), and the strategic learning approach with Conscientiousness (.55 and .62). Despite this conceptual and empirical overlap, learning styles added to the prediction of educational attainment. Komarraju, Karau, Schmeck and Avdic (2011) found that learning styles explained an additional 3% of GPA variance on top of the 14% already explained

by the FFM (see also Rosander and Bäckström, 2012). More important, several studies have shown that specific learning styles mediate the relationship between the Big Five and examination grades (Komarraju et al., 2011, for Openness, and Diseth, 2013b, for Openness, Neuroticism and Conscientiousness), suggesting they are better conceived as characteristic adaptations resulting from more basic cognitive and non-cognitive tendencies.

4.3. Malleability

A number of observational and intervention studies provide evidence on malleability of the proposed constructs.

Observational studies

Research demonstrating that personality traits show substantial plasticity and continue to develop in adulthood is now widely available (e.g., Helson, Kwan, John, and Jones, 2002; Srivastava et al., 2003; for a review, see Roberts, Walton, and Viechtbauer, 2006). Extensive observational studies have demonstrated that throughout early and middle adulthood, many people increase in what has been called *psychosocial maturity*, that is, they increase in conscientiousness and agreeableness, and they decrease in negative affect. However, much of that research has been conducted only at the broad level of the Big Five domains; recent research including facets (Soto and John, 2014) suggests that one developmental pattern does not always hold for all the facets in a domain. For example, among the conscientiousness facets, self-discipline increased substantially from age 20 to 60 whereas orderliness did not increase much at all.

Observational research on naturally occurring personality development in children and adolescents is just beginning to hit its stride (see the recent special issue of the *European Journal of Personality*, edited by Denissen, 2014). This slower start occurred, in part because research on children faces even greater hurdles than longitudinal studies of personality change in adults (see the final section of this report for a discussion of some of those issues). With younger children, self-reports cannot be used as an efficient method to collect data; assessment instruments have to be made age-specific, so that they are appropriate for the age-specific emotional and behavioural repertoire of the child and then the adolescent, making it more difficult to compare developmentally as instruments change; and changes occur more rapidly than in adults, necessitating yearly assessments whereas in adult life-span studies assessments may be limited to every 5 or even 10 years (George, Helson, and John, 2011). Again, most of the available studies of children have focused on only the Big Five domains (see Roberts et al., 2006, for a review), thus cannot speak to the more differentiated facet structure of the framework considered here.

One recent study has used the same instrument (the BFI) from late childhood (age 10) to late adolescence (age 20), thus making mean-level comparisons across these age groups easier; age differences were examined for both Big Five domains and facets, in a large sample recruited via the internet (Soto, John, Gosling, and Potter, 2011). Clear evidence emerged for substantial age differences throughout this difficult-to-measure development period, with a curvilinear pattern. After age 11, the data showed that on average, socio-emotional functioning was challenged by the onset of adolescence; however, by age 15, both girls and boys had begun to recover from the “Sturm und Drang” of early adolescence and showed positive age trends on the facets related to psychosocial maturity (higher agreeableness and conscientiousness) all the way to age 20.

Gender differences also developed during this adolescent period. At age 10, boys and girls did not differ in negative affect but then girls quickly increased to the elevated levels typical of young adult women (by age 15). In contrast, boys stayed stable overall and increased somewhat in self-confidence (i.e.,

lower anxiety levels) all the way to age 20. On the other hand, gender differences in agreeableness were already apparent in late childhood, with girls scoring higher than boys in each age studied. One needs to be careful before generalizing from a single (if large) sample study in one country. However, similar effects are being found in a cross-cultural sample assessed starting from age 12, with the gender difference in negative affect emerging again only in adolescence, and replications are under way, for example, in Brazil.

The critical question for future research, however, involves individual differences in change trajectories, which can only be studied in a longitudinal design. Specifically, we expect that not all adolescents will show the age-typical deterioration in socio-emotional functioning; worse, not all who did show that deterioration will recover from it as quickly as the normative data suggest, as problems such as juvenile delinquency emerge during this time and can potentially create a longer negative developmental dynamic.

The OECD Study on Skill Development will be longitudinal, and should include at least yearly measurement points to be able to capture such non-linear patterns of rapid change during the school years. The study will be designed to learn about social-emotional skill development trajectories and how these relate to a broad range of outcomes, some of which emerge during that age period. At the same time, developmental contexts during these trajectories will be prospectively studied, with the individual countries in which the studies are conducted serving as macro-environmental factors. The primary questions of this investigation will hence be centred on within-country comparisons across time.

Intervention Studies

One of the reasons why the Longitudinal Study of Social and Emotional skills in Cities is needed is that, at this point, there is no compelling research available where the development of social-emotional skills is studied longitudinally in a natural context, across a substantial time-interval, using adequate measures for which measurement equivalence has been demonstrated, and relying on multiple informants. The reason for the paucity of systematic research is, as we noted earlier, the lack of consensus on how to define social-emotional skills and construct a consensual taxonomy representing their features and content.

In contrast, there are a multitude of studies examining the impact of different kinds of school-based interventions to enhance students' social and emotional learning. These programs usually aim to either increase particular socio-emotional skills (e.g., peaceful conflict resolution) or influence a specific subset of the outcomes to be targeted in the Longitudinal Study of Skill Development in Cities, including positive social behaviours, conduct problems, emotional distress, psychological well-being, physical health, and academic performance.

A number of impressive meta-analyses have been conducted (Durlak, Weissberg, Dymnicki, Taylor, and Schellinger, 2011; Park-Higgerson, Perumean-Chaney, Bartolucci, Grimley, and Singh, 2008; Sklad, Diekstra, De Ritter, Ben, and Gravesteijn, 2012) examining the impact of such interventions, with special attention for important moderators. Durlak and colleagues conducted an impressive meta-analysis on the impact of school-based social-emotional learning programs published before 2007, summarizing findings obtained from kindergarten to high school and reporting on a total sample of N=270,034. As moderators they included whether programs were (a) run by expert/consultants or by teachers themselves, (b) organized at the level of the classroom only versus at the classroom and the school level, (c) developed according to SAFE (sequenced, active, focused and explicit) or non-SAFE criteria, and finally whether (d) implementation problems for the programs were reported or not. Overall, small to moderate intervention effects were reported for attitudes, positive social behaviour, conduct problems, emotional distress, and academic performance, with moderate effects reported for the malleability of social-emotional skills.

Across these criteria, programs implemented by teachers showed significant impact, suggesting that teacher-based programs are effective, and that one does not necessarily need external personnel/consultants in the classroom to achieve results. Only programs designed according to SAFE criteria, this means an active involvement of pupils, following a sequenced, focused and explicit program, demonstrated effectiveness. Finally, only interventions for which no implementation problems were reported or not mentioned turned out to be effective. A constraint of this meta-analysis was that 53% of the source of the outcome data were child-reported.

A second major meta-analysis on the subject has been conducted by Sklad and collaborators (2012), reporting effects of 75 universal school-based intervention programs for which the data were published between 1995 and 2008 with an average reported intervention sample size of N=543 (range 13 to 8280). Skald et al.'s meta-analysis provides an excellent follow-up on Durlak et al.'s review, because they also included 16 non-American based studies (21% of the total meta-analysis) and investigated immediate and delayed outcomes. The majority of the reviewed studies had a post-test between 0 to 6 months (73.3%), 36% of the studies had a follow-up between 7 and 18 months and for 21.3% follow-up data were available after 19+ months. Again here, the outcome measurement relied chiefly on self-reports (60% of the programs) and for 73.3% of the programs, no intervention manual was available, making it difficult to really study the content of interventions.

Sklad and colleagues (2012; Table 4.4) found substantial evidence indicating an improvement but the effect sizes varied by domain targeted for intervention; *d* effect size estimates for immediate effects were .70 for socio-emotional skills, .46 for positive self-image, .46 for immediate academic achievement, -.43 for antisocial behaviour, .39 for prosocial behaviour, -.19 for mental disorders, and -.09 for substance abuse. In other words, socio-emotional skills were most malleable in intervention contexts, whereas mental disorders and substance abuse were least affected by the interventions. As one would expect, effect sizes at a later follow-up decreased substantially for all outcomes, with effect sizes reduced to.26 for academic achievement, -.20 for antisocial behaviour, -.10 for mental disorders, .07 for positive self-image, .12 for prosocial behaviour, .07 for social-emotional skills and -.18 for substance abuse. The authors concluded from these data that, despite large immediate gains, long-term effects were small, with the average program participant still outperforming the average non-participant by 5%.

Additional key findings were that programs with a duration of less than a year had more impact on social skills than those that had a longer time-frame; also a smaller number of sessions (less than 20 sessions) turned out to be more effective. Intervention impact on social skills was equally large in primary and secondary school, whereas effectiveness to reduce antisocial behaviour was strongest in primary school. These findings suggest that antisocial behaviours are better tackled early on at school, whereas there is equal room for improvement of social skills across both primary and secondary school. Teachers in Sklad's analysis further turned out to be as effective as non-teachers to run programs, confirming Durlak's (2011) conclusion that teachers can successfully implement these programs. Finally, interventions' impact on social skills seems to be equal in North-American samples versus studies conducted outside of North America, suggesting that malleability generalizes across societies.

Besides these meta-analyses targeting a broad range of outcomes, there is also a wide range of studies, including randomized control trials, on reducing aggressive behaviours (e.g. Park-Higgerson, et al., 2008), and focusing on antisocial personality (Scott, Briskman, and O'Connor, 2014), oppositional defiant disorder (ODD) (Scott, et al., 2014), and conduct disorder (CD). There are effect evaluations examining broad and intensive clinical programs, often also working with parents (Scott, et al., 2014), broad and intensive community versus clinical programs (Kolko et al., 2009), short (reduced) programs) and the effect of organizing booster sessions (Lochman et al., 2014) to maintain long-term effects of interventions.

5. COHERENCE WITH OTHER FRAMEWORKS AND EDUCATIONAL GOALS

In this final review section, we examine how the framework proposed here maps onto the three major domains of functioning specified by the OECD (Managing emotions; Working with others; Achieving goals). In addition, we show links between the Big Five derived measurement framework and other approaches that aim to define the major *goals or objectives of education*, including the approach proposed by the *Collaborative for Academic, Social, and Emotional Learning* (CASEL), which is closely related to the socio-emotional learning approaches by Durlak, Elias, and colleagues reviewed above, the integrative proposal for educational goals by the Center for Curriculum Redesign (CCR), and finally the framework adopted by the *Knowledge is Power Program* (KIPP) schools.

Table 5.1 provides an integration of different frameworks that have been suggested previously for the social-emotional skill domain. The table is not meant to be exhaustive, and primarily focuses on frameworks that have been reviewed earlier in this report. It starts from the common framework offered for the OECD's Longitudinal Study of Skill Development in Cities (OECD, 2015), proposing the skill domains of 'Managing emotions', 'Working with others', and 'Achieving goals'. Such broad grouping of social-emotional skills is particularly useful for conceptual and communicative purposes to link skills with potential outcomes, though these domains are too general to be applicable at an operational level to track development from childhood to young adulthood. The Longitudinal Study of Skill Development in Cities exactly needs a more fine-grained proposal representing the commonality across constructs suggested in different social-emotional skill frameworks, but also needs to cover those skills and constructs that are necessary to understand developmental trajectories and explain outcomes. In addition, the suggested system should be sensitive to cultural differences and also embrace constructs that may help to explain trajectory and outcome variation across cultures.

As described earlier, the currently best-researched taxonomy of individual differences that comes closest to the domain of social-emotional skills is the Five-Factor taxonomy of personality. Personality psychologists now agree that five main dimensions represent the core qualities underlying personality differences. Although consensus at the lower-order facet level in the Five-Factor Model hierarchy has not yet been established, Table 4.3 shows that there is starting convergence also at that level. At the same time, these five dimensions also describe how people interact with and adapt to the environment they live in. It should hence not come as a surprise that these five dimensions also have a cardinal position in an analysis of social-emotional or 21st century skills. Emotional stability directly taps into the domain of 'Managing one's emotions', Extraversion and Agreeableness describe how we 'get along, engage, and work with others', whereas Conscientiousness and Openness to experiences are respectively about 'getting things done/achieving goals', and 'being explorative and innovative'. These different adaptive functionalities map well onto the OECD's framework (column 1).

Table 5.1. Mapping Different Frameworks Proposed for Socio-Emotional Skills and for Goals of Education

OECD	Skill equivalents here	Proposed constructs here	CASEL	CCR	KIPP Schools
Managing emotions	Emotion awareness skills	Stress resistance		Mindfulness (Q)	
	Emotion acceptance skills	Self-esteem	Self-management		
		Emotional control			Self-control
	Emotion reappraisal skills	Self-compassion			
	Emotion modification skills	Self-confidence			
		Fear-of-happiness			
Working with others	Assertiveness skills	Assertiveness		Courage (Q)	
	Presentation skills	Enthusiasm			Zest/optimism
	Social contact skills	Social approach & connection			
	Leadership skills			Leadership (Q)	
		Interdependent self-construal			
	Collaboration skills	Compassion	Relationship skills	Collaboration (S)	
		Trust			
		Relationship harmony			Gratitude
	Communication skills	Respect for others	Social awareness	Communication (S)	Social intelligence
				Ethics (Q)	
Achieving goals	Responsible decision-making skills	Responsibility	Responsible decision-making		
	Goal setting skills	Goal orientation		Resilience (Q)	Grit
	Task engagement skills	Task initiation			
		Self-discipline			
		Organization			
	Creativity skills	Creative Imagination		Creativity (S)	
		Intellectual Curiosity		Curiosity (Q)	Curiosity
	Appreciation skills	Aesthetic interests			
	Self-reflective skills	Self-reflection/awareness	Self-awareness		
	Critical thinking skills	Autonomy/Independence		Critical thinking (S)	

5.1. Extensions to Other Frameworks

To further illustrate the FFM's dimensions' status as constructs referring to the core qualities represented in the amalgam of skills, learning objectives and attitudes, that are proposed as standards of aspiration in the 21st century, we classified the constructs proposed by two key players in the social-emotional learning field in Table 5.1, in the columns 4 and 5.

The influential Collaborative for Academic, Social, and Emotional Learning (CASEL; www.casel.org) from the University of Illinois at Chicago has suggested five competency areas that should be advocated in social-emotional learning programs, i.e. Self-awareness, Social-Awareness, Self-management, Relationship skills, and Responsible decision-making. Three of these competency areas are content-wise consistent and align unequivocally with specific FFM dimensions. Social awareness (understanding and empathy, taking others' perspectives) and Relationship skills (working in teams, positive relationships, reciprocity, conflict handling) are clearly related to the Agreeableness domain, describing individual differences in the quality of social interactions with persons. Self-awareness (recognizing one's emotions, self-confidence, accepting limits, but also recognizing strengths) is without doubt related to the FFM Emotional Stability dimension.

Self-management (being able to control and regulate emotions to achieve goals, conscientiousness and perseverance) is a more hybrid competency domain from an FFM perspective, primarily related to Emotional Stability, but also to Conscientiousness. Finally, Responsible decision-making (assessing risks and making deliberate decisions, respecting others) is more at the intersection of Conscientiousness and Agreeableness, so we have classified this skill domain in the Conscientiousness area, but also closely to the Agreeableness dimension. The classification of the CASEL framework within Table 5.1, illustrates its strong background in positive psychology, whereas it seems to lack competency areas that have to do with extraverted qualities and openness to experience, two other key dimensions to engage with the world outside.

The integrative set of skills and qualities proposed by the Center for Curriculum Redesign (CCR, 2015) are described in column 5. These concepts can be classified within the FFM framework when considered at the level of their overarching labels. Mindfulness (e.g., self-awareness, tranquillity) aligns with Emotional stability, whereas Courage (e.g., bravery, energy) and Leadership (e.g., charisma, assertiveness, responsibility) refer to "getting ahead" and are thus conceptually related to Extraversion. Ethics (e.g., benevolence, compassion, honesty) associates with Agreeableness, whereas Resilience is defined by CCR in terms of perseverance, grit, and self-discipline, and thus maps primarily onto Conscientiousness. Curiosity importantly taps into the Openness to experience domain. Where do the "Four Cs" in the CCR concepts, namely Collaboration, Communication, Creativity, and Critical thinking fall? Collaboration and communication belong in the Agreeableness domain, whereas creativity and critical thinking belong in the Openness domain. Overall, it appears that when evaluated from the perspective of our OECD measurement framework, the CCR concepts are more comprehensive than, for example, the CASEL framework because they provide a better coverage of the Openness domain.

The final framework shown in Table 5.1 is the Knowledge is Power Program (KIPP). In association with Dr. Duckworth (www.kipp.org/our-approach/character), they proposed seven constructs in their character strength approach, including zest, grit, optimism, self-control, gratitude, social intelligence and curiosity. As can be seen in the last column of Table 5.1, their constructs can be conceptually classified across the five dimensions defined by the Big Five.

By extension, similar conceptual classifications can be made for other frameworks that have been developed by scholars and interest groups in the social-emotional learning area. The Partnership for 21st

Century Skills (www.p21.org/our-work/p21-framework) distinguished among Life and Career Skills, Learning and Innovation Skills, and Information, Media, and Technology Skills. Life and Career Skills include "flexibility and adaptability", "initiative and self-direction", "social and cross-cultural skills", "productivity and accountability", and "leadership and responsibility". The Learning and Innovation Skills group encompasses "creativity and innovation", "critical thinking and problem solving", and "communication and collaboration". Again, this skill set sorts across four of the five FFM categories, except for the Emotional Stability domain. P21 further distinguishes Information, Media and Technology Skills, including 'information literacy', 'media literacy', and 'ICT literacy'. Although part of the skills in this group (e.g., flexibility and critical thinking) may be related to openness, they also tap into more functional skill domains, related to specific technologies.

In association with the OECD, the Swiss-led DeSeCo Project (Definition and Selection of Key Competencies) aimed to build a competency framework for measuring the competence level of young people and adults across different countries. The DeSeCo Model groups its key competencies into three broad categories: Using tools interactively, Interact in heterogeneous groups, and Act autonomously. In addition, they underscore the importance of reflectiveness, defined as the ability to deal with change, learn from experience and critical thinking. Using tools interactively is further split into: 'use language, symbols and texts interactively', 'use knowledge and information interactively', and 'use technology interactively'; Interacting with heterogeneous groups includes 'relate well to others', 'co-operate and work in teams', and 'manage and resolve conflicts'; Acting autonomously refers to 'act within the big picture', 'form and conduct life plans and personal projects', and 'defend and assert rights, interests, limits and needs'.

Finally, the Strive Partnership (www.strivetogether.org) distinguishes among the social-emotional competencies of 'Academic self-efficacy', 'Growth mindset and mastery orientation', 'Grit or perseverance', 'Emotional competence', and 'Self-regulated learning and study skills'. The factors specified by these theories and others in the literature, such as Self-Determination Theory (Deci and Ryan, 2002), Goleman's Emotional Intelligence (1995), or Blair's (2002) work on social and emotional competencies are represented by the facet constructs, either singly or in combination, proposed in this framework.

5.2. Cultural Sensitivity

As discussed earlier, Table 5.1 also includes several constructs hypothesized to be important to understand cultural differences, such as 'Fear-of-happiness' on Emotion Regulation, as well as 'Independent self-construal (on Openness) and Interdependent self-construal' (on Agreeableness). These constructs are prominent in cross-cultural research on individual differences but are rarely considered in the social-emotional learning literature, given that most of the educational frameworks for social-emotional skills were developed within single cultures.

5.3. Trait Building Blocks and Skills

Finally, the second column in Table 5.1 illustrates how most of these FFM qualities can be translated into skill constructs. Facets of the Emotional Stability domain can be transformed into emotion-regulation skills. For example, the Emotion-Regulation Skills Questionnaire (ERSQ; Berking & Znoj, 2008) describes a set of emotion regulation skills, including among others, emotion awareness, emotion acceptance, emotion reappraisal and emotion modification skills. Likewise, the lower-level Extraversion constructs of assertiveness, activity/enthusiasm, and sociability (initiating contact and connect to others) easily translate into assertiveness, presentation, and social contact skills, whereas these latter skills sets are necessary for developing leadership skills. The facets of Agreeableness form the heart of collaboration and

communication skills, whereas the FFM facets orderliness, self-discipline, task engagement and achievement orientation culminate into task engagement and goal setting skills. Finally the domain of Openness to experience refers to qualities that have to do with being creative and open to innovation, demonstrating eagerness to learn and curiosity, appreciating beauty and developing aesthetic sensitivity, being open and reflect on your own feelings, and having independent thought (autonomy). These trait qualities form the basis of creativity, appreciation, self-reflective and critical thinking skills.

The framework suggested in Table 5.1 further connects with other classification and taxonomic schemes developed for related areas, such as labour market and human resources competencies (see De Fruyt, Wille & John, 2015). The skills grouped in the 'Managing emotions' set are primarily intrapersonal competencies, whereas the skills related to extraversion and agreeableness are typically grouped in competency models under the header of interpersonal competencies. The agreeableness by conscientiousness skill area refers to skills that relate to Morality and Character, such as being reliable and respecting, including responsible decision-making and being accountable for decisions. Finally, the skills related to the openness by conscientiousness dimensions tap into Learning and Achieving competencies, reflected in the notion of life-long learning advocated in all educational frameworks nowadays.

6. VALIDATION STRATEGY

This section outlines a strategy to validate the social and emotional framework presented in this report. The validation process, which will take place during the feasibility study (2015-18) of the Longitudinal Study of Skill Development in Cities, involves the following major steps:

- Identification, adaptation and development of a range of appropriate measurement instruments;
- Measurement of a comprehensive set of children's social and emotional skills across grades 1-12 in participating countries/cities; and
- Content, external, structural and cross-cultural validity tests.

The core underlying principles of the validation strategy are as follows:

Start with best existing measures: There is already a wealth of information on measurement instruments and their measurement properties for a range of social and emotional skills constructs. This study will build on the past research, and focus on areas where more work is needed to improve existing measures.

Employ multiple methods and scale adjustments: Given that none of the existing instruments provide precise measures of social and emotional skills, the optimal strategy is likely to include multiple sources of data and triangulate them to reduce measurement errors and biases (see also Annex 2). In addition to the core data sources, namely reports by self, teachers and parents, we suggest considering other types of measures, such as performance tests, behavioural indicators (e.g., truancy) and situational judgment tests. There is also a need to consider adjusting rating scales by using anchoring vignettes and forced-choice methods. This will help reduce various biases that plague rating scales such as cross-cultural differences, social desirability, reference group bias and response style bias.

Ensure cross-cultural relevance, comparability, and invariance: The measurement instruments must provide good measures of the latent socio-emotional constructs of interest. Moreover, the scaling of these instruments must also be comparable across participating countries/cities and population groups within a country/city. Cross-cultural invariance might be improved using performance tests, behavioural measures, anchoring vignettes and forced choice, though measurement equivalence will also need to be demonstrated for these methods.

Malleability and age relevance: Some may argue that social and emotional development can be highly age-dependent, and that some of the social and emotional skills may not have developed sufficiently before a child reaches certain ages. For instance we may not anticipate "interdependent self-construal" (e.g., I feel my fate is intertwined with the fate of those around me) to have developed among many children during the early grades. The feasibility study may provide relevant information on the 'starting grade' and 'frequency' in which each of the social and emotional skills may usefully be measured. The former can be assessed by evaluating from what age children start understanding and developing a particular socio-emotional concept. The latter can be indirectly tested by assessing to what extent do social and emotional skills vary across different grades and children (after taking into account some of the individual differences such as demographic and socioeconomic background of children and parents).

Stakeholder consultation: It is of vital importance that the conceptual framework use labels of social and emotional skills (both at the facet and factor levels) that are well recognised by education stakeholders. After extensive psychometric analysis is performed, some of the labels used in the current framework could be validated or adjusted as necessary through consultation with superintendents, teachers, parents, employers and education officials.

7. POLICY QUESTIONS

The proposed Social and Emotional Skills Framework is a guiding principle for developing instruments to measure social and emotional skills. Together with the instruments to measure learning contexts (family, school and community) and outcomes (e.g. tertiary attainment, labour market and health), they will help address the following questions that are considered pertinent for policy-makers, teachers, school administrators and parents:

- **Which socio-emotional skills of children predict their cognitive, educational, labour market and social outcomes?**
- **Which family learning contexts predict children's social and emotional development?**
  - Parenting activities (e.g., involvement in childcare and childrearing tasks and time spent with children)
  - Attitudes towards parenting (parenting stress, parental satisfaction)
  - Family resources (e.g., education, income, employment, benefit dependency)
- **Which school learning contexts predict children's social and emotional development**?
  - School environment (e.g., classroom environment, violence)
  - Curricular and instructional practices (e.g., contents, delivery methods)
  - Teacher characteristics (e.g., age, gender, years of experience)
  - Extra-curricular activities (e.g., contents, objectives, delivery methods)
  - School resources (e.g., infrastructure, class materials, class size, child-to-staff ratio)
- **Which community learning contexts predict children's social and emotional development**?
  - Community learning environment (e.g., availability of civic and cultural activities)
  - Community safety (e.g., violence, quality of life)
  - Community characteristics (e.g., peer's socioeconomic background and values)
  - Community resources (e.g., public services, welfare regimes)
- **How malleable are social and emotional skills**? Are similar patterns observed across participating cities?
- **What are the social and emotional skills gaps by children's gender and parental SES**?
- **To what extent do these skills gaps vary across cities**? Are cities with narrower skills gaps those that provide strong support to parents and children?
- **Do these skills gaps grow over time?**
- **Do children's learning contexts prior to formal schooling (e.g., parenting, pre-school) predict children's skill development during the first years of formal schooling,** even after accounting for families' socio-economic background?

The effectiveness of addressing these questions will depend on the constructs and quality of instruments employed to measure learning contexts and outcomes. The developmental work on learning contexts and outcomes is scheduled to start in 2017. Moreover, the choice of empirical models used to identify the dynamics of social and emotional skill formation will also affect the precision in which these questions can be addressed. The latest OECD report*: 'Skills for Social Progress: The powers of social and emotional skills'* (OECD, 2015), presents an innovative method that takes into account: (a) skills today depends on the previous skills and investment in skills made during the previous year, (b) previous skills affect the productivity of mobilising new investments in skills, (c) pervious socio-emotional skills affect the development of both socio-emotional and cognitive skills, (d) independent effects as well as interactive effects of skills (e.g. the impact of literacy on depression is larger the higher the self-esteem) on education, labour market and social outcomes. The longitudinal data structure and the range of measurement

instruments to be proposed in this study will permit application of such elaborate empirical models in order to better shed light on the key questions for policy-makers and educators.

ANNEX A1. ALPHABETICAL LIST OF 21ST CENTURY CHARACTERISTICS

Ability to quickly acquire and apply new knowledge	Diligence	Integrity	Reverence
Abnegation	Discipline	Interconnectedness	Risk taking
Abstract problem solving	Diversity	Interdependency	Self-actualization
Acceptance	Efficiency	Justice	Self-awareness
Accountability	Effort	Kindness	Self-care
Adaptability	Empathy	Leadership	Self-compassion
Altruism	Energy	Leading by example	Self-control at school
Applying technology	Engagement	Learning from mistakes and failures	Self-control in relationships
Appreciation	Enthusiasm	Listening to others	Self-direction
Appreciating beauty in the world	Equanimity	Living in harmony with nature	Self-discipline
Appreciating others	Equity	Living in harmony with others	Self-esteem
Appreciating what I have	Ethics	Load management	Self-kindness
Assertiveness	Excitement of creating something new	Love	Self-reflection
Authenticity	Executing plans, follow through	Loyalty	Self-respect
Balance	Existentiality	Mental flexibility	Selflessness
Belonging	Exploration	Mentorship	Sensibility
Benevolence	Fairness	Mercy	Sharing
Bravery	Feedback	Mindfulness	Social awareness
Camaraderie	Feeling awe	Modesty	Social intelligence
Care	Flexibility	Motivation	Social perspective
Charisma	Focus	Negotiation	Socialization
Charity	Followership	Observation	Speaking out, taking a stand
Cheerfulness	Following	Oneness	Spirituality
Citizenship	Forgiveness	Open-mindedness	Spontaneity
Civic-mindedness	Fortitude	Optimism	Sportsmanship
Commitment	Generosity	Organization	Spunk
Common humanity	Genuineness	Passion	Stability
Compassion	Goal orientation	Patience	Tackling tough problems
Confidence	Grace	Perseverance	Teamwork
Conscientiousness	Gratitude	Persistence	Tenacity
Consciousness	Grit	Playfulness	Timeliness
Consideration	Growth	Precision	Tinkering / inventing
Consistency	Happiness	Presence	Tolerance
Cooperation	Helpfulness	Problem solving	Toughness
Courage	Heroism	Productivity	Tranquillity
Critical thinking	Honesty	Professionalism	Trustworthiness
Cross-cultural awareness	Honour	Project management	Truthfulness
Curiosity	Humaneness	Prudence	Verve
Dealing with ambiguity	Humbleness / humility	Public speaking	Vigour
Decency	Humour	Receptivity	Virtue
Decisiveness	Inclusiveness	Reliability	Vision
Decorum	Initiative	Resilience	Willingness to try new ideas
Delegation	Innovation	Resourcefulness	Wonder
Dependability	Inquisitiveness	Respect for others	Work ethic
Determination	Insight	Responsibility	Zeal
Devotion	Inspiration	Results orientation	Zest

Note. Items derived from Trilling & Fadel (2009) and Fadel (March 2014). Table adapted from John & Mauskopf (2014).

ANNEX A2. FEASIBILITY STUDIES: INITIAL METHODOLOGICAL CONSIDERATIONS

We have discussed various issues related to assessment and research design throughout this report, including the need to supplement self-report data with information collected from other data sources, like parents, teachers, or potentially coaches (e.g., in the US high school system, they tend to know the kids involved in athletics better than most teachers do), as well as data from school records (e.g., lateness) and possibly situational judgment tests, like the MSCEIT. In this final section, we offer some initial thoughts to stimulate consideration and further discussion by the expert group.

One critical set of more technical issues involves the minimum sample size required to test and validate the items to be used for self-reports, as well as for teacher and parent ratings. Other issues involve the grade levels that should be studied and what kind and how much data needs to be obtained from each student, at multiple occasions, at what grade levels children can start providing meaningful self-reports with acceptable levels of internal consistency and differentiation among the concepts rated, and how many items children at particular ages can be expected to complete in any one testing session.

Age ranges for self-reports by children and adolescents

We have commented earlier that adults have tended to underestimate the capacity of children to provide reliable personality ratings. Using a puppet interview technique, Measelle, John, Ablow, Cowan, and Cowan (2005) found that children as young as age 5-7 were able to report on their emerging self-concept in terms of the Big Five domains, with modest but significant internal consistency, temporal stability, discriminant validity among the Big Five domains, and external validity as assessed with teacher and parent ratings. An individually administered puppet interview is certainly not a feasible assessment device for the proposed OECD study. However, one could conceivably develop a computer-based measure consisting of 50 short animations showing relevant behaviours or emotions that are inspired by the content of the puppet interview and use a similar “is like me” or “is not like me” response format.

In terms of late childhood and early adolescence, age 10 has been the youngest age studied using existing Big Five measures. Using the BFI, Soto, John, Gosling, and Potter (2008) found that in the US, by age 14, self-reports by adolescents were essentially indistinguishable from those of adults; moreover, when problems with acquiescence and rating scale usage were controlled, many kids as young as age 10 could provide self-reports with reasonable psychometric characteristics, such as internal consistency and discriminant validity. The BFI was designed to have a 5th grade reading level; the starting ages for the slightly more difficult NEO PI-R items tend to be closer to age 12.

Moreover, the very large and economically diverse data set collected in schools by Primi and colleagues (2014, submitted) in Brazil generally replicated the US findings; when variation in rating scale usage was controlled, the Big Five factors could be identified reliably in self-reports of children as young as age 10-12. However, some caveats apply. The Brazilian self-report items had been carefully tailored and pilot-tested to conform to the language used by these children, and the number of items given in one assessment session had been limited to a maximum of 65 short items for the younger groups (ages 10-13) and 95 for the older age groups (14-18). Moreover, these findings apply only to measures of the broad Big Five domains; we know much less about facet level traits. Future development work needs to examine whether discriminations among facets within Big Five domains can be made reliably before age 14; for example, Soto et al.’s (2008) analyses suggest that the facets of extraversion may not cohere into one unified domain as early as agreeableness and conscientiousness.

Sample size considerations

In general, research suggests that sample sizes of N=500 will generate stable factor structures and reliability estimates in kids' self-reports for up to 100 items. Thus, one might suggest minimum samples of 500 girls and 500 boys, to analyse them separately and check that results generalize, so total N=1000. This should also be enough to examine effects of SES (lowest 20% = 100 boys and 100 girls, vs. the rest) and test the social-perceptual and language skills needed to make judgments about self and other (e.g., as assessed by vignettes requiring children to rate hypothetical others on the same items). These sample size estimates assume there are no critical minority groups, for whom separate estimates or analyses have to be run. If there are, larger samples would have to be drawn.

It will be critical to ensure that the items are easy enough to understand and do apply to kids in grade 7 (expected ages 12-13) but are not too childish and thus still useful for kids up to grade 12 (age 17-19). One design to test for measurement equivalence at reasonable age intervals would be to assess 500 boys and 500 girls in grade 7, another in grade 9, and another in grade 11. This would also permit an initial test of the expected plasticity effects, where age differences should be apparent at least in the facets related to agreeableness and conscientiousness (and gender differences should appear and increase in magnitude in the negative-affect facets).

Number and type of items to be completed

If previous research is any guide, the 11th and 12th graders will be able to respond to 100 BFI-type items in about 20-30 minutes (depending on their prior experience with rating tasks), thus leaving time to administer vignettes to obtain ratings of others. So, if schools can provide 60 min of assessment time, that would permit the administration of an initial unrefined item pool of 200 items plus one vignette set rating for each of the Big Five domains (but not the for each of the facet constructs). However, if the starting point is 300 initial candidate items to select the final (let's say 90) items, either more time (or multiple sessions) will be needed, or more participants are needed to permit a partial-administration design.

These estimates are likely to hold with non-minority students in developed countries. In Brazil, our colleagues Primi and Santos (see Primi et al., 2014) concluded from their extensive pilot testing that they should *not give more than 100 items per session even to 12 graders*, given the particular limitations in reading, concentration, and classroom set-up. This becomes a bigger issue in the younger groups, with 40 - 60 items doable for ages 10-11; if starting with age 12 in 7th grade, K=100 items may be feasible per session but one can expect more than 100 items to try out in the Round 1 of measure development. Thus, it may be necessary to double or triple the above estimates of sample sizes, so that different kids can do smaller but different subsets of the items.

Applicable for multi-source ratings

Paraphrasing Wim Hofstee (1994) in the tile of this section, a key question is: “Who is the best informant on children’s developing social-emotional skills”? Most approaches to personality assessment agree that appraisals of an individual must be based on multi-informant designs, rather than a single source of data (Wiggins, 1973). In the present context, multiple different stakeholders, such as the children and adolescents themselves as well as their parents and their teachers, can provide unique perspectives on children’s standing and development of skills, in addition to common variance.

The evaluation of social-emotional skills is usually done via self- or observer ratings on how children and adolescents typically behave or how well they can do so in general or in more specific

situations. Alternatively, more situational judgment type of assessments, where children are presented a situation description and have to select the most appropriate response from a predefined set of reactions, are used infrequently and have their own shortcomings. Major problems are that they usually span a very small range of constructs relative to the number of items and time necessary to measure a construct reliably. Besides, these methods identify if people are able to select that particular answer for which subject matter experts agree that this is the most appropriate response, which is very different from explaining how people will react in daily life.

Children can provide reliable and valid self-descriptions on personality and social-emotional skill descriptive items on average starting from the age of 10 onwards (Soto, John, Gosling, and Potter, 2011). This capacity is dependent on a series of critical factors, including language proficiency, but also cognitive and social development. First, children need to have acquired a certain vocabulary and a basic reading level to be in a position to administer the assessment. Simplicity and clarity in language is anyway an important requirement for skill descriptive items, because assessments not only have to be completed by children and adolescents themselves, but often also by parents of different socio-economic classes, and teachers who will have to rate multiple children in their class. These constraints require grammatically streamlined and short items, an easily understandable response scale format, and clear instructions. Many of these principles have been used when constructing the Big Five Inventory (BFI; John, Donahue, and Kentle, 1991) or the Hierarchical Personality Inventory for Children (HiPIC; Mervielde and De Fruyt, 2002). Explicit guidelines for the item-writing process are provided by Hendriks, Hofstee and de Raad (Hendriks, Hofstee, and De Raad, 1999).

Probably more important is that children also need to have developed some first self-reflective and social-comparison skills. According to Barenboim (1981), children first make behavioural comparisons (e.g. “Ricardo runs faster than Daniel”), and start to actively use trait terms thereafter (e.g. “Trudy is shy”). Furthermore, children’s person perception skills need to develop into a multidimensional scheme, to a point where they have a notion that multiple independent trait attributes may apply to themselves or another person. During development, children first associate a single individual with one characteristic [see for example the figures portrayed in children’s books and comics that are even named after a single trait, e.g. the different smurfs, each with their typifying characteristics, Asterix (small but smart) and Obelix (raw power), gnome “Lui” (lazy), …], and this perspective needs to progress into a multidimensional space of person-perception that can be used to describe differences between, but also within persons. The evidence available right now suggests that this is achieved by age 10-11, in line with the emergence of formal-operational thinking.

Besides self-reports, observer reports of skills and personality are also frequently used and form a necessary amendment to self-descriptions. For children in primary school, parents and teachers often act as first informants, whereas in secondary school parents and teachers often complement adolescent self-ratings. Research shows that all perspectives have unique and valuable viewpoints on individual differences, with father and mother ratings correlating around .60 to .70, teachers with parents correlating between .30 and .60, and parent and teacher ratings correlating around .30 with children/adolescents’ self-ratings (Mervielde and De Fruyt, 2009). The magnitude of these correlations suggests that all perspectives share some variance, but also have their own specific and informative viewpoint. Teachers, for example, have experience with children in the more structured context of the classroom, and are in a good position to observe more interpersonal and task-oriented skills, whereas parents provide ratings relying on the home-context. In addition, teachers rely on a much broader frame-of-reference to describe pupils’ characteristics, because they accumulate professional experience with 20 new pupils in their classroom each year, whereas the scope of parents is usually much smaller and more idiosyncratic. De Los Reyes and colleagues (2013), for example, suggested that parents and teachers may have different “decision thresholds” for concluding

that behaviour is problematic. Finally, Rescorla et al. (2014) have conducted one of the most impressive studies on parent-teacher agreement on children's problem behaviours across 21 societies. There were striking similarities across societies: parents reported higher problem scores than teachers across societies (with some differences in magnitude across cultures), and similar age and gender differences were observed. Rescorla et al. (2014) also found, that within and across societies, parents and teachers agreed strongly on items that received low, medium, and high ratings.

In sum, the overall recommendation across many studies is to advocate cross-informant assessment because it brings shared but also complementary perspectives to the study of personality, social-emotional skills or problem behaviours. Informant discrepancies should hence not be considered by default as measurement error, though these viewpoints bring substance to the discussion. In addition, a multi-informant design has many psychometric advantages, including better ways to deal with common method variance to explain criteria and enhanced possibilities to estimate the variables at stake.

Ease of administration

Finally, constructs and accompanying measures will have to be relatively easy to administer to all informants (via an electronic assessment platform). Assessments will have to be completed by children and adolescents themselves, parents of different socio-economic classes, and teachers who will have to rate multiple children in their class. These constraints require grammatically streamlined and short items, an easily understandable format for the response scale to be used, and clear instructions. Guidance for such an approach can be found in several examples, such as the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) or the Hierarchical Personality Inventory for Children (HiPIC; Mervielde & De Fruyt, 2002).

Assessing short-term "transient error": Need for multiple sessions

Factor structure and reliability involve internal consistency across the test items, which can be assessed within a single testing occasion (or time), so those basic issues can be addressed in the design sketched out so far. However, another critical unknown when aiming to assess *change* is what has been called "transient error" (Chmielewski and Watson, 2009)—that is, low temporal reliability of the items within the same individual over different occasions (e.g., a week or a month at most). If that temporal error variation is too high, and we don't know that in advance, we will not be able to separate true, lasting, long-term change within the individual (e.g., change over 2 years) from short-term error variations within the same individual. Existing studies and data sets do not have data on kids, especially for comparing grades 7 and 12, and even adult data are very rare because repeated measurements are required. But kids are expected to attend school 5 days per week, thus it will be possible for the OECD study to obtain these kinds of data. So, transient error is a critical issue to examine in the feasibility studies. For example, the data may show that *more items* are needed in grade 7 to reach acceptable levels of short-term stability than in the later grades.

Context effects on measurement

Likely of lesser importance than transient error is that researchers do not know much about effects of context in schools on assessments, such as day of the week (e.g., Monday vs. Friday?) and time of day (e.g., early morning vs. before lunch vs. after lunch?). One might think that such effects are small but if they do not get examined in the feasibility studies, then context effects may have to be held constant in the longitudinal study itself (e.g., all assessment have to be completed before lunch). In longitudinal college studies, it can make a big difference whether assessments are done in the beginning or end of the semester, or close to finals, etc. Similar considerations would seem to apply here.

Teacher and parent ratings

Similar considerations apply to the younger cohorts in grades 1 to 6, for teacher and parent ratings. However, as these ratings are made from an external perspective, they tend to be less differentiated (more internally consistent across items) and less variable (more stable) over time. So, especially teacher's ratings can be obtained using fewer items, and assessing "transient error" (or short-term temporal stability) could be done in a subset of the teachers providing ratings (50% or even only 33% of the teachers or parents would have to be asked to provide their ratings a second time).

Mapping the anticipated measurement transition from grades 6 to 7

The issue of mapping the different data sources (e.g., teacher ratings in grade 6 and self-reports in grade 7) onto each other should be addressed by obtaining both data sources in either grade 6 or 7 (or preferably in both). That is, ideally the entire pilot sample from grades 6 and 7 would provide teacher or parent ratings as well as self-reports. This information will be critical in making sure that one can developmentally map the information collected from grades 1 to 6 with the information collected from grades 7 to 12. Again, the convergence of the 3 data sources ought to be tested separately for boys and girls, and possibly for ethnic minority groups, or even for high vs. low SES subgroups.

A related issue is how many "age specific" or "age appropriate" measures would have to be introduced. There is reason to assume that the same measure could be used from Grade 7 to 12. Recall that Soto et al. (2008, 2011) used the same BFI items (with their 5th grade readings levels) for ages 10 to 18. However, it would seem likely that the younger kids in grades 5 and 6 and maybe 7 (ages 10-13) might do better with a measure specifically designed for their age group, and that the more advanced conceptual language appropriate for older kids might best be limited to the high school grades (9-12). The feasibility studies will have to explore those possibilities.

REFERENCES

Allport, G. W. (1958). What units shall we employ? In G. Lindzey (Ed.), Assessment of human motives (pp.238-260). New York: Rinehart.

Ananiadou, K., and M. Claro (2009). 21st century skills and competencies for New Millenium Learners in OECD countries. Paris, France: Centre for Educational Research and Innovation (CERI) - New Millenium Learners.

Barenboim, C. (1981). The development of person perception in childhood and adolescence - from behavioral-comparisons to psychological constructs to psychological comparisons. *Child Development, 52*(1), 129-144. doi: 10.2307/1129222

Benet-Martinez, V., and S. Oishi (2008). Culture and personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 542-567). New York: Guilford.

Benet-Martinez, V., and O. P. John. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*(3), 729-750. doi: 10.1037/0022-3514.75.3.729

Berking, M., and H. Znoj (2008). Entwicklung und Validierung eines Fragebogens zur standardisierten Selbsteinscha¨tzung emotionaler Kompetenzen [Development and validation of a self-report measure for the assessment of emotion-regulation skills]. *Zeitschrift fur Psychiatrie, Psychologie und Psychotherapie*, 56, 141–152.

Bipp, T., and K. Van Dam (2014). Extending hierarchical achievement motivation models: The role of motivational needs for achievement goals and academic performance. *Personality and Individual Differences, 64*, 157-162. doi: 10.1016/j.paid.2014.02.039

Blair, C. (2002). School readiness. Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. American Psychologist, 57 (2), 111-127.

Briley, D. A., Domiteaux, M., and E. M. Tucker-Drob (2014). Achievement-relevant personality: Relations with the Big Five and validation of an efficient instrument. *Learning and Individual Differences, 32*, 26-39.

Catanzaro, S. J., and J. Mearns (1990). Measuring generalized expectancies for negative mood regulation: Initial scale development and implications. *Journal of Personality Assessment, 54*, 546-563.

CCR (2015). Character education for a challenging century: What should students learn? Downloaded from www.curriculumredesign.org

Cheung, F. M. et al. (2001). Indigenous Chinese personality constructs - Is the five-factor model complete? *Journal of Cross-Cultural Psychology, 32*(4), 407-433. doi: 10.1177/0022022101032004003

Cheung, F. M. et al. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology*(27), 181-199.

Chmielewski, M., and D. Watson. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology, 97*(1), 186-202. doi: 10.1037/a0015618

Church, A. T. (2010). Measurement issues in cross-cultural research. In G. Walford, E. Tucker and M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 151-177). Los Angeles: Sage.

Church, A. T. et al. (2011). Are Cross-Cultural Comparisons of Personality Profiles Meaningful? Differential Item and Facet Functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology, 101*(5), 1068-1089. doi: 10.1037/a0025290

Connelly, B. S., and D. S. Ones (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092-1122. doi: 10.1037/a0021212

Costa, P. T., and R. R. McCrae, (1992). *Revised NEO Personality Inventory and Five-Factor Inventory Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Coutinho, S. A. (2007). The realtionship between goals, metacognition, and academic success. *Educate, 7*(1), 39-47.

De Bolle, M. et al. (2012). General personality and psychopathology in referred and nonreferred children and adolescents: An investigation of continuity, pathoplasty, and complication models. *Journal of Abnormal Psychology, 121*(4), 958-970. doi: 10.1037/a0027742

Deci, E. and R. Ryan (2002). *Handbook of Self-Determination Research.* Rochester, NY: The University of Rochester Press.

De Fruyt, F. et al. (2009). Assessing the universal structure of personality in early adolescence with The NEO-PI-R and NEO-PI-3 in 24 Cultures. *Assessment, 16*(3), 301-311. doi: 10.1177/1073191109333760

De Fruyt, F. et al. (2006). Police interview competencies: Assessment and associated traits. *European Journal of Personality, 20*(7), 567-584.

De Fruyt, F., et al. (2006). Five types of personality continuity in childhood and adolescence. Journal of Personality and Social Psychology, 91(3), 538-552.

De Fruyt, F., and B. De Clercq. (2014). Antecedents of personality disorder in childhood and adolescence: Toward an integrative developmental model. *Annual Review of Clinical Psychology, 2014*(10), 449-476.

De Fruyt, F., and J. P. Rolland (2013). *Personality for Professionals Inventory: PfPI - Handleiding*. Amsterdam: Pearson.

De Fruyt, F., and K. Van Leeuwen (2014). Advancements in the field of personality development. [Article]. Journal of Adolescence, 37(5), 763-769. doi: 10.1016/j.adolescence.2014.04.009

De Fruyt, F., and B. Wille (2013). Cross-cultural issues in personality assessment. In N. D. and Tett Christiansen, R. P. (Ed.), Handbook of Personality at Work (pp. 333-354). New York.

De Fruyt, F., Wille, B., and O. P. John, (2015). Employability in the 21st century: Complex (interactive) problem solving and other essential skills. *Industrial and Organizational Psychology*, *8*, 2, 276-281.

De Haan, A. D. et al. (2013). Developmental personality types from childhood to adolescence: Associations with parenting and adjustment. *Child Development, 84*(6), 2015-2030. doi: 10.1111/cdev.12092

De Los Reyes, A. et al. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology, Vol 9, 9,* 123-149. doi: 10.1146/annurev-clinpsy-050212-185617

Debusscher, J., Hofmans, J., and F. De Fruyt (2014). The curvilinear relationship between state neuroticism and momentary task performance. *Plos One*, *9*(9). doi: e10698910.1371/journal.pone.0106989

Decuyper, M. et al. (2013). Latent personality profiles and the relations with psychopathology and psychopathic traits in detained adolescents. *Child Psychiatry and Human Development, 44*(2), 217-232. doi: 10.1007/s10578-012-0320-3

Decuyper, M. et al. (2009). A meta-analysis of psychopathy-, antisocial PD- and FFM associations. *European Journal of Personality, 23*(7), 531-565. doi: 10.1002/per.729

Denissen, J. J. A. (2014). A Roadmap for further progress in research on personality development. European Journal of Personality, 28, 213-215.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, *91*, 6, 1138-1151.

Diener, E. (2013). The remarkable changes in the science of subjective well-being. *Perspectives on Psychological Science, 8*(6), 663-666. doi: 10.1177/1745691613507583

Diseth, A. (2003a). Personality and approaches to learning as predictors of academic achievement. [Article]. *European Journal of Personality, 17*(2), 143-155. doi: 10.1002/per.469

Diseth, A. (2013b). Personality as an indirect predictor of academic achievement via student course experience and approach to learning. *Social Behavior and Personality, 41*(8), 1297-1308. doi: 10.2224/sbp.2013.41.8.1297

Duckworth, A. L., and M. E. P. Seligman (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*(1), 198-208.

Duckworth, A. L., et al. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101. doi: 10.1037/0022-3514.92.6.1087

Durlak, J. A. et al. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405-432. doi: 10.1111/j.1467-8624.2010.01564.x

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. New York: Psychology Press.

Elias, M. J. et al. (Eds.). (1997). *Promoting social and emotional learning: Guideliness for educators*. Alexandria, VA: Association for Supervision and Curriculum Development.

English, T. and O. P. John (2013). Understanding the social effects of emotion regulation: The mediating role of authenticity for individual differences in suppression. *Emotion, 13,* 314-329.

Entwistle, N. J., and H. Tait (1995). The revised approaches to studying inventory. Edinburgh: Centre for Research on Learning and Instruction.

Epstein, M. H. (2004). Behavioral and Emotional Rating Scale-2: A strength-based approach to assessment. Austin, TX: PRO-ED.

Erikson, E. H. (Ed.). (1950). *Childhood and society*. New York: Norton.

Fanous, A. H. et al. (2007). A longitudinal study of personality and major depression in a population-based sample of male twins. *Psychological Medicine, 37*(8), 1163-1172. doi: 10.1017/s0033291707000244

Fischer, S., Smith, G. T., and M. A. Cyders (2008). Another look at impulsivity: A meta-analytic review comparing specific dispositions to rash action in their relationship to bulimic symptoms. *Clinical Psychology Review, 28*(8), 1413-1425. doi: 10.1016/j.cpr.2008.09.001

Flavell, J. H. (1979). Meta-cognition and cognitive monitoring - new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906-911. doi: 10.1037/0003-066x.34.10.906

Føllesdal, H., and K. A. Hagtvet (2009). Emotional intelligence: The MSCEIT from the perspective of generalizability theory. *Intelligence, 37,* 94–105.

Friedman, H. S., and M. L. Kern (2014). Personality, Well-Being, and Health. In S. T. Fiske (Ed.), *Annual Review of Psychology, Vol 65* (Vol. 65, pp. 719-742). Palo Alto: Annual Reviews.

Furlong, M.J., et al. (2007). Cross-Validation of the Behavioral and Emotional Rating Scale-2 Youth Version: An Exploration of Strength-Based Latent Traits. Journal of Child and Family Studies. DOI 10.1007/s10826-006-9117-y

Furnham, A. (1996). The FIRO-B, the learning style questionnaire, and the five-factor model. *Journal of Social Behavior and Personality, 11*(2), 285-299.

Furnham, A., and J. Taylor (2004). *The dark side of behaviour at work: Understanding and avoiding employees leaving, thieving and deceiving.* Hampshire: Palgrave MacMillan.

George, L.G., Helson, R., and O. P. John (2011). The "CEO" of women's work lives: How Big Five Conscientiousness, Extraversion, and Openness predict 50 years of work experiences in a changing sociocultural context. *Journal of Personality and Social Psychology, 101,* 812-830.

Gerlach, G., Herpertz, S., and S. Loeber (in press). Personality traits and obesity: a systematic review. *Obesity Review*. doi: doi: 10.1111/obr.12235

Goldberg, L. R. et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96.

Goleman, D. (1995). Emotional Intelligence: Why It Can Matter More Than IQ, New York, Bantam

Gratz, K. L., and L. Roemer (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the Difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment, 26*, 41–54.

Halverson, C. F. et al. (2003). Personality structure as derived from parental ratings of free descriptions of children: The inventory of child individual differences. *Journal of Personality, 71*(6), 995-1026. doi: 10.1111/1467-6494.7106005

Hampson, S. E., John, O. P., and L. R. Goldberg (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. Journal of Personality and Social Psychology, 51, 37-54.

Hawes, D. J., Price, M. J., and M. R. Dadds (2014). Callous-unemotional traits and the treatment of conduct problems in childhood and adolescence: A comprehensive review. *Clinical Child and Family Psychology Review, 17*(3), 248-267. doi: 10.1007/s10567-014-0167-1

Helson, R. et al. (2002). The growing evidence for personality change in adulthood: Findings from research with personality inventories. *Journal of Research in Personality, 36*, 287-306.

Hendriks, A. A. J., Hofstee, W. K. B., and B. De Raad (1999). The Five-Factor Personality Inventory (FFPI). *Personality and Individual Differences, 27*(2), 307-325.

Hills, P., and M. Argyle (2001). Emotional stability as a major dimension of happiness. *Personality and Individual Differences, 31*(8), 1357-1364. doi: 10.1016/s0191-8869(00)00229-4

Hoekstra, H. A., and E. Van Sluijs (Eds.). (2003). *Managing competencies: Implementing human resource management*. Nijmegen: Royal Van Gorcum.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8*(3), 149-162.

Hofstee, W. K. B., Deraad, B., and L. R. Goldberg (1992). Integration of the Big-5 and Circumplex Approaches to Trait Structure. Journal of Personality and Social Psychology, 63(1), 146-163.

Hogan, J., and B. Holland (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*(1), 100-112.

Honey, P., and A. Mumford (1982). The Manuals of Learning Styles. Maidenhead: Honey Press.

Hurtz, G. M., and J. J. Donovan (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85*(6), 869-879.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66-100). New York: Guilford Press.

John, O. P. et al. (1994). The "Little Five": Exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development, 65*, 160-178.

John, O. P., and J. Eng (2014). Three approaches to individual differences in affect regulation: Conceptualization, measures, and findings. In J.J. Gross (Ed.), *Handbook of emotion regulation* (2nd ed.). New York: Guilford.

John, O. P., and J. J. Gross (2007). Individual differences in emotion regulation. In J.J. Gross (Ed.), *Handbook of emotion regulation* (pp. 351-372). New York: Guilford.

John, O. P., and J. J. Gross (2004). Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development. *Journal of Personality, 72*, 1301-1333.

John, O. P., and S. Mauskopf (2015, June 4). Self-reported socio-emotional qualities: Five factors for 21st century skills? Poster presented at the Bi-annual Meetings of the Association for Personality Research, Saint Louis, Missouri, USA.

John, O. P., and S. Srivastava (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, and O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). New York: Guilford.

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

John, O. P., Naumann, L., and C. J. Soto (2008). Paradigm shift to the integrative Big Five trait taxonomy: Discovery, measurement, and conceptual issues. In O.P. John, R.W. Robins, and L.A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114-158). New York: Guilford.

Joseph, D.L., and D. A. Newman (2010). Emotional intelligence: An integrative meta-analysis and cascading model. Journal of Applied Psychology, 95, 54-78.

Joshanloo, M. et al. (2014). Cross-cultural validation of fear of happiness scale across 14 national groups. *Journal of Cross-Cultural Psychology*, *45*(2), 246-264. doi: 10.1177/0022022113505357

Judge, T. A. et al. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*(3), 621-652.

Kautz, T. et al. (2014), "Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success", *OECD Education Working Papers*, No. 110, OECD Publishing, http://dx.doi.org/10.1787/5jxsr7vr78f7-en.

Kern, M. L., and H. S. Friedman (2008). Do conscientious individuals live longer? A quantitative review. *Health Psychology*, *27*(5), 505-512. doi: 10.1037/0278-6133.27.5.505

Klimstra, T. A. et al. (2014). Personality and externalizing behavior in the transition to young adulthood: the additive value of personality facets. *Social Psychiatry and Psychiatric Epidemiology*, *49*(8), 1319-1333. doi: 10.1007/s00127-014-0827-y

Kohnstamm, G. A. et al. (1998). *Parental descriptions of child personality. Developmental antecedents of the Big Five?* . Mahwah, NJ: Lawrence Erlbaum Associates.

Kolb, D.A. (1984). Experimental Learning. Englewood Cliffs, New York: Prentice-Hall.

Kolb, D.A. (1976). Learning Style Inventory: Techinical Manual. Boston: MA: McBer.

Kolko, David J. et al. (2009). Community vs. clinic-based modular treatment of children with early-onset ODD or CD: A clinical trial with 3-year follow-up. *Journal of Abnormal Child Psychology, 37*(5), 591-609. doi: 10.1007/s10802-009-9303-7

Komarraju, M. et al. (2011). The Big Five personality traits, learning styles, and academic achievement. *Personality and Individual Differences, 51*(4), 472-477. doi: 10.1016/j.paid.2011.04.019

Kotov, R. et al. (2010). Linking "Big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*(5), 768-821. doi: 10.1037/a0020327

Kwan, V. S. Y., Bond, M. H., and T.M. Singelis (1997). Pancultural explanations for life satisfaction: Adding relationship harmony to self-esteem. *Journal of Personality and Social Psychology, 73*(5), 1038-1051. doi: 10.1037/0022-3514.73.5.1038

Kyllonen, P. C., and J. P. Bertling (2013). Innovative Questionnaire Assessment Methods to Increase Cross-Country Comparability. Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis, 277.

Laceulle, O. M. et al. (2014). A test of the vulnerability model: temperament and temperament change as predictors of future mental disorders - the TRAILS study. *Journal of Child Psychology and Psychiatry, 55*(3), 227-236.

Landy, F. (2005). Some historical and scientific issues related to research on emotional intelligence. *Journal of Organizational Behavior, 26*, 411–424.

Lievens, F., and P. R. Sackett (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97*(2), 460-468. doi: 10.1037/a0025741

Lievens, F., De Corte, W., and E. Schollaert (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*(2), 268-279. doi: 10.1037/0021-9010.93.2.268

Lochman, J. E. et al. (2014). Does a booster intervention augment the preventive effects of an abbreviated version of the coping power program for aggressive children? *Journal of Abnormal Child Psychology, 42*(3), 367-381. doi: 10.1007/s10802-013-9727-y

Loehlin, J. C. et al. (1998). Heritabilities of common and measure-specific components of the Big Five personality factors. *Journal of Research in Personality, 32*, 431-453.

Lopes, P. N. et al. (2005). Emotion regulation ability and the quality of social interaction. *Emotion, 5*, 113 – 118.

Markus, H. R., and S. Kitayama (1991). Culture and the self - implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224-253. doi: 10.1037//0033-295x.98.2.224

Martel, M. M. et al. (2014). Longitudinal prediction of the one-year course of preschool ADHD symptoms: Implications for models of temperament-ADHD associations. *Personality and Individual Differences, 64*, 58-61. doi: 10.1016/j.paid.2014.02.018

Martel, M. M. et al. (2010). The structure of childhood disruptive behaviors. *Psychological Assessment, 22*(4), 816-826. doi: 10.1037/a0020975

Mayer, J. D., and Y. N. Gaschke (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology, 55*, 102–111.

Mayer, J. D., Salovey, P., and D. R. Caruso (2002). *MSCEIT users manual.* MHS: Toronto.

McCrae, R. R., and P. T. Costa (1996). Toward a new generation of personality inventories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), The five-factor model of personality: Theoretical perspectives (pp. 51-87). New York/London: Guilford Press.

McCrae, R. R., and O. P. John (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality, 60*, 175-215.

McCrae, R. R., and A. Terracciano (2008). The Five-Factor Model and its correlates in individuals and cultures. In F.J.R. Van de Vijver, D. A. Van Hemert and Y. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 249-283). Mahwah, NJ: Erlbaum.

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

Measelle, J. R. et al. (2005). Can children provide coherent, stable, and valid self-reports on the Big Five dimensions? A longitudinal study from ages 5 to 7. *Journal of Personality and Social Psychology, 89*, 90-106.

Mervielde, I., and F. De Fruyt (2002). Assessing children's traits with the Hierarchical Personality Inventory for Children. In B. De Raad and M. Perugini (Eds.), *Big Five Assessment* (pp. 129-146). Gottingen: Hogrefe and Huber Publishers.

Mervielde, I., and F. De Fruyt (1999). Construction of the Hierarchical Personality Inventory for Children (HiPIC). In I. Mervielde, I. Deary, F. De Fruyt and F. Ostendorf (Eds.), *Personality Psychology in Europe, Proceedings of the Eight European Conference on Personality Psychology* (pp. 107-127). Tilburg, The Netherlands: Tilburg University Press.

Mervielde, I., De Fruyt, F., and B. De Clercq (2009). *Hiërarchische Persoonlijkheidsvragenlijst voor Kinderen [Hierarchical Personality Inventory for Children]: Handleiding*. Amsterdam: Hogrefe Publishers.

Milfont, T. L., and C. G. Sibley (2012). The big five personality traits and environmental engagement: Associations at the individual and societal level. *Journal of Environmental Psychology, 32*(2), 187-195. doi: 10.1016/j.jenvp.2011.12.006

Minbashian, A., and D. Luppino (2014). Short-term and long-term within-person variability in performance: An integrative model. *Journal of Applied Psychology, 99*(5), 898-914. doi: 10.1037/a0037402

Minbashian, A., Wood, R. E., and N. Beckmann (2010). Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology, 95*(5), 793-806. doi: 10.1037/a0020016

Morris, A. S. et al. (2007). The role of the family context in the development of emotion regulation. *Social Development, 16*, 362-388.

Munafo, M. R., Zetteler, J. I., and T. G. Clark (2007). Personality and smoking status: A meta-analysis. *Nicotine and Tobacco Research, 9*(3), 405-413. doi: 10.1080/14622200701188851

National Academy of Sciences (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. Washington, D.C.: National Academies Press.

Neff, K. D., Kirkpatrick, K. L., and S. S. Rude (2007). Self-compassion and adaptive psychological functioning. Journal of Research in Personality 41 (2007) 139–154

Nel, J. A. et al. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality, 80*(4), 915-948. doi: 10.1111/j.1467-6494.2011.00751.x

Nigg, J. T. et al. (2002). Big Five dimensions and ADHD symptoms: Links between personality traits and clinical symptoms. *Journal of Personality and Social Psychology, 83*, 451-469.

Obschonka, M. et al. (2013). The regional distribution and correlates of an entrepreneurship-prone personality profile in the United States, Germany, and the United Kingdom: A socioecological perspective. *Journal of Personality and Social Psychology, 105*(1), 104-122. doi: 10.1037/a0032275

Obschonka, M., Schmitt-Rodermund, E., and A. Terracciano (2014). Personality and the gender gap in self-employment: A multi-nation study. *Plos One, 9*(8). doi: e103805

Obschonka, M., Silbereisen, R. K., and E. Schmitt-Rodermund (2012). Explaining entrepreneurial behavior: Dispositional personality traits, growth of personal entrepreneurial resources, and business idea generation. *Career Development Quarterly, 60*(2), 178-190. doi: 10.1002/j.2161-0045.2012.00015.x

OECD (2014). Skills for social progress. Paris, France: OECD-Centre for Education Research and Innovation.

OECD (2015). Draft proposal: OECD Longitudinal Study of Social and Emotional skills in Cities. Paris: OECD.

OECD (2015). ESP Longitudinal Study of Skill Dynamics in Cities. Paris: OECD.

Oh, I. S., Wang, G., and M. K. Mount (2010). Validity of observer ratings of the Five-Factor Model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*(4), 762-773. doi: 10.1037/a0021832

Omoto, A. M., Snyder, M., and J. D. Hackett (2010). Personality and motivational antecedents of activism and civic engagement. *Journal of Personality, 78*(6), 1703-1734. doi: 10.1111/j.1467-6494.2010.00667.x

Oswald, A. J., Proto, E., and D. Sgroi (2014). Happiness and productivity. Warwick: University of Warwick.

Papachristou, H. et al. (2013). Higher levels of trait impulsiveness and a less effective response inhibition are linked to more intense cue-elicited craving for alcohol in alcohol-dependent patients. *Psychopharmacology, 228*(4), 641-649. doi: 10.1007/s00213-013-3063-3

Park-Higgerson, H. K. et al. (2008). The evaluation of school-based violence prevention programs: A meta-analysis. *Journal of School Health, 78*(9), 465-479. doi: 10.1111/j.1746-1561.2008.00332.x

Poropat, A. E. (2014a). A meta-analysis of adult-rated child personality and academic performance in primary education. [Article]. *British Journal of Educational Psychology, 84*(2), 239-252. doi: 10.1111/bjep.12019

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

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. *Psychological Bulletin, 135*(2), 322-338. doi: 10.1037/a0014996

Primi, R. et al. (2014, submitted). The development of a nationwide inventory assessing social and emotional skills in Brazilian youth.

Prinzie, P., Onghena, P., and W. Hellinckx (2005). Parent and child personality traits and children's externalizing problem behavior from age 4 to 9 years: A cohort-sequential latent growth curve analysis. Merrill-Palmer Quarterly-Journal of Developmental Psychology, 51(3), 335-366.

Rescorla, L. A. et al. (2014). Parent-teacher agreement on children's problems in 21 societies. *Journal of Clinical Child and Adolescent Psychology, 43*(4), 627-642. doi: 10.1080/15374416.2014.900719

Roberts, B. W., Walton, K. E., and W. Viechtbauer (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*(1), 1-25. doi: 10.1037/0033-2909.132.1.1

Roberts, B. W., Wood, D., and J. L. Smith (2005). Evaluating Five Factor Theory and social investment perspectives on personality trait development. [Proceedings Paper]. Journal of Research in Personality, 39(1), 166-184. doi: 10.1016/j.jrp.2004.08.002

Rosander, P., and M. Backstrom (2012). The unique contribution of learning approaches to academic performance, after controlling for IQ and personality: Are there gender differences? *Learning and Individual Differences, 22*(6), 820-826. doi: 10.1016/j.lindif.2012.05.011

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

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

Salgado, J. F. et al. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*(3), 573-605.

Salgado, J. F. et al. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*(6), 1068-1081.

Salovey, P. et al. (1995). Emotional attention, clarity, and repair: Exploring emotional intelligence using the trait meta-mood scale. In J. W. Pennebaker (Ed.), *Emotion, disclosure, and health* (pp. 125–154). Washington, DC: American Psychological Association.

Santos, D. and R. Primi (2014). Promoting social and emotional skills for societal progress in Rio de Janeiro. Paris, France: OECD-Centre for Education Research and Innovation.

Saucier, G., and F. Ostendorf (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology, 76*, 4, 613-627.

Schmitt, D. P. et al. (2007). The geographic distribution of big five personality traits - Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*(2), 173-212. doi: 10.1177/0022022106297299

Schnittker, J. and J. R. Behrman (2012). Learning to do well or learning to do good? Estimating the effects of schooling on civic engagement, social cohesion, and labor market outcomes in the presence of endowments. *Social Science Research, 41*(2), 306-320. doi: 10.1016/j.ssresearch.2011.11.010

Schraw, G., and R. S. Dennison (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460-475. doi: 10.1006/ceps.1994.1033

Scott, S., Briskman, J., and T. G. O'Connor (2014). Early prevention of antisocial personality: Long-term follow-up of two randomized controlled trials comparing indicated and selective approaches. *American Journal of Psychiatry, 171*(6), 649-657. doi: 10.1176/appi.ajp.2014.13050697

Seligman, M. E. P. et al. (2005). Positive Psychology Progress: Empirical Validation of Interventions. *American Psychologist, 60, 5,* 410–421 DOI: 10.1037/0003-066X.60.5.410.

Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. *Personality and Social Psychology Bulletin, 20*(5), 580-591. doi: 10.1177/0146167294205014

Sklad, M. et al. (2012). Effectiveness of school-based universal social, emotional, and behavioral programs: do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools, 49*(9), 892-909. doi: 10.1002/pits.21641

Soto, C. J. et al. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*, 330-348.

Soto, C. J. et al. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*, 718-737.

Soto, C. J., and O. P. John (in press). Traits in transition: The structure of parent-reported personality traits from early childhood to early adulthood. *Journal of Personality*.

Soto, C. J., and O. P. John (2015). *Conceptualization, development, and validation of the BFI-2: The next version of the Big Five Inventory.* Manuscript submitted for publication.

Soto, C. J., and O. P. John (2012). Development of Big Five domains and facets in adulthood: Mean-level age trends and broadly versus narrowly acting mechanisms. *Journal of Personality, 80,* 881-914.

Soto, C. J., and O. P. John (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality, 43,* 84-90.

Spengler, M. et al. (2013). Personality is related to educational outcomes in late adolescence: Evidence from two large-scale achievement studies. *Journal of Research in Personality, 47*(5), 613-625. doi: 10.1016/j.jrp.2013.05.008

Srivastava, S. et al. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84,* 1041-1053.

Stautz, K. and A. Cooper (2013). Impulsivity-related personality traits and adolescent alcohol use: A meta-analytic review. *Clinical Psychology Review, 33*(4), 574-592. doi: 10.1016/j.cpr.2013.03.003

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin, 133*(1), 65-94. doi: 10.1037/0033-2909.133.1.65

Steel, P., Schmidt, J., and J. Shultz (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*(1), 138-161. doi: 10.1037/0033-2909.134.1.138

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401-426. doi: 10.1016/j.intell.2006.09.004

Sutin, A. R. et al. (2011). Personality and obesity across the adult life span. *Journal of Personality and Social Psychology, 101*(3), 579-592. doi: 10.1037/a0024286

Tackett, J. L. (2006). Evaluating models of the personality-psychopathology relationship in children and adolescents. *Clinical Psychology Review, 26*(5), 584-599.

Tackett, J. L. et al. (2013). Delineating personality traits in childhood and adolescence: Associations across measures, temperament, and behavioral problems. *Assessment, 20*(6), 738-751. doi: 10.1177/1073191113509686

Tackett, J. L. et al. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology, 122*(4), 1142-1153. doi: 10.1037/a0034151

Tamir, M. et al. (2007). Implicit theories of emotion: Affective and social outcomes across a major life transition. *Journal of Personality and Social Psychology, 92,* 731-744.

Taylor, N., and G. P. De Bruin (2005). Basic Traits Inventory: Technical manual. Johannesburg: Jopie van Rooyen and Partners, SA.

Terracciano, A. et al. (2013). Personality and resilience to Alzheimer's disease neuropathology: a prospective autopsy study. [Article]. *Neurobiology of Aging, 34*(4), 1045-1050. doi: 10.1016/j.neurobiolaging.2012.08.008

Thomas, A. et al. (1963). *Behavioral individuality in early childhood*. New York: New York University Press.

Trapnell, P. D., and J. D. Campbell (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology, 76*(2), 284-304. doi: 10.1037/0022-3514.76.2.284

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

Tsai, J. L., Knutson, B., and H. H. Fung (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology, 90*(2), 288-307. doi: 10.1037/0022-3514.90.2.288

Valchev, V. H. et al. (2014). Beyond agreeableness: Social-relational personality concepts from an indigenous and cross-cultural perspective. *Journal of Research in Personality, 48*, 17-32. doi: 10.1016/j.jrp.2013.10.003

Van de Vijver, F., and K. Leung (1997). *Methods and data analysis for cross-cultural research*. Thousands Oaks, CA: Sage.

Van den Akker, A. L. et al. (2013). The Development of Personality Extremity From Childhood to Adolescence: Relations to Internalizing and Externalizing Problems. *Journal of Personality and Social Psychology, 105*(6), 1038-1048. doi: 10.1037/a0034441

Van den Akker, A. L., Dekovic, M., and P. Prinzie (2010). Transitioning to adolescence: How changes in child personality and overreactive parenting predict adolescent adjustment problems. *Development and Psychopathology, 22*(1), 151-163. doi: 10.1017/s0954579409990320

Van Hiel, A., Mervielde, I., and F. De Fruyt (2006). Stagnation and generativity: Structure, validity, and differential relationships with adaptive and maladaptive personality. *Journal of Personality, 74*(2), 543-573.

Vandenberg, R. J., and C. E. Lance (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for roganizational research *Organizational Research Methods*(3), 4-70.

Vecchione, M. et al. (2014). Fakability of Implicit and Explicit Measures of the Big Five: Research findings from organizational settings. *International Journal of Selection and Assessment, 22*(2), 211-218. doi: 10.1111/ijsa.12070

Versteeg, H. et al. (2012). Type D personality and health status in cardiovascular disease populations: a meta-analysis of prospective studies. *European Journal of Preventive Cardiology, 19*(6), 1373-1380. doi: 10.1177/1741826711425338

Wang, L., et al. (2009). Assessing Teamwork and Collaboration in High School Students: A Multimethod Approach. C*anadian Journal of School Psychology* 2009; 24; 108 DOI: 10.1177/0829573509335470.

Weinberg, A., and E. Klonsky (2009). Measurement of emotion dysregulation in adolescents. *Psychological Assessment, 21*, 616–621.

Wiggins, J. S. (1973). *Personality and prediction*. Jossey-Bass Publishers.

Wille, B., De Fruyt, F., and M. Feys (2013). Big Five traits and intrinsic success in the new career era: A 15-year longitudinal study on employability and work-family conflict. *Applied Psychology-an International Review-Psychologie Appliquee-Revue Internationale, 62*(1), 124-156. doi: 10.1111/j.1464-0597.2012.00516.x

Wille, B., De Fruyt, F., and M. Feys (2010). Vocational interests and Big Five traits as predictors of job instability. *Journal of Vocational Behavior, 76*(3), 547-558. doi: 10.1016/j.jvb.2010.01.007

Woods, S. A. et al. (2013). Personality across working life: The longitudinal and reciprocal influences of personality on work. *Journal of Organizational Behavior, 34*, S7-S25. doi: 10.1002/job.1863

ANNEX B.

ANCHORING VIGNETTES REDUCE BIAS IN NONCOGNITIVE RATING SCALE RESPONSES

Patrick Kyllonen and Jonas Bertling, Educational Testing Service, Princeton, New Jersey, USA

TABLE OF CONTENTS

Executive Summary

The current method of choice for measuring noncognitive skills (social-emotional skills, self-management skills, personality, attitudes, etc.) is the rating scale. Rating scales have the student rate him or herself on some behavior, trait, or attribute (e.g., "I work hard") typically on a 4- or 5-point agreement or frequency scale (e.g., "strongly agree" to "strongly disagree," or "never" or "seldom" to "often" or "always"). Alternatively rating scales are completed by teachers or parents who rate students. The rating scale is more popular, and the one about which more is known than any other method. For example, many studies on noncognitive skills in psychology, education, and economics are based on data from rating scales. This supports a proposal to use the rating scale methodology for the OECD Longitudinal Study of Skill Dynamics.

Despite their popularity, and their support in the scientific literature, rating scales have limitations. An important limitation is that rating scale responses are subject to certain biases, most notably response style biases and reference group effects. The most well-known biases are as follows:

a. Extreme Response Style – the tendency to respond using the extremes of the response scale (e.g., "strongly agree," "strongly disagree");
b. Midpoint Response Style – the tendency to respond using the midpoint of the response scale (e.g., "neither agree nor disagree");
c. Acquiescence Response Style – the tendency to agree with statements regardless of the statement (i.e., respond "agree" or "strongly agree" to most or all of the items);
d. Socially Desirable Response Style – the tendency to respond in a socially desirable way so as to make oneself look good;
e. Modesty Response Style – the tendency to avoid boasting or exaggeration in responses;
f. Reference Group Effects – the effect on responses of reference group comparisons, for example, to indicate low absolute achievement due to comparing oneself with relatively high achievers;
g. Halo and Horn Effects – the tendency for others to fail to differentiate targets on a set of traits, and instead to vote them uniformly high or low on all traits.

These biases are important generally, but are particularly important in cross-cultural comparisons—because of large cross-cultural differences in response styles and reference groups it is difficult to compare responses in one jurisdiction with responses given in another. This has been an issue in PISA and other large-scale international assessments.

There are several methods to address these biases and to increase cross-cultural comparability. *Forced-choice formats* are effective, but take longer to administer than rating scales. *Ratings by others* reduce many of the biases, and can supplement self-ratings, but are subject to halo and horn effects. *Situational judgment tests* are also useful, but are time consuming and require higher development costs.

Anchoring vignettes are a new method for addressing biases and increasing cross-cultural comparability. The method was used successfully in PISA 2012 to increase data quality and cross-cultural comparability. The method has also been used in other studies for measuring a variety of constructs, ranging from socioeconomic status to political efficacy. Anchoring vignettes are ideally suited for measuring the noncognitive skills that will be measured in the OECD Longitudinal Study of Skill Dynamics.

Anchoring vignettes involve students (or teachers or parents) reading a set of vignettes (e.g., 3), where each vignette describes a hypothetical person. The hypothetical persons are designed to be either high, medium, or low on the trait being measured. For example, the trait might be conscientiousness, and one of the hypothetical persons will be described as being organized, consistently meeting deadlines, and working hard (an example of a "high" trait individual). Another might be described as sometimes putting in a good effort, but also as being somewhat disorganized and often late for meetings (an example of a "low" trait individual). The student is asked how they would rate each hypothetical person on a trait, such as "conscientiousness." The student then rates himself on measures of conscientiousness (e.g., "I work very hard," "I am organized," "I meet deadlines"). The self-ratings are recoded to align with the anchors (detailed description is provided in the main paper, paragraphs 11-17).

The recoded, or anchoring-vignette-adjusted ratings, have been shown to be superior to the unadjusted ratings in the following ways:

a. The correlation between noncognitive skills and achievement is usually higher (for a given country);
b. Countries with the higher noncognitive skills also have the highest achievement (this was not true for the unadjusted scores, where the highest scoring countries had the lowest noncognitive scores).

It is not necessary to have anchoring vignettes for every scale—if that were necessary anchoring vignettes would be a very time consuming. Instead, it is possible to correct several response scales (e.g., dependability, organization, industriousness), using responses to a set of vignettes that only measures one of those response scales (e.g., industriousness only). Often, adjustments made based on a related scale (e.g., organization vignettes) correct, say, industriousness items, as well as adjustments made based on the scale itself (i.e., industriousness vignettes).

There are methods that can be used to study the quality and effectiveness of anchoring vignettes. They are as follows:

a. The best vignette sets are *regular* ones in which the highest vignette is rated higher than the medium vignette, which is rated higher than the lowest vignette. Vignette sets in which the high and medium vignettes are given the same rating (or the medium and low vignettes) (these are called "ties"), or in which a medium vignette is rated higher than a high vignette (these are called "violations"), do not produce data that is as good in quality as regular ones do. So we can look at the percentage of ties and violations as a diagnostic for the quality of vignettes.
b. The best vignette sets produce a roughly even distribution into the new score categories. With anchoring vignette scoring, the original scale (e.g., 4-point rating scale) is transformed into a new scale (e.g., 7-point adjusted scale). For good vignette sets, there are responses in each of the 7 categories. Poor vignette sets will have uneven response frequencies in the different categories. The distribution of responses across categories is also a good diagnostic for the quality of vignettes.

Although anchoring vignettes have been developed for various projects in K-12, including PISA 2012 and PISA 2015, and have proven useful, there still is a need for pilot testing of anchoring vignettes to ensure that they will provide the highest quality data when used for scoring rating scale responses for noncognitive skill assessment.

Nevertheless it is clear that anchoring vignettes are a low-risk, high-payoff method designed and proven to improve data quality in general, to reduce biases associated with rating scale responses, and to improve cross-cultural comparability of noncognitive skill assessments. Also, anchoring vignettes take less than one minute to administer, and probably only one to three need to be administered for the purposes of enabling score adjustments for all measures in a survey. For this reason it is recommended that anchoring vignettes be included in any K-12 noncognitive assessments that rely on rating scale responses.

Background

By far the most common method for measuring noncognitive skills, including personality, attitudes, values, and subjective norms, is the rating scale, also commonly known as the Likert scale (1932). Numerous meta-analyses of the relationship between personality and school or workforce outcomes have been conducted based exclusively on rating scale responses (e.g., Barrick & Mount, 1991; Poropat, 2009; Salgado & Tauriz, 2014). With the rating scale method respondents rate themselves, or others, by indicating their level of agreement with a set of descriptive statements through choosing an ordered category that best characterizes their agreement.

For example, the respondent might be asked to, "indicate your level of agreement with the following statement: 'I work well with others,'" and be presented the response categories, "strongly disagree, disagree, agree, strongly agree." Alternatively respondents could be asked to indicate frequency (e.g., "often, sometimes, rarely"), feelings of regard (e.g., "very much, somewhat, a little, not at all"), quality judgments (e.g., "poor, average, above average, outstanding, truly exceptional") or other qualities in response to a question or statement.

Rating scales are widely used because they provide more information than true-false judgments (as in Guttman and Thurstone scales), and are more efficient than preference judgments. For example, the NEO-PIR, a widely used commercial Big 5 personality inventory presents 243 items in 35 minutes (on a 1 = "strongly disagree" to 5 = "strongly agree" scale), a rate of approximately 7 items per minute. This rate is consistent with other personality inventories. Alternatives to ratings scales, such as preference judgments, take longer. For example, creating a scale from preference judgments (e.g., "which do you agree with more, 'I work well with others,' vs. 'I work hard'), typically requires collecting judgments on $n \times (n-1) / 2$ statement pairs rather than on n statements. Thus 243 statements paired would take over 460 hours and each item (pair) requires reading two statements rather than one as in the single statement format. (Devising strategies to reduce this number, such as sampling pairs, is a way to make the preference judgment strategy more feasible, but the general point that preferences are less efficient than statement ratings is still true.)

For the reason of efficiency, rating scales have been the method of choice for measuring noncognitive qualities, and are used in the questionnaires of all international large scale assessments, including OECD's Program for International Student Assessment (PISA), Program for the International Assessment of Adult Competencies, the Teaching and Learning International Scale, as well as the U.S. Department of Education's National Center for Education Statistics National Assessment of Educational Progress (NAEP), IES's The International Mathematics and Science Survey (TIMSS), PIRLS, ALS, and others.

The Bias Problem

An assumption in interpreting rating scale responses is that there is a common understanding among participants in what the response categories mean, and that respondents use the rating scale in the same way. If two respondents both "strongly agree" with a statement, then an assumption is that the two

respondents (or the targets of the ratings, in the case of others' ratings) have a similar level of the trait the item is designed to assess.

However, it has been known for over 60 years (e.g., Cronbach, 1946) that there are response style effects or response biases that threaten the validity of this interpretation, meaning that respondents commonly present construct-irrelevant response patterns. Among the best documented response style biases are the following:

a. *Acquiescence* or "yea-saying:" the tendency to respond positively to a statement (e.g., to agree with everything);
b. *Extreme response style*: the tendency to use the endpoints of the response scale, such as "extremely" or "never;"
c. *Midpoint response style*: the tendency to use the middle points or midpoint of the response scale, such as "neutral" or "neither disagree nor agree;"
d. *Socially desirable response style*: the tendency to select a response that is more socially acceptable or reflects more favorably on the respondent, regardless of how well it characterizes the respondent;
e. *Modesty response style*: the tendency to respond in a way indicating a desire to avoid bragging, boasting, expressiveness, or exaggeration, and to present humility, cautiousness, and modesty (this is sometimes related to midpoint response style; and is often associated with Confucian Asian culture) (Culpepper, et al., 2012);
f. For ratings of others, there is the *halo or horn response style*, which is the tendency to rate the respondent in a consistent way across items (e.g., favorably or unfavorably) regardless of the specific differences between items. This response style results in high correlations across items (e.g., all scales are correlated around r = .70 with all other scales) and a lack of differentiation between dimensions underlying the items. A factor analysis of such a matrix will result in reduced dimensionality, or even a single factor, when without the response style influence on responses there might be multiple factors.

There are other construct-irrelevant influences on ratings, which operate like response styles in contributing bias to ratings measures, such as context effects and item position effects.

One of the most important of these is reference group effects, which affect ratings through the influence of different comparison (or reference) groups for different respondents. For example, if students are asked how strong their mathematical skills are (e.g., "very strong, strong, weak, or very weak"), they might respond differently depending on their classmates' abilities. An illustration of a reference group effect is that in one study of PISA Marsh and Hau (2004) found little variation in academic self-concept across schools, despite substantial variation across schools in achievement itself. Van de gaer, et al., (2012) noted with PISA 2006 data, with the focus on science, that academic self-concept and achievement were positively related among students from the same school and country. But the correlation between these variables was negatively correlated at the level of schools or countries, that is, countries (and schools) with high average self-concept tended to have low average achievement. They attributed this finding to reference group effects where the high standards and norms that characterize the high achieving schools and countries tended to decrease academic self-concept, with low performing schools and countries doing the opposite. They supported this hypothesis with the finding that for countries with tracked schools, or more selective schools the between-school correlation was larger—selective schools and tracking distorts the reference group, thus lowering high ability (in high track) students' academic self-concept and increasing low ability (in low track) students' self-concept.

The Attitude-Achievement Paradox

Kyllonen and Bertling (2013) pointed out that the phenomenon of high positive within-group correlations between a noncognitive factor obtained through ratings and an achievement score on the one hand, and a negative between-group correlation between the same two factors on the other hand, is not limited to countries and schools but is seen with other subgroups, and sometimes is referred to as the attitude-achievement paradox. The cause could be reference group effects, but more generally it could be due to any number of response biases or response style effects as listed above. The primary issue is that respondents from different groups are not using the response categories in a common way, either due to differing understandings of what the response categories mean, or other response style influences.

Description of Anchoring Vignettes

If the problem in interpreting rating-scale responses across individuals and groups is that individuals differ in the way they interpret the response categories in rating scale items, then one way to address this issue is through the use of anchoring vignettes (King et al., 2004; King & Wand, 2007). The anchoring vignette technique presents a set of vignettes describing hypothetical individuals (or situations). Individuals (or situations) are designed to vary on the trait being evaluated. The respondent rates the vignettes using the same rating scale on which the respondent rates him or herself.

For example, PISA 2012 included the following set of anchoring vignettes to measure the construct *Student-Teacher Relations*:

ST61

01 Below you will find descriptions of three mathematics teachers. Read each of the descriptions of these teachers. Then let us know to what extent you agree with the final statement.

(Please check only one box on each row.)

	Strongly agree	*Agree*	*Disagree*	*Strongly disagree*
a) Ms. Anderson assigns mathematics homework every other day. She always gets the answers back to students before examinations. **Ms. Anderson is concerned about her students' learning.**	☒ 1	☐ 2	☐ 3	☐ 4
b) Mr. Crawford assigns mathematics homework once a week. He always gets the answers back to students before examinations. **Mr. Crawford is concerned about his students' learning.**	☐ 1	☒ 2	☐ 3	☐ 4
c) Ms. Dalton assigns mathematics homework once a week. She never gets the answers back to students before examinations. **Ms. Dalton is concerned about her students' learning.**	☐ 1	☐ 2	☐ 3	☒ 4
02 My teacher lets students know they need to work hard.	☐ 1	☒ 2	☐ 3	☐ 4

In this example, the item numbered "01" is the set of anchoring vignettes describing different teachers. Boxes are checked to indicate how a student might rate these boxes. The item number "02" is the self-rating. In PISA, self-ratings are done in sets (typically 4-6 ratings) to measure a scale, so there would be additional self-ratings here, but only one is shown to illustrate the concept. In this example, the student rates exemplar "a" "1: Strongly agree," exemplar "b" "2: Agree," and exemplar "c" "4: Strongly disagree." The student rates himself or herself as "2: Agree" to the statement "My teacher lets students know they need to work hard." The vignettes and the self-rating statement are all designed to reflect the construct, *Student-Teacher Relations*.

In nonparametric Anchoring Vignette scoring, the student's response is recoded to a new scale that has $2k + 1$ categories, where k is the number of vignettes. In this case, with three vignettes, the new scale would have 7 categories; let's call the original response R, and the anchoring vignette recoded response R*.

Vignettes are written to suggest either low (L), medium (M), or high (H) levels of the trait. In the example above, "a" is a high vignette, "b" is a medium vignette, and "c" is a low vignette.

In the *regular* case, that is when the respondent's vignette ratings follow L < M < H (respondent rates the lowest vignette the lowest, the highest vignette the highest, and the medium vignette somewhere in between) for a set of vignettes for a particular trait (we reverse code here to make the example easier to follow), then an anchoring-vignette score for an item that measures the trait is computed as follows:

$$\text{If } R < L, \text{ then } R^* = 1 \quad (1)$$
$$\text{If } R = L, \text{ then } R^* = 2 \quad (2)$$

$$\text{If } L < R < M, \text{ then } R^* = 3 \quad (3)$$
$$\text{If } R = M, \text{ then } R^* = 4 \quad (4)$$
$$\text{If } M < R < H, \text{ then } R^* = 5 \quad (5)$$
$$\text{If } R = H, \text{ then } R^* = 6 \quad (6)$$
$$\text{If } R > H, \text{ then } R^* = 7 \quad (7)$$

where R is the self-rating on an item that measures that trait, and R* is the anchoring vignette adjusted score for that item. So in the example, R = M (i.e., the respondent's "agree" response to item 02 is the same response as he or she gave to the Medium, or M exemplar), and so R* = 4. Response (1) can be interpreted as the respondent rating his or her own teacher as lower than he or she rated any of the vignette teachers; response (7) can be interpreted as the respondent rating his own teacher as higher than any of the vignettes, response (3) can be interpreted as the respondent rating his own teacher as between the low and medium vignette teachers, and so forth.

In cases of *ties*, either L = M < H, L < M = H, or L = M = H, there is partial indeterminacy in the recoding. For example, in the latter case (L = M = H), if the respondent rates him or herself as below his vignette ratings, that is R < L = M = H, then R* = 1, as in the *regular* case, and if he rates himself as above, R > L = M = H, then R* = 7, as in the *regular* case, but if he rates himself as R = L = M = H, then R* in principle could take any of the values 2 to 6. In PISA data, we tried various approaches to the indeterminacy problem, such as (a) let R* be the lowest possible value it can take, (b) let R* be the highest possible value it can take, (c) let R* be the median value it can take, and found that following rule "a" led to a better psychometric properties for R* (e.g., reliability).

In cases of *model violations*, either L > M, L > H, or M > H, then there are several possible treatments. (a) One is to treat the data for items showing violations as missing. (b) Another is to fix the

violation by changing an inequality to an equality (if L > M, let L = M; if M > H, let M = H; if L > H, then let L = H, and that might create another model violation, e.g., L = H > M, then fix that violation in the same way, L = H = M) then treating the changed as ties in the way ties are treated. When an inequality is changed to equality, then the resulting value must be established, in the same was as in 16. We found in PISA data that assuming the lower of the two values tended work well. These are obviously atheoretical, data-based solutions, and so these issues can be revisited with new data.

Applications of Anchoring Vignette Scoring

Anchoring vignettes have been used for various purposes. King et al. (2004) presented an example of a one-item vignette to measure political efficacy.

> *[Moses] lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future. How much say does [Moses] have in getting the government to address issues that interest him?*

Following this, respondents were asked to judge the amount of political efficacy they had themselves, and King and Wand (2007) reported that 40% of Chinese and 16% of Mexican respondents rated themselves as having less political efficacy than Moses, even though the Chinese paradoxically rated themselves as having more political efficacy than Mexicans according to the unadjusted rating. This example illustrates the usefulness of anchoring vignettes in increasing cross-cultural comparability, and also illustrates that even a single vignette can help improve measurement. In PISA 2012 and 2015 and in other applications we use three vignettes, but more or fewer can be used. Mottus et al. (2012) used 30 vignettes to adjust for respondent's Conscientiousness ratings!

There is some evidence that even without anchoring vignette scoring, there is benefit to presenting anchoring vignettes prior to participants performing their self-ratings (Hopkins & King, 2010). Vignettes may serve to anchor respondents' use of the response scale to decrease response style bias and to better communicate the meaning of the question. (Hopkins and King also recommend not combining vignettes and self-assessments because it has the opposite effect in that it reduces response consistency and response informativeness.)

Personality-Based Anchoring Vignettes

We developed anchoring vignettes for Personality-related constructs that are currently tested in the PISA 2015 field trial (Bertling & Kyllonen (2012a). Vignettes for Test Anxiety, Motivation, and Organization were developed (see following table for an example).

Table 3. . Example Anchoring Vignette from PISA 2015 Field Trial Questionnaire

Anchoring Vignette – Conscientious student		Please read the descriptions about the following three students. Based on the information provided here, how much would you agree with the statement that this student is organized
Low	1	Tom often delays before starting on his homework and sometimes turns in assignments late.
Medium	2	George likes to make detailed to-do lists but sometimes does things at last minutes.
High	3	Andrew works consistently throughout the term and keeps detailed notes for all subjects.

Source: Bertling & Kyllonen (2012a)

We further developed anchoring vignettes specifically for the Big Five that were designed to be used either in an online survey or a phone interview (Bertling & Kyllonen, 2012b). In addition, several ETS projects currently investigate the validity of the anchoring vignettes methodology in K-12 and Higher Education Assessments (Bertling & Almonte, 2014; Bertling, Olivera-Aguilar, Petway, & Robbins, in preparation). See the following tables for examples.

Table 4. Anchoring Vignettes Developed for Assessing the Big Five Personality Variables

no.	Domain	Level	Alpha-Version	Beta-Version	Statement
1	O	Low	[X] dislikes reading. She/he seems puzzled when presented with complex ideas. [x] is most comfortable in familiar settings.	[X] finds it hard to understand why people get emotional. She/he dislikes learning and does not like to speculate about things.	[X] is open-minded.
	O	Med	[X] likes to travel and visit new places. She/he likes to solve complex problems but prefers not to participate in philosophical discussions.	[X] does not enjoy reading. But she/he can easily handle a lot of information and sometimes asks questions that nobody else does.	[Y] is open-minded.
	O	high	[X] is constantly seeking explanations for things. She/he enjoys learning and discussing books with others.	[X] appreciates all forms of art. She/he constantly looks for opportunities to learn and is valued by others for her/his objectivity.	[Z] is open-minded.
2	C	Low	[X] is often late does little planning. She/he quickly lose interest in the tasks she/he starts and gives up easily.	[X] often makes a mess of things. She/he puts little time and effort into her/his work and often leaves things unfinished.	[X] is conscientious.
	C	Med	[X] likes to make lists and tries to avoids mistakes. But sometimes she/he rushes into things.	[X] makes careful choices but tries to avoid responsibilities. Mostly she/he sticks with what she/he decides to do.	[Y] is conscientious
	C	high	[X] always makes and effort and starts tasks right away. She/he wants every detail taken care of.	[X] demands perfection in her/himself and others. She/he makes plans and sticks to them. She/he never gives up.	[Z] is conscientious
3	E	Low	[X] avoids crowded places and loud music. She/he enjoys silence and rarely starts conversations. She/he does not like talking about her/himself.	[X] is quiet around strangers. She/he prefers to be alone and does not like to share her/his opinions with others. She/he avoids dangerous situations.	[X] is extraverted.
	E	Med	[X] enjoys sharing jokes and cheering people up. She/he likes to be part of a group but carefully discloses her/his intimate thoughts.	[X] is known for her/his sense of humor. She/he likes the company of others but also enjoys spending time by her/himself.	[X] is extraverted.
	E	high	[X] enjoys parties and is very talkative. She/he frequently laughs out loud. She/he likes to be in the center of attention.	[X] seeks excitement. She/he loves to talk and often shares her/his feelings with others. She/he enjoys being reckless.	[X] is extraverted.
4	A	Low	[X] often takes advantage of other people and speaks ill of others. She/he wants to have more power than other people.	[X] frequently Insults people. She/he lies to get her/himself out of trouble and tries to avoid doing favors for others.	[X] is an agreeable person.
	A	Med	[X] usually tries to trust people and is good with working with a group. But sometimes she/he likes to complain.	[X] likes to criticize others but is usually polite to strangers. She/he is a good listener but hard to convince.	[X] is an agreeable person.
	A	high	[X] is very trusting of other people. She/he does what others want her/him to do and always acknowledges others' accomplishments.	[X] sometimes lets other people take credit for her/his work. She/he trusts what people say and is good with working with a group.	[X] is an agreeable person.
5	ES	Low	[X] often feels desperate. She/he tends to get annoyed at the slightest irritation and hardly knows where his/her life is going.	[X] often snaps at people. She/he has a dark outlook at the future and gets overwhelmed by emotions easily.	[X] is emotionally stable.
	ES	Med	[X] always controls her/his emotions but sometimes feels it hard to get going. She/he needs approval of others.	[X] usually remains calm under pressure but sometimes feels sad. She/he worries about what other people thinks of her/him.	[Y] is emotionally stable.
	ES	high	[X] is always in a good mood. Rarely does she/he get annoyed. She/he faces danger confidently and stays calm even in stressful situations.	[X] readily overcomes setbacks and can handle stress very well. She/he hardly ever cried in her/his life.	[Z] is emotionally stable.

Note. O=Openness; C=Conscientiousness; E=Extraversion; A=Agreeableness; ES=Emotional Stability; each vignette is represented by three descriptions (low, med, high); there are two versions for each vignette.

Source: Bertling & Kyllonen (2012b)

Table 5. Example Anchoring Vignettes for Computer Familiarity

Linda often uses apps to talk to her friends or to play games. She does most of her homework on a computer and knows how to write and format papers on a computer and how to create complex tables or charts. She also created a few presentations using a computer. Linda can type pretty accurately using ten fingers when looking at the computer keyboard.

On a scale from 0-10 where zero is not at all familiar and 10 is very familiar, **how familiar with computers and digital technology do you think Linda is?**

Source: Bertling & Almonte (2014)

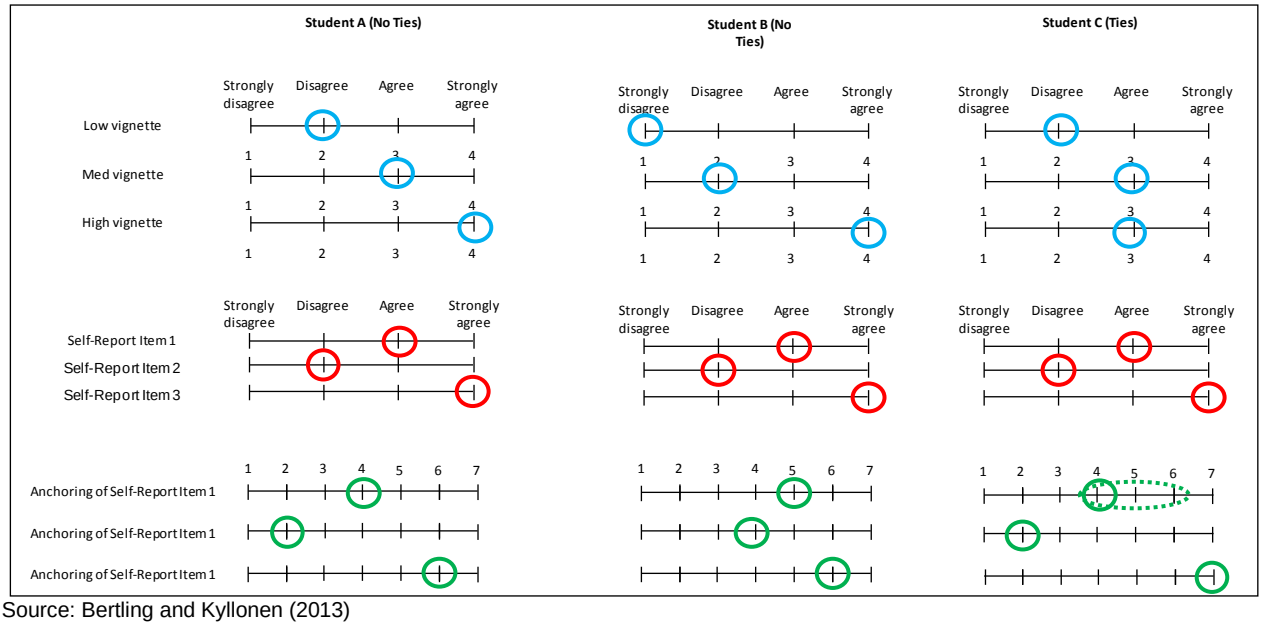
Likert Scale Adjustments Based on Anchoring Vignettes in PISA 2012

In PISA 2012 we extended the original nonparametric anchoring vignette method so that multiple items could be anchored based on the same set of anchoring vignettes. When items were scored based on vignettes, numerical values for student responses were not assigned based on the concrete response option chosen (e.g., the value 4 for "strongly agree" and 3 for "agree") but based on the self-report answer relative to the personal standard captured by the respondent's individual rating of the three vignettes that form one set. Regardless of where on the 4-point agreement scale a student places the vignettes, a student's self-report can be scored relative to his/her rating of low, medium and high for the vignettes. Based on this approach, in PISA 2012, students' responses on the original 4-point agreement scale were re-scaled into a 7-point scale representing all possible relative rank comparisons of students' self-reports and their rating of the vignettes. On this 7-point scale, the value one represents a rating lower than the low vignette, the value two represents a rating at the level of the low vignette, the value three represents a rating higher than the low but lower than the medium vignette, and so forth. The maximum score, seven, is assigned when a student's self-reported response is higher than the rating of the high vignette. In other words, low values are assigned when a self-report rating is relatively low compared to the evaluation of the vignettes, and high values are assigned when a self-report rating is relatively high compared to the evaluation of the vignettes. The following table shows the possible values for original and anchored item responses in PISA 2012.

Responses to question as presented in questionnaire	Strongly disagree	Disagree	Agree	Strongly Agree			
	1	2	3	4			
Anchored responses	lower than low vignette	Same as low vignette	In between low and medium vignette	Same as medium vignette	In between medium and high vignette	Same as high vignette	Higher than high vignette
	1	2	3	4	5	6	7

In this way, the three vignettes are used to anchor student judgments, providing context for the ratings on other questions sharing the same response scale. Scoring is applied on the individual student level using each student's responses to the vignettes as an anchor for this student's self-reported responses to various Likert-type questions. A graphical illustration of the scoring procedure based on vignettes for three examples with and without ties is given in **Error! Reference source not found.**. The three hypothetical students in this example provided exactly the same responses to the three self-reported items shown, but differ in their responses to the vignettes. As a result, scores on the anchored items also differ between the three students.

Figure 3. Illustration of scoring based on vignettes for three hypothetical students



Source: Bertling and Kyllonen (2013)

Enhanced Cross-country Comparability in PISA 2012

Two sets of anchoring vignettes were included in the PISA 2012 Student Questionnaire to allow for alternative scoring of self-report items based on students' defined standards when using the 4-point agreement scale (strongly agree – agree – disagree – strongly disagree). As the following graph shows evaluations of the three vignettes for both sets differ considerably across countries. While there is a clear order of low, medium, and high vignettes overall (left side), absolute ratings on the four-point Likert scale differ between countries (right side).

Source: Bertling & Kyllonen (2013)

Field trial and main survey analyses consistently showed that within-country correlations with achievement tended to be higher for anchored scales, and correlations at the between-student within-country level and the between-country level did not show the inconsistencies found for unanchored scales. No "paradoxical" correlations were found for any anchored index, but correlations on the between-country level were similar to student-level correlations, both within countries and for the pooled sample. The absolute values of the between-country correlations tended to be larger than the correlations at the between-student within-country level. The following table shows changes to the correlations with achievement when items are scored based on vignettes. The strong negative correlation based on country means flipped into moderate to strong positive correlations when items were scored based on vignettes. Only scoring of vignettes showed consistent results within countries, based on country means, and based on the total sample.

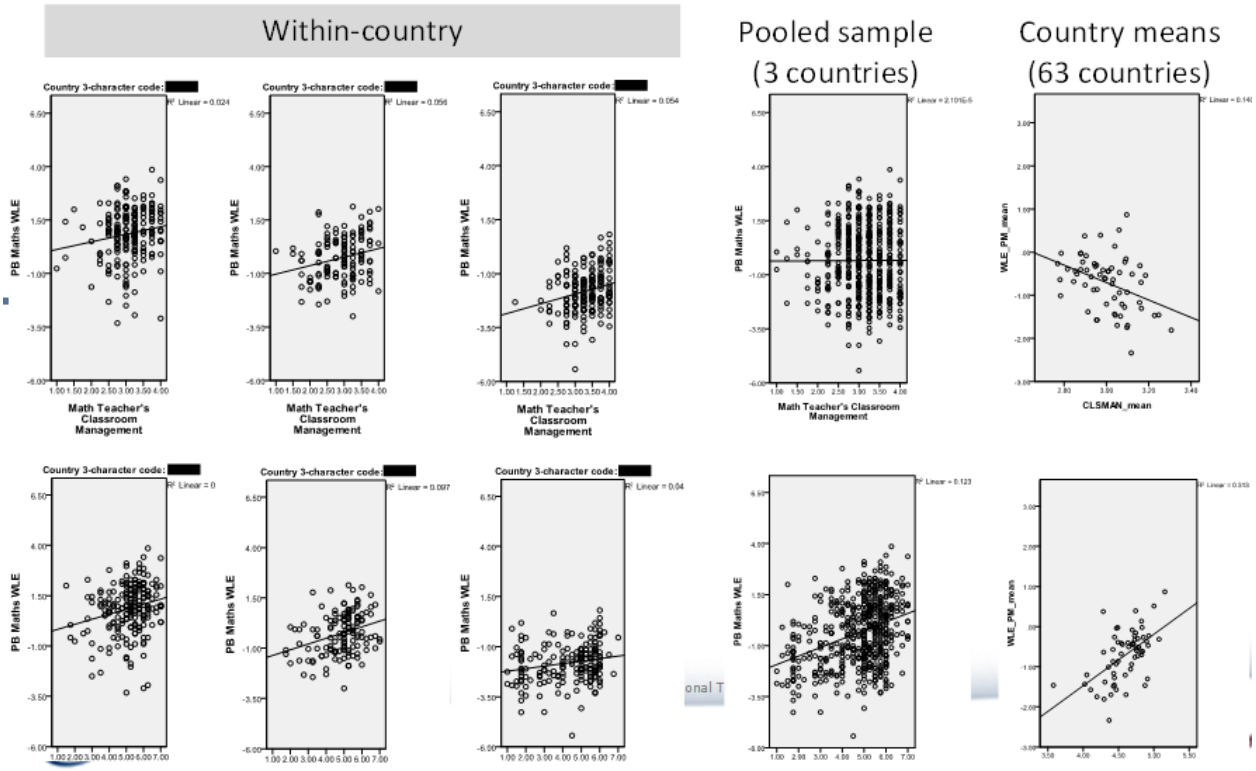
		Correlations with Mathematics Proficiency			Correlations with Response Style	
		Average Within-Country	Correlation based on ountry Means	Pooled (total) sample	ARS	ERS
Teacher Support	Likert-scale based index	**0.04**	**-0.41**	**-0.03**	**0.48**	**0.25**
	Scoring based on vignettes	**0.13**	**0.22**	**0.15**	**0.15**	**0.05**
Classroom Management	Likert-scale based index	**0.07**	**-0.38**	**0.03**	**0.21**	**0.14**
	Scoring based on vignettes	**0.20**	**0.56**	**0.23**	**0.01**	**0.01**

Source: Bertling & Kyllonen (2013)

The following graph further illustrates these findings. On the left side it is shown that the validity of the classroom management index, indicated by a positive relationship with mathematics achievement,

holds within countries when the anchoring adjustment is applied. Here, no large effects on the correlations with achievement are found. However, as shown on the right side, the validity of self-report indices is substantially improved when pooled data for all countries or country means are investigated. While results on different data aggregation levels disagree for Likert-scale based indices, correlations align when anchoring is used.

Comparisons of the two sets of anchoring vignettes showed very similar results with no major differences in the pattern of correlations between scales and achievement. Further, in comparison with the original indices, anchored indices showed smaller degrees of DIF and smaller correlations with indicators of acquiescensce or disacquiesence response styles (Bertling & Kyllonen, 2013).



Source: Bertling & Kyllonen (2013)

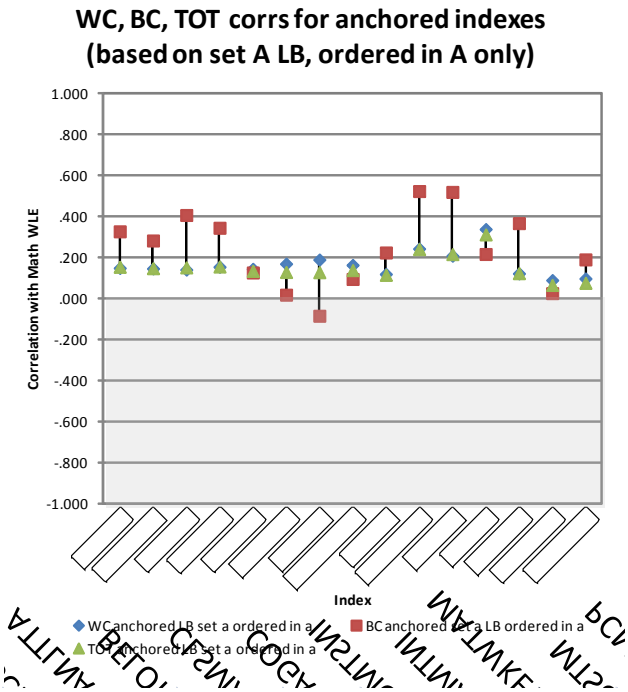
Validity Improvement within Countries in PISA 2012

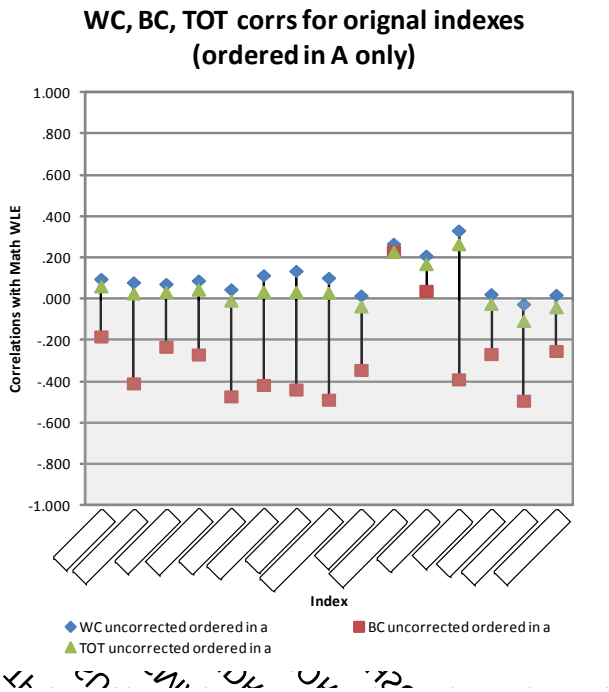
In addition to validity improvements on the country level, findings from PISA 2012 analyses demonstrate that validity can as well be improved within countries, for instance, by removing response style variance from students' self-report ratings. We looked at relationships between anchoring and response styles in several different ways: First, correlations between response style scores and vignette ratings were investigated. High vignette ratings correlated positively with indicators of Acquiescense Response Style and negatively with Disacquiescense response style. As shown in the table above Likert-scale based indices show moderate to strong correlations with achievement but indices based on vignette scoring do not show these correlations. In order to investigating further what the unique contribution of anchored scores to the prediction of achievement, and what the contribution of response style analyses is, we specified stepwise multiple regression models for four constructs. In a first step, achievement was

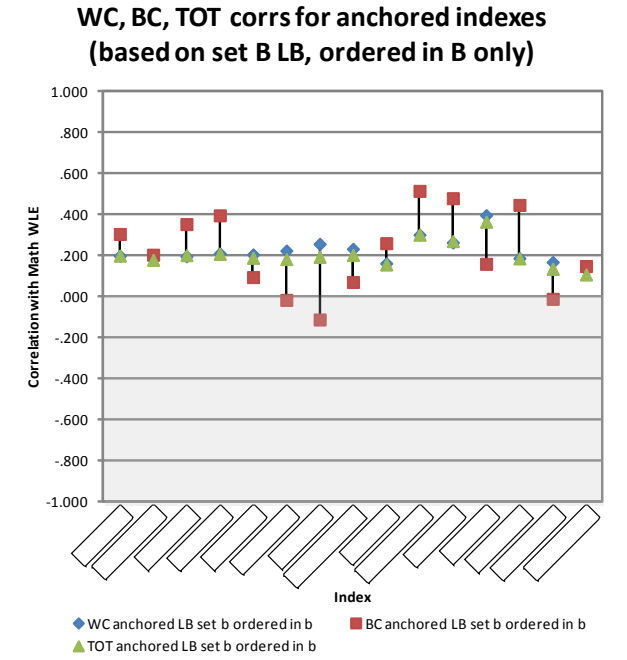
predicted based on uncorrected scales only. In a second step, response style was added as a predictor. As a third step, anchored scores for the index included in step one were added to the model. We thereby could estimate that contribution to better measurement of vignettes in addition to statistical response style correction. In a first step, achievement was predicted based on student-reported teacher classroom management only. The predictive power of this model is very poor with less than one percent of explained variation in mathematics proficiency scores. In a second step, response style was added to the model. Only acquiescence was included here. Previous analyses showed that inclusion of other response styles did not change the models considerably. Performance of the model significantly increased with a value of R2 a bit higher than two percent based on the inclusion of the additional two response style indicators. In a third step, adjusted self-report scores based on nonparametric anchoring were included, yielding another significant increase in the prediction of the model. R2 changed from .022 to .084 in the classroom management model (change in R2 was between .046 and .069 for the different models). Standardized regression weights indicated that anchoring-adjusted self-reports had the highest impact on mathematics proficiency compared to the other predictors in the model. Furthermore, partial correlations indicated that the unique contribution of response styles almost vanished when anchoring was considered. However, the effects of ARS not disappear completely which might be an indication that response styles are not fully captured by the vignettes. It is also possible that students do not use the same standards when evaluating their own teachers or themselves vs. hypothetical individuals presented in the vignettes. In sum, these analyses show that Anchoring Vignettes can successfully adjust Likert scale response for response styles and can enhance measurement over and beyond what can be done by simple statistical corrections without vignettes.

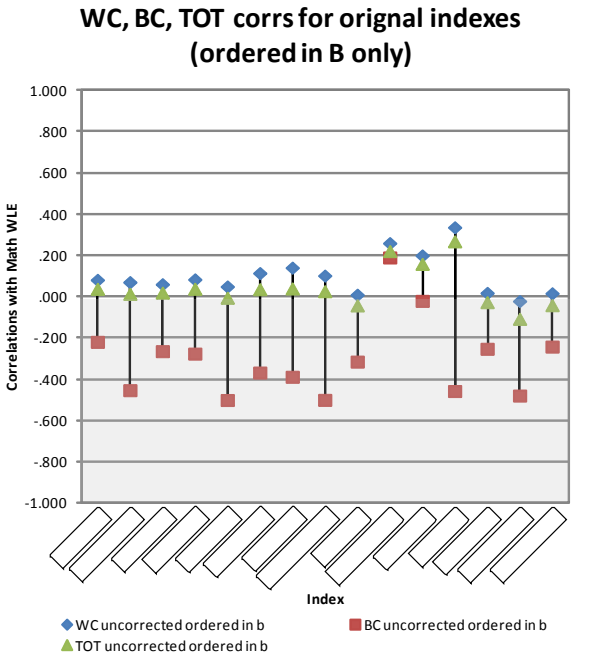
Anchoring Adjustments across Multiple Items and Scales

Anchoring vignettes were designed to correct bias in questionnaire indexes, not in stand-alone questions as the examples given above. A new approach of implementing anchoring vignettes for complete scales was developed and tested in the PISA 2012 field trial. Here, the same anchors were applied to all self-report responses using the same scale, i.e. for a 5-item scale all five item responses would be rescaled based on the same set of anchors. Aligned with this different goal, vignettes were written to capture broader constructs as measured with several Likert-type items, not only the content of one specific item. For example, one set of vignettes combined several teacher behaviors that were identified as valid indicators of teacher support (here: assigning homework, giving feedback, timeliness of feedback). Similar behaviors were then also captured in the self-report items that are combined to build the reflective index. A third assumption is made for the current application of using one set of vignettes to a larger number of self-report items (especially in scenario B where up to 15 indexes could be adjusted) in addition to the two assumptions described above (i.e., vignette equivalence, response consistency). That is, it is assumed that evaluative standards are invariant across self-report items as long as the same response format is used. The following figures show the success of this approach. On the left the alignment of correlations for 15 anchored indices is shown. In the first figure all indices are anchored based on the firt set of vignettes with the same vignettes applied to all 15 indices. In the second figure all indices are anchored based on a second set of vignettes. On the right results for original responses are displayed. It can be seen that (a.) correlations for all 15 indices closely align when anchoring is applied but not for original scales, and (b.) results for the two sets of vignettes are almost identical. These findings support that general response style adjustments are possible across multiple indices without including one or several vignettes for each index.





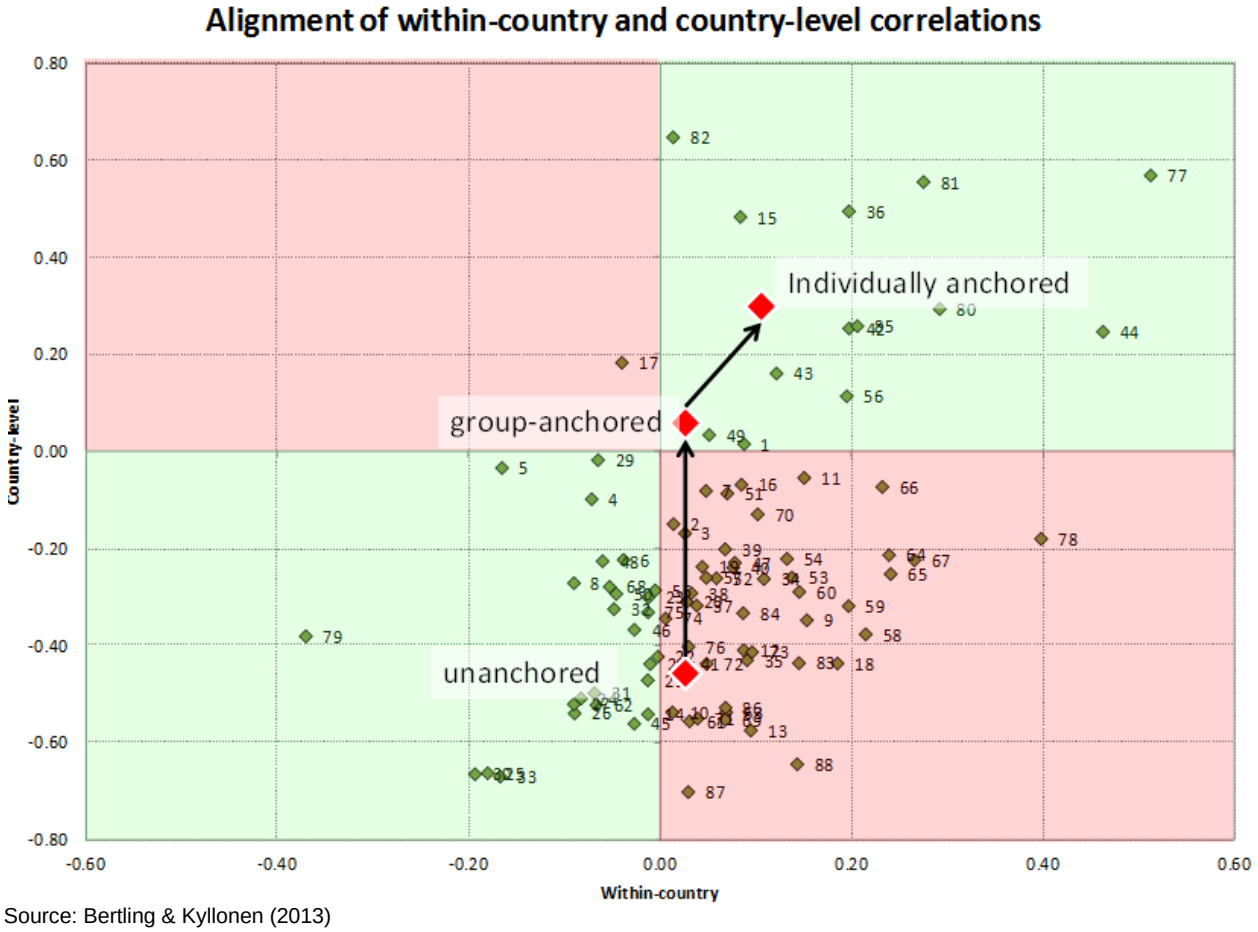




Source: Bertling & Kyllonen (2013)

Group-based Anchoring

Based on PISA 2012 field trial data we investigated whether typical evaluations for the vignettes in each country (indicated by the most popular response for each vignette in a given country) could lead to the same validity improvement. Results indicated that the validity could be improved but the results fell short of the improvements if individual-level adjustments were applied. The following graph shows how the within and between country correlations changed based on group-anchoring versus individual anchoring. On the horizontal axis the average correlation within the PISA countries is shown. On the vertical axis the correlation based on country means is shown. Every point represents an index. Points in the green quadrants signal that the sign of the two correlations is the same. Points in red quadrants signal that the signs of the two correlations are in conflict.



Source: Bertling & Kyllonen (2013)

Variability in Anchoring Vignette Quality

In studies, such as PISA 2012, there are clear indications that some vignettes are more successful than others in yielding better response data. Better can be defined in psychometric qualities such as higher reliability, and higher correlations with external variables, such as achievement.

h. For example, in PISA 2012 there were two vignette sets (Teacher Support and Classroom Management). They differed in the degree to which corrections based on them changed the correlations.
i. ETS researchers developed four sets of vignettes to measure various noncognitive skills in community college students. The sets were developed to correspond with four clusters of constructs. Vignette sets differed substantially in the percentage of ties (9%, 19%, 20%, 47%) and violations (4%, 7%, 9%, 22%, respectively). The vignette set with the fewest ties and violations (set 1) turned out not only to be the best correction for the constructs it was designed for (i.e., the two set 1 constructs), but also turned out in many cases to provide better corrections (better defined as correlations with external variables) for constructs it was not designed for (i.e., set 2 constructs). That is, set 2 constructs adjusted by set 1 anchoring vignettes were better than set 2 constructs adjusted by set 2 anchoring vignettes.

There are a set of diagnostics that can be used to help evaluate vignette quality. These are:
(a) distributing recoded scores across the 2k + 1 categories—vignette sets that lead to a more normal or even distribution of recoded scores (R*) can be considered a good vignette set; and

(b) minimizing the number of ties and violations. Vignettes sets that produce a minimal number of ties (e.g., L = M) and violations (e.g., L > M) can be considered a good vignette set.

(c) These two diagnostics also tend to be correlated with other psychometric measures of recoded scores such as reliability and correlations with external variables. That is, a prediction is that holding all else constant, vignette sets that distribute scores evenly across recoded categories, and minimize the number of ties and violations are likely to yield scores that have higher reliability and higher correlations with external (criterion) variables, such as achievement scores.

Using Anchoring Vignettes to Measure Growth

Although to our knowledge growth measurement per se has not been an application area for anchoring vignettes, there is promise in using them this way. Consider the one-anchor vignette (paragraph 18). Tracking percentages of respondents rating themselves above the anchor over time (e.g., in the years 2010 vs. 2020) would seem to be a useful approach to gauge growth in political efficacy in the society. In the realm of noncognitive measurement vignettes might serve as concrete anchors against which to monitor growth. For example, to measure responsibility, “Charlie is able to find the classroom without parental assistance” might be a high anchor in first grade, but a low anchor by third grade.

In general, vertical scaling techniques could be used to measure growth over time. In vertical scaling and equating, a subset of common test items (i.e., standardized cognitive test items) is administered in different grades (e.g., 3rd grade and 4th grade) to enable equating scores collected in different grades, and putting such scores on a common scale across grades. This enables comparison of students in different grades. In the same way, a vertical scaling technique could be applied to anchoring vignettes, which would allow evaluating individual student noncognitive skills growth from year to year on a common scale, comparing students in different grades with respect to their noncognitive skill proficiency, and so forth. A means for doing this with anchoring vignettes would be to use common anchors across school grades, introducing new anchors with higher grades as the common anchors become easier (or lower). This would permit evaluations such as determining growth in the percentage of students rating themselves higher than a common anchor given in two different grades. The interpretability of this kind of statistic would depend on common interpretation across school grades in a manner similar to the assumption of common

interpretation across individuals and groups with anchoring vignettes in general. Some thought and experimentation might be necessary to develop anchoring vignettes that would operate effectively to measure growth across grades, but there is no reason in principle why this method should not be successful and useful.

Comparison of Anchoring Vignettes vs. Other Corrections

There are several alternatives to anchoring vignettes that can be used to correct response style and reference group effects, and other response biases.

One of these can be called response style pattern corrections. Unlike anchoring vignettes, these methods use existing response data from rating-scale responses, and recode them, reweight them, or simply correlate them with external variables, such as country, or achievement, or personality. He's (2014) dissertation on this topic related response styles to external variables, and found correlations with personality (where personality was measured with the forced-choice technique so as to eliminate response style effects) (see also, Smith & Fischer, 2008). Khorammdel and Von Davier (2014) presented a new multidimensional item-response theory method to detect trait-unrelated response styles reliably, particularly, extreme- and midpoint- response styles, and found that response styles were consistent across Big 5 trait scales, and that there were reliable cultural differences in their use. They also had suggestions for correcting for response styles. Buckley (2012) applied pattern correction methods to PISA 2009 data and showed that doing so affected country rank orderings on various noncognitive scales. While all these methods are somewhat effective (although they might not affect validity, Ones et al, 1996), they are fundamentally limited in that there is always some arbitrariness in determining what part of a response is attributable to response style and what part is attributable to the construct. If a person responds with a "strongly agree" how much of that is due to actual strong agreement, and how much is due to the person displaying an extreme response style?

Another method is forced-choice, or rankings more generally. Bartram showed that a forced-choice version of a personality inventory (e.g., "Select the one that better matches you---(a) I work hard, or (b) I work well with others") provided more sensible and interpretable correlations with country-level indicators (such as United Nations Quality Index than did rating scale responses. We (Kyllonen & Bertling, 2013) also demonstrated that PISA forced-choice pairs showed better psychometric properties than PISA rating scales for the statements in those forced choice pairs.

Conclusions

At this time the most common and efficient way to measure many noncognitive skills, including personality, attitudes, and values, is through the use of rating scales. But such scales also introduce response bias, including reference group, social desirability, and response style bias. This problem is important in comparing schools and subgroups within a country, and is particularly important in cross-cultural comparisons. There are several ways to address this problem, including the use of pattern correction methods, and the use of forced-choice item administration. Pattern correction methods are useful, but they suffer from arbitrariness in partitioning a response into trait-relevant and trait-irrelevant components. Forced-choice methods are very promising and are increasingly being applied to measure noncognitive skills cross-culturally. We can expect continued development in this area. However, although methods are improving, this approach has the drawback of being inefficient in that many more paired-comparisons are required to get measurement comparable to what is given by ratings of statements.

At this time anchoring vignettes appear to be the most promising means for correcting responses from noncognitive assessments. Results from PISA 2012 seemed to indicate that the use of anchoring vignettes resulted in much more comparable measures across countries, resulting in more interpretable findings comparing countries on noncognitive scales. The additional amount of time required to administer anchoring vignettes is fairly minimal—it appears that it is not necessary for respondents to rate a set of anchoring vignettes for each trait scale. Instead it appears that just a set or two of anchoring vignettes can be administered then used to adjust responses across a wide range of self-rating scales. Anchoring vignettes also seem promising as a means to evaluate growth over time, or over grades, using a fixed benchmark. This is particularly important in a project that seeks to evaluate longitudinally measured noncognitive skills growth across primary and secondary education years.

REFERENCES

Barrick, M. R. and M. K. Mount, (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Bertling, J. P. and D. E. Almonte (2014). Anchoring Vignettes to improve the measurement of students' familiarity with computers and digital technology in the National Assessment of Educational Progress. Internal memo. ETS, Princeton.

Bertling, J. P. and P. C. Kyllonen (2012a). Domain-general student attitudes and behaviors: Constructs and Items for PISA 2015 (Module 10). Technical report submitted to PISA Questionnaire Expert Group.

Bertling, J. P. and P. C. Kyllonen (2012b). Big Five Assessment Plan: Multi-method assessment based on Likert type items, anchoring vignettes, and forced choice items. Technical memo submitted to OECD.

Bertling, J. P. and P.C. Kyllonen (2013), Using anchoring vignettes to detect and correct for response styles in PISA questionnaires, in: Prenzel, M. (Chair), The Attitudes-Achievement-Paradox: ko

How to Interpret Correlational Patterns in Cross-Cultural Studies, Invited Symposium at the EARLI 2013, Munich, Germany.

Bertling, J. P. et al. (in preparation). Assessing noncognitive factors in higher education: A comparison of traditional self-report measures and anchoring vignettes.

Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf.

Cronbach, L. J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.

Culpepper, R., Zhao, L. and J. Balenger (2012). Investigating the use of Likert-type, variable-response items in cross-cultural research: The influence of cultural-based response bias. In 7th European Conference on Research Methodology for Business and Management.

He, J. (2014). The psychological meaning of survey response styles from a cross-cultural perspective (Unpublished doctoral dissertation). Tilburg University, the Netherlands.

He, J. et al. (2014). Response Styles and Personality Traits: A Multilevel Analysis. Journal of Cross-cultural psychology. 0022022114534773

Hopkins, D. and G. King (2010). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. Public Opinion Quarterly: 1-22. Copy at http://j.mp/jVFIVg

Khorramdel, L. and M. von Davier (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49(2),* 161-177.

King, G. and J. Wand (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis 15*, 46-66.

Kyllonen, P. C. and J. P. Bertling (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski, (eds.) *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis.* Boca Raton: CRC Press.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology 140.

Marsh, H. W. and K. T. Hau (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. Journal of Educational Psychology, 96(1), 56-67.

Mottus, R. et al. (2012). The effect of response style on self-reported conscientiousness across 20 countries. Personality and Social Psychology Bulletin, 38, 1423-1436.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin, 135(2), 322-338. doi: 10.1037/a0014996.

Salgado, J.F. and G. Tauriz (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23(1),* 3-30.

Smith, P. B. and R. Fischer (2008). Acquiescence, extreme response bias and culture: A multilevel analysis. In F. J. R. van de Vijver, D. A. van Hemert & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 285-314). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Van de Gaer, E. et al. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. Journal of Cross-Cultural Psychology, 43(8),1205-1228.

Van Vaerenbergh, Y. and T. D. Thomas (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*. doi: 10.1093/ijpor/eds021