Organisation for Economic Co-operation and Development

EDU/WKP(2024)11

Unclassified

English - Or. English

4 June 2024

DIRECTORATE FOR EDUCATION AND SKILLS

Cancels & replaces the same document of 4 June 2024

Innovative Tools for the Direct Assessment of Social and Emotional Skills

OECD Education Working Paper No 316

By Adriano Linzarini* and Daniel Catarino da Silva*, OECD

Adriano Linzarini, OECD, Adriano.LINZARINI@oecd.org
Daniel Catarino da Silva, OECD Daniel.CATARINODASILVA@oecd.com .

* The authors contributed equally to this work

JT03545185

OECD EDUCATION WORKING PAPERS SERIES

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Acknowledgments

The authors would like to thank Clemens Lechner, Ingrid Schoon and Francesco Avvisati for their valuable reviews of this working paper.

The authors would also like to thank Natalie Foster and Shivi Chandra for their insightful inputs during the writing process, Mykolas Steponavičius for his contributions during early stages of the literature review process, and Sasha Ramirez-Hughes, for his dedicated support in styling and formatting the online data table associated with this Working Paper.

Editing and administrative support was provided by Jessica Bouton and Federico Bolognesi.

The development of the working paper was guided by Andreas Schleicher and Tia Loukkola. This is the second of two working papers on social and emotional skills, produced under the oversight of Noémie Le Donné.

This working paper was developed under the oversight of the OECD's Centre for Educational Research and Innovation (CERI) Governing Board.

Abstract

Social and emotional skills (SES) are important for various life outcomes, such as academic achievement, mental health, job performance or civic engagement. The assessment of these skills in children and adolescents, however, currently relies heavily on the use of self-reported questionnaires. As such, there is an urgent need for more direct measurement approaches of SES, which look at behaviours, actions and choices, in order to diversify the current portfolio of available assessments. The aim of this working paper is, thus, to map and review innovative assessment tools as well as technological approaches, aimed at the direct assessment of SES. Firstly, the paper documents almost 60 different behavioural tools, namely tasks and digital games. These instruments are reviewed according to a set of criteria, including their reliability, construct and ecological validity, and feasibility. Secondly, the paper identifies technological approaches, such as biophysiological measures, virtual reality or different artificial intelligence applications. Many of these technologies have the potential of being transversally integrated into different tasks and game, enriching the quality of SES assessment, albeit bringing new challenges. Lastly, the paper promotes a dialogue between the different types of innovative assessments, identifying comparative strengths and challenges.

Acronyms

AI – Artificial Intelligence

AR – Augmented Reality

EEG – Electroencephalography

EI – Emotional Intelligence

fMRI – functional Magnetic Resonance Imaging

fNIRS – functional Near-Infrared Spectroscopy

HRV – Heart Rate Variability

NPC – Non-Playable Character

PISA – Programme for International Student Assessment

SEL – Social and Emotional Learning

SES – Social and Emotional Skills

SJT – Situational Judgement Test

ToM – Theory of Mind

VR – Virtual Reality

Assessment Tools' acronyms:

FERET – Facial Emotion Recognition and Empathy Test

GERT – Geneva Emotion Recognition Test

GERT-S – Geneva Emotion Recognition Test Short version

IAT – Implicit Association Test

LEAS – Levels of Emotional Awareness Scale

LEAS-C – Levels of Emotional Awareness Scale - Children

MET – Multifaceted Empathy Test

MSCEIT – Mayer–Salovey–Caruso Emotional Intelligence Test Version 2.0

MSCEIT-YV – Mayer–Salovey–Caruso Emotional Intelligence Test–Youth Version, Research Version

PERC (Task) – Persistence, Effort, Resilience, and Challenge-Seeking Task

SAT-MC – Social Attribution Task - Multiple Choice

TTCT – Torrance Test for Creative Thinking

V-SEIP – Video-Social-Emotional Information Processing

VESIP – Virtual Environment for Social Information Processing

Table of contents

FIGURES

TABLES

1. Introduction

Social and emotional skills (SES) play a critical role in education and society, and there is widespread agreement regarding their importance for so many dimensions of our individual and collective lives (Jones and Kahn, 2017[1]). Defined as characteristics that manifest in maximal behaviours rather than typical ones (see Box 1.1), research shows SES are predictive of a wide range of key life outcomes, such as academic achievement and mental health, but also job performance and civic engagement. Critically, SES are also proven teachable through specific educational interventions, which grants schools a pivotal role in their development (Steponavičius, Gress-Wright and Linzarini, 2023[2]). These findings justify an increasing incorporation of SES into the curricula of many education systems around the world, targeting students from kindergarten through the end of secondary school (Cipriano et al., 2023[3]; OECD, 2020[4]).

Box 1.1. Definition of social and emotional skills, as defined by the OECD

Social and emotional skills: Individual characteristics that are:

- expressed in repeatable patterns of thoughts, feelings and behaviours;
- manifested in maximal behaviour, or maximal capabilities, more than typical behaviour (and therefore distinct from personality traits);
- dependent on situational factors (e.g. motivation, task context, fatigue);
- subject to developmental change and genetic predispositions;
- teachable/responsive to intervention;
- predictive of key life outcomes;
- conceptually distinct from foundational cognitive processes (e.g. visual processing, attention, memory retrieval) and academic skills (e.g. literacy, numeracy)

Source: (Steponavičius, Gress-Wright and Linzarini, 2023[2]), "Social and emotional skills: Latest evidence on teachability and impact on life outcome", *OECD Education Working Papers*, No. 304, OECD Publishing, Paris, http://dx.doi.org/10.1787/ba34f086-en

However, many social and emotional learning (SEL) programmes and interventions are implemented without any associated robust assessment methods that measure their effectiveness and their impact on specific SES. A seminal meta-analysis reviewed more than 200 SEL programmes and identified that a significant proportion lacked reliable and valid outcome measurements (Durlak et al., 2011[5]). Even after more than a decade later, a recent survey shows that implementers of SEL programmes state the absence of adequate evaluation methods as the second most relevant challenge (OECD, 2023[6]). Identifying and designing scientifically solid assessment methods for measuring SES and, thus, the impact of SEL interventions is therefore an urgent necessity. Current methods to assess SEL programmes or to measure these skills at an individual level rely heavily on the use of self-reports or teacher-reports (Cipriano et al., 2023[3]). Although offering insightful information, these indirect measurements carry multiple biases and are limited in providing comparable and objective information (Abrahams et al., 2019[7]). Consequently, there is a need to expand and diversify the portfolio of available assessment measures, to include direct measurements for assessing SES, namely looking at behavioural and performance

data. With a wider portfolio of validated direct assessments, it will be possible to identify the most promising ones and, from those, develop a large-scale assessment of SES. Such a tool will be instrumental to assess SEL interventions and compare their efficacy across different cultural contexts, or to understand the most important skills to target in various countries.

The purpose of this paper is to identify and review the existing innovative assessment tools and technological approaches focused on direct measurements of SES. Such assessments and approaches tap on the manifestation of these skills on behaviour, including actions, choices and performances, rather than relying on subjective judgements from students themselves, teachers, parents or peers (Abrahams et al., 2019[7]). Thus, direct measurement tools include behavioural assessments, which comprise tasks and digital games. Additionally, technological approaches that can transversally improve the quality and scope of different assessments tools are also discussed. These include virtual reality (VR) and augmented reality (AR) technologies, biophysiological measures, digital footprints and more advanced artificial intelligent (AI) applications.

Although this paper covers all these instruments and techniques, the limitations and practical constraints of some of them led us to focus our analysis on behavioural assessments, which are presently more promising. We conducted a comprehensive review, mapping almost 60 different behavioural assessments, between tasks and digital games, for which we present descriptive information as well as a comparative analysis based on several validity criteria. Collectively, these innovative assessment tools tend to offer higher levels of immersiveness and engagement, in a variety of interface and gameplay designs (Emihovich, Arrington and Xu, 2019[8]; Fulya Eyupoglu and Nietfeld, 2019[9]). The contextualised social and emotional stimuli these tools allow results in a manifestation of more spontaneous and authentic behaviours in children, beyond what classical reports can offer (Ren, 2019[10]). Such benefits result in increased ecological validity and objectivity of SES assessment and expands the portfolio of current assessment approaches.

The paper is structured in the following way. Section 2 discusses the contrast between the benefits and disadvantages of self-reports and direct measurements for assessing SES and discusses Situational Judgement Tests (SJTs) as a step towards more direct measurement approaches. Section 3 introduces the methodology used to conduct the thorough mapping of the existent behavioural assessment tools and the technological approaches for direct assessment, including the presentation of various criteria used to analyse the different tools. Then, Section 4 presents an ample coverage of innovative behavioural assessment tools, mapping the existent tasks and digital games used to assess SES. Section 5, on the other hand, focuses on technological approaches that can transversally apply to different tasks and games to improve assessment, including biophysiological measures, VR and AR, AI applications, as well as digital footprints and behavioural data from videogames. Finally, Section 6 summarises and discusses the reviewed assessment tools and transversal technological approaches, promoting a dialogue between different types of tools, as to identify the strengths and challenges associated with each of them. In particular, the contrasts between game-based and task-based behavioural assessments are extensively discussed.

2. Expanding beyond indirect assessment

2.1. Comparison between Self-reports and Direct measurements

Self-reported questionnaires offer a wide range of available measurements regarding the breath and specificity of the skills assessed, which results from the easy adaptability of the

wording of the questionnaire items and the many decades of testing and validation. Self-reports present the advantage of allowing children and adolescents to introspectively access much more information, and more nuanced, about themselves than what other respondents can provide, such as parents or teachers (Wigelsworth et al., 2010[11]). This might be particularly beneficial to assess skills that are less dependent on interactions with others, such as Self-awareness, Self-efficacy or Curiosity. However, such an ability to introspect and to develop a self-concept are subject to neurocognitive development, and younger children, when compared to adolescents, are likely less able to provide accurate responses about themselves (Wigelsworth et al., 2010[11]).

Also, it is widely recognised that self-reports suffer from significant levels of biases and limitations, including social desirability bias, immediacy bias, acquiescence bias, memory bias and reference bias. Social desirability bias occurs when children provide answers they consider to be socially desirable. For example, when asked whether they consider themselves hardworking or whether they care about others, children might be tempted to opt for a higher level of agreement with such statements as a response to social pressure or to align with perceived social expectations of themselves or of others (West et al., 2016[12]). The immediacy bias refers to children having an appreciation of their behaviour and their abilities based on recent events rather than a more accurate judgement of cumulative experiences (Wigelsworth et al., 2010[11]). Further, the acquiescence bias is the tendency of some individuals to consistently respond to items in a questionnaire in either a positive (agreement) or negative (disagreement) way, regardless of their actual content. Of note, younger children, when compared to adolescents and even more to adults, and those with lower educational attainment, are much more prone to suffer from this particular bias (Primi et al., 2020[13]).

Researchers also point to possible mismatches between self-reported behaviour and actual behavioural actions, since questionnaires rely on memories for past events, which can obviously suffer from reliability and consistency issues. Moreover, different internal understandings of what constitutes good or poor behaviours, for example in statements "I am good at…", might lead children to under or overestimate their own abilities or traits (Snow et al., 2016[14]). Related to this is the reference bias, regarded as the different implicit standards of comparison that influence individual responses. For example, when confronted with the statement "I am a hardworking student", children must conceptualise that ideal representation of a hardworking person in order to place themselves against that image. Children with high standards, those raised in strict households or those studying in elite schools might have very different practical understandings of what it means to work hard, regarding homework completion, academic performance or hours of study. Such frameworks are, of course, dependent on internal and external pressures and different children will have different frameworks, and different responses, when responding to the same questionnaire, even if their real behaviour is in fact similar (West et al., 2016[12]). It is worth noting that the reference bias can occur at several levels, such as at a country level, for example with different references across different cultures, but also at a classroom level, as students typically assess their skills relatively to those of their classmates. When analysing self-reported questionnaires, this can artificially inflate differences across students who have objectively similar skills.

Lastly, self-reports usually focus on typical behaviours even though questionnaire items can be worded to tap into maximal capabilities (Wigelsworth et al., 2010[11]). The maximal behaviours, which better align with the concept of skill (Steponavičius, Gress-Wright and Linzarini, 2023[2]), benefit from a direct assessment paradigm that complements and expands beyond self-perceived and self-report abilities to focus on actual actions, choices, behaviours and performances of individuals, manifested in practical and contextualised situations. It should be said that, unlike questionnaires, these direct measurements on

behaviours and skills usually require more time and human resources to be produced, administered and scored. Also, for those assessments requiring a correct course of action, the optimal choice often relies on the establishment of consensus from a group of experts, which can result in cultural biases (Wigelsworth et al., 2010[11]).

Direct measurement approaches can rely on different strategies. Behaviour-based assessments constitute one major subset, using SJTs, tasks and games to present test-takers with contextualised social and emotional situations in increasing levels of audiovisual stimulation and ecological validity, to assess their SES in contexts that elicit more authentic behaviours. Rather than asking children whether they consider themselves empathetic, creative or collaborative, these assessments present them with performance tests, hypothetical contexts or real-life situations where they can contextually express their capabilities. In the most complex behavioural approach, assessing students through gamified simulation-based experiences allows them to enact actions, choices and behaviours in an immersive way that is engaging and motivating, potentially alleviating the burden of test anxiety. However, assessing children with stimuli and contexts that are very specific, without presenting a variety of situations, might limit our ability to extrapolate date and generalise conclusions regarding children's behaviours and SES (De Klerk, Veldkamp and Eggen, 2015[15]; Kyllonen and Kell, 2018[16]).

Other direct measurement approaches involve collecting biophysiological measures, which can derive from a variety of sources, such as heart activity, saliva or blood, eye movements and attention, skin conductance as well as multiple manifestations of brain activity, ranging from simple to cutting-edge technology. Biophysiological measurements reflect spontaneous actions which are less prone to conscious alterations than self-reports, and have evolved to show somewhat valid correlations with some psychological measurements (Abrahams et al., 2019[7]). However, biodata is also subject to many variables, so there are still issues regarding how direct and linear extrapolations about psychological states and SES can be made. Similarly, rather than asking people how they feel and think, it becomes possible to analyse their speech, facial expressions, and gestures with unprecedent sophistication, detecting subtle differences which are informative of their internal states (Beyan et al., 2021[17]) (Westera et al., 2019[18]). Various AI applications are now capable of collecting subtle and humanely undecipherable information on individuals, enriching our portfolio of direct data for analysis, surpassing many biases of self-reports. With all the foreseeable progress, there are, of course, many other biases and ethical concerns that arise and that are the centre of societies' current debates and decisions (Tuomi, 2022[19]).

These direct measurement approaches for SES illustrate that the scientific progress aligned with recent technological advances has resulted in the development of more sophisticated and ecologically valid tools to capture more authentic behaviours. These approaches and tools counter many of the biases self-reports carry, albeit bringing new issues and challenges that will require rigorous efforts to be overcome (Abrahams et al., 2019[7]). Nevertheless, there are new paradigms for assessment which diversify the current portfolio and future improvements and technologies will certainly help consolidate their potential. Direct measurements offer, thus, complementary benefits to those of self-reports, allowing these two broad assessment approaches to collectively shed light on the same reality, human behaviour.

2.2. Situational judgement tests

Self-reports ask participants to report on their feelings and intentions, whereas behavioural assessments directly look into people's behaviours, actions and choices and, from those, infer the appropriate SES. SJTs can be considered a self-report as they explicitly ask participants to report on feelings, intentions and decisions, rather than having their practical

behaviour effectively assessed. Nonetheless, SJTs can also be framed as an evolution from traditional self-reports, composed of simple statements and Likert scales, towards a more direct assessment of SES, as they present hypothetical, contextualised, complex and, thus, more realistic social situations. Moreover, in these scenarios, participants are often expected to analyse the behaviours of others and select courses of action other social agents should follow to achieve certain outcomes (Zhuang et al., 2008[20]; Patterson et al., 2012[21]). Therefore, SJTs are presented in this paper as an incremental step towards more direct measurement approaches.

These behavioural tests are widely used in professional settings, namely for recruitment purposes, and, as mentioned above, allow the assessment of more subtle and complex judgement processes than typical self-reported questionnaires (Zhuang et al., 2008[20]; Patterson et al., 2012[21]). Test-takers are usually asked to select the most appropriate choice from a pre-determined pool of options, regarding the correct interpretation of the situation and the most prosocial course of action to overcome a certain problem. Alternative response modalities include ordering the options according to their effectiveness or individually ranking their effectiveness on Likert-type scales (Corstjens, Lievens and Krumm, 2017[22]). Such a description bears similarities with tasks and games (discussed in Section 4) which tackle the skills Perspective-taking and Social problem-solving. However, the judgment of social situations in the more innovative assessment tools either allows for open-ended responses or is integrated into digitalised audiovisual tasks, or even gamified into narrative-driven immersive experiences. As such, the SJTs described and discussed in this section cover more classical, analogic and paper-based assessment tools, where the social scenarios are presented in written vignettes without any audiovisual stimuli. For this reason, and since the aim of this paper is to focus on more innovative assessment tools with higher ecological validity, namely tasks and digital games, only three SJTs are presented in this subsection, to give readers a flavour of their different contents and designs.

These classical SJTs carry an advantage since they have been extensively used and, thus, their reliability and validity metrics are more consolidated, when compared to newly developed innovative tasks and games (Corstjens, Lievens and Krumm, 2017[22]). A clear downside is a certain compromise of ecological validity since the social situations presented are all descriptive, lacking nuances related to speech and tone and even social perception of non-verbal communication, such as facial and body expressions, which other immersive tools can offer. However, it should be noted that recent efforts have attempted to incorporate multimedia elements to SJTs, including 3D animations and motion-capture techniques. These provide increments to the realism and the fidelity of stimuli of the social situations being presented, in a greater alignment with the task- and game-based assessments described later in this paper (Weekley et al., 2015[23]).

The ***Situational Test of Emotional Understanding (STEU)*** describes social and emotional situations, occurring in specific personal life and workplace contexts or presented in a more abstract and context-reduced setting, that can occur to any individual (MacCann and Roberts, 2008[24]). For each scenario, it offers test-takers five possible emotional words asking which the most probable emotion felt in such circumstances. Overall, the 42 different scenarios in *STEU* cover 17 different discrete emotions, which can be considered a measurement for **Perspective-taking** skills. One example of such scenario is: "An unwanted situation becomes less likely or stops altogether. The person involved is most likely to feel: (a) regret, (b) hope, (c) joy, (d) sadness, (e) relief", with the last option being considered the correct answer.

The paired ***Situational Test of Emotional Management (STEM)***, on the other hand, presents brief social and emotional situations and participants are asked which of four options could best manage the emotions of the people in the scenario, leading to a more

positive outcome, which corresponds to **Social problem-solving** skills (MacCann and Roberts, 2008[24]). In this test, the 44 items also cover different workplace and personal-life contexts, while focusing only on three different and socially tense emotions, anger, sadness and fear. An example is: "Lee's workmate fails to deliver an important piece of information on time, causing Lee to fall behind schedule also. What action would be the most effective for Lee? (a) Work harder to compensate, (b) Get angry with the workmate, (c) Explain the urgency of the situation to the workmate, (d) Never rely on that workmate again", with C being established as the most prosocial course of action. Both these tests have been extensively validated in adult and adolescent populations, in various countries, with acceptable to good internal consistency metrics, as well as solid convergent validity with other psychological measurements on similar constructs, and consequential validity given by the correlation with higher levels of self-reported psychological well-being and life satisfaction, as well as higher academic achievement (MacCann and Roberts, 2008[24]; da Motta et al., 2021[25]; Dirzyte et al., 2021[26]; Lea et al., 2023[27]).

Another SJT is the ***Social Relationship Competence–Ability Measure (SRC-AM)***, which looks at friendship contexts, presenting participants with several dilemmas, where a hypothetical friend has done a certain act that could be negatively impactful for the relationship (Persich, Krishnakumar and Robinson, 2020[28]). For each scenario, participants are presented with four different courses of action, which can lead to positive or negative outcomes, and each possible action to address the conflict must be rated according to its perceived effectiveness. The scenarios presented cover multiple interpersonal dimensions that inhabit real-life friendships, such as the ability to form and maintain high-quality and long-lasting friendships, to express concern and provide explicit social support, and to respond to tension and conflict in constructive ways. For all this, the social relationship competence targeted in this test covers the skills **Empathy** and **Social problem-solving**. As an example, one scenario is: "Randy notices that his friend seems to be ignoring him. Rate the effectiveness of the following ways that Randy could deal with the situation: a) Confront the friend, b) Ask his friend why he is being ignored, c) Ignore the friend in return and d) Convince himself that the friend is not ignoring him on purpose".

The test shows acceptable levels of internal consistency as compared to other SJTs (for 10 scenarios requiring 40 ratings, Cronbach's alpha coefficient = 0.7), and researchers have found that the performance in the test positively correlated with self-reports on commitment to friendships, perception of mutual care and support, and quality and intimacy level of said relationships (Persich, Krishnakumar and Robinson, 2020[28]). The test also correlated with measures of conflict-resolution inclination, with high-performance individuals being better at suppressing negative impulses and being more direct and open when approaching friends to overcome tensions (Robinson, Persich and Irvin, 2022[29]). Additional validation was obtained from reports by informants, close friends of those being tested, which found correlations between participants' real-life behaviours and attitudes towards the respective friendships and their performance in the *SCR-AM* (Persich, Krishnakumar and Robinson, 2020[28]). Moreover, those who did well in the test were also found to benefit from better psychological well-being and higher levels of positive emotion and life satisfaction, and to manifest widespread prosocial behaviours, beyond relationship contexts (Robinson, Persich and Irvin, 2022[29]).

3. Review of direct assessment tools and innovative approaches

This section presents the methodology used to review the literature and search for existing innovative direct assessment tools and technological approaches to measure SES. This section also details the criteria used to describe and review each tool. The comprehensive

list of tools reviewed is presented in an online table (Behavioural Assessment Tools for SES), and a large part of these instruments are described in section 4 and section 5.

Since the purpose of this work is to identify and map new tools and assessment approaches, self-report or other-report questionnaires were not reviewed, and SJTs were not reviewed consistently (see section 2.2). The review focuses extensively on Behaviour-based assessment tools, but also covers transversal technological approaches for innovative assessments, including Biophysiological measurements, VR and AR, AI applications, and analysis of Digital footprints and behavioural measures from videogames.

These broad categories have been chosen for practical reasons/readability, and we acknowledge that some tools and approaches contain elements that would fit in several of them. Indeed, the categorisation presented here should not be considered as a formal taxonomy regarding the organisation of assessment types. Also, a vast literature was identified for each of these assessment approaches, along with significant differences regarding their current scientific reliability and validity and practical applicability. Therefore, in section 4, we opted to present in detail a number of tools within the behavioural-based assessments, while section 5 gives a larger overview of the state of the art for the remaining technological approaches.

3.1. Review methodology

Firstly, to map the state of the art of the field and identify research trends, we conducted a general search, using Google Scholar (https://scholar.google.com/) as our primary academic search engine. We looked for combinations of the terms "social and emotional skills" OR "life skills" OR "social-emotional competencies" with general types of assessment or approach, such as "behaviour-based assessments", "biophysiological data" (including "heart activity", "cortisol", "eye tracking"), "virtual reality", "augmented reality", "artificial intelligence", "digital footprint" and "behavioural residue". Initially, we looked for these combinations in the articles' titles and abstracts. We identified and reviewed relevant articles relating SES to each assessment tool type and each approach.

Given the large number of results in favour of behaviour-based assessments, we conducted a second round of search, more refined and systematic, to identify behaviour-based assessments of all individual SES, using Google Scholar. The search strings were designed by combining two components, the names of the individual SES followed by iterations of the different subtypes of assessments. The first string component, the search terms for the individual SES, derived primarily from the framework of the OECD Survey on Social and Emotional Skills (Steponavičius, Gress-Wright and Linzarini, 2023[2]), to which synonyms were added from the Concept notes of Future of Education and Skills 2030 (OECD, Forthcoming[30]) and from Harvard's ExploreSEL online comparative tool (http://exploresel.gse.harvard.edu/compare-terms/) (for the full list of search keywords, see Annex A: Search terms for SES). The second search string component, the assessment type, started with a first search round involving a general reference to "behavioural assessment" or "behaviour-based assessment", after which searching iterations were run for specific behavioural approaches using "game", "game-based assessment", "task" and "task-based assessment". As an example, the search strings for "Self-control" were all possible combinations of: Self-control / self-discipline + behavioural assessment / behaviour-based assessment / game / game-based assessment / task / task-based assessment.

Aside from Google Scholar, additional behaviour-based assessments were identified through the SEL assessment compendia RAND Assessment Finder (https://www.rand.org/education-and-labor/projects/assessments/tool.html) and using a snowballing approach.

Given the focus on innovative assessments underlying this paper, a filter was applied to locate articles published in 2015 or afterwards. Of note, tools developed before 2015 were included and reviewed only when more recent scientific publications used, and sometimes even adapted, those same tools. The 2015 cutoff is a way to ensure that the tools included in this review have still been used in relatively recent years and, therefore, still relevant. Titles and abstracts were screened in order to assure that the papers introduced and tested a new behavioural assessment tool, rather than relying solely on indirect questionnaire-based reports. Articles that included assessment tools specifically designed for clinical purposes, targeting neurodivergent populations or populations that suffer from neuropsychiatric conditions, such as autism, depression or attention deficit hyperactivity disorder, were excluded, unless the tool was also validated on a neurotypical control population. Articles that included assessment tools only applicable to preschoolers (< 6 years old) were excluded. Also, we focused primarily on original research papers, although review articles that included references to original assessment tools were also initially kept for further extraction. Finally, given that the purpose of this work was to map new assessment tools in the literature, we excluded task- or game-based behavioural assessments already developed and validated by the OECD. They include the Xandar Task, on Collaborative problem solving, used in PISA[1] 2015 (OECD, 2017[31]), or the Creative Thinking Assessment, applied in PISA 2022 (OECD, 2022[32]).

The final selection of skills and keywords was driven by the need to cover the literature on direct assessment tools of SES as thoroughly as possible. The objective is to provide a broad mapping of the state of the art on a large number of skills. This work does not aim to impose a taxonomy of SES. Debates regarding the classification of certain constructs, such as whether they should be classified as skills or traits, or whether they fall under the category of SES, are briefly mentioned in the relevant sections. However, a detailed analysis of these debates is beyond the scope of this work. To ensure that the constructs reviewed were consistent with the broad OECD definition of SES (see Glossary), we defined each skill by incorporating contributions from the OECD framework (Steponavičius, Gress-Wright and Linzarini, 2023[2]), along with other definitions from the literature. Additionally, we included the operational definitions of each skill used by the developers of the assessment tools. For this reason, we established an equivalence between the construct evaluated by the tool and our terminology, based on the content of their respective definitions. For example, if the purpose of a tool is to assess "Impulsivity Control", we code that skill as "Self-control", after verifying the definition provided in the publication and the construct being assessed by the tool effectively align the OECD framework.

We identified and reviewed 57 behavioural assessment tools and paradigms, including, 34 task-based assessments and 20 game-based assessments. We also identified and included information from several articles that more extensively tested existing assessment tools, developed either by the same research group or other groups. These articles expanded the applicability of the respective tools, by modifying or modernising their design or by further validating them with different populations. The full list of assessment tools reviewed can be consulted in Annex B: Full list of identified behavioural assessment tools, where DOIs links for the respective articles are provided.

3.2. Criteria for review

The tools were described and reviewed according to a set of criteria. While these criteria and their associated ratings were originally defined to allow an informed comparison of

[1] OECD's PISA: Programme for International Student Assessment

these tools in the context of the development of an international large-scale assessment of SES, they can also be used to better understand the current state of the assessment field for the techniques and for the skills reviewed.

These criteria are intended to help illustrate the strengths and limitations of the assessment tools reviewed. They are not intended to be a formal comparison of tools or an endorsement of any tool. The ratings for the criteria are qualitative in nature, not quantitative. They were created relatively to the initial purpose of this review (that is, to select the most promising tools for the creation of a large-scale international direct assessment of SES). The criteria and the ratings are defined in more detail hereunder, along with the full list of descriptors, available in the online table (Behavioural Assessment Tools for SES). The online table allows for the filtering and ranking of the tools based on different characteristics reflecting different priorities, be it age coverage, quality of validation process or type of skills covered, for example. We encourage readers to prioritise which criteria are most important for use in their particular context.

3.2.1. Descriptors

On top of the criteria used to compare the tools, available descriptors in the online table include:

- Name of the tool;
- Brief description of the tool and variables used;
- Source(s);
- Type of assessment tool (task, task paradigm, game, mixed battery);
- OECD skill equivalent (Based on the definition of the skill(s) given in the publication, each tool is associated with one or several OECD skill equivalents from the list of skills reviewed - see Annex A: Search terms for SES).
- Original skill(s) and respective definition(s);
- Target age range information (as the content of the measure must be developmentally appropriate, it is important that emotional/social stimulus material is properly validated on the target population (Mehlsen et al., 2019[33]). Therefore, the target age group for each tool is indicated as: Children (6-12); Adolescents (12-18); Adults (>18). Finer age information is provided when available);
- Countries and languages (in which tools have been tested).

3.2.2. Ecological validity

Ecological validity is understood as a measure of the representativeness of the instrument; in other words, the correspondence between the form and context of the assessment and the situations in natural contexts (Burguess et al., 2006[34]). It is particularly important in the context of this review of very heterogenous tools, for which there are strong differences in the presentation of content. The ecological validity of each tool is rated according to several sub-criteria:

- Openness of action: Restricted freedom of exploration, with pre-determined order of activities / Possible exploration of different activities (0/1)
- Type of response: Fixed choice (e.g., based on multiple choices, or pressing button quickly) / Open-ended (includes writing full answers) OR free choice (e.g., choose whether to move an avatar, choose to switch between tasks) (0/1)

- Richness of stimuli: Only written content with no audiovisual stimuli / Inclusion of either visual (e.g., illustrations, facial pictures) or audio (e.g., vocalisations of emotions, voice recording for characters in the story) stimuli / Inclusion of multisensory stimuli, audio and visual (0/1/2)
- Adaptivity: The assessment has a fixed script / The assessment script changes according to the player's actions or choices (0/1)

3.2.3. Reliability and validity

Reliability refers to whether an instrument measures the skill in a consistent way across respondents, over time, or across raters. Validity refers to whether an instrument measures what it is intended to measure, and whether the inferences drawn from an instrument are appropriate (Cox, Foster and Bamat, 2019[35]). Each tool was screened for reliability (internal consistency and test-retest reliability), construct validity (generalisability and convergent validity), criterion validity (concurrent or predictive validity), and fairness (Socioeconomic, ethnic, and/or cross-cultural group comparisons).

As a limitation of this review, we acknowledge that information on the reliability, validity and fairness measures of each tool is limited and should be taken with caution. The ratings are given for information only, to give an overall qualitative view of each tool. Due to the mapping purpose of this review, a thorough analysis of these estimates is beyond the scope of this work. In addition, the wide range of approaches and paradigms reviewed makes direct quantitative comparison of estimates difficult. For example, some measures of reliability and validity estimates may not be applicable to certain paradigm designs. And even when these measures are used, there are no absolute thresholds for reliability and validity estimates, so acceptable levels may vary depending on the type of assessment paradigm.

- Internal consistency – the degree to which different test items that probe the same construct produce similar results: The tool has not been tested for internal consistency or no sufficiently solid internal consistency found / The tool has been tested for internal consistency (but low estimates of internal consistency are flagged)[2] (0/1).
- Test-retest reliability – whether the results of two consecutive administrations of the same test to a group of individuals are highly correlated: The tool has not been tested for test-retest reliability / The tool has been tested for test-retest reliability (but low estimates of test-retest reliability are flagged) (0/1).
- Construct validity – whether the score from the instrument correlates with scores from different modes of measurement (e.g., self-reports) of the same skill (generalizability) or with scores from other similar instruments measuring similar skills (convergent validity) (Cox, Foster and Bamat, 2019[35]). Due to the lack of a gold standard, there is no clear consensus in the literature as to whether or not the validity of behavioural or biophysiological measures should be estimated in relation to self-report measures or measures from other assessment approaches (Degner and Wentura, 2008[36]; Mehlsen et al., 2019[33]): The assessment tool has not been

[2] By convention, a Cronbach alpha (α) of .65–.80 is often considered "adequate" for a scale used in human dimensions research (Vaske, Beaman and Sponarski, 2016[218]). Other internal consistency measurements include for example split-half methods.

compared with at least one other established measurement for the same construct / The assessment tool has been compared with at least one other established measurement for the same construct (but non-significant correlations are flagged) (0/1).

As a limitation, it should be noted that the review of construct validity did not include divergent/discriminant validity.

- Criterion validity – whether the score from the instrument predicts current (concurrent validity) or future (predictive validity) real-life performance or other external outcomes of interest (e.g., academic success, reported prosocial/aggressive behaviours, subjective well-being three years later): Behavioural results from the assessment tool have not been correlated with external outcomes / Behavioural results from the assessment tool correlate with external outcomes (0/1)
- Fairness – whether an instrument is not biased against specific socioeconomic or cultural or ethnical subgroups of individuals. Information that could support this component of validity includes statistical tests showing that scores from the measure function similarly across all subgroups (Cox, Foster and Bamat, 2019[35]): The assessment tool has not been tested for differences between diverse socioeconomic or cultural or ethnical subgroups / The assessment tool has been tested for differences between diverse socioeconomic or cultural or ethnical subgroups, at a country or international level (0/1).

  As a limitation, it should be noted that Fairness did not include a review of measurement invariance analyses, nor gender differences.

3.2.4. Feasibility, costs, and licence

This criterion encompasses information on practical aspects of the tool, related to:

- Duration – mean time to perform the assessment: More than 30 minutes per skill / Less than 30 minutes per skill (0/1)
- Administration: Staff is required to administer the assessment / Staff is not required to administer the assessment (0/1)
- Grading: Grading of participants' responses is not automatic and requires an evaluator / Automatic grading of responses (0/1)

Information on availability is reported when available (including equipment needed, such as headset for VR, or license). Since the vast majority of the tools included are digital, the review assumes minimal requirements: hardware devices; internet connection.

4. Behaviour-based assessment tools

As an alternative to asking a student or teacher to report on behaviour, it is possible to observe behaviour through various tasks and games, where students make choices and actions from which we can infer aspects of their behaviour and, by extension, their SES. A behavioural task is essentially a situation that has been designed to elicit meaningful differences in behaviour of a certain kind. Observing students in the identical contrived situation eliminates the possible confound of variation in the base rates of certain types of situations (Galla and Duckworth, 2015[37]).

Behaviour-based assessments include tasks as well as digital games, with different levels of complexity, richness of audiovisual stimuli, immersiveness, interactivity, and,

ultimately, ecological validity. The next subsections will introduce a great variety of assessment tools for various SES, with a description of their rationale, design and gameplay mechanics, punctuated by qualitative notes regarding their validity and reliability psychometrics. While acknowledging it might not constitute a fully comprehensive review, this section gives a solid overview of the available tools and the current state of the scientific literature on this topic. Particularly, for task-based assessments, there is a broad illustration of assessment tools for each of the SES within the OECD framework (Steponavičius, Gress-Wright and Linzarini, 2023[2]), whereas for game-based assessments the review goes into more detail only for a few selected games.

For a clear understanding of their distribution across the SES, the tools identified for each individual skill can be visualised in Table 4.1 below (alternatively, the full list of reviewed assessment tools can also be consulted in Annex B: Full list of identified behavioural assessment tools). The comprehensive review and ranking of all the assessment tools, based on the criteria defined above, can be found in the online table (Behavioural Assessment Tools for SES).

Table 4.1. List of assessment tools per skill

	Collaboration			Open-mindedness		
	Empathy	**Trust**	**Co-operation**	**Tolerance**	**Curiosity**	**Creativity**
Games	• Hall of Heroes • Zoo U		• Circuit Runner • ENACT • Hall of Heroes • Physics Playground • Zoo U		• Questions Worlds	• Physics Playground
Tasks	• Facial Emotion Recognition and Empathy Test • Multifaceted Empathy Test	• Pizzagame • Prisoner's Dilemma • Public Goods Game • Trust Game	• Intergroup Prisoner's Dilemma • Pizzagame • Prisoner's dilemma • Public Goods Game • Trust Game	• Extrinsic Affective Simon Task • Intergroup Prisoner's Dilemma	• Faculty Game	• CREA • Divergent Thinking Task • Minecraft Task • Torrance Test of Creative Thinking
Mixed batteries						

	Task performance				Emotion regulation			Engaging with others		
	Self-control	**Responsibility**	**Persistence**	**Achievement Motivation**	**Stress Resistance**	**Emotional Control**	**Optimism**	**Energy**	**Assertiveness**	**Sociability**
Games	• Hall of Heroes • Rumble's Quest • Zoo U		• Physics Playground	• Posterlet	• Simulation game	• Hall of Heroes • Rumble's Quest • Zoo U			• ENACT	• Game for Social Anxiety • Hall of Heroes • Vox Populi • Zoo U
Tasks	• Academic Diligence Task • EMOTICOM • Emo. Go/Nogo Task • Emo. Stroop Task		• Mirror Tracing Frustration Task	• EMOTICOM		• Beach Balls Task • Laboratory coping and emotion regulation task	• Future Expectations Task			
Mixed batteries	• SELweb EE					• MSCEIT • SELweb LE				

	Supplementary skills					
	Metacognition / Self-Awareness	**Self-efficacy / Self-esteem**	**Critical thinking**	**Perspective-taking**	**Social problem-solving**	**Grit**
Games	• Crystal Island	• Athenea • Virtual Environment for Social Information Processing	• Noah Kingdom • Seaball	• Emodiscovery • Virtual Environment for Social Information Processing • Vox Populi	• Emodiscovery • Virtual Environment for Social Information Processing	
Tasks	• Levels of Emotional Awareness Scale - Children			• Assessment of Social Perspective-taking Performance • Combined Stories Test • EMOTICOM • Emotion Recognition Index • Emotional Literacy Test with Hypothetical Scenarios • Facial Emotion Recognition and Empathy Test • Faux Pas Recognition Test • Hinting Task • Interpersonal Perception Task • Levels of Emotional Awareness Scale – Children • MSCEIT • Multifaceted Empathy Test • Reading the Mind in the Eyes Test - Child Version • Short version of the Geneva Emotion Recognition Test • Social Attribution Task-Multiple Choice • Video-Social-Emotional Information Processing	• Video-Social-Emotional Information Processing	• Academic Diligence Task • The Persistence, Effort, Resilience, and Challenge-Seeking Task
Mixed batteries	• MSCEIT			• MSCEIT • SELweb EE • SELweb LE	• MSCEIT • SELweb EE • SELweb LE	

4.1. Tasks

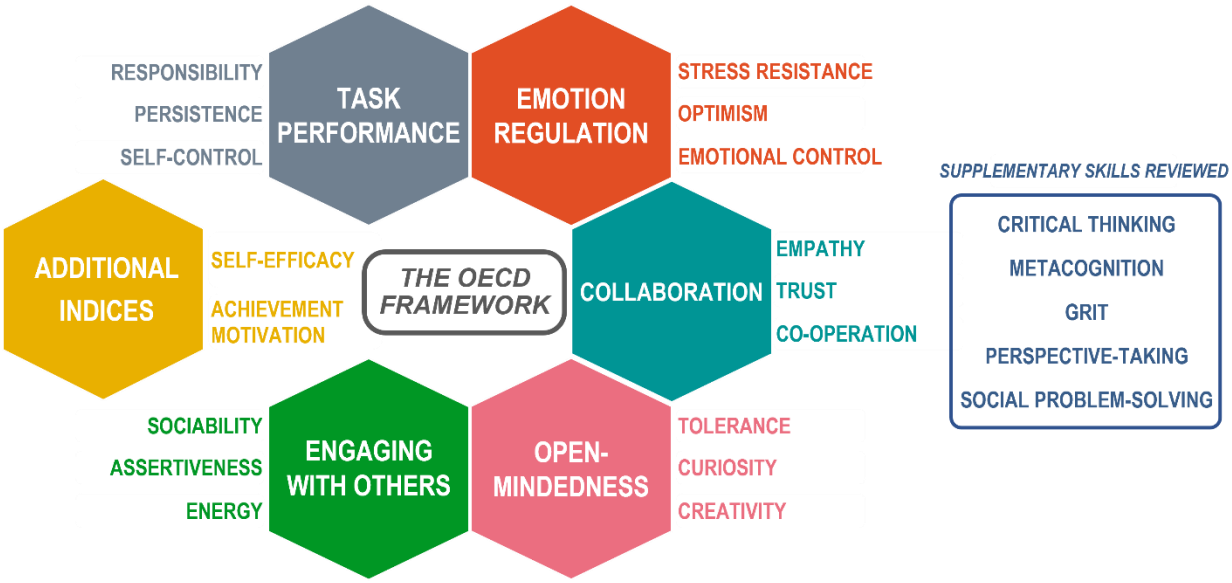
Task-based assessments measure specific SES using all sorts of quantitative and qualitative behavioural data, ranging from nature of choices to number of responses to reaction times. But unlike game-based assessments, task-based assessments present individuals with direct instructions that are not contextually embedded in interactive storytelling.

The use of task-based assessments to measure SES has the clear advantage of providing observable, quantifiable behaviours that reflect an individual's ability to perform in specific situations. These tasks are designed to elicit meaningful differences in behaviour under controlled conditions, thus providing a direct measure of an individual's ability in a particular domain without relying on their subjective judgements, thus avoiding the challenge of accurately self-reporting complex internal states. Task-based assessments are also advantageous because they limit common pitfalls of self-report measures, such as social desirability bias and reference bias. Finally, task-based assessments can be more sensitive to subtle changes over time compared to self-reports, making them valuable for tracking progress and assessing the impact of SEL interventions (Abrahams et al., 2019[7]; Duckworth and Yeager, 2015[38]).

On the other hand, task-based assessments also have significant limitations (Abrahams et al., 2019[7]; Duckworth and Yeager, 2015[38]). Compared with paper-and-pencil questionnaires, they can be logistically difficult to administer and require carefully controlled conditions that may not reflect a student's typical environment. This artificiality may limit the generalisability of results to real-world settings, where students may use strategies to avoid or cope with challenging situations that are not available in a test setting. Furthermore, tasks may suffer from 'task impurity', where the behaviour observed is influenced by factors irrelevant to the construct being measured, such as a student's physical coordination, familiarity with the test format (e.g., digital literacy for computer-based assessments), or more general cognitive functioning. In addition, these tasks are susceptible to practice effects, where repeated exposure to the task affects performance independently of the trait being assessed. Finally, reliance on a single or small set of tasks to represent complex and multifaceted constructs can lead to oversimplification, overlooking the variability and richness of students' SES as manifested in different contexts and situations.

For the sake of clarity, due to the large number of tasks reviewed, this section is organised by presenting the available assessment tools according to the five OECD domains of SES (Kankaraš and Suarez-Alvarez, 2019[39]), and then per skill (see Figure 4.1). Skills for which strong similarities in assessment tools were found are discussed together. A sixth category of skills, referred to as 'Self-reflection', includes self-reflective skills that were (see Annex A: Search terms for SES) for the description of the domains and the full list of skills and keywords reviewed). Importantly, this organisation should not be considered a proposed taxonomy for SES, which falls outside the scope of this work.

Figure 4.1. The OECD framework and supplementary skills reviewed



4.1.1. Task performance

Self-control

Self-control, defined as the ability "to avoid distractions and sudden impulses and focus attention on the current task in order to achieve personal goal" (Chernyshenko, Kankaraš and Drasgow, 2018[40]; Kankaraš and Suarez-Alvarez, 2019[39]), is closely linked to Attention control and Inhibitory control, representing the cognitive aspects of Self-regulation (http://exploresel.gse.harvard.edu/compare-terms/, accessed 10 January 2024). In cognitive psychology and neuropsychology, Inhibitory control is an executive function described as the ability to control our behaviour, emotions, and cognitions in order to adapt to our natural and social environment (Musek, 2017[41]). However, the relationship between Self-control, Self-regulation, Inhibitory control, and other executive functions is not uniformly defined, with some disagreement in the literature (Steponavičius, Gress-Wright and Linzarini, 2023[2]). Inhibitory control encompasses a broader range of cognitive processes, not limited to SES. To distinguish SES from foundational cognitive processes (see the OECD definition of SES in see Glossary), tasks reviewed for evaluating Self-control include emotional or social aspects, ensuring they assess more than just cognitive abilities.

Most of the "emotional" tasks for Self-control identified in this review are modified versions of traditional neuropsychological tests that evaluate Inhibitory control (Bland et al., 2016[42]). Standard neuropsychological tests measuring Inhibitory control through motor or response inhibition include the ***Stroop Task, Go/No-Go Task, Simon Task, Flanker Task, Antisaccade Tasks*** and ***Stop-Signal Tasks***. These tasks focus on the individual's ability to override natural responses for goal-oriented behaviours. In these foundational paradigms, Inhibitory control is gauged by comparing the response time or quality of response to relevant stimuli versus the reaction time to irrelevant stimuli, essentially measuring the delay in automatic response to a stimulus containing both relevant and irrelevant information. To infuse an emotional dimension, these seminal paradigms are modified by substituting neutral stimuli with emotionally charged ones, like words signifying emotions or emotionally charged terms, such as 'war' or 'holidays', or visual stimuli like expressive faces or images eliciting strong emotions, like disgust or sadness

(see for example Figure 4.2). As in the original tasks, in these emotionally adapted tasks, Inhibitory motor control is gauged by measuring the reaction time difference in responding to relevant versus irrelevant stimuli (Bland et al., 2016[42]). While these paradigms are relatively simple, and have been thoroughly tested, validated and used to assess very diverse populations, it is debatable whether these tests actually measure SES, or rather more basic, fundamental cognitive processes that may be constituents/components of the SES but not the full SES themselves.

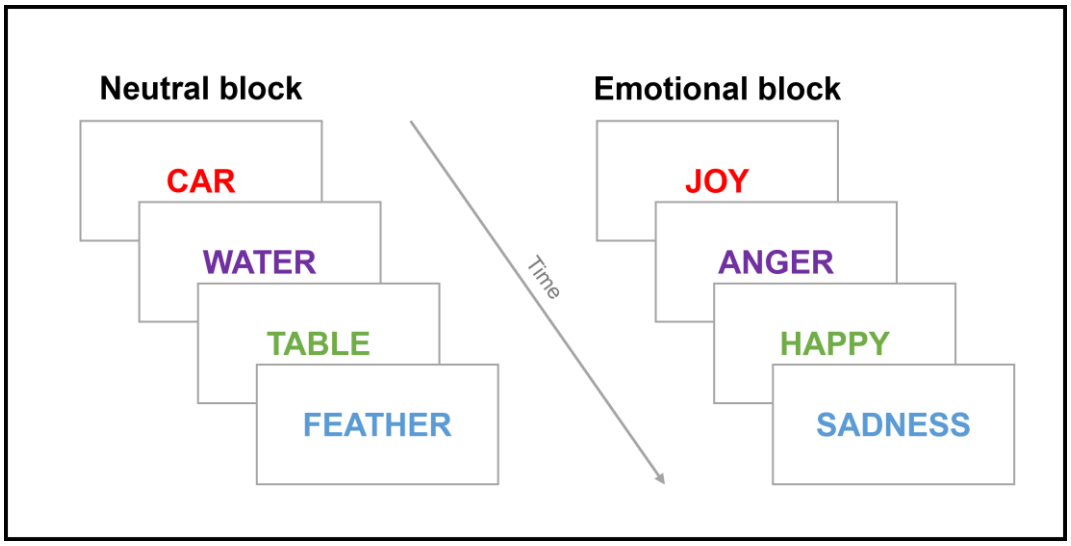
Another aspect of Self-control relates to the ability to delay gratification, which is measured in neuropsychology by another set of Inhibitory control paradigms, the ***Delay of Gratification Tasks***. In these tasks, participants have to choose between small but short-term rewards or larger rewards to be received later.

Reflecting these two aspects of Self-control, the ***EMOTICOM*** battery, made of 16 neuropsychological tasks destined to assess various facets of four core domains of affective cognition (Emotion processing, Motivation and reward, Impulsivity, and Social cognition) (Bland et al., 2016[42]), distinguishes between two types of Impulsivity measures (Impulsivity being very similar to Self-control as defined by the OECD). It proposes to measure Waiting Impulsivity, described as a measure of action inhibition or motor inhibitory control, and Delay and Probability Discounting, a measure of preference for immediate gratification or ability to delay gratification. Waiting Impulsivity is measured through the *Four-Choice Serial Reaction Time Task*. This task assesses visual attention, and the ability to respond to unpredictable targets while inhibiting automatic motor response. Participants have to indicate as fast as possible, from four choices, the box in which a target symbol has briefly appeared but withhold their answer for other non-target symbols. Delay and Probability Discounting is measured through a *Discounting task*. In such task, participants choose between smaller immediate rewards and larger delayed ones, such as deciding between $5 now or $10 in a week. These choices help measure the discount rate, which reflects how much a person devalues future rewards based on the wait time. High discount rates show a preference for immediate gratification, suggesting impulsivity, while lower rates indicate patience and a tendency towards future-oriented decisions. Finally, the *EMOTICOM* battery also comprises an emotional version of the *Go/No-Go Task*. While the *Go/No-Go Task* is traditionally a seminal task for assessing Inhibitory control, it is used to assess Emotion recognition in the *EMOTICOM* battery. The *Face Affective Go/No-Go Task* involves presenting participants with a series of faces displaying different emotions (e.g., happy, sad, angry) and instructing them to respond (typically by pressing a button) only to faces showing a specific target emotion (the "Go" condition) while refraining from responding to faces showing non-target emotions (the "No-Go" condition). Importantly, although the *EMOTICOM* battery has the advantage of exploring different facets of skills using different tasks, it currently lacks robust validation of internal consistency and construct and criterion validity. It has been however tested with a sample of participants with ethnic characteristics representative of the UK demographics.

Another neuropsychological test adapted into an emotional variant is the ***Emotional Stroop task*** (Aïte et al., 2018[43]; Bouhours et al., 2021[44]). This test exists in various adaptations and explores how emotional content affects a person's reaction time. In the simplest, colour-naming version of the task, participants are shown words in different colours and asked to name the writing colour, rather than reading the word itself (see Figure 4.2). Some of these words are emotionally charged (e.g., "anger", "joy"), while others are neutral. Participants take longer to name the colour of the ink for words with emotional content compared to neutral words. This delay is thought to occur because the emotional significance of the word captures the participant's attention, interfering with the task of colour naming. Other versions of this paradigm include for example pictures of expressive faces. The *Emotional Stroop Task* is used to study attentional biases towards emotional information, which can

vary among individuals and be influenced by cognitive abilities, age, gender, or psychological conditions such as anxiety or depression.

Figure 4.2. Emotional Stroop Task



Note: In one version of the Emotional Stroop task, participants are asked to respond to the writing colour of a series of items presented without considering the meaning of these words. Performance is measured as the difference in response times between a block of neutral items (left) and a block of emotional items (right).

Responsibility

No task assessing **Responsibility** defined as the desire and ability to *"follow through with promises to others"* (Kankaraš and Suarez-Alvarez, 2019[39]) were identified.

Persistence, Achievement motivation, and Grit

Very little research was found on direct measurements of **Persistence**. Defined as the ability to persevere in tasks and activities until they get done (Chernyshenko, Kankaraš and Drasgow, 2018[40]; Kankaraš and Suarez-Alvarez, 2019[39]), Persistence is a very task-specific skill. As such, it is for example tested by the ***Mirror Tracing Frustration Task*** (Meindl et al., 2019[45]) as the proportion of time that participants allocate between a difficult and frustrating task (which requires tracing the outline of a shape by looking at its reflection through a physical mirror) and a "distractor", more enjoyable, task (looking at entertaining videos and playing games). Despite its short duration (the test lasts 5 minutes), this measure was correlated with self-reported measures of similar constructs, including Self-control and Grit, as well as real-world academic achievement outcomes (Zamarro et al., 2020[46]) (Meindl et al., 2019[45]). Internal consistency estimates of the measure are also reported. More generally, Persistence can also be inferred from process data or paradata (e.g., time spent on a task, or total number of actions) in assessments that target completely different constructs. Therefore, Persistence measures could also be built from some other behaviour assessment tasks presented in this review.

Achievement motivation is a skill conceptually related to Persistence. Achievement motivation is understood as the ability to set high standards for oneself and work hard to meet them by putting in consistent effort and being highly productive (Kankaraš and Suarez-Alvarez, 2019[39]). The ***EMOTICOM*** battery (Bland et al., 2016[42]) proposes an adapted *Monetary Incentive Reward Task* and an adapted *Progressive Ratio Task* to assess

Incentive motivation, a facet of Motivation and reward. Incentive motivation tests measure how much effort an individual is prepared to exert to gain reward. The monetary incentive reward task assesses effort to avoid punishment and gain reward, and the progressive ratio task identifies the maximum effort that a participant will expend in order to receive a reward (Bland et al., 2016[42]). In terms of reliability and validity, these two neuropsychological tasks are initially destined to assess clinical conditions of patients with neuropsychiatric disorders. The tasks of the *EMOTICOM* battery (Bland et al., 2016[42]) were also tested on volunteers with no self-reported previous or current psychiatric disorders, but as a control condition. There is the need to assess the discriminant power of such tasks in the broader population. Preliminary data presented in the publication showed low test/retest reliability of the monetary incentive task.

A third highly related concept is **Grit**. In the literature, Grit is defined as "perseverance and passion for long-term goals [and] not just resilience in the face of failure, but also having deep commitments that [one] remain[s] loyal to over many years" (Duckworth et al., 2007[47]). In contrast to Self-control, Grit is distinguished by its emphasis on effort and interest sustained over months and years rather than minutes and hours (Galla et al., 2014[48]). Grit is commonly defined as a complex trait that encompasses two key components: "Perseverance of effort" and "Consistency of interest" (Credé, Tynan and Harms, 2017[49]). While "Perseverance of effort" seems to relate to the Persistence skill of the OECD framework, "Consistency of interest" seems to be aligned with the Achievement motivation skill of the OECD framework (Steponavičius, Gress-Wright and Linzarini, 2023[2]). This two-dimensional structure of Grit is evident in the design of one of the main self-assessment tools used to evaluate it: the *Grit Scale* (Duckworth et al., 2007[47]) and the *Short Grit Scale* (Duckworth and Quinn, 2009[50]). These aspects are also present in assessments using compound measures for evaluating Grit. Sutter and colleagues (Sutter, Untertrifaller and Zoller, 2022[51]), for example, combine three different choices of children on (i) perseverance as an ability to work hard, (ii) the willingness to challenge oneself by choosing voluntarily a more difficult task when given the choice, and (iii) the likelihood to follow through with that task until completion.

However, the literature is divided regarding the usefulness of conceptualising Grit as a higher-order construct characterised by two lower-order facets. Some research underscores: a lack of strong methodological evidence for a higher-order factor structure in Grit; a general loss in predictive power when combining perseverance and consistency scores; Grit's modest correlation with academic performance, which is weaker compared to established predictors like cognitive ability or study habits; and finally, Grit's strong overlap with Conscientiousness and Self-control. This suggests that Grit might merely be a rebranding of existing constructs rather than a unique predictor of performance (Credé, Tynan and Harms, 2017[49]). A recent systematic review (Fernández-Martín, Arco-Tirado and Hervás-Torres, 2020[52]) indicates that research on Grit needs more systematic and rigorous methodologies, and although there is evidence supporting Grit as a predictor of success in education, profession, and personal life (Lechner, Danner and Rammstedt, 2019[53]), the limited number and/or quality of studies hinder definitive causal conclusions. The review suggests the need for stronger, more comparable evidence in the field of Grit research.

The ***Academic Diligence Task*** was created to assess Academic diligence defined as "working assiduously on academic tasks which are beneficial in the long-run but tedious in the moment, especially in comparison to more enjoyable, less effortful diversions" (Galla et al., 2014[48]). The authors of the task relate this construct to **Grit** and **Self-control**. The *Academic Diligence Task* is a behavioural measure designed to assess the allocation of time and effort between a beneficial yet monotonous math skill-building activity and engaging distractions such as video games or YouTube clips (e.g., music videos, movie trailers). This

20-minute task, aimed at mirroring real-world conflicts between academic work and digital distractions, presents participants with a split-screen interface. They can choose to solve single-digit subtraction problems, enhancing math skills, or divert to leisure activities like watching short YouTube videos or playing Tetris. Although participants can switch between tasks, they are limited to one activity at a time. The task emphasises the utility of practicing basic math for improving problem-solving skills, while also allowing the freedom to engage in entertainment, thereby presenting a realistic scenario of academic diligence versus digital distractions. It is very similar to the *Mirror Tracing Frustration Task*, except the difficult task is here presented as more constructive and beneficial, rather than a pure challenge (Meindl et al., 2019[45]). The *Academic Diligence Task* has been tested for sufficient internal consistency, for construct validity with other self-control and grit measures (although the strength of the correlations with the task's performance scores were small), for criterion validity and for fairness (Galla et al., 2014[48]).

The ***Persistence, Effort, Resilience, and Challenge-Seeking (PERC) Task*** was developed to assess Mastery behaviours in primary, middle and secondary school children (Porter et al., 2020[54]). Mastery behaviour is defined as: "Seeking out challenging tasks and continuing to work on them despite difficulties" (Dweck and Leggett, 1988, p. 256[55]). Based on work from developmental psychology, four key components of mastery behaviours are identified in children and adolescents: Challenge-seeking (choosing difficult tasks), Effort (commitment to learning), Persistence (ongoing engagement with tough tasks), and Resilience (recovering from failure). Collectively, these components are very similar to the common definitions of **Grit**. However, as they are tested through a very short-term task, they also appear strongly related to individual skills like **Persistence**, **Self-control**, and **Achievement motivation**. In the *PERC Task*, participants are presented with four series of Raven's matrices[3]. The initial set comprises relatively easy puzzles (80-90% accuracy) to facilitate success without reliance on pre-set feedback, enabling the evaluation of Challenge-seeking through subsequent choices between easier and harder puzzles. The second series provides a measure of Effort by featuring medium-difficulty puzzles (40-50% accuracy) and including feedback and optional tips. The third series assesses Persistence based on the duration of engagement with challenging puzzles (15-25% accuracy). The final series, mirroring the difficulty of the first, measures Resilience through performance after experiencing failure. This task has been tested for internal consistency but not for test-retest reliability, convergent, divergent (although it has not been controlled for correlation with general IQ) and criterion validity, and on different populations (in the USA and in South Africa) (Porter et al., 2020[54]; Porter et al., 2020[56]; Porter et al., 2020[57]).

Interestingly, all of these assessment approaches use very short-term tasks. It remains to be seen whether skills like Grit or Achievement motivation, which are characterised by the intention to exert effort over a long period of time based on long-term goals, in contrast to shorter-term and task-oriented skills such as Self-control and Persistence, can be assessed by individual tasks of this kind.

4.1.2. Emotion regulation

Emotion regulation is measured using two main methods: self-report questionnaires, which focus on typical Emotion regulation but may not accurately reflect one's actual

[3] Raven's matrices are multiple choice items that involve pattern recognition and are designed to measure a person's ability to form perceptual relations and to reason by analogy independent of language and formal schooling. In each item, the task is to complete the missing part of a pattern in a matrix of geometric designs. The patterns increase in complexity and difficulty, challenging the test-taker's problem-solving abilities and cognitive processing.

ability to regulate emotions; and experimental approaches using experiential or performance measures (Mehlsen et al., 2019[33]).

Experimental studies have predominantly inferred Emotion regulation from changes in experiential (subjective mental state/level of emotional arousal) and/or physiological emotional states (e.g., changes in skin conductance or in heart rate activity – see Biophysiological measures) following use of Emotion regulation strategies (Mehlsen et al., 2019[33]). While in most experiential studies, individual differences in regulation performance are not compared with other key behavioural and psychological indicators (Mehlsen et al., 2019[33]), some studies find correlations between effective strategy use, particularly cognitive reappraisal[4], with outcomes such as stress levels, depressive symptoms, well-being, and general cognitive functioning (McRae et al., 2012[58]). Because they provide deeper insights than self-reports by examining regulatory effectiveness of different strategies in relation to psychological functioning and individual differences, these studies highlight the importance of adding a performance-related component to experiential measures in Emotion regulation research. The varying effectiveness of different regulation strategies also highlights the importance of context and individual mastery in determining outcomes. A meta-analysis further supports this by demonstrating the differential effects of various Emotion regulation strategies, with cognitive change strategies, such as reappraisal and perspective taking, having the greatest impact on emotional outcomes (Webb, Miles and Sheeran, 2012[59]).

Notably, many experimental approaches to Emotion regulation assessment focus on specific strategies put in place by the participant, such as cognitive reappraisal, perspective-taking, or distraction regulation strategies, which limits their usage as assessment tools for a general Emotion regulation skill. Indeed, research shows that the adaptiveness of Emotion regulation varies with individual factors (e.g., age, gender, personality), situation (e.g., emotion type, intensity, context), and chosen strategies (e.g., cognitive demands, availability of support) (Ng et al., 2022[60]). While current assessments of Emotion regulation provide insight into regulatory efforts, more work is needed to capture the full picture of regulatory success. One Emotion regulation strategy does not fit all scenarios and more comprehensive assessments are needed. In addition, a recent literature review shows that while most assessments of Emotion regulation focus on negative emotions, few assessments explore the regulation of positive emotions (joy, interest, pride). Understanding how students regulate positive emotions is also important for their academic and personal development (Ng et al., 2022[60]). Overall, tailoring assessments to specific emotions and contexts can better inform targeted interventions.

Stress resistance

No task-based assessments were found specifically for **Stress resistance**, defined as the ability to effectively modulate anxiety and stress and the ability to solve problems calmly (Chernyshenko, Kankaraš and Drasgow, 2018[40]; Kankaraš and Suarez-Alvarez, 2019[39]). This skill is typically more easily assessed using biophysiological measures (see Biophysiological measures).

[4] Cognitive reappraisal involves changing how one thinks about or appraises a given (generally emotionally charged) situation. For example, it requires taking a step back and viewing a provoking event objectively, rather than immersing oneself in angry feelings and thoughts, as to control the outburst of emotions (Denson and Fabiansson Tan, 2023[219]). Cognitive reappraisal can apply to many different emotions and feelings.

Emotional control

Emotional control is defined in the OECD framework as the ability to implement effective strategies to regulate temper, anger and irritation, in the face of frustrations (Chernyshenko, Kankaraš and Drasgow, 2018[40]; Kankaraš and Suarez-Alvarez, 2019[39]). Interestingly, while Emotional control could also be understood as the ability to upregulate positive emotions (focusing on happy memories when feeling a bit depressed, for example), or downregulate positive emotions in inappropriate social contexts (do not overtly show pride for a successful school test in front of a friend who failed, for example), most task-based assessments focus on downregulation of negative emotions. Another important element to take into consideration when assessing Emotional control is emotional intensity, which is a measure of a person's emotional response force. Studies show that lower emotional intensities often lead to adaptive regulation strategies, aiding in emotion processing, whereas higher intensities can result in maladaptive strategies, causing disengagement from emotions (Watanabe, Motomura and Saeki, 2022[61]).

The ***Beach Balls Task*** measures frustration tolerance (Jiménez-Soto et al., 2022[62]) .Aimed at children aged 6-10, the task measures Frustration tolerance by requiring participants to select the smallest of four beach balls on screen within five seconds. The test comprises three sets, with Sets 1 and 3 identical to gauge baseline performance. Set 2, designed to induce frustration, features minimal size differences between balls, thus challenging the children's ability to perform under increased difficulty. Frustration tolerance is measured as the performance difference between Set 3 (after the "frustrating" phase) and Set 1. A limited drop in performance is considered a hallmark of higher Frustration tolerance. In the publication, Frustration is defined as the subjective emotion of dislike associated with high levels of effort and low levels of success in a task, and while Frustration tolerance is described as a *"process dependent on emotional regulation"* (Jiménez-Soto et al., 2022[62]), the task is very similar to Persistence tasks. Moreover, this task does not seem to allow for disentangling Emotional control of irritation from boredom.

The ***Laboratory Coping And Emotion Regulation Task*** assesses **Emotion regulation** in children and adolescents defined as a set of conscious, controlled processes that aim to regulate emotions, thoughts, behaviours, and physiological responses in the face of stressors (Bettis et al., 2019[63]). Participants are assessed on secondary control coping strategies like acceptance, cognitive reappraisal, and distraction, through a task involving viewing images depicting parental sadness and irritability. The task, designed to reflect real-life family stress, included images of negative parental emotions and neutral images. Participants were instructed to either reappraise these images positively, distract themselves, or simply react, to gauge their coping and emotion regulation in response to familiar stressors (Bettis et al., 2019[63]). This test has been controlled for internal consistency (but not test-retest reliability), construct validity and criterion validity.

Optimism

Optimism is the ability to have positive and optimistic expectations for self and life (Kankaraš and Suarez-Alvarez, 2019[39]). Optimism can be defined as a bias characterised by positive assessments of self-risk relative to the average other (Hennefield and Markson, 2022[64]). In adults, this manifests as overestimating positive outcomes and underestimating negative events, even when only known probabilities should guide predictions.

Both of the identified task-based assessments of Optimism, the ***Future Expectations Task*** (Bamford and Lagattuta, 2020[65]) and the ***Story Task*** (Hennefield and Markson, 2022[64]), use a similar approach. Participants are usually presented with scenarios (written and sometimes illustrated) and asked to choose between optimistic and pessimistic outcomes

and then rate the likelihood of their prediction. Often a distinction is made between Optimism trials and Wishful thinking trials, the latter contrasting ordinary positive outcomes with highly unlikely positive ones, to understand participants' bias towards probable or desirable future events. Some instruments also vary the focus between Optimism for oneself and Optimism for others, exploring the comparative optimism bias, where self-related predictions are found to be more optimistic than those for peers. This makes it possible to explore how participants perceive and predict future events, balancing probability and desirability.

4.1.3. Engaging with others

The literature review did not identify task-based assessments focusing specifically on the skills related to Engaging with others, that is **Sociability**, **Assertiveness**, and **Energy**. Sociability is defined as the ability to approach others, both friends and strangers, and to initiate and maintain social connections. Assertiveness is defined as the ability to confidently voice opinions, needs, and feelings, and exert social influence, and Energy is defined as the ability to approach daily life with energy, excitement, and spontaneity (OECD, 2021[66]).

4.1.4. Collaboration

Empathy, Perspective-taking and Emotion recognition

In psychology and related sciences, **Empathy** is commonly conceptualised in two facets: cognitive and affective (Drimalla et al., 2019[67]; Thompson, van Reekum and Chakrabarti, 2022[68]). The first facet, Cognitive empathy, describes a person's ability to infer and understand the emotional states of others. The second facet, Affective empathy (sometimes referred to as Emotional empathy or Emotional contagion), is defined as an observer's emotional response to the emotional state of another individual (the ability to be sensitive to and to vicariously experience emotions of others). In other words, Affective empathy is the ability to experience another person's feelings while Cognitive empathy, the capacity to understand such feelings (Quinde-Zlibut et al., 2021[69]). This conceptualisation of Empathy into two facets is supported by behavioural, neuroimaging, and neurological findings (Drimalla et al., 2019[67]). Recent evidence now adds a third important component to the multifaceted nature of Empathy, Empathic concern. Empathic concern can be described as the subsequent compassionate response to the perceived or felt emotions of others (Quinde-Zlibut et al., 2021[69]; Watanabe, Motomura and Saeki, 2022[61]). Empathic concern involves feeling sympathy and compassion towards others' experiences, distinct from Affective empathy, which is about sharing the same emotions. It might involve actionable expressions of that concern.

Cognitive empathy is conceptually related to **Perspective-taking** (also called **Theory of Mind** – ToM) and **Emotion recognition** (Drimalla et al., 2019[67]). Understanding someone else's emotion (Cognitive Empathy), especially a complex or subtle emotion in a rich context, requires the knowledge of this emotion and its related behavioural expressions (Emotion recognition), and the ability to attribute mental states – beliefs, intents, desires, emotions – to others and take their perspective (Perspective-taking). Depending on the definitions, these concepts overlap more or less, especially Perspective-taking and Cognitive empathy. For the sake of clarity and parsimony, Perspective-taking skills, defined as the ability to accurately perceive the thoughts and experiences and feelings of others and how these might differ from one's own (OECD, Forthcoming[30]), are hereunder considered synonymous to Cognitive empathy (Cerniglia et al., 2019[70]).

Task-based assessments of Empathy and related skills can be organised based on the number of aspects of Empathy they focus on. Some tests simply assess the ability to recognise an emotion presented in simple stimuli, such as a picture of a facial expression or a brief written description (Emotion recognition), some tests assess the ability to infer emotions of some characters based on more complex social or non-social situations (Cognitive empathy), and some tests also add questions on the subjective feeling of the participant regarding the situation described (Affective empathy).

Many task-based assessments of Perspective-taking and Cognitive empathy (including Emotion recognition) exist. Here only a few are described, selected based on their diversity and on the quality of their validity/reliability metrics. Assessments of **Emotion recognition** include very minimalistic tests such as the ***Reading the Mind in the Eyes Test - Child Version*** (conceptually derived from the adult version) (Rosso and Riolfo, 2020[71]; Vogindroukas, Chelas and Petridis, 2014[72]), that requires children to interpret mental states from photographs of adult eyes, and select between one of four descriptive words. Some computer-administered tests use multimodal stimuli to evaluate Emotion recognition. This can be done either separately, such as the ***Emotion Recognition Index*** (Scherer and Scherer, 2011[73]) that has two subscales for facial and vocal emotion recognition, or together as the ***Geneva Emotion Recognition Test*** **(*GERT*)** (Schlegel, Grandjean and Scherer, 2014[74]) and its short version (***GERT-S***) (Schlegel and Scherer, 2016[75]), which assess the recognition of a range of emotions from short video clips with sound (in which facial, vocal, and bodily cues are presented simultaneously). In consequence, the *GERT* has arguably better content validity than existing tests that largely focus on static facial expressions and basic emotions.

By increasing the complexity of the stimuli presented (particularly by introducing contextual and/or social elements), other assessments are better equipped to test **Perspective-taking**/**Cognitive empathy** abilities. Some tests evaluate how individuals interpret and analyse social scenarios presented in text. The ***Assessment of Social Perspective-taking Performance*** (Kim et al., 2018[76]) measures both simple (articulating how actors think, feel, or are inclined to behave) and complex (contextualising the position actors take, with consideration of their roles, circumstances, experiences, and motivations) Perspective-taking abilities. Participants give written responses to open-ended questions based on hypothetical school-related dilemmas (such as discovering a friend having cheated on an exam). The ***Combined Stories Test*** (Achim et al., 2012[77]) evaluates a wide range of mental states such as beliefs, intentions, and emotions. Participants analyse interactions between characters, focusing particularly on second-order ToM, which involves understanding a character's mental state about another's mental state (e.g., a character's intention or false belief about another character's action or belief), as opposed to first-order ToM that deals with perceiving mental states about physical world conditions. Similarly, the ***Faux Pas Recognition Test*** focuses on more advance ToM capabilities, by testing whether children and adolescents are able to identify socially inappropriate comments from one character about another, in written vignettes (Baron-Cohen et al., 1999[78]; Hayward and Homer, 2017[79]). Test-takers are expected to understand both states of mind, that the person committing the faux pas does not have malicious intent, yet says something which makes the other person feel confused or uncomfortable. While the ***Emotional Literacy Test with Hypothetical Scenarios*** (Watanabe, Motomura and Saeki, 2022[61]) currently lacks extended validation for reliability and validity, this test uses open-ended questions to assess the ability not only to identify and explain the rationale for characters' emotions, but also to assign intensity to these emotions.

Finally, some assessments use audiovideo film stimuli instead of written or illustrated narrative vignettes. For example, the ***Social Attribution Task - Multiple Choice*** (***SAT-MC***) (Johannesen et al., 2013[80]), assesses implicit social attribution using an animation of

geometric shapes enacting a social drama (like a domestic fight between two larger triangles representing the parents and a smaller triangle representing the child, for example). It reduces verbal and cognitive demands (and to a certain extent, culturally dependent environmental information) and presents multiple-choice questions about the actions and emotional intents depicted in the animation. The ***Video-Social-Emotional Information Processing*** (***V-SEIP***) (Coccaro et al., 2017[81]) is a clinical psychology task that uses clips of socially ambiguous situations featuring aversive actions, either overt or relational aggression, to test participants' ability to read the situation and correctly assess the intentions of the depicted characters (in the general population and in patients with identified violent behaviour[5]). Participants rate the actions of a character with whom they identify. They summarise the clips and then answer questions assessing hostile, benevolent and instrumental attributions and their own emotional reactions. They also rate their agreement with different response strategies shown in subsequent videos, which can be considered a measure of **Social problem-solving** (see Social problem-solving).

Finally, some task-based assessments focus on **Affective empathy**. The ***Facial Emotion Recognition and Empathy Test*** (***FERET)*** (Coskun, 2019[82]) and the ***Multifaceted Empathy Test*** (***MET***) (Dziobek et al., 2008[83]) are two computerised tasks that assess both Cognitive and Affective empathy in response to a series of emotionally charged facial expressions. The design of both tests is quite similar. Participants are presented with a series of pictures and tested on their ability to (1) identify the emotion depicted (Emotion recognition), (2) display an appropriate empathic response by sharing the same feeling (Affective empathy). However, while the *FERET* assumes that facial Emotion recognition/Cognitive empathy is a precursor to Affective empathy, this is not the case for the *MET* (which originally allowed for the differentiation of deficits in Cognitive empathy from typically functioning Affective empathy in children with autism spectrum disorders).

In the *FERET*, participants first identify an emotion in a picture (coloured drawing of a face), then imagine how they would feel if a classmate displayed that emotion, and finally choose one of three emotional response options provided (coloured drawings of faces). This assessment has been tested for internal consistency, but not for construct or criterion validity, on 7–10-year-old children. The *MET* and the ***MET-J*** were originally designed to differentiate empathy components in adolescents aged 10-17 years with autism spectrum disorder or conduct disorder (Poustka et al., 2008[84]). The items consist of photographs depicting individuals of different genders and ages in emotionally charged situations with rich contexts (e.g., a hospital room, a bicycle race) in which other individuals are also seen. Participants are asked to identify the emotion presented (Cognitive empathy), but also to rate their level of emotional arousal in response to each stimulus using a Likert scale (implicit measure of Affective Empathy). The MET has been tested for reliability and construct and criterion validity (to measure of social behaviour and social engagement), but the validity correlations were weak and internal consistency estimates were low. While the *MET* and *MET-J* are mostly used with clinical populations, they use photorealistic pictures in ecologically rich contexts compared to the *FERET*. These tests are more ecologically valid (especially the *MET* and *MET-J*, as they use realistic stimuli, including complex emotions and contexts) and mitigate potential social desirability biases carried by self-report measures of empathy, as they do not rely as heavily on the level of insight an individual has into their own emotions (Drimalla et al., 2019[67]).

[5] These constructs fit into a theory of aggression that posits that cognitive processes such as attribution of others' intentions and response evaluation of possible response options influence whether an individual will behave aggressively in a given situation (Coccaro et al., 2017[81]).

Social problem-solving

Social problem-solving, or conflict resolution, is the ability to identify and enact solutions to social life situations in an effort to resolve interpersonal problems, conflicts and/or one's relation to these (Kankaraš and Suarez-Alvarez, 2019[39]). The review did not identify any task-based assessment specifically testing social-problem solving, although the ***Mayer–Salovey–Caruso Emotional Intelligence Test Version 2.0* (*MSCEIT*)** and the ***Mayer–Salovey–Caruso Emotional Intelligence Test–Youth Version, Research Version* (*MSCEIT-YV*)** include a task for such skill (see Mixed batteries).

Trust and Co-operation

Trust can be defined as "a psychological state comprising the intention to accept vulnerability based upon the positive expectations of the intentions or behavior of another" (Rousseau et al., 1998[85]). Recently, an important distinction has been made between situational Trust (a transient state depending on the context or the task) and dispositional Trust or propensity to Trust (an enduring trait) (Evans and Bond, 2020[86]). In this sense, Trust understood as a skill should be assessed as transient state Trust, so as the ability of a person to trust effectively in certain situations, and not as a natural tendency to trust (more related to a personality trait). This review focused on tasks defining Trust as such.

Co-operation can be defined as the act or process of working together to get something done, to achieve a common purpose or mutual benefit, either for an individual being co-operative or acting cooperatively (Tyler, 2010[87]). Interestingly, some recent economic studies have used tasks to distinguish active and reactive co-operation (i.e., non-exploitation versus non-retaliation) – two aspects of co-operative behaviour associated with different basic personality traits (the traits Honesty-Humility and Agreeableness in the HEXACO[6] model, respectively) (Thielmann, Hilbig and Niedtfeld, 2014[88]).

Trust and Co-operation, as well as other prosocial behaviours, have been extensively studied in economics and other disciplines using economic games. While the number of existing tasks is enormous, most of them come from a small set of paradigms that have been continuously adapted over the years. Because they change the dynamics of the task, but also the ultimate purpose of the game, these adaptations (sometimes very small changes) have an important impact on the nature of the behaviour and the associated skill that they are designed to assess. Furthermore, these paradigms have been used extensively with groups of participants from all demographic backgrounds, in-game decisions and behaviours have been correlated with real-life behaviours, and these tasks are usually easy to implement both in reality and virtually. However, the often-simplistic efficiency of these tasks is a limitation when it comes to distinguishing between specific prosocial skills such as Trust and Co-operation. This could limit the potential applicability of these tools for in-depth assessment of specific skills, unless the design is adapted to specifically differentiate between skills. As most economic games are based on analysing the participant's decision to collaborate and therefore to trust or not trust the other player(s), they can also be considered as an assessment of Co-operation. Since the task paradigms presented hereafter are paradigms tested and adapted many times in many different contexts, specific information regarding reliability and validity metrics is not applicable.

Probably the most famous economic game is the ***Investment Game***, or ***Trust Game*** (see (Thielmann et al., 2021[89]), for a review of the methodological aspects of this and other

[6] HEXACO is a six-dimensional model of human personality, whose factors include honesty-humility (H), emotionality (E), extraversion (X), agreeableness (A), conscientiousness (C), and openness to experience (O)

Economic games). The *Trust Game* is a sequential game involving two players, a trustor and a trustee, each initially endowed with an equal number of tokens. The trustor decides how many tokens to transfer to the trustee, which are then multiplied and added to the trustee's endowment. The trustee then decides how many tokens to return to the trustor. The *Trust Game* simplifies real-life situations of unilateral dependence, like online transactions, and has been used to understand the dynamics of trust and co-operative behaviour in different contexts. In a one-shot game, the trustor should theoretically send nothing, anticipating that the trustee has no incentive to return anything. It allows for the expression of beliefs about prosociality and motives like altruism, fairness, or greed. While the game's simplicity is sometimes valued as a way to measure **Trust** in a purely abstract way, it has been widely adapted and implemented with elements such as communication, acquaintance, and repeated interaction to more closely approximate corresponding real-life situations (Thielmann et al., 2021[89]).

Other economic games used to assess **Trust** and **Co-operation** include social dilemmas such as the ***Prisoner's dilemma*** and the ***Public Goods Game***. Social dilemmas are characterised by conflicts between immediate self-interest and long-term collective goals and are critical in understanding co-operation development. The *Prisoner's dilemma* presents a scenario where two individuals, acting independently, must each decide whether to cooperate with the other or to act selfishly (defect) (Thielmann et al., 2021[89]). The dilemma arises because the optimal outcome for each individual (defecting while the other cooperates) leads to a worse collective outcome than if both had cooperated. If both cooperate, they receive a moderate reward; if one defects and the other cooperates, the defector gets a high reward while the cooperator gets a low or no reward; if both defect, they both get a low reward. This setup illustrates the conflicts between individual interests and collective well-being. Various adaptions exist, such as the ***Intergroup Prisoner's Dilemma*** which looks at co-operative behaviours across groups (Thielmann et al., 2021[89]). Here, each group decides whether to cooperate with or defect against the other. Mutual co-operation leads to moderate benefits for both, mutual defection results in worse outcomes, and one group defecting while the other cooperates gives the defecting group a high benefit but disadvantages the co-operative group. This paradigm is used to study group dynamics and intergroup relations, highlighting how collective interests and biases (toward the other group) impact decision-making and often make intergroup Co-operation more challenging. In that sense, this variation can be used to assess aspects of Open-mindedness such as **Tolerance**, on top of Trust and Co-operation (see Open-mindedness and Tolerance in the following section).

In *Public Goods Game* paradigms, players decide how much of their initial endowment to keep or contribute to a public good, which is then multiplied and equally redistributed, with individual outcomes depending on both their and others' contributions (Keil et al., 2017[90]). ***Pizzagame*** (Keil et al., 2017[90]) is an example of *Public Goods Game*. This computer task developed for children and adolescents simulates playing with peers online, though interactions are with computer-generated players. The participants face three conditions based on the other virtual players' strategies: co-operative strategy, selfish strategy, and divergent co-operative–selfish strategies. The task can thus measure the change in strategy of the participant based on the other's strategies (conditional Co-operation).

4.1.5. Open-mindedness

Tolerance

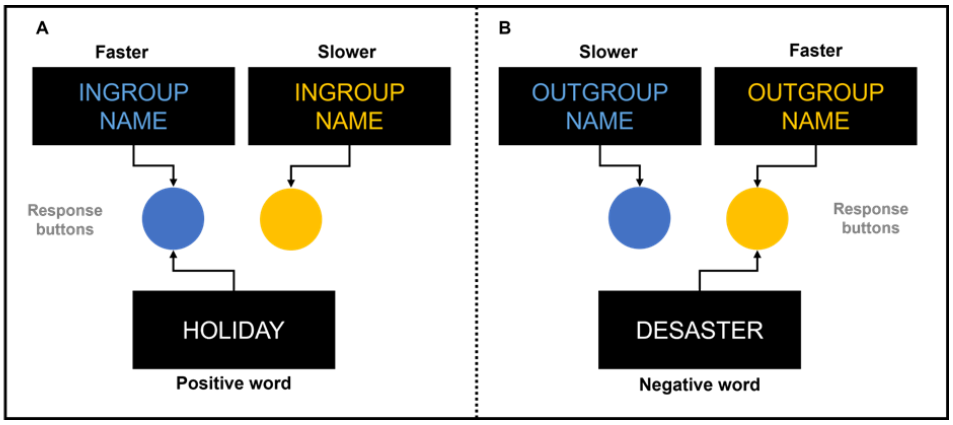
Tolerance is defined as the ability to be open to different points of view, to value diversity, and to be appreciative of foreign people and culture (OECD, 2021[66]).

As discussed in the previous section (see Trust and Co-operation), the ***Intergroup Prisoner's dilemma*** is used in Economics research to evaluate intergroup behaviours (Thielmann et al., 2021[89]). Players are divided into two groups. Each player receives a set number of tokens and decides privately how many to contribute to their group's pool. These contributions are multiplied and redistributed equally among in-group members but also negatively impact the out-group's payoff. The collective optimum (highest sum of points of all players) is achieved when no one contributes. However, contributing maximizes the in-group's welfare relative to the out-group. This dilemma creates a conflict between individual, in-group, and collective interests, thus creating a conflict between helping their group and the overall good of all players (Thielmann et al., 2021[89]).

In recent decades, research on social cognition, particularly attitudes and beliefs about social groups, has shifted from traditional self-report measures to indirect measures of mental content, such as reaction times and unconscious responses (Kurdi et al., 2019[91]). Applications of these methods (adapted from cognitive psychology) examined semantic associations, showing faster responses to mentally related word pairs. This approach was extended to uncover representations of social categories, revealing biases in responses to racially or stereotypically related stimuli (Kurdi et al., 2019[91]). The ***Implicit Association Test*** (***IAT***) (Greenwald, McGhee and Schwartz, 1998[92]) exemplifies this shift, using response speed and accuracy to infer implicit evaluations, or spontaneous likes and dislikes, by measuring how quickly participants associate categories (e.g., local vs foreigner) with attributes (e.g., good vs bad). While implicit measures were initially assumed to assess stable individual differences, a large chunk of literature posit that they reflect context-dependent processes, which raises the question of the link between IAT-like test results and real-life behaviours – see for example (Elder, Wilson and Calanchini, 2023[93]).

However, several other tools have been developed, inspired by this approach, and use a proxy to infer implicit/unconscious prejudice biases, especially towards outer groups. For example, Degner and Wentura (2008[36]) used an adaptation of the ***Extrinsic Affective Simon Task*** to assess behaviour towards groups of people perceived as ingroups or outgroups (De Houwer, 2003[94]). In this type of reaction time task (see Figure 4.3. Image representing an Extrinsic Affective Simon Task), participants are presented with white words to classify on the basis of stimulus valence (for example, positive words to blue button and negative words to yellow button) and coloured words to classify on the basis of colour (blue words to blue button, yellow words to yellow button). On trials where the word refers to a positive target concept (in this case, stereotypical names of the participant's culturally similar in-group), performance is superior when the words are in the colour associated with the positive button colour. The opposite is true for trials in which the word represents a negative target concept (in this case, stereotypical names of the participant's culturally distinct out-group). Their results showed that stereotypical names of a culturally distinct out-group were judged more negatively than stereotypical names of a culturally similar in-group, and that several measures of task performance significantly correlated with measures of explicit prejudice. However, this correlation was moderated by participants' motivation to control their expression of prejudice (Kurdi et al., 2019[91]).

Figure 4.3. Image representing an Extrinsic Affective Simon Task



Note: Participants are (A) faster to respond to implicitly positive-coloured words when the answer button is associated to the positive words in the white condition, and (B) faster to respond to implicitly positive-coloured words when the answer button is associated to the negative words in the white condition.

Curiosity

Only one task-based assessment was found to assess the skill of **Curiosity**. Defined simply as *"the desire to know"* (Schutte and Malouff, 2020[95]), this ability is better tested in game contexts, where the free decision of the players are informative of their preferences and of their unsolicited and genuine tendency to look for knowledge.

In a study aimed to explore how Curiosity can be initiated, supported, and assessed in a digital environment, Sher, Levi-Keren and Gordon (2019[96]) designed a novel app destined to university applicants. The ***Faculty Game*** allows users to freely explore and learn about various topics without predefined tasks, capturing their exploration patterns and the depth and breadth of their interests. Through detailed behavioural measures extracted from the app's usage data – such as the start time of interaction, the number of facts explored, and the users' exploration patterns – the study proposed a way to quantify curiosity. These measures provided insights into the participants' specific and diverse interests, as well as their exploratory behaviour, offering a comprehensive view of Curiosity. Additionally, the study looked into how providing users with the autonomy to cease their exploration at will affected their behaviour and learning. This setup aimed to mirror natural curiosity-driven learning environments closely.

Creativity

In past OECD work, **Creativity** as the ability to generate novel ways to do or think about things through tinkering, learning from failure, insight, and vision (Chernyshenko, Kankaraš and Drasgow, 2018[40]). In the direct assessment literature, Creativity is defined as the ability to produce tangible objects and ideas that are (1) novel, original, unexpected and (2) appropriate, useful and adaptive concerning task constraints (Perry and Karpova, 2017[97]). These two components of Creativity, novelty and usefulness, should be taken into consideration for its assessment (Shaw, 2022[98]).

The direct assessments of Creativity primarily revolve around two main approaches: process-oriented assessments and product-oriented assessments (Rafner et al., 2023[99]). Process-oriented assessments typically involve standardised tests that focus on creative processes like idea generation and refinement. This approach is valued for its scalability, allowing for widespread application across various contexts. However, they often face criticism for their lack of real-world applicability and motivational elements. On the other

hand, product-oriented assessments involve evaluating actual creative products, particularly in fields that hold significant relevance and complexity, such as music and design. This approach is lauded for providing data with high ecological validity because the combination of expert assessment and relevant, complex scenarios is regarded as highly effective. However, it is also considered much more costly in time and resources, and less suited for scalability. For this reason, we believe that exploring methods that integrate the scalability of process-oriented assessments with the ecological validity of product-oriented assessments, such as game-based assessments, could significantly improve the overall construct validity of creativity assessment tools (see Digital games).

The majority of Creativity assessment tasks are process-oriented. Sometimes known as divergent thinking tests, most of them ask participants to generate different ideas in response to specific stimuli, often pictorial or verbal (Clapham, 2011[100]). Common response formats in these tasks include creating drawings from incomplete figures, writing questions for hypothetical scenarios, or listing potential uses for objects. Typically, divergent thinking tests evaluate several features of these responses, such as fluency (the ability to develop large numbers of ideas), flexibility (the ability to produce ideas in numerous categories), originality (the ability to produce unusual or unique ideas) and elaboration (the ability to adapt abstract ideas into realistic solutions).

The ***Divergent Thinking Task*** developed by An, Song and Carr (2016[101]) and tested on a population of Korean students is a paradigmatic example of a divergent thinking test, where participants are asked to produce as many of their own creative hypotheses as possible to account for a real-world problem. Experts then score the divergent thinking task using three criteria: fluency, flexibility, and originality. This task has been controlled for internal consistency (but not test-retest reliability) and tested for construct and criterion validity. However divergent thinking scores did not correlate with self-reported measures of creativity, and divergent thinking scores were predicted significantly by external measures of creative personality.

A seminal test of divergent thinking is the ***Torrance Test for Creative Thinking*** (***TTCT***) (Torrance, 1966[102]). The *TTCT* has been widely used and extensively tested in different populations, age groups and cultures, and has reached its current form through repeated revisions since its first publication in 1966 (Yoon, 2017[103]). The revisions have mainly concerned the scoring system, while the content and form have remained unchanged. What distinguished the *TTCT* from other creativity or divergent thinking tests was not only how Ellis Paul Torrance defined creativity, but also how he made the test fun, easy to use, and applicable to different populations and cultures (Alabbasi et al., 2022[104]). Another important difference from other divergent thinking tests was that Torrance added measures of other creative expressions to the list of creative strengths, such as humour, storytelling and boundary breaking.

This expanded the scope of the *TTCT* beyond divergent thinking (Alabbasi et al., 2022[104]). The *TTCT* consists of two subtests, the Verbal battery and the Figural battery, which both have two parallel forms, A and B. The Figural battery is generally known to be less influenced by cultural biases or the subject's linguistic ability than the Verbal battery (Yoon, 2017[103]). The Verbal battery of the *TTCT* consists of six tasks relying heavily on writing (Alabbasi et al., 2022[104]). Each task of the Verbal battery presents participants with a picture as a stimulus for verbal exercises to which they respond in writing. The first three tasks, Asking, Guessing causes and Guessing consequences (collectively known as Ask-and-Guess), explore the link between curiosity and creativity. The fourth task, Product improvement, focuses on improving existing products as opposed to creating entirely new ones. In the fifth task, named the Unusual uses task, participants think of alternative uses for common objects, such as wheels. Finally, the sixth task named Just suppose challenges

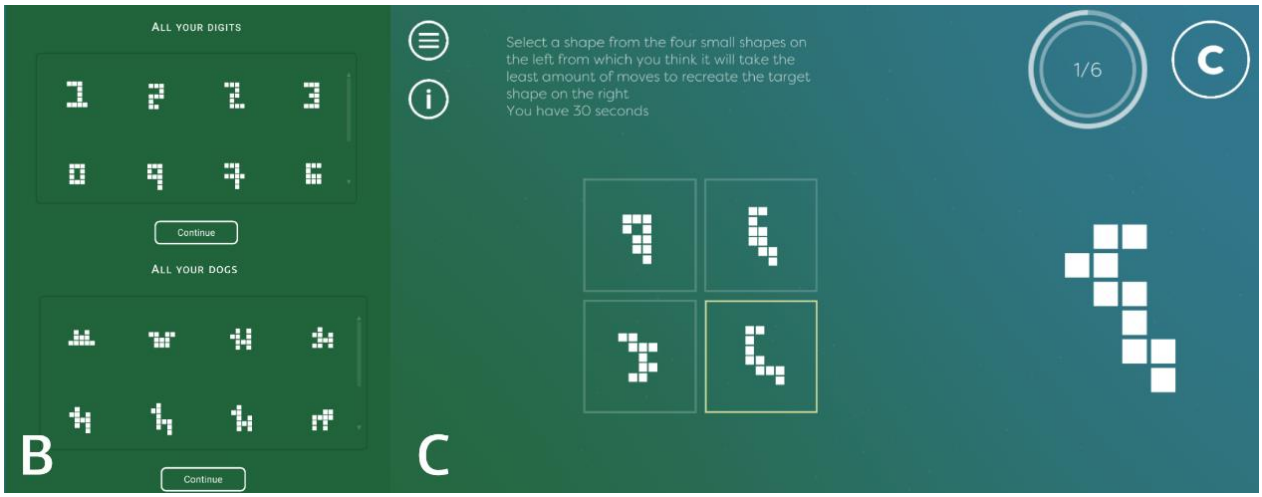
individuals to hypothesise about unlikely scenarios, testing their tolerance for and playfulness with unusual situations. This battery assesses fluency, flexibility and originality within a 45-minute timeframe. The Figural battery of the *TTCT* consists of three tasks that require very little writing (Alabbasi et al., 2022[104]). The Picture construction task requires the subject to draw a picture based on a given stimulus. The Picture completion task requires the subject to draw a picture using incomplete figures and to title each drawing. The Repeated line (or circle) task requires the subject to draw a picture using pairs of lines or circles given as stimuli. The Figural Battery assesses students' fluency, originality, abstractness of titles, elaboration, resistance to premature closure, and the Checklists of Creativity Strengths in 30 minutes. While the evidence for the reliability of the *TTCT* and similar tests is fairly strong, the predictive and discriminant validity (i.e., the lack of correlation with irrelevant measures) of divergent thinking tests is still debated in the literature (Rafner et al., 2023[99]).

The ***CREA*** suite is a more recent and more holistic approach to the direct assessment of creative skills (Rafner et al., 2023[99]). Still under development, this digital assessment portfolio consists of a series of tests designed to assess divergent and convergent thinking in a variety of ways to capture different aspects of Creativity. While divergent thinking is defined as the ability to come up with many different solutions to a prompt, convergent thinking is defined as the ability to find the single best (or correct) option within a space of solutions. The *CREA* suite includes *Crea.tiles* and *Crea.blender* – non-verbal tests of both divergent and convergent thinking, and *Crea.ideas* – an adaptation of the *Alternative Uses Task*, a standard test of divergent thinking. The CREA suite includes another tool, *Crea.logic*, a non-verbal test designed to test general abstract reasoning, thus ensuring that the creative abilities being measured are distinct from general reasoning (controlling for divergent validity).

Crea.tiles is an adaptation of the Creative Foraging Game. In this non-verbal assessment, participants create shapes on a grid of ten squares, aiming to design aesthetically pleasing forms within a set time (see Figure 4.4). The game features two modes: Divergent thinking and Convergent thinking. In the divergent thinking mode, players form various shapes across different categories within a limited time, while the convergent thinking mode challenges them to transform a starting shape into a target figure in the fewest steps and shortest time, with challenges of increasing difficulty.

In contrast to *Crea.tiles*, *Crea.blender* is a co-creative (human-machine) task that allows players to blend or merge existing images into new ones using AI (Rafner et al., 2020[105]). It works on the principle of constraint-based combinational creativity, visually allowing for creative outcomes. When mixing images in *Crea.blender*, participants use sliders to indicate how much each image should contribute to the resulting image. Apart from the underlying differences in the game mechanics, the task and prompt design have been retained as much as possible to maintain contextual homogeneity between the two sub-challenges. The game has three modes: Creatures Mode, where players have five minutes to create as many "animal-like" creatures as possible from six images; Challenge Mode, where players have three minutes to identify and recreate a target image from a set of source images; and Open Play Mode, which asks the players to create and save during five minutes any image they find interesting from the same image sources as in Creatures Mode. This versatile platform is designed to provide a playful, engaging environment for assessing creativity, facilitating both divergent and convergent thinking studies.

Figure 4.4. Screenshot from CREA



Note: An example of crea.tiles, from CREA. B) In the Divergent Thinking mode, participants are asked to produce different original shapes of dogs from a limited number of tiles. C) In the Convergent Thinking mode, participants are asked to select, from the four options, the shape which will require less moves to be transformed into the target shape on the right.
Source: Adapted from (Rafner et al., 2023[99]), "Towards Game-Based Assessment of Creative Thinking", *Creativity Research Journal*, Vol. 35/4, pp. 763-782, http://dx.doi.org/10.1080/10400419.2023.2198845

Finally, in *Crea.ideas*, participants are asked to come up with as many creative uses as possible for three common objects (a tyre, a brick and a paper clip). The instruction is for participants to be 'creative' in their thinking. Participants have two minutes to complete each prompt (Rafner et al., 2023[99]).

In terms of reliability and validity of these tools, Divergent thinking measures in *Crea.ideas* and in *Crea.tiles* correlate, supporting convergent validity. No correlation was found with self-reported measures of Creativity. However, this discrepancy between self-report and direct measurements of Creativity have been reported by other tools as well. While *Crea.tiles* and *Crea.blender* are very innovative and promising approaches to creativity measurement, data is still lacking regarding their reliability, criterion validity and applicability to various populations. However, these tools are still being developed.

Finally, it worth mentioning that the famous videogame Minecraft has been used to assess *Creativity* (Shaw, 2022[98]). Participants were asked to create original and functional houses in the 3D environment of the game. The creations were then rated independently by a jury of experts for novelty and usefulness. Correlations were found between self-report scores of Creativity and novelty scores based on in-game creations, and novelty and usefulness scores had different significant relationships with the Big Five personality traits.

Critical thinking

No task-based assessments of Critical thinking were found, defined as ability to question, to analyse and evaluate information as a basis for beliefs and actions (Kankaraš and Suarez-Alvarez, 2019[39]).

4.1.6. Self-reflection

Metacognition

Despite the widening debate about what **Metacognition** actually is and how it should be assessed, it is generally understood as a multifaceted ensemble of processes related to awareness, knowledge and understanding of inner processes and subjective experiences, such as thoughts and feelings (Gascoine, Higgins and Wall, 2017[106]). Understood as a skill, Metacognition is the ability to reflect on, articulate and deliberately control such experiences (Steponavičius, Gress-Wright and Linzarini, 2023[2]). Metacognition has been approached by many disciplines and is conceptually related to several concepts such as executive function, Self-control and Emotional control (Gascoine, Higgins and Wall, 2017[106]), or Perspective taking and Cognitive empathy. While the debate around the exact nature of Metacognition goes beyond the scope of this work, it is important to acknowledge that many of its facets are generally associated with foundational cognitive processes. Various approaches exist for assessing these cognitive processes; however, the focus of this research was to identify tools that assess aspects of Metacognition with an emotional feature.

The ***Levels of Emotional Awareness Scale*** (***LEAS***) (Lane and Smith, 2021[107]) and the ***Levels of Emotional Awareness Scale – Children*** (***LEAS-C***) (Bajgar et al., 2005[108]) are two tasks designed to assess adults' and children's ability to recognise and describe their own emotions (**Emotional awareness**, **Metacognition**), and the emotions of others (**Perspective-taking**, **Empathy**). Developed over thirty years ago, the *LEAS* has been extensively tested for validity and reliability (for a review see for example (Siegling, Saklofske and Petrides, 2015[109]) and adapted into a digitalised tool. The scale's effectiveness and reliability have been supported by extensive research across healthy and clinical populations, demonstrating its significance in understanding emotional self-regulation, social adaptation, and overall mental and physical health. This assessment involves presenting individuals with a series of hypothetical but relatable scenarios covering a range of emotional contexts, such as social interactions, achievements, losses, and ethical dilemmas (Lane and Smith, 2021[107]). Each scenario is a brief description of a situation that involves interpersonal interactions or personal experiences likely to generate emotions. The key aspect of the *LEAS* is that it does not directly ask respondents to identify or choose emotions from a given list. Instead, it requires them to describe in their own words what they would feel in the situation described, as well as what any other person involved in the scenario might feel. The open-ended responses are then scored based on the level of emotional awareness they reflect, according to a specific scoring system. The scoring focuses on the structure and differentiation of the emotion words used rather than their content or appropriateness. The total *LEAS* score is calculated by summing the scores across all scenarios, providing a measure of the individual's overall level of emotional awareness. Higher scores indicate greater Emotional awareness, including the ability to recognise and describe complex and nuanced emotional states in oneself and others. A large amount of research using this tool exists, making it a well-validated tool. *LEAS* shows good estimates of internal consistency and test-retest reliability. Scores in the assessment have been correlated with scores from other tools measuring similar constructs, as well as with various outcomes (criterion validity).

Self-efficacy and Self-esteem

Self-efficacy is defined as "the strength of individuals' beliefs in their ability to execute tasks and achieve goals" (Kankaraš and Suarez-Alvarez, 2019[39]). The literature search pointed out to a related concept, namely **Self-esteem**, defined as "the perception of oneself

and a personal evaluation one makes about oneself". While both Self-efficacy and Self-esteem reflect self-evaluation, self-efficacy is more task-specific and malleable and self-esteem is more general and stable (Chen, Gully and Eden, 2004[110]). No tasks assessing neither Self-efficacy nor Self-esteem were identified, although literature exists on assessments of task-specific Self-efficacy, such as in mathematics (Siefer, Leuders and Obersteiner, 2021[111]). It is worth noting that the nature of Self-esteem and Self-efficacy makes it questionable whether it can be considered a skill according to the OECD definition.

4.2. Mixed batteries

Mixed batteries are a collection of assessment tools that includes two or more measurement types, namely tasks, SJTs and indirect reports (self-reports from students, but also teacher, parent and peer reports). For this reason, usually mixed batteries have a larger scope of skills being assessed, covering multiple skills simultaneously with the potential benefit of offering complimentary validation for the same skill by different measurement tools. In other instances, although several skills are assessed through the battery, each individual skill is only assessed by one type of assessment tool. A clear advantage for the use of this collective tool lies with its scope and overlapping validation, which makes it an efficient method for assessing multiple SES (Walton et al., 2022[112]).

SELweb is a well-validated mixed battery assessment tool, which includes different tasks and SJTs. Covering various skills, two versions have been developed, one for early elementary school pupils (EE), from ages 4 to 9 (McKown, 2019[113]), and one for late elementary school students (LE), ages 9 to 12 (McKown, Russo-Ponsaran and Karls, 2023[114]). The ***EE SELweb*** version tests the skill **Emotion recognition**, through a task displaying facial emotion expressions which requires pupils to directly identify those emotions. Additionally, *EE SELweb* separately tests **Perspective-taking** and **Social problem-solving** skills, through the use of SJTs. In these, illustrated written vignettes are presented to the pupils and they are asked questions about the words and actions, as well as the true intentions of the characters in the story. They are also asked how they want the situation to turn out and which actions they would choose. Finally, EE pupils are tested for **Self-control** skills using a simple computerised task involving frustration and gratification delay, with the use of digital rockets and geometrical shapes.

The design is slightly different for LE students, under a principle of increasing difficulty according to their respective age. In ***LE SELweb*** students are tested for complex Emotion recognition, being presented with illustrated and narrated stories of complex scenarios which elicit mixed or complex emotions, such as pride, embarrassment or guilt. This is in clear contrast with the identification of simpler facial expressions of happiness or sadness that characterise the EE version, so we consider the assessment of complex Emotion recognition as, in fact, an assessment of **Perspective-taking** skills. Then, like the version for younger students, *LE SELweb* also tests for Perspective-taking and **Social problem-solving** skills using situational judgement tests. Using written vignettes that this time present students with more ambiguous and socially challenging situations, they are asked to make accurate inferences about the story characters' mental state, choose a preferred course of action and how would they deal with the consequences of those choices. Finally, LE students were tested for **Emotional control** skills, by written vignettes asking them to imagine undesirable emotions and which strategy would they use to deal with those emotional states effectively. While infective strategies were presented, such as "punching a pillow", effective emotion regulation strategies included actions such as "taking a few deep breaths", "walking away from a situation", or "thinking about the situation in a way that is not upsetting".

Both these mixed batteries have been validated in thousands of children, contemplating gender and ethnicity comparisons, with high reliability, including internal consistency and test-retest metrics, for all different tests used. Moreover, construct validity has been found for both versions and, in particular, *EE SELweb* results has also been positively tested for criterion validity by correlating with academic outcomes (McKown, 2019[113]; McKown et al., 2023[115]; McKown, Russo-Ponsaran and Karls, 2023[114]).

The ***MSCEIT*** [Mayer, Salovey, & Caruso, 2002] and the ***MSCEIT-YV*** (Rivers et al., 2012[116]) are batteries developed to assess Emotional Intelligence (EI) in adults, and in children and adolescents, respectively. The conceptualisation of EI is a highly debated topic in the scientific literature and the exact scope of abilities it captures is not consensual (Anglim et al., 2020[117]; Vaida and Opre, 2014[118]). The authors of this seminal battery have defined it as "the ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions" (Salovey and Mayer, 1990[119]). According to this definition, EI can be framed as complex psychological construct which encompasses multiple dimensions and skills that can be dispatched across the OECD framework of SES. In that sense, it is possible to decompose EI into Metacognition, Emotion recognition, Perspective-taking, Emotional control, and Social problem-solving skills.

The overall scale of EI is divided in four branches of abilities grouped in two EI areas (Experiential EI includes Perceiving emotions and Facilitating thought; Strategic EI includes Understanding emotions and Managing emotions). Each branch score, in turn, is made up of two individual tasks. ***MSCEIT-YV*** is slightly different from the adult version. One of its tasks involves identifying emotions in photographed faces of youth, covering **Emotion recognition** abilities. A different task complexifies what is requested from participants, by presenting definitions of emotions or describing social situations. Then, through a multiple-choice method, test-takers are asked to interpret the definitions, the social context, or the causes of emotional states in others, in order to identify the correct emotion term that fits the situation, which aligns both with **Metacognition** and **Perspective-taking**. Another task asks test-takers to imagine which physical sensations they associate with certain emotions, for example by describing a scenario and then asking how much the corresponding feeling of anger corresponds with each of these words: "hot, red, relaxed, or heavy". Such a task requires the ability to understand own's thoughts and emotions and create connections with subjective appreciation of other concepts, requiring **Metacognition** skills. Lastly, test-takers are asked to rank the effectiveness of alternative actions to address social and emotional situations involving other people, which taps into **Social problem-solving**.

The *MSCEIT* batteries remain the flagship test of EI in adults. Even if most research has focused on the adult version of *MSCEIT*, literature also supports that the *MSCEIT-YV* is both objectively scoreable and reliably measurable. Additionally, EI scores show convergent validity with similar measures and divergent validity with assessments measuring similar but distinct constructs (Peters, Kranzler and Rossen, 2009[120]; Rivers et al., 2012[116]). However, convergence between *MSCEIT* scores and other measurements of EI is still affected by the differences in the conceptualisation of EI and the subsequent mismatch between scales of different instruments (Windingstad et al., 2011[121]). Moreover, criterion validity has also been confirmed by correlations of between EI scores with positive outcomes, such as academic achievement, prosocial behaviour and well-being/health (Rivers et al., 2012[116]).

4.3. Digital games

Game-based assessments are a step forward in the way we assess behaviours, performances and skills, in children and adults. Through the generation of immersive, interactive and adaptive environments, which are engaging and often attempt to reproduce realistic scenarios, with storylines, dialogues and actions that mimic real-life interactions, videogames emerge as the new generation of assessment tools for SES (Kim and Ifenthaler, 2019[122]; Oranje et al., 2019[123]; Shute and Ke, 2012[124]). In our review, we collected more than 20 game-based assessments for SES, games which not only vary immensely in their design, narrative and gameplay mechanics, but also in how they extract data for assessing skills. In this paper, we distinguish games from tasks as the former having an imbued narrative, a story arc with character interactions that players must pursue in order to conclude the game. Alternatively, when a narrative is not present, we categorise games as having interactive elements where the player has some freedom to explore the virtual world. Tasks, on the other hand, display simpler interfaces and are absent in narrative, requiring a more direct and explicit set of actions from players, without story or exploratory elements. Unlike in the previous subsection describing task-based assessments, where descriptions and discussions were organised by skill to offer an ample presentation of a large number of tasks, this section will be more focused. Given the greater complexity each game entails, in the following paragraphs we will present only some examples of game-based assessments. This will offer a balance between an overview of the diversity of these immersive assessments and more detailed descriptions of how some of these games work to reach their goal.

The game-based assessments benefit from a great variety of narratives and visual designs which contribute to more immersive and engaging experiences for children and adolescents. Also, having the different skills tested in scenarios where tangible social and emotional elements can be incorporated, including with audiovisual elements, often based in human voice recordings, confers extra realism and, thus, increases ecological validity. The rich environments and simulations games can deliver are particularly useful to assess complex "hard-to-measure" skills, like co-operation or creativity, which traditional methods struggle to capture. Narrative-driven games also contribute to the paradigm of stealth assessment, where data is collected on children's performances and choices without the overt feeling they are being evaluated, which counters self-presentation biases and test anxiety, better assessing authentic behaviours. Beyond these aspects, game-based assessments can usually be easily and independently administered in laptops or tablets, without the need for one-on-one staff guidance and often with automatic scoring of students' behavioural choices and performances (Buckley et al., 2021[125]; Rafner et al., 2022[126]; Ren, 2019[10]).

On the negative side, visually and narratively complex games can be costly, due to the diversity of expertise and workload required to develop and polish all its aspects (Buckley et al., 2021[125]; Ren, 2019[10]). Furthermore, due to the investment in ecological complexity where characters and narratives are built to tell a specific story, these games often require a significant time to be administered while only assessing one individual skill. On the other hand, increased ecological complexity and realism of stimuli might create more noise in the data collected, making it harder to isolate variables and measure specific, clearly defined skills, which can taint the assessment of SES skills (Clauser, Margolis and Clauser, 2015[127]). Moreover, compared to more classical and established personality or skill measurements, many of these games are recently developed tools with not yet acceptable reliability, or sufficiently robust construct validity and criterion validity metrics, requiring improvement of some narrative or mechanical elements. Some games have shown partially acceptable psychometrics but are still limited regarding international comparability or

relation to external outcomes, such as mental health or academic performance, due to their recent development.

4.3.1. ZooU – Emotional control / Self-control / Empathy / Co-operation / Sociability

Although most games are developed with narratives and mechanisms focused on one discrete skill, some have been designed to explore complex social scenarios, so that multiple skills can be exercised, and thus evaluated, while playing the same game. Such is the case of ***ZooU*** (DeRosier, Craig and Sanchez, 2012[128]). Published in 2012, *ZooU* is a game targeted at elementary school students (3rd and 4th graders) that attempts to measure 6 different SES, corresponding to 5 of the skills in the OECD framework (Steponavičius, Gress-Wright and Linzarini, 2023[2]): **Emotional control**; **Self-control**; **Empathy**; **Co-operation**; and **Sociability**. In this game, children embody an avatar of a student in a virtual school where they learn to become zookeepers and must interact with other non-playable characters (NPCs), such as classmates, teachers and animals. With its narratives, the game is designed to create analogous social situations to those lived by children in their daily schooldays, which include engaging in conversations with others, performing fun playground activities, attending to classmates in need or attentively taking care of animals.

The choices children make, regarding their actions, the NPCs with whom they interact, their dialogue options, or the time they decide to spend in certain activities, are the core elements for the assessment of each SES. Regarding practical gameplay, children interact with objects and with NPCs by clicking on them, initiating pre-scripted action or dialogue options which can also be heard through actor-recorded audio, an element that helps the game become more immersive. How does *ZooU* measure different skills? In each individual scene, children encounter a specific social problem that needs to be solved. For example, in a scene assessing Self-control, the child needs to decide whether to feed the animals before going to recess or, specifically considering the dialogue options, the child can also decide whether to ask the teacher questions about the animals' names or to be immediately allowed to go to recess. The objects present throughout the game also constitute key elements for assessment, as they can offer critical information, such as a clipboard with instructions on how to feed the animals, or they can work as distractors which are not task related. Even though children are given freedom to explore the virtual world, their scores will decrease if they diverge from their main task for a long time. Much like Self-control, all other five skills have their own specificities in terms of the exact type of data that is extracted from the game and used for assessment. As the authors explain, in social scenes tapping into Self-control, Co-operation and Sociability, the scores attributed to children depend largely on their behaviours, such as the objects and people with which they interact and the time spent during, and between, certain behaviours. Alternatively, for skills such as Empathy the key data emerges from the sequence of dialogue choices when interacting with classmates and teachers NPCs.

Naturally, it is important that games collect multiple psychometric elements that help validate their findings. One way to achieve that is to compare the results extracted from the behavioural game data with indirect reports on the same skills, assessed either by the students themselves, or their parents, teachers and peers. In the case of *ZooU*, the researchers reported statistically significant correlations between the data obtained from the game and social skills measurements based on the teachers' reports on their students (DeRosier, Craig and Sanchez, 2012[128]). Additional validation for this game came from a subsequent study by the same group, validating *ZooU* with a population of Japanese students, while respecting the necessary language adaptations for appropriate cultural validity (Craig, Derosier and Watanabe, 2015[129]). In this study, Japanese children performed better than American children in Emotional control and Co-operation skills,

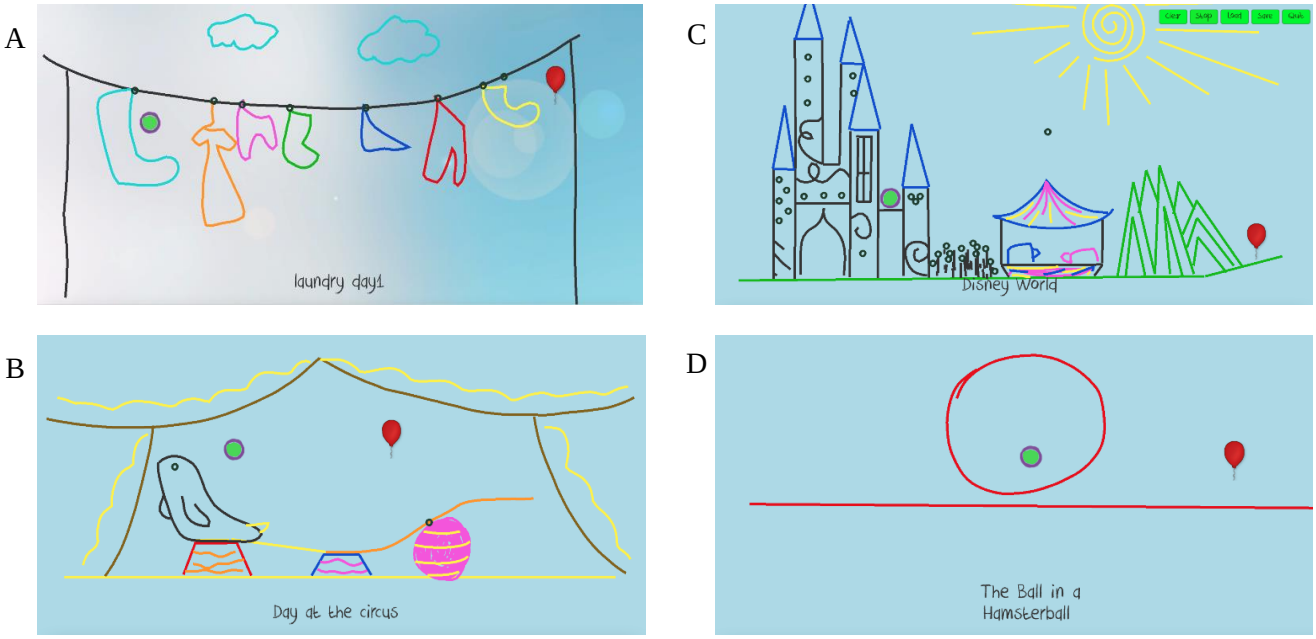
while the American students outperformed their Japanese counterparts in both Empathy and Sociability skills. This difference in social skills' performance observed through *ZooU*, reflected documented cultural differences between these countries, further validating this tool as an assessment able to capture cultural variations reinforcing its utility as a measurement tool for comparability. Further validation came from a more recent study associating the performance in SES with social competence outcomes, in terms of children's behaviours and academic performance in school environments (DeRosier and Thomas, 2018[130]). Researchers show that children with more developed SES, as measured by *ZooU*, displayed more prosocial behaviours and academic adjustment, "above and beyond demographic influences", when compared to colleagues who performed poorly in the game. This detailed description, covering both the game mechanisms and some validity metrics, serves as an example of how a game can be designed to specifically target individual skills, correlating the data with external psychometric assessments which attest for the quality of the game-based assessment being employed.

4.3.2. Physics Playground – Persistence / Creativity / Co-operation

Through very simple design and mechanisms, games like ***Physics Playground*** can benefit from a great deal of flexibility in the way they assess SES. Initially oriented towards testing **Persistence** (Ventura and Shute, 2013[131]), based on a principle of overcoming increasingly difficult tasks, this game has more recently been used to test entirely different skills, such as Creativity (Shute and Rahimi, 2021[132]) and Co-operation (Sun et al., 2022[133]). *Physics Playground* was built around the premise of teaching the basic principles of physics by having players draw different mechanisms, like ramps, levers and pendulums, to move certain objects to certain places, in order to complete puzzles and proceed to the next level (Shute, Ventura and Kim, 2013[134]). As these challenges are increasingly difficult, children must be persistent in their attempts, creating and perfecting different devices while relying on gravity and the laws of physics to do the rest of the work. Researchers have taken advantage of this design to measure Persistence, by collecting the time spent on each unsolved and solved problem by middle school students aged 13 to 15. The performance data extracted from the game correlated well with the results from a different task to measure Persistence, even after controlling for videogame experience and pretest physics knowledge, although it did not meet correlations with self-reported levels of persistence. This discrepancy was interpreted as self-reports on Persistence being inadequate measurements, since students might perceive their skill in a manner which does not necessarily correspond to their real behaviour (Ventura and Shute, 2013[131]).

As mentioned, ***Physics Playground*** can also be used to measure **Creativity** skills, when different sets of data are collected (Shute and Rahimi, 2021[132]). For example, researchers can assess how original a mechanical device to solve a puzzle is, by measuring how it differs from the other players' solutions, and also how flexible it is, measured by the design of multiple and effective physics mechanisms to solve the puzzle, rather than drawing simplistic ones. In order to prompt maximal Creativity, players are, in fact, incentivised to create the "most awesome" solution. The researchers also assessed players on their ability to design their own creative levels from scratch (Figure 4.5). Similarly to what was observed with the Persistence skill, Creativity measurements, as defined by performance in the game, correlated well with these students' performance in other creative tasks but not with their self-reports.

Figure 4.5. Screenshots from Physics Playground (Creativity)



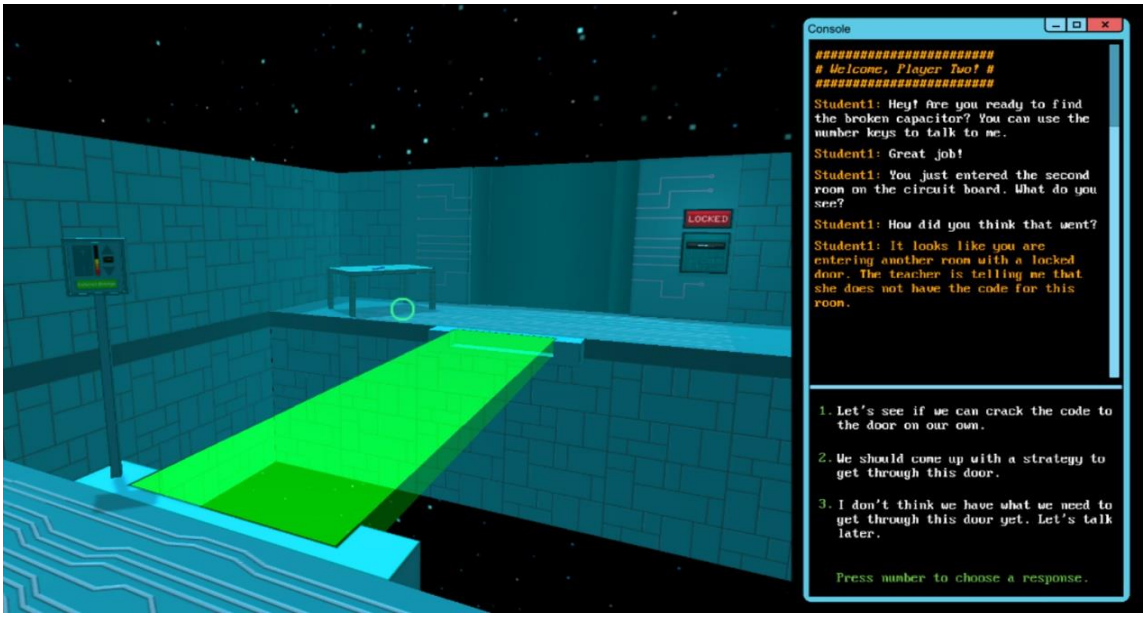
Note: In this version of Physics Playground, participants are prompted to use a level editor to generate their own creative designs. The creativity of these designs is assessed by their relevance, originality, aesthetics, humour/surprise, and elaboration. Solving the puzzles requires that the green ball hits the red balloon, by interacting with drawn structures and obeying the laws of physics. The participants enjoy complete freedom when choosing the themes and the titles of each of their designs. Images A-C were considered, by expert raters, as creative designs, whereas image D was considered not creative.
Source: Adapted from (Shute and Rahimi, 2021[132]), "Stealth assessment of creativity in a physics video game", *Computers in Human Behavior*, Vol. 116, http://dx.doi.org/110.1016/j.chb.2020.106647

Co-operation is another skill that the flexibility of ***Physics Playground*** design has allowed to be tested (Sun et al., 2022[133]). In this context, teams of three people work together to design effective mechanical solutions and solve the puzzles presented by the game. Working in groups of three, the players collaborate via videoconferencing in order to exchange ideas through spoken language rather than through chatbox messaging. One player is randomly assigned the role of the controller, responsible for the mouse interactions and the actual design, while the other two, the contributors, observe the gameplay live and provide contributions. Then, the role of controller changes so that all players have the opportunity to play the different roles. Since all team members received joint rewards in the game for their progress, there is a clear incentive to effective teamwork and collaborative behaviour, based on interdependence. In fact, under the umbrella of Co-operation, multiple behavioural dimensions are being tapped, from maintaining active and constructive communication throughout the gameplay and understanding other's perspectives, to encouraging others, solving possible conflicts, and negotiating solutions. In the case of this experiment, researchers observed convergent validity between the collective self-reported co-operative behaviour of the team members and the co-operative behaviour displayed while playing the game, which translated in better performance. Overall, *Physics Playground* is an example of how data from the same game can be flexibly collected to assess different SES, depending on how the gameplay mechanisms are designed and on the exact goal to which children are prompted.

4.3.3. Circuit Runner – Co-operation

As detailed above, *Physics Playground* assessed Co-operation by relying on the interaction between three students, playing different roles in turns, to achieve a common goal. Other games, such as ***Circuit Runner*** (Stoeffler et al., 2020[135]), have attempted to assess **Co-operation** by designing a narrative-driven world, where the player interacts and collaborates with virtual characters. In *Circuit Runner*, a game tested in middle school students (Polyak, von Davier and Peterschmidt, 2017[136]) and in adults (Stoeffler et al., 2020[135]), the interface is a 3D virtual world, where the player must enter a maze and the virtual agent narratively stays a base location, in possession of key information and resources for the game to proceed. This interdependence is crucial to navigate the map and overcome the challenges and, thus, requires Co-operation skills, or collaborative problem-solving skills, as the authors frame it. While navigating the maze, the players face problems, such as barriers or doors, that requires specific dialogue interactions or strategic transfers of power with the virtual agent, via a chatbox, for the solutions to be obtained (see Figure 4.6).

Figure 4.6. Screenshot from Circuit Runner



Note: In Circuit Runner, players interact with virtual agents and are prompted to collaborate with them via a chatbox in order to navigate a maze and solve challenges. These virtual agents are in possession of key information and resources which are indispensable to the game progression, requiring specific dialogue choices and transfers of power so that solutions are effective. The players' actions and dialogue choices can be extracted and analysed to inform about their collaborative skills.
Source: Adapted from (Stoeffler et al., 2020[135]), "Gamified performance assessment of collaborative problem solving skills", *Computers in Human Behavior*, Vol. 104, http://dx.doi.org/110.1016/j.chb.2019.05.033

It is the dialogue choices and the pattern of actions the player chooses to do based on the information provided by the virtual agent, that data can be collected on sub-skills, such as perspective-taking, goal-oriented behaviour or strategy, which collectively form a score on the co-operation skill of the player. Regarding validation metrics, the game shows high internal consistency and some relevant convergent validity with external measurements. This includes statistically significant positive correlations, even if low level, between the player's behaviours in the game and both their self-reports and performance in SJTs, assessing their co-operative behaviours (Stoeffler et al., 2020[135]).

4.3.4. Questions Worlds – Curiosity

Dealing with unexpected events and, thus, being open to different experiences, ideas and behaviours are leveraging qualities in a multicultural world where globalised communication and digitalisation have bridged people together and multiplied the opportunities for interaction and communion. Being curious about what surrounds us is the first step to deal with all that novelty and, to that end, ***Questions Worlds*** was designed to assess the skill **Curiosity** for children between the ages of 11 and 15, based on the rationale that more curious children tend to ask distinct types of questions, that are more specific (Tor and Gordon, 2020[137]). In the game, children encounter alien worlds that they can explore according to their own interests, by freely interacting with the aliens, indigenous plants or pieces of technology present in that world (see Figure 4.7).

Figure 4.7. Screenshots from Questions Worlds

Note: In Questions Worlds, students visit five different alien worlds, each with new environments, creatures, plants and objects. They are prompted to click on these different agents, which pops up pre-set questions they can ask, in order to learn more about them. These questions include "How does it work?" or "What is it made of?" and, importantly, students are limited by time or by the type of question they can ask in any given world. The rationale for assessment is that the sort of questions students decide to ask reflects their curiosity skills. Source: Adapted from (Tor and Gordon, 2020[137]), "Digital interactive quantitative curiosity assessment tool: Questions worlds", *International Journal of Information and Education Technology*, Vol. 10/8, pp. 614-621, http://dx.doi.org/10.18178/ijiet.2020.10.8.1433

When selecting these objects, children can then choose which questions to ask, such as "How does it work?"; "What is it made of?" or "Why is it here?". The objects and respective questions are part of a specific underlying story arc that, then, leads to the appearance subsequent questions children can ask or new objects with which children can interact. A critical element of the game is that children are given specific limitations which both incentivise Curiosity and condition them to steer it in a productive way. For example, players face different restrains in the five different worlds they visit, sometimes they are given only 60 seconds to explore the world and its objects, other times they can only ask 5 questions or ask multiple questions but of a single type (only questions starting with "why" or "how", but not both). This design prompts children to be judicious in the sort of questions they ask and stimulates an active and productive Curiosity. Regarding how the skill Curiosity is assessed, the researchers established different parameters for each question available in game, such as Breath (number of answers to that question), Depth (number of new questions which can arise from the given answer) and Specificity (how many other

questions lead to the same answer). This allows to value questions differently and generate a measure of Curiosity. A critical element of the game mechanics was the decision not to add any external reward, like a trophy or point, as to guarantee that Curiosity was driven by a sense of internal reward, an intrinsic drive to know regardless of compensation.

Interestingly, the researchers (Tor and Gordon, 2020[137]) found that level of Curiosity of the students was independent of their age, gender, and perceived intelligence. However, one clear limitation of this study is its demographic population, a group of "talented and gifted middle schoolers", which, as the authors recognise, are expected to have high levels of Curiosity, which conditions the generalisation of any conclusions to the wider student population. This study did not contrast the game performance with self-reports on perceived Curiosity but collected questionnaires from teachers as an external validation. A correlation was found between Specificity of the questions asked, as assessed by the game, and the teacher's reports of their students' Curiosity. Within this group of high achieving students, those who are perceived by teachers as highly curious chose highly specific questions in the later stages of the game.

4.3.5. VESIP – Perspective-taking / Social problem-solving / Self-efficacy

Navigating the social world in a successful way requires the ability to understand those around us, their particular perspectives, thoughts and emotions, which can be ambiguous or even contradictory. In social scenarios, we are often asked or expected to appease tensions, help solve a conflict or choose the most prosocial course of action (Doesum, Van Lange and Van Lange, 2013[138]). Thus, it is critical to "read the room" or, in other words, to recognise, disentangle and appreciate the different contexts that underlie each person's intentions and behaviours, so that sound judgements can be made. This ability for **Social problem-solving** has been explored in the game ***VESIP*** (Virtual Environment for Social Information Processing), aimed at children between the ages of 8 and 12 (Russo-Ponsaran et al., 2018[139]). In the game, children embody a fully customisable avatar and are confronted with challenging social situations, which are presented in a 3D virtual school environment where other virtual child characters interact, in settings such a classroom, a cafeteria or a playground (see Figure 4.8). Children face ten different social situations, which include ambiguous provocation, bullying, compromise or social initiation scenarios, where the other characters are voiced by age-appropriate actors to provide extra realism and increase engagement. When presented with these different social scenarios, players can choose between different options, regarding how they perceive the problem they observe and the intent they attribute to the actions of other children in the scene (taps into **Perspective-taking**). Then, they must choose how they want the situation to turn out, ranging from antisocial to more positive, prosocial behaviours, which assesses their Social problem-solving skills. For example, in a bullying scenario, they can decide, in ascending order of prosocial behaviour, whether they prefer retribution, avoidance, mediation from others or a direct and constructive conversation, to deal with the situation. They are also asked the extent to which they believe then can actually go through with the course of action they chose, which taps into the skill **Self-efficacy**, in a scale from "not at all sure" to "very sure".

Figure 4.8. Screenshot from VESIP

Note: In VESIP, players embody the avatar of a student in a school, who engages in several socially challenging situations throughout two virtual school days. Players are asked to interpret the intention of other characters, decide on a course of action to resolve conflicts and state how confident they feel about their own ability to pursue these solutions. Students indicate their responses either through multiple choice or slider-scale response options. Each option plays out as an animated visualization of the option. Each dialogue choice or course of action is associated with different scores for social information processing.
Source: Adapted from (Russo-Ponsaran et al., 2021[140]), "Psychometric properties of Virtual Environment for Social Information Processing, a social information processing simulation assessment for children", *Social Development*, Vol. 30/3, pp. 615-640, http://dx.doi.org/10.1111/sode.12512

For analysis, children's performance is scored according to their ability to perceive the actions and intentions of others correctly and the preference for prosocial behaviours when faced with challenging situation, providing an overall score for their Social problem-solving abilities. Regarding the game's psychometrics, internal consistency and test-retest reliability were considered good. Moreover, convergent validity was also positively verified by the correlations between the performance in the game and the results from other measurements on Social problem-solving skills, as well as from teacher reports on students' social behaviours. Additionally, researchers found a positive correlation between Social problem-solving performance in the game and better academic competences, while noting that estimations from teachers were used to indicate academic competences rather than more objective academic test scores (Russo-Ponsaran et al., 2021[140]).

4.3.6. Adaptable games and future directions

Some games do not fit the typical paradigm for assessing SES, while they nevertheless offer interesting concepts or mechanics which can be used as inspiration, or even directly adapted, for more applicable purposes. ***Simoland*** is such an example. It is a simple 2D game initially designed to assess real-life relationship satisfaction (Schönbrodt and Asendorpf, 2011[141]), but that has been subsequently modified to assess attachment anxiety and attachment avoidance (Schönbrodt and Asendorpf, 2012[142]), as well as feelings of loneliness (Luhmann et al., 2015[143]). Such an expansion of evaluated psychological dimensions is demonstrative of the game's significant level of adaptability in its design and mechanics. In this game, the player embodies a blob, a very simple organism, and is prompted to interact with a similar creature, who, in the context of the game, represents a romantic partner. Towards this virtual partner, players can select various actions and verbal requests, whenever they desire and how often they desire. These options are not dichotomous, so that players can explore a great variability of behaviours, contributing to

a more realistic and spontaneous setting, as the partner character also reacts accordingly to each action. Other modifications of the game have introduced more complex narrative elements, such as separation scenes, where the partner leaves for an indefinite amount of time, or conflict scenes, where the partner accuses the player of infidelity. All this to test how the players modify their behaviours and interactions with the partner character, therefore testing anxiety and loneliness feelings and attachment patterns. In fact, researchers have found positive correlations between psychological traits and in-game behaviours, for example, by observing that self-reported anxious individuals displayed more negative emotional behaviours when interacting with their virtual partners in the different socially tense scenarios (Schönbrodt and Asendorpf, 2012[142]).

As with other games, *Simoland* is not an operational assessment tool for children in its current design since the narrative presentation is targeted at people in current real-life relationships. However, as the modifications briefly described above show, due to its simple design and customisable dialogue and action options, the game can be easily adapted to younger populations, using appropriate narratives. For example, by presenting the other creature as a friend or a new colleague in the school, and not as a partner, the skill Sociability can be assessed by observing whether the child initiates social contact and how it chooses to proceed with the conversation. One can also create specific narratives where the other character is in need of help, to assess Empathy, or create certain visual or behavioural modifications in the other character so that Tolerance can be measured.

The several digital games presented in this section provide a representative overview of the inventiveness and diversity of conceptual designs and gameplay mechanics that are currently available to assess SES. Collectively, they offer an immersive, contextualised, interactive and engaging experience through which more authentic behaviours manifest. As some of these games are new and experimental, research efforts must continue to consolidate them as valid and reliable assessment tools. Nonetheless, this review shows the last decade of research has undoubtedly harnessed the potential of digital technologies to develop the next generation of innovative assessment tools.

5. New technological approaches for assessment

5.1. Biophysiological measures

Data on human behaviours and abilities can also be obtained through other technology-based methods, which measure biological and physiological information. Such biophysiological measures include data on heart rate activity, cortisol levels in the saliva and in the hair, eye movements, skin conductance, brain activity, which can be collected to inform of variations in the internal psychological and emotional states of individuals. This biophysiological data can then be used as a complementary method for more directly assessing individuals' behaviours and skills, as this information is spontaneous and derives from implicit processes people cannot easily control, thus also limiting social desirability biases for example (Abrahams et al., 2019[7]).

The use of assessment tools based on these measurement technologies has expanded and diversified as they have become more affordable and portable. However, even though accessibility has increased, the complexity and required expertise associated with the setup, calibration and synchronisation of these instruments have remained, challenging their implementation at large scale. Another challenge lies with the ecological validity of these instruments or, in other words, the application of these technologies in authentic settings where imponderable variables abound, distant from the tightly controlled laboratory settings, where most of these technologies have been developed and applied. Other authors

have shown that physiological reactions in laboratory settings greatly differ from those occurring in real-life contexts, namely in situations dealing with stress and emotional stimuli. Therefore, a systematic application of these technologies outside the laboratory requires solid validation before conclusions regarding psychological states can be firmly taken. Moreover, a practical concern associated with the use of biophysical technologies is their inherent noise in the data, which then requires a significant number of trials so that a strong signal can be extracted. Such need for repeated measurements implicates additional time requirements and tests the ability and patience of young children to remain engaged with the experimental tasks. Finally, the need to obtain the appropriate consent for health data collection from children might constitute an additional practical difficulty. All these issues further limit the feasibility of an at-scale collection of biophysiological data, particularly in young populations (Dahlstrom-Hakki, Asbell-Clarke and Rowe, 2019[144]; Larradet et al., 2020[145]).

We recognise There exists of biodata relying on electrodermal activity, galvanic skin response and facial expressions, as well as various methods to directly test brain activity, such as functional near-infrared spectroscopy (fNIRS), functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) (Nebel and Ninaus, 2019[146]). Moreover, the use of wearable smart devices should also be noted, which introduces practicality and portability to the same techniques mentioned above, collecting measurements on heart rate or brain activity, for example. These electronics are equipped with wireless sensors and can be integrated into clothes or accessories, such as smart glasses, smartwatches and necklaces, but also head bands, arm bands or chest bands. They are particularly applicable to the assessment of anxiety, stress responses and stress recovery, thus related to skills like Stress Resistance or Emotional control (González Ramírez et al., 2023[147]; Hickey et al., 2021[148]). However, given the multitude of available technologies, some yet exploratory or with limited practical applicability outside the laboratory, the next paragraphs will focus only on a few examples. Thus, three methods for measuring biophysiological data will be introduced and discussed, namely using heart activity, eye tracking and cortisol levels, with examples of games which have integrated them to assess SES.

5.1.1. Heart activity

Data on heart activity is one of the ways biophysiological measures can inform us about the internal states of individuals. Even though studies have found blood pressure variations to be good predictors of behavioural perseverance measurements (Chauntry, Williams and Whittaker, 2019[149]), heart rate variability (HRV) is usually singled out as the most reliable indicator (Dormal, Vermeulen and Mejias, 2020[150]). The concept of heart rate relates to the number of heart beats per minute, whereas HRV, on the other hand, is a more dynamic measure of the time interval variations occurring between consecutive heart beats. The heart pumping mechanisms are complex and in constant change, in order to allow for rapid cardiovascular adjustments to any internal or external stimuli. HRV can thus be considered a valid indicator for assessing internal states and, in particular, the workings of the autonomic nervous system. Critically, a dynamic HRV with a wide amplitude has been correlated with more efficient emotional regulation mechanisms, stress coping and resilience. In practice, this means a cardiovascular system which can efficiently reduce or accelerate heart rate in response to changing environments. In contrast, those with limited variations in HRV have been associated with stress vulnerability and even with more severe psychiatric disorders, such as depression or generalised anxiety (Dormal, Vermeulen and Mejias, 2020[150]).

Some authors have directly associated HRV with the concept of resilience, defining resilience a multidimensional construct and a dynamic mental process where an individual displays a functional adaptation to stressful stimuli in order to preserve its integrity and stability. Indeed, these authors report diverse evidence that an ample HRV is consistently associated with a better capability to regulate and adapt psychophysiological responses. In particular, higher HRV was associated with better abilities in reappraisal and emotional regulation strategies when people were confronted with stress-inducing tasks. Additionally, a dynamic HRV was also related to more functional neurocognitive strategies when dealing with self-control tasks involving stressful stimuli. HRV could, thus, be considered a reliable biomarker for mental resilience, stress resistance and emotional regulation (Perna et al., 2020[151]). HRV data can also be used in a more advanced way in order to assess and improve stress resistance abilities, with the application of biofeedback (Dormal, Vermeulen and Mejias, 2020[150]). Through biofeedback, a non-invasive technique, individuals can be informed about their physiological state and, by making a dynamic use of that immediate information, they can exert voluntary control over their own autonomic activity. This way, they can regulate their breathing and enhance heart-brain synchronisation, which leads to a more efficient mental health state and an improvement in performance.

Some games have thoroughly integrated such biofeedback mechanisms into their gameplay and, despite not being appropriate, in their current design and/or mechanics, to assess SES in children, they offer inventive ideas which can also be adapted for future assessments. Such is the case of ***Nevermind***, a game with horror elements designed to test **Emotion regulation** in adults, using a principle of heart rate biofeedback (Lobel et al., 2016[152]). In this game, using first-person perspective, players embody a therapist from the future, who enters the minds of patients suffering from psychological distress and the goal is to travel through the mind of the patient and complete the level by finding the cause of the trauma or anxiety. The distinct element of this game lies with its biofeedback mechanism, since the horror-themed visual setting of the game is designed to be disturbing to players, causing their heart rate to increase, which is used as a physiology input in the game itself. Therefore, as players get stressed and their heart rate increases, the game system reacts by becoming more hostile and difficult, namely by matching the player's stress level with a screen increasingly obscured by static or having elements in the environment attacking the player's character. The way to overcome this hostility and proceed with solving the puzzles in the game is to actively exercise emotional regulation, remaining calm in the face of these horror-themed stimuli, so the player's own heart rate is lowered which, in turn, forces the game to readjust and become less hostile to the player. In fact, beyond a simple measurement of HRV, researchers considered more subtle indicators of emotional regulation, by considering the speed of recovery from high to lower cardiac frequencies or the frequency of strong drops in heart rate variability. These can be more informative about each individual's reactions to the stimuli and their own strategies to overcome the hostility in the game. The researchers also found a correlation between an in-game emotion regulation strategy of reappraisal and real-life behaviours, where players who actively employ this emotional strategy in their daily lives were more able to counteract the game's hostile environments (Lobel et al., 2016[152]). Such finding offers an important hint for construct validity. For other strategies, such as seeking active resolution, rumination or suppression, no behavioural correlations were found.

As it is self-evident, this game is not appropriate for children nor is it enough developed to constitute a solid assessment tool for emotional regulation, as other psychometric validations are missing (for these reasons, it is not part of our main review). However, the principle of biofeedback, having the game incorporating and actively reacting to the player's internal states, prompting the player to readjust in turn, is an immersive feature that contributes to a sense of realism where emotional skills can be maximally measured.

Assuming the obvious need for adjustments so that content is appropriate for children and adolescents, without such disturbing stimuli, future game-based assessments could adapt this biofeedback principle to induce ethically acceptable levels of frustration or stress. The ability of children to exert control over these negative states, and thus influence the game, could then be used to measure skills such as Emotional control, yes, but also Self-control in less emotionally charged settings, or Stress resistance in time-pressured challenges, for example.

5.1.2. Eye tracking

Eye movements and fixations can be considered indicators of cognitive and emotional processes underlying spatial and visual analysis, attention allocation, engagement, as well as metacognition or emotional states, such as anxiety or fear. In more detail, eye tracking strategies involve collecting data on different dimensions of gaze, such as eye position and eye movement, which correspond, respectively, to measures on fixation and saccade. As suggested by the name, fixations represent very brief pauses in eye movement, allowing for the eye to properly percept a certain object. A practical example of its use includes associating variations in fixation duration of participants with their focus on more relevant visual information and, with that, better performance in a certain laboratory task. Saccades, on the other hand, refer to rapid eye movements between different fixation points. For example, eye tracking headsets which can detect eye movement and fixation points during certain tasks can be used to infer information about processing difficulty or attention, which can, in turn, be a proxy for motivation and engagement. These eye tracking technologies have been also applied to children to understand cognitive abilities and learning processes and have proved able to capture subtle changes in gaze occurring over the course of a brief training session. An advantage of this method is that eye tracking can be simultaneously informative about what is being observed and when, with a high level of spatial and temporal accuracy. Even though these devices have become increasingly portable, data synchronisation still limits their application outside laboratory-controlled conditions. It is worth mentioning that eye tracking technologies have been also applied to children to understand cognitive abilities and learning processes and have shown this technology captures subtle changes in gaze which happens over the course of a brief training session (Dahlstrom-Hakki, Asbell-Clarke and Rowe, 2019[144]; Fletcher-Watson and Hampton, 2018[153]; Rappa et al., 2022[154]).

Interesting experiments with eye tracking have included a study exposing participants to affective image and video stimuli and using their gaze behavioural responses to infer and predict personality traits, by using machine learning algorithms (Berkovsky et al., 2019[155]). Eye activity included multiple sub-indicators, such as the aforementioned number of eye blinks, saccades and fixations, but also pupillary dilation measures. With the important limitation of a small sample size, researchers reported that using these ocular physiological responses to affective visual stimuli alone all personality traits were predicted with an accuracy between 80% and 90%. These personality traits include those present in the Big Five and the HEXACO models of personality, namely Agreeableness, Conscientiousness, Extraversion, Honesty, Resiliency and Openness (Berkovsky et al., 2019[155]).

Eye tracking has also employed to assess SES. It should be mentioned that a great number of studies using eye tracking to predict SES focus on Emotion Recognition, and do so by focusing on populations suffering from neuropsychiatric disorders, such as autism spectrum disorder (Black et al., 2017[156]; Wieckowski and White, 2017[157]). Beyond that, researchers have also found a relation between Curiosity and eye movements, albeit with a very small sample of only women. In this experiment, people who self-reported to be more

curious showed different eye movement patterns when engaging with reading materials, namely by displaying different gaze dynamics when anticipating new knowledge (Baranes, Oudeyer and Gottlieb, 2015[158]). Similar findings were observed for the relationship between Creativity and eye movements, by having one hundred adult participants solving a figural creativity test while wearing eye tracking devices. The researchers found that the more fixations recorded on particular areas of interest, and the longer those fixations, the higher the participants' scores on the creativity test were. Thus, simple information on eye movements and fixations was able to predict the scores obtained by participants in a creative task (Jankowska et al., 2018[159]).

Eye tracking technologies have been incorporated into game-based assessments to provide additional information on players' internal states and behaviours, including specific skills. An example is ***Crystal Island***, a game initially designed to teach middle school students about varied science content, as well as scientific reasoning and problem-solving abilities in an engaging way (Lester et al., 2014[160]). Since its initial development, research efforts have multiplied the number of variations around the game and of experiments to understand the learning process underlying the exploration of this world, with studies mostly in young adult populations. In one of these variations, researchers created a complex narrative regarding a mysterious disease which has spread throughout this fictional island (Taub et al., 2017[161]). In this exploratory 3D game, the player is tasked to collect clues in order to uncover the source of the disease, by exploring the island, visiting different buildings, talking to people and interacting with objects. Critically, these objects include papers and posters, sometimes with complex texts, and food items which can be tested in the laboratory, all to understand the microbiological source of the disease. After collecting the information they had to fill out matrices with questions to reach a conclusion. To assess the **Metacognition** skill of the players, the game tracked the players' choices regarding the number and the specific virtual books they choose to interact and read before giving their answers. Importantly, they complement this data by incorporating eye tracking mechanisms to collect data on specific fixation points and fixation durations associated with the reading material necessary to solve the mystery. This way, the exact information players look for when reading the books and papers and portions they choose to reflect on could be verified.

Researchers found that participants who have higher fixation periods both when reading the books for information and when filling out the matrices with questions show poorer performance. In turn, lower proportions of fixations and going back to each book more frequently was associated with better performance, when having to conclude about the origins of the disease. The researchers' interpretation of these results was that participants who decided to spend less time reading the entire book, and instead strategically opted to scoop for the relevant information needed to complete the matrices, show better metacognitive abilities. So, going back and forth between the questions and the source of information needed to answer them, quickly identifying the more relevant pieces of text, reflects a prioritising of quality over quantity, and is framed as a more efficient metacognitive ability. The advantage of this game in combining the assessment of metacognition skills in a challenging and immersive learning context has led to various other papers exploring additional dimensions of metacognition and agency, by employing both eye tracking and facial expression analysis (Dever and Azevedo, 2019[162]; Dever et al., 2020[163]; Emerson et al., 2020[164])

5.1.3. Cortisol

Cortisol is a hormone physiologically associated with stress and stress responses, usually rising in reaction to various psychological and physical stressors. It can thus be used a

biological marker for different internal skills such as stress resilience or emotional control. In fact, ineffective regulation of cortisol responses to stress or emotional pressure shows correlations with lower levels of emotional well-being (de Vries, van de Weijer and Bartels, 2022[165]; James et al., 2023[166]; Walker et al., 2017[167]). Additionally, it has been suggested cortisol can indeed be used as biomarker for psychiatric disorders such as anxiety or depression, including in young populations (Špiljak et al., 2022[168]). Although cortisol runs in the blood, a major advantage of this method is that its collection can be made using non-invasive sampling, with the resort to salivary cortisol, which correlates well with the correspondent blood levels (Hellhammer, Wüst and Kudielka, 2009[169]). However, it should be pointed out that, even though higher levels of salivary cortisol are usually associated with poorer emotional and behavioural responses, there is not a linear correlation between these two variables and the associations are far more complex. As such, other relevant indicators include awakening cortisol levels and daily slopes, which allow a better mapping of the correlations with the physiological state of the individuals. It is important to note that cortisol levels are also subject to secretion cycles dependant on internal and individual circadian patterns[7], unrelated to the impact of stressors, which can confound conclusions regarding the effect of stressful stimuli. Additionally, individual variations are also affected by age and sex, as well as eating, sleep and physical activity factors, all of which must be considered when planning the most reliable and harmonised periods for sample collection, in order to guarantee consistency across individuals (Golub et al., 2019[170]; Keil, 2012[171]).

Some research groups have also looked at the comparison between salivary cortisol and cortisol levels in hair follicles, that in turn reflect accumulated cortisol over periods of weeks or months, and found correlations between these two indicators to be inconsistent. Although hair follicle cortisol offers the advantage of giving a larger picture of cortisol secretion profiles of individuals and how they manage stress over time, it does not allow to test stress responses to acute stress-inducing events or stimuli like salivary cortisol does (Golub et al., 2019[170]; Joseph, Jiang and Zilioli, 2021[172]; Zhang et al., 2018[173]). Constituting a rather simple collecting procedure while allowing for multiple and consecutive sample collections, salivary cortisol has been widely used in both adult and child research to understand patterns of cortisol secretion in response to stress-inducing laboratory experiments but also in non-clinical environments. Stress-inducing stimuli usually activate the neuroendocrine system to secrete cortisol, with its peak levels arising 15 to 30 minutes after the initial activation, a time dimension which should be considered when collecting samples for analysis. Another positive aspect for its feasibility is that salivary cortisol samples remain stable and reliable after collection for up to 7 days of storage, with no need for freezing (Katz et al., 2016[174]; Keil, 2012[171]).

Together, the different biophysiological measurement techniques presented in this section, that can be incorporated into tasks and games, offer multiple avenues for innovation and for enriching the data associated with SES assessment. However, they also pose new challenges, regarding the limited scope of skills they currently assess, the costs associated with some of the equipment, the complexity inherent to biodata for SES inferencing, and the limited applicability of these technologies for large-scale implementation.

[7] Circadian rhythms are any periodic variation in physiological or behavioural activity that repeats at approximately 24-hour intervals (Source: American Psychological Association, Dictionary of Psychology)

5.2. Virtual reality and augmented reality

The technologies sustaining Virtual Reality (VR) and Augmented Reality (AR) have greatly expanded over the last decade and the increasingly affordable costs associated with these tools have paved way for higher accessibility and more diverse applications, both for entertaining and learning purposes. In educational contexts, for instance, the immersive and interactive worlds these technologies offer to users allow them to explore virtual field trips or visually manipulate 3D scientific concepts like human cells or the solar system. These interactive learning experiences, otherwise unattainable in the real world or limited to two-dimensional representations on paper or screen, leverage the engagement and motivation in students (AlGerafi et al., 2023[175]; Papanastasiou et al., 2019[176]).

VR is a technology which immerses people in a completely virtual and digital 3D environment, whereas AR fuses digital information with the physical environment individuals see, allowing for a fluid interaction between real and digital elements, hence augmenting the real-world experience. In education contexts, AR allows students to interact with observable objects in their classroom which are blended with digital annotations or animations, amplifying real-world visualisation with contextual information. VR, on the other hand, fully immerses people in environments in which variables are fully controlled to display a very precise experience. Together, these two technologies contribute to diversify learning opportunities for students, in more engaging, immersive and creative ways (AlGerafi et al., 2023[175]; Papanastasiou et al., 2019[176]).

The literature review has shown the educational research on VR and AR is much more prolific regarding the use of these technologies for learning rather than for assessment. However, these two dimensions are not mutually exclusive, as the learning interventions implemented in VR often also comprise an assessment component, often formative assessment, suggesting the possibility of a relatively easy adaption. Therefore, the use of VR and AR in learning interventions will be briefly discussed here. A recent metanalysis has demonstrated the use of VR in education results in better learning outcomes for both knowledge and skill development, when compared with the impact of non-immersive digital technologies or even with typical school lectures. Importantly, the authors found this technology particularly benefits primary and lower secondary students, when compared to post-secondary school learners (Wu, Yu and Gu, 2020[177]). Other work has corroborated the advantage of immersive VR for learning outcomes, especially for highly complex scientific content and problems demanding more visual-spatial cognitive demands (Hamilton et al., 2021[178]). Similarly, a large body of research has validated that the inclusion of AR in pedagogical practices, across different countries and school levels, also leads to improved learning outcomes and high levels of student engagement (Lampropoulos et al., 2022[179]). On the other hand, a metanalysis looking in particular at the use of VR for social skills development found that, even though VR shows effectiveness, it is not necessarily better than non-immersive digital technologies (Howard and Gutworth, 2020[180]). Such finding, however, might only reflect that existing games and programmes do not yet fully explore the potential of VR. Since this technology is under constant development, both from graphical and interactive perspectives, future virtual training programs will perhaps prove more effective in developing social skills.

One of the main advantages of the immersiveness of VR is precisely to foster a greater "sense of presence", as it is often described, which better mimics real-world experiences (Wilkinson, Brantley and Feng, 2021[181]). For example, research has shown immersive environments can more effectively stimulate cognitive function, resulting in improved memory performance when performing tasks in VR rather using classical digital tools (Ventura et al., 2019[182]). In the context of this work, it is particularly useful to understand

the potential benefit of these technologies in relation to SES, although evidence is not as abundant as the one that can be found for the digital tasks and games, discussed in previous sections. Many studies have also taken advantage of VR immersiveness to improve the ecological validity of tasks in research contexts. VR allows the presentation of 360-degree realistic audio-visual stimuli, that can be shown to participants in dynamic ways, allowing them to react to these in an interactive way rather than being passive responders to static stimuli. Moreover, the realistic all-around world that VR can evoke allows for the complex stimuli to be contextually embedded in an environment designed to provide extra social information, which conditions the players' actions and interpretations, as it happens in real life (Parsons, 2015[183]).

Some research has focused on enhancing audiovisual stimuli in VR in search of a more ecologically relevant emotion elicitation paradigm, for a better understanding of human affection. A recent literature review has identified a major increase in publications relating VR and emotion since 2015 and has pointed to the use of VR to heighten the emotional salience of emotions such as anxiety, fear, stress or arousal (Marín-Morales et al., 2020[184]). The effect of these emotions is assessed through individual or combined biophysiological techniques integrated into the VR experience, such as HRV, electrodermal activity or eye tracking. However, this literature review shows that even though there are dozens of studies studying emotion in VR contexts, as of now very few attempted to find validation by contrasting results between in VR and in non-immersive environments. When validation was attempted, results point toward immersive VR eliciting more authentic emotional states, confirmed through self-reports and physiological biomarkers correlations (Higuera-Trujillo, López-Tarruella Maldonado and Llinares Millán, 2017[185]; Marín-Morales et al., 2018[186]). Two recent studies also showed differential brain activity when emotions are elicited in VR, using EEG techniques. People who were immersed in VR showed significantly stronger neural activations to both positive and negative emotions, than when those emotions were elicited by observing 2D screens, especially in brain areas related to emotional and sensory information processing (Schubring et al., 2020[187]; Xie et al., 2023[188]).

A few tasks and games have made use of VR technologies to assess SES. A group (Geraets et al., 2021[189]) has used **Emotion Recognition** tasks, coupled with eye tracking, to assess the ability to identify specific emotions, contrasting photos, videos with real actors, and immersive VR environments with virtual characters, as the sources of the stimuli. Of note, Emotion recognition success rates were similar for the three variants, with VR environments leading to better recognition of emotions such as anger and surprise, when compared to the videos. To assess **Perspective-taking** and **Empathy**, another research group (Parra Vargas et al., 2022[190]) immersed adult participants in a VR environment mimicking social workplace contexts, where players embodied the role of a new worker in a company. Integrating different technologies, the researchers used VR for immersiveness, while employing complex eye-tracking analysis embedded into the VR headset and using machine learning to analyse behavioural data derived from decisions taken in the game. This physiological and behavioural data was then contrasted with self-reports for both skills, and eye-tracking data revealed to be the strongest predictor of both Perspective-taking and Empathy levels, allowing for a stronger discriminant power when compared with the decision-making behavioural data. Gaze is, thus, a spontaneous and reliable way to infer about people's abilities to identify, distinguish and interpret the intentions and emotions of others.

Another example is ***Athenea***, a VR-based game which attempts to assess **Self-esteem** and **Self-efficacy**, amongst other psychological dimensions (Giglioli et al., 2021[191]). It immerses players in narrative-driven story arc where they embody an astronaut in a spaceship, trying to find new habitable planets. In this setting, players interact with other

virtual characters, who have pre-established personality traits and competences occupying explicit roles in the ship. The game has multiple situations, each composed of several episodes and corresponding tasks. Crucially, at the end of each situation, the players are asked to assess how they feel about themselves, by also observing how the other virtual members of the team report about their emotional state, as to assess Self-esteem. Self-efficacy, on the other hand, is assessed by asking players to evaluate their perception of efficacy regarding their own game performance, contrasting this data with actual performance indicators, such as reaction times or number of completed tasks. Researchers found the in-game responses were strong predictors of the self-reported Self-efficacy but low predictors of self-reported Self-esteem.

In conclusion, VR and AR technologies increase engagement, can improve learning outcomes and enhance cognitive and emotional engagement by presenting audiovisual stimuli in contextualised environments where higher immersion contributes to greater ecological validity. However, although affordability has increased, challenges with costly equipment that guarantee equitable access are still relevant. The issue with costs and feasibility is especially pertinent when considering VR and AR implementation for thousands of students in school settings, in order to enforce large-scale assessments. Moreover, structural problems of using these technologies, such as cybersickness or the high cognitive load required to familiarise oneself with the equipment and the immersive digital stimuli, still require attention (AlGerafi et al., 2023[175]; Rappa et al., 2022[154]). Also, specifically regarding the assessment of SES, no clear studies seem to compare the difference of a game-based assessment, in its VR and non-VR forms.

Therefore, it is at this point difficult to conclude whether behaviours in immersive virtual environments constitute better representations of daily real-life actions and choices, when compared to the behaviours in simpler game-based assessments, which can be played on a two-dimensional screen. Benefiting from an intense technological advancement, VR and AR technologies hold great promise for the coming years, however current limitations and less consolidated scientific validity still condition their use to assess SES.

5.3. Artificial intelligence applications

The integration of AI into the assessment of SES could represent a significant leap forward from traditional methods. Many major recent breakthroughs in data-driven AI technologies were enabled by advances in Deep Learning. These breakthroughs in various AI fields, including machine learning, natural language processing, computer vision, robotics and knowledge representation and reasoning offer innovative ways to analyse personal characteristics (Tuomi, 2022[19]). These technologies analyse complex verbal and written communication, facial expressions, and body language to identify social cues and emotions, and attempt to infer emotional states, personality traits, or SES (Beyan et al., 2021[17]). In the context of assessments, three main AI use cases can be distinguished:

- In **assessment design**, generative AI technologies can assist in creating interactive and immersive tasks and games, by producing text, audio, images, or even video. This can make assessments more responsive to user input and more realistic, for example, through chatbots and avatars, but also to reduce costs and time related to the assessment development.
- In **assessment analytics**, predictive AI can be used to infer emotional states, skills, or personality traits, from data collected in dedicated assessments.
- AI can be used for **skills inferencing**, which involves inferring skills from vast and unstructured data sources that may not have been initially designed for that purpose.

This is an emerging category, and although a growing number of commercial providers are already offering such services (for example, Beamery, Techwolf, Empath, or Workera), it is still a very early field and may not be mature and developed for SES measurement.

However, there are currently many limitations, and the scientific literature appears divided on issues related to the reliability and ethics of these tools. This is reflected in the large number of organisations calling for a comprehensive review of their use in different educational sectors, as well as the increasing number of policies enforced by governments on this issue, especially in education. For example, the European Commission's AI Act (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, consulted 5 February 2024) is the first major step forward at regulating this new era at an intergovernmental scale. This regulatory proposal addresses the risks of AI, by providing AI developers, implementers and users with clear requirements and obligations regarding specific uses of AI. Critically, the regulatory framework defines four levels of risk in AI and identifies as high-risk AI technologies used in education or training that can determine someone's access to education and career progression (e.g., scoring exams). The proposal is part of a broader artificial intelligence package that will ensure the safety and fundamental rights of people and companies in AI.

5.3.1. Methodologies and applications

Through advanced algorithms, AI systems can sift through vast datasets to identify patterns and correlations that human analysts might overlook. Various strands of research, known collectively as affective computing, have been developed to enable computational systems to recognise, interpret and respond to human emotions (Aranha, Correa and Nunes, 2021[192]). The tools developed both for assessment analytics and skills inferencing, include a wide range of technologies, such as emotion recognition techniques, facial expression analysis, gesture recognition, voice recognition or text analysis. They are used to interpret facial expressions, body language, voice frequency and intensity, or written content and style, which are essential indicators of an individual's emotional states (Westera et al., 2019[18]). These technologies are also used for social signal processing. Social Signal Processing focuses on the automatic detection and interpretation of social signals, which are non-verbal cues that can convey socially relevant information (e.g., gestures, eye gaze, posture, vocal characteristics, interpersonal distances, facial expressions) (Beyan et al., 2021[17]). For example, from a social AI perspective, the contraction of a series of facial muscles producing a smile can be coded as evidence with a certain probability that a person is likely to be experiencing the emotion of happiness or displaying a friendly attitude. These approaches enhance the processing of socially and emotionally relevant information, potentially improving the assessment of SES.

Additionally, machine learning models are increasingly employed for personality detection (Mehta et al., 2020[193]). Automatic personality detection is a rising field (in terms of academic publications) given the large number of industrial applications. As for emotions and social interactions, methods for automatic detection of personality include text processing, visual processing such as analysis of facial features, audio processing, and multimodal processing – which combines inputs from several modalities. Other recent and more exotic approaches include, for instance, the use of AI to analyse colour features of photos posted on social media to predict personality traits (Khorrami, Khorrami and Farhangi, 2022[194]). Visual processing is the most common method and is considered the most accurate unimodal prediction. Audio processing is mostly used in combination with the visual method. While bimodal predictions remain the most popular, multimodal methods have been shown to be so far the most accurate in predicting personality (Mawalim

et al., 2023[195]; Mehta et al., 2020[193]). Even though these machine learning models are currently applied to personality traits, there are explorations of these models to detect SES, such as Emotion recognition (Dai and Ke, 2022[196]; Terhürne et al., 2022[197]).

AI tools developed from research in affective computing and social signal processing are being applied in various contexts. For instance, in educational settings, AI has been utilised to analyse learning analytics. The interactions of students with learning platforms and the resulting trace data have been used to infer the Big Five personality traits of students (Mehta et al., 2020[193]). This data-driven approach also helps to identify individual learning strategies, enhancing personalised learning experiences (Kim and Kim, 2020[198]). The application of AI in education could foster the continuous monitoring of SES and other competencies, including motivation, emotion, and social interaction skills, offering a holistic view of student development (Tuomi, 2022[19]). Other examples of AI application to infer SES outside of educational contexts include automated assessments in profession recruitment processes, namely by using interactive scenarios. Many tools have been developed to automatically assess communication skills and behavioural aspects effectively, offering insights into personality, dominance, and leadership under various contexts (Rasipuram and Jayagopi, 2020[199]). These applications underscore AI's ability to process large-scale behavioural data, a task traditionally challenging for human annotation abilities due to its time-consuming and expensive nature.

Although automatic skill and personality prediction have been extensively studied in various scenarios, there are still several open research questions that need to be addressed (Rasipuram and Jayagopi, 2020[199]), starting with the question of whether these new AI applications measure skills, traits, or both. Further research should also explore the connections between personality, communication and social skills, dominance, or leadership in different contexts, test the most effective scenarios for predicting individual personality traits or SES, and identify the independent factors that explain good performance in different skill sets. Do models developed in laboratory situations can be generalised to the real world, where there is diversity in terms of people, culture, sensing, and computing platforms? These unanswered questions have significant implications for the use of these tools on a larger scale, particularly in sensitive contexts such as education.

5.3.2. Challenges, limitations and ethical considerations

The deployment of AI in SES assessment is not without its challenges. Indeed, despite its potential, the AI-driven assessment of SES faces challenges concerning accuracy, reliability, cultural sensitivity, and privacy. The effectiveness of automatic analysis hinges on data quality, data annotation for model training, and the integration of information from multiple modalities (Rasipuram and Jayagopi, 2020[199]). It also strongly depends on the quality and diversity of the data used for the training of the AI model (Chen, 2023[200]). This has several implications.

First, AI's interpretation of emotional and social cues can vary widely, raising questions about the consistency and reliability of such assessments. For example, reported prediction accuracy rates of tools for automatic emotion recognition in children range from 65% in some publications to over 80% in others (Dai and Ke, 2022[196]), which can be considered both high and very low accuracy, depending on the stakes of the assessment and the use of the data. Cultural and contextual variations pose additional complexities, requiring AI systems to be adaptable and nuanced in their analysis. Ethnicity (Corrigan et al., 2023[201]), gender (Gross, 2023[202]), socioeconomic status and other sociodemographic-related biases in recognition and performance predictions (Khan, 2023[203]), due to biases in training data pools among others, have been repeatedly flagged in tools used in various contexts, from human resources and recruitment (Chen, 2023[200]) to healthcare (Timmons et al., 2023[204])

to education (Tuomi, 2022[19]). These challenges highlight the need for AI systems that are both adaptable and nuanced, capable of accommodating cultural and contextual differences in emotional and social expressions. Another challenge is the limitation of the field's advancement by proprietary data constraints, making it difficult to compare methodologies or validate data against rigorous psychological standards through peer-reviewed publication processes (Suman et al., 2022[205]). Finally, the collection and handling of sensitive personal data requires strict privacy and security measures to protect individuals' rights and confidentiality, especially when AI tools are applied to data collected from sensitive populations (such as students and children) or in sensitive contexts (such as schools).

5.3.3. Conclusion

AI's integration into the assessment of SES may open a new era of insight, offering great opportunities to support individuals' development more effectively. However, addressing challenges in accuracy, cultural sensitivity, and privacy is essential for harnessing AI's full potential. As AI technologies continue to evolve, constant dialogue between ongoing research in academia and the private sector, as well as strong policies to regulate AI use, will be essential in harnessing their full potential for the betterment of educational, workplace, and healthcare outcomes.

5.4. Digital footprints and behavioural measures from videogames

In today's digital landscape, individuals leave behind a vast richness of data through their interactions on social media and within videogames. This data has the advantage of being readily available, saving the cost of creating an assessment tool, and of being collected over a longer period of time and in a broader context than what a specific assessment would allow.

Digital footprints encompass the traces individuals leave across various digital environments, including social networks, learning platforms, massive online courses, and digital information systems, such as forums and blogs (Buitrago-Ropero, Ramírez-Montoya and Laverde, 2023[206]). These footprints can be passive, comprising data accessible online without deliberate action, or active, created intentionally often for specific audiences. While some digital footprints are generated involuntarily, others result from deliberate actions such as publishing messages or exchanging information. Recent approaches have used digital footprints for psychometric modelling, to infer personality traits (Li et al., 2022[207]) or social-related skills such as leadership (Buitrago-Ropero, Ramírez-Montoya and Laverde, 2023[206]), for example.

Another source of data comes from behavioural measures extracted from commercial off-the-shelf videogames, not initially developed for assessment nor educational purposes. In recent years, the recreational videogame industry has embraced a more narrative-driven, choice-based format (Burrus, Rikoon and Brenneman, 2022[208]). Videogames such as Life Is Strange (2015), Detroit: Become Human (2019), or Baldur’s Gate 3 (2023) exemplify this shift, presenting players with intricate social interactions and dilemmas that directly impact gameplay and characters well-being. Players are empowered to make consequential choices that change the narrative of the game. Some videogames even allow players to revisit choices, encouraging reflection and exploration of alternative paths. This format has the potential to enhance immersive, contextualised and realistic SES assessment experiences within gaming environments. These existing commercial games have been used in scientific research to assess SES in a stealthy and engaging way, by extracting selective data on players’ performance and decisions. Their level of complexity allows for

the collection of a wide range of measures, such as game behaviour (choices and strategies), cognitive performance, motor performance, social behaviour and derived affect relevant to SES such as Creativity, Social problem-solving, Co-operation, Assertiveness or Empathy, among others (Mandryk and Birk, 2019[209]).

Games like ***Poptropica*** (Dicerbo, 2013[210]), for example, rely on the difficulty of tasks in order to assess the skill Persistence. While playing this game, elementary- and middle school children (6- to 14-year-olds) visit "islands" with various themes and are expected to persevere through overarching quests, with multiple challenging tasks, in order to reach completion. Using data logs covering the time spent on quest events, the number of quest events completed or, more specifically, the maximum time spent on an individual quest event successfully completed, researchers can infer about the persistence levels of each player. However, unlike with other games discussed in previous sections, researchers have not used external measurements to assert the validity of the assessment. This weakens firmer conclusions about whether Persistence is exactly the skill being measured, as differences in individual performance could be due to cognitive abilities or differential motivation to play this particular game.

Overall, digital footprints and behavioural data from videogames offer an interesting opportunity to assess SES, providing insights into authentic behaviours and interactions in digital environments. However, while promising, these approaches face similar challenges to the use of AI (see section 5.3 Artificial intelligence applications). In addition to obvious privacy concerns, particularly with passive footprint data mining, they face challenges related to representation bias and interpretation, as well as the challenges of accurately interpreting these digital behaviours, which require a careful approach. Given that not all demographic groups are equally represented, and that individuals' online personas may not fully reflect their real-world selves, it is critical for researchers to implement strategies that mitigate these biases and contextualise their findings within the limitations of the data. Furthermore, the complexity of deciphering this data without oversimplifying or misinterpreting it highlights the importance of interdisciplinary collaboration. Drawing on diverse expertise can enhance the analytical process, ensuring that interpretations are nuanced and deeply informed by an understanding of the intricacies of digital interactions.

While these current limitations need to be carefully addressed, analysing digital footprints and behavioural residues from videogames holds significant potential for advancing our understanding of SES in the digital age.

6. Review discussion

6.1. What different benefits tasks and games have to offer?

The lasts two sections presented a large number of direct assessment tools as well as several technological approaches to measure SES, moving beyond self-reports and other-reports, to collect more objective data on the authentic behaviours and skills of individuals. The diversity of innovative assessment tools mapped and reviewed greatly amplifies the portfolio of available assessment methods, which can supplement current, more traditional practices. The instruments mapped for this paper also facilitate the exploration of promising avenues for assessing SES in the context of an international, summative assessment of school-age children. Collectively, task- and game-based assessments combine a wide variety of designs with a significant level of immersiveness and interactivity, covering almost the full panoply of SES contemplated in the OECD framework. Moreover, they allow stealth assessment paradigms that are able to counter various biases and contribute to flesh out more authentic behaviours (Shute et al., 2016[211]).

Within behavioural assessments, however, tests and digital games have different benefits and, naturally, not all tasks nor all games present the same assessment quality. A fine-grained analysis is, therefore, needed. Table 6.1. Comparisons of strengths and challenges for tasks and digital games presents summarised information regarding the distinct strengths and challenges associated with tasks and with digital games, in the context of SES assessment. It includes aspects of ecological validity and immersiveness, reliability and validity psychometrics, scope of skills assessed and considerations on feasibility and costs.

Table 6.1. Comparisons of strengths and challenges for tasks and digital games

	Strengths	Challenges
Tasks	• Wide variety of assessment designs • Collectively offer a comprehensive coverage of SES • Offer more time-efficient measurements, with reduced variables and noise, for a quicker and cleaner extraction of data • Many tasks provide robust reliability and validity psychometrics to constitute an assessment alternative to self-reports, although they have often not been scaled up to international implementation	• The focus on direct instructions and simple interfaces reduces subtilties of social stimuli, narratives and contexts, which compromises ecological validity • Reliance on a single task to represent complex and multifaceted constructs can lead to oversimplification, reducing its validity • Can overlook the richness of students' SES as manifested in different contexts, moments and situations. • Cognitive load might create more noise in the data, making it harder to isolate variables and measure clearly defined SES
Digital Games	• Wide variety of assessment designs • Some can assess multiple SES simultaneously • Tend to have higher levels of ecological validity, with contextualised storytelling, richer audiovisual elements, and greater immersiveness and interactivity • Data can be extracted through multiple ways, including implicit behaviours under stealth assessment paradigms	• Often require a significant time to be administered • Most games have merely been tested in small populations with limited representativeness. • Complexity of stimuli and cognitive load might create more noise in the data, making it harder to isolate variables and measure clearly defined SES • Game development is more complex and costly

In this paper, games are distinguished from tasks for having a conducting narrative usually anchored in role-playing, or for their non-linear gameplay, whereas tasks present players with more direct instructions not contextually embedded in interactive storytelling. This distinction allows for the same skills to be assessed in different ways, that reflect their complementary facets, as the next examples can demonstrate. Empathy, when assessed through the game *ZooU* (DeRosier, Craig and Sanchez, 2012[128]), is measured as Empathetic concern, through actionable choices, for example when choosing to attend to a colleague who appears sad instead of engaging in a fun activity during recess. These actions occur in an interactive environment with observable impacts in the other characters. On the other hand, tasks give straightforward instructions which allow for a direct extraction of data on decisions and behaviours, although in a non-interactive environment where the social and emotional stimuli are static. For example, when using the MET task (Dziobek et al., 2008[83]), Empathy is assessed in its Cognitive and Emotional empathy facets, by presenting participants with emotional facial expressions and asking them to identify the emotion and then whether they share the same feeling. Similarly, Self-control in the same game, *ZooU*, is assessed as a response to social, external stimuli, through multiple actionable choices related to controlling aggressive behaviours, or to do responsible actions instead of giving in to distracting impulses, all in contextualised social environments. In

tasks, assessments usually measure Self-control by expecting participants to focus on certain tasks, by dealing with internal impulses and avoiding distractors, with non-complex playable interfaces.

These two skills exemplify how games can offer a socially contextualised assessment of SES, with improved ecological validity and higher complexity, whereas tasks offer more time-efficient measurements, with reduced variables and noise, for a quicker and cleaner extraction of data. These two dimensions are not an automatic advantage or disadvantage, as each individual skill might in fact be better assessed by one method or the other. There is indeed a trade-off between control/standardisation and complexity/realism, as the more realistic and complex the context and stimuli in the assessment, the less controlled the assessment becomes, and the harder it might be to isolate individual skills from those assessments (Clauser, Margolis and Clauser, 2015[127]). Skills which are usually employed in social contexts, such as Empathy, Collaboration, Perspective-taking or Social problem-solving, are perhaps more robustly assessed in their different subtilties when measured through games, which offer higher levels of complexity and immersive social stimuli, mimicking real-life situations. On the other hand, skills less dependent on interpersonal interactions, such as Self-control, Creativity or Curiosity, might be more efficiently assessed with directed tasks which offer more focus and have fewer confounding variables by eliminating unnecessary social elements. However, the great majority of the reviewed tasks and games have been validated individually, so a direct contrast between them is not available for us to firmly conclude which method is the most effective for each skill. What is clear is that tasks and games can be framed in a spectrum of simple to complex ecological validity, presenting different advantages and disadvantages, which are differently relevant for the distinct skills we consider.

Another critical aspect is the strength of the scientific reliability and validity, which constitute the safeguard that assessments are effectively and consistently testing the respective skill, while also relating to relevant outcomes. Several tasks, on one hand, have been used for a long time, sometimes decades, meaning they have been polished and validated in multiple circumstances and populations. Such robustness is reflected on the high scores for the various validation criteria we considered (Behavioural Assessment Tools for SES). However, even highly validated task-based assessments are usually tested in relatively small samples and would benefit from having a large-scale deployment in an international context for a wider validation. Game-based assessments, on the other hand, tend to be more recent and often experimental. Although promising in their innovative design and gameplay, many of them have merely been tested in small populations with limited representativeness. Only a few exceptions have simultaneously been validated for convergence with other more established measurements, for a relation with academic outcomes and prosocial behaviours, for international contexts, and for diverse populations, from ethnical and socioeconomic perspectives.

A notable finding from the research work is that assessments, considering both games and tasks, are not evenly distributed across all skills. Skills present in our framework, such as Responsibility, Stress Resistance, Tolerance, Assertiveness or Curiosity have none, or only one, associated behavioural assessment. In contrast, skills such as Self-control, Emotional control, Empathy, Co-operation, Creativity, Perspective-taking and Social problem-solving benefit from having a significant number of different tasks and games to assess them. A first-order explanation could be that all skills are not easily measurable and, thus, it is difficult to rigorously design behavioural assessments for them, resulting in this observed discrepancy. However, it could also be the case that different skills have different levels of perceived relevance to society, to research and to education, which leads to a differential attraction of researchers and funding towards them. The unevenness in the distribution of available assessment tools per skill is evident. Whether this finding reflects differential

easiness of measurement, subtle cultural and intellectual biases of interest from researchers, or contrasting social perceptions of the relevance of these skills, is difficult to judge.

6.2. The (limited) potential of transversal technological approaches to assess SES

Transversal technologies which can transversally be implemented in the context of task- and game-based assessments, for greater richness and quality of SES assessment, were also explored in this paper. Table 6.2 contrasts and summarises the strengths and challenges associated with the different technological approaches which can be used to assess SES. The use of biophysiological measures offer, by principle, less biased data compared to self-reports or even behavioural tests, since our heartrate, hormone levels or spontaneous gaze are not under our direct control (Ketonen et al., 2023[212]; Richardson et al., 2020[213]). Technologies such as VR and AR can potentiate immersiveness and ecological validity, when compared to playing digitalised tasks or games in front of a 2D screen (Giglioli et al., 2017[214]). AI technologies are widely regarded as a hallmark of the future and its permeation into education appears indisputable (Chiu et al., 2023[215]; Cope, Kalantzis and Searsmith, 2021[216]). Therefore, their use in educational assessment cannot be ignored, as AI manifests its potential in different dimensions, from assessment design to a more efficient and sophisticated analysis of various types of data.

Even though all these technologies offer complementary benefits to SES assessment, each of them also introduces additional challenges and limitations. From the non-linear relationship between biodata and skills or behaviours, to the need for specialised staff and equipment for its collection, which limits the large-scale application of biophysiological measures for assessment. Although VR and AR potentially offer more immersive and realistic experiences, they lack the upper hand when contrasted with tasks and digital games that, even though come in classical 2D format, have been subject to much more validation scrutiny. Moreover, even though AI for assessment analytics and for skills inferencing hold many promises and their incorporation into education and assessment is realistically unavoidable, at this point their unconsolidated state, limited validation studies and ethical concerns form sufficient obstacles to limit their clear recommendation.

Table 6.2. Strengths and challenges of transversal technological approaches

	Strengths	Challenges
Biophysiological measures	• Several techniques can be easily adapted and integrated into existing tasks and games • Data is spontaneous, less subject to conscious alteration • Offers more fine-grained information about internal states than externalised behaviours	• Biodata is complex and correlation with specific emotions and SES is not linear • Strongly affected by multiple factors such as stress levels, sleep, food intake or innate circadian rhythms • Requires specialized staff and equipment for data collection and analysis of results • Usually used to assess skills related to stress and attention, restraining their potential as a comprehensive method for assessing SES
VR and AR	• Offer greater levels of immersiveness, with contextualised social information and richer audiovisual stimuli, when compared to tasks or games in 2D screens • Participants report strong levels of engagement and a high sense of presence	• Current use of VR and AR is targeted at learning purposes much more than at assessment • Research is yet to conclusively show that VR and AR provide measurable and definite construct validity gains when compared to 2D screen experiences • Equipment costs limit feasibility of these technologies for large-scale assessment
AI applications and digital footprints	• Generative AI can improve assessment design while reducing costs and time with development • AI can quickly and accurately analyse large amounts of data from text, gaze, facial expressions, gestures or speech • Processing of participants' behaviours and choices, when playing games and tasks, potentially allows the identification of patterns and the profiling of their SES	• Applications are still being developed and clear validation studies are not yet available • AI development mostly comes from private industry, resulting in limited published data and compromised peer-reviewed validation standards • Ethical concerns, due to the still limited regulatory constraints, on the use of AI to process data and on its algorithm designs • Unresolved biases related to culture, ethnicity, gender, socioeconomic status and other sociodemographic characteristics

7. Conclusion

An evaluation of all direct assessment tools and approaches, their pros and cons and their current state of development, allows the conclusion that behavioural assessments, namely task- and game-based assessments, currently offer the highest potential for SES assessment. However, from the mapping of around 60 different tools, it is not possible to single out one instrument which alone offers robust ecological, construct and criterion validity to assess SES, while tested on a sufficiently representative sample, and thus recommendable for international large-scale assessment. Nonetheless, collectively various high-quality tasks and games can inspire the development of a more comprehensive and scientifically validated assessment tool, which can assess multiple SES.

Given all the distinct benefits of tasks and games and the uneven assessment coverage of the different skills, what assessment strategy could be designed to effectively assess multiple SES simultaneously? How can objective and comparable data be obtained to measure SES and attest the quality of SEL interventions? Beyond presenting children with various disconnected tasks, not harmonised in their interface or narrative, a solution might be to attempt a unified experience, where the advantages of both tasks and games are combined (Allen et al., 2023[217]). A game anchored in a social narrative, with scenarios, characters and storytelling, while populated with brief tasks, but still contextualised in the

gaming experience. This way, the narrative arcs could cover certain skills while the embedded direct tasks could cover others. This approach could provide a strategy to cover a wide range of SES, increase scientific validity and preserve the stealth assessment paradigm, through an immersive and engaging experience. Leveraging on the scientific advances documented in this paper, a unified behaviour-based assessment is a possible solution to offer an innovative assessment method for SES, by presenting contextualised, realistic and interactive social and emotional stimuli. By eliciting more authentic behaviours in children, it will be possible to infer their SES more objectively and expand on the limited knowledge current questionnaires can provide.

References

Abrahams, L. et al. (2019), "Social-Emotional Skill Assessment in Children and Adolescents: Advances and Challenges in Personality, Clinical, and Educational Contexts", *Psychological Assessment*, Vol. 31/4, pp. 460-473, https://doi.org/10.1037/pas0000591. [7]

Achim, A. et al. (2012), "Mentalizing in first-episode psychosis", *Psychiatry Research*, Vol. 196/2-3, pp. 207-213, https://doi.org/10.1016/j.psychres.2011.10.011. [77]

Aïte, A. et al. (2018), "Adolescents' inhibitory control: Keep it cool or lose control", *Developmental Science*, Vol. 21/1, https://doi.org/10.1111/desc.12491. [43]

Alabbasi, A. et al. (2022), "What do educators need to know about the Torrance Tests of Creative Thinking: A comprehensive review", *Frontiers in Psychology*, Vol. 13, https://doi.org/10.3389/fpsyg.2022.1000385. [104]

AlGerafi, M. et al. (2023), *Unlocking the Potential: A Comprehensive Evaluation of Augmented Reality and Virtual Reality in Education*, Multidisciplinary Digital Publishing Institute (MDPI), https://doi.org/10.3390/electronics12183953. [175]

Allen, K. et al. (2023), *Using Games to Understand the Mind*, https://doi.org/10.31234/osf.io/hbsvj. [217]

An, D., Y. Song and M. Carr (2016), "A comparison of two models of creativity: Divergent thinking and creative expert performance", *Personality and Individual Differences*, Vol. 90, pp. 78-84, https://doi.org/10.1016/j.paid.2015.10.040. [101]

Anglim, J. et al. (2020), "Predicting trait emotional intelligence from HEXACO personality: Domains, facets, and the general factor of personality", *Journal of Personality*, Vol. 88/2, pp. 324-338, https://doi.org/10.1111/jopy.12493. [117]

Aranha, R., C. Correa and F. Nunes (2021), "Adapting Software with Affective Computing: A Systematic Review", *IEEE Transactions on Affective Computing*, Vol. 12/4, pp. 883-899, https://doi.org/10.1109/taffc.2019.2902379. [192]

Bajgar, J. et al. (2005), *Development of the Levels of Emotional Awareness Scale for Children (LEAS-C)*, https://doi.org/10.1348/026151005X35417. [108]

Bamford, C. and K. Lagattuta (2020), "Optimism and Wishful Thinking: Consistency Across Populations in Children's Expectations for the Future", *Child Development*, Vol. 91/4, pp. 1116-1134, https://doi.org/10.1111/cdev.13293. [65]

Baranes, A., P. Oudeyer and J. Gottlieb (2015), "Eye movements reveal epistemic curiosity in human observers", *Vision Research*, Vol. 117, pp. 81-90, https://doi.org/10.1016/j.visres.2015.10.009. [158]

Baron-Cohen, S. et al. (1999), *Recognition of Faux Pas by Normally Developing Children and Children with Asperger Syndrome or High-Functioning Autism*. [78]

Berkovsky, S. et al. (2019), *Detecting personality traits using eye-tracking data*, Association for Computing Machinery, https://doi.org/10.1145/3290605.3300451. [155]

Bettis, A. et al. (2019), "Laboratory and Self-Report Methods to Assess Reappraisal and Distraction in Youth", *Journal of Clinical Child and Adolescent Psychology*, Vol. 48/6, pp. 855-865, https://doi.org/10.1080/15374416.2018.1466306. [63]

Beyan, C. et al. (2021), "Personality Traits Classification Using Deep Visual Activity-Based Nonverbal Features of Key-Dynamic Images", *IEEE Transactions on Affective Computing*, Vol. 12/4, pp. 1084-1099, https://doi.org/10.1109/taffc.2019.2944614. [17]

Black, M. et al. (2017), *Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography*, Elsevier Ltd, https://doi.org/10.1016/j.neubiorev.2017.06.016. [156]

Blanch, A. (ed.) (2019), "Grit (effortful persistence) can be measured with a short scale, shows little variation across socio-demographic subgroups, and is associated with career success and career engagement", *PLOS ONE*, Vol. 14/11, p. e0224814, https://doi.org/10.1371/journal.pone.0224814. [53]

Bland, A. et al. (2016), "EMOTICOM: A neuropsychological test battery to evaluate emotion, motivation, impulsivity, and social cognition", *Frontiers in Behavioral Neuroscience*, Vol. 10/FEB, https://doi.org/10.3389/fnbeh.2016.00025. [42]

Bouhours, L. et al. (2021), "How does social evaluation influence Hot and Cool inhibitory control in adolescence?", *PLoS ONE*, Vol. 16/9, https://doi.org/10.1371/journal.pone.0257753. [44]

Buckley, J. et al. (2021), *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, OECD Publishing, Paris, https://doi.org/10.1787/589b283f-en. [125]

Buitrago-Ropero, M., M. Ramírez-Montoya and A. Laverde (2023), "Digital footprints (2005–2019): a systematic mapping of studies in education", *Interactive Learning Environments*, Vol. 31/2, pp. 876-889, https://doi.org/10.1080/10494820.2020.1814821. [206]

Burguess, P. et al. (2006), "The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology", *Journal of the International Neuropsychological Society*, Vol. 12/2, pp. 194-209, https://doi.org/10.1017/s1355617706060310. [34]

Burrus, J., S. Rikoon and M. Brenneman (2022), *Assessing Competencies for Social and Emotional Learning*, Routledge, New York, https://doi.org/10.4324/9781003102243. [208]

Cerniglia, L. et al. (2019), *Intersections and Divergences Between Empathizing and Mentalizing: Development, Recent Advancements by Neuroimaging and the Future of Animal Modeling*, Frontiers Media S.A., https://doi.org/10.3389/fnbeh.2019.00212. [70]

Chauntry, A., S. Williams and A. Whittaker (2019), "Blunted cardiovascular responses to acute psychological stress predict low behavioral but not self-reported perseverance", *Psychophysiology*, Vol. 56/11, https://doi.org/10.1111/psyp.13449. [149]

Chen, G., S. Gully and D. Eden (2004), "General self-efficacy and self-esteem: toward theoretical and empirical distinction between correlated self-evaluations", *Journal of Organizational Behavior*, Vol. 25/3, pp. 375-395, https://doi.org/10.1002/job.251. [110]

Chen, Z. (2023), "Ethics and discrimination in artificial intelligence-enabled recruitment practices", *Humanities and Social Sciences Communications*, Vol. 10/1, https://doi.org/10.1057/s41599-023-02079-x. [200]

Chernyshenko, O., M. Kankaraš and F. Drasgow (2018), "Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills", *OECD Education Working Papers*, No. 173, OECD Publishing, Paris, https://doi.org/10.1787/db1d8e59-en. [40]

Chiu, T. et al. (2023), "Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education", *Computers and Education Artificial Intelligence*, Vol. 4/3, https://doi.org/10.1016/j.caeai.2022.100118. [215]

Cipriano, C. et al. (2023), "The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions", *Child Development*, Vol. 94/5, pp. 1181-1204, https://doi.org/10.1111/cdev.13968. [3]

Clapham, M. (2011), "Testing/Measurement/Assessment", in *Encyclopedia of Creativity*, Elsevier, https://doi.org/10.1016/B978-0-12-375038-9.00220-X. [100]

Clauser, B., M. Margolis and J. Clauser (2015), "Issues in Simulation-Based Assessment", in *Technology and Testing: Improving Educational and Psychological Measurement*, Taylor and Francis, https://doi.org/10.4324/9781315871493-3. [127]

Coccaro, E. et al. (2017), "Social emotional information processing in adults: Development and psychometrics of a computerized video assessment in healthy controls and aggressive individuals", *Psychiatry Research*, Vol. 248, pp. 40-47, https://doi.org/10.1016/j.psychres.2016.11.004. [81]

Cope, B., M. Kalantzis and D. Searsmith (2021), "Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies", *Educational Philosophy and Theory*, Vol. 53/12, pp. 1229-1245, https://doi.org/10.1080/00131857.2020.1728732. [216]

Corrigan, C. et al. (2023), *AI Ethics in Higher Education: Insights from Africa and Beyond SpringerBriefs in Ethics*, Springer, https://doi.org/10.1007/978-3-031-23035-6. [201]

Corstjens, J., F. Lievens and S. Krumm (2017), "Situational Judgement Tests for Selection", in *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention*, Wiley Blackwell, https://doi.org/10.1002/9781118972472.ch11. [22]

Coskun, K. (2019), "Development of facial emotion recognition and empathy test (FERET) for primary school children", *Children Australia*, Vol. 44/1, pp. 23-31, https://doi.org/10.1017/cha.2018.51. [82]

Cox, J., B. Foster and D. Bamat (2019), *A review of instruments for measuring social and emotional learning skills among secondary school students.*, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. [35]

Craig, A., M. Derosier and Y. Watanabe (2015), "Differences between Japanese and U.S. Children's Performance on "zoo U": A Game-Based Social Skills Assessment", *Games for Health Journal*, Vol. 4/4, pp. 285-294, https://doi.org/10.1089/g4h.2014.0075. [129]

Credé, M., M. Tynan and P. Harms (2017), "Much ado about grit: A meta-analytic synthesis of the grit literature", *Journal of Personality and Social Psychology*, Vol. 113/3, pp. 492-511, https://doi.org/10.1037/pspp0000102. [49]

da Motta, C. et al. (2021), "Rasch Measurement of the Brief Situational Test of Emotional Management in a Large Portuguese Sample", *Journal of Psychoeducational Assessment*, Vol. 39/1, pp. 112-127, https://doi.org/10.1177/0734282920936936. [25]

Dahlstrom-Hakki, I., J. Asbell-Clarke and E. Rowe (2019), *Showing Is Knowing: The Potential and Challenges of Using Neurocognitive Measures of Implicit Learning in the Classroom*, Blackwell Publishing, https://doi.org/10.1111/mbe.12177. [144]

Dai, C. and F. Ke (2022), "Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review", *Computers and Education Artificial Intelligence*, Vol. 3/8, https://doi.org/10.1016/j.caeai.2022.100087. [196]

De Houwer, J. (2003), "The Extrinsic Affective Simon Task", *Experimental Psychology*, Vol. 50/2, pp. 77-85, https://doi.org/10.1027/1618-3169.50.2.77. [94]

De Klerk, S., B. Veldkamp and T. Eggen (2015), "Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example", *Computers and Education*, Vol. 85, pp. 23-34, https://doi.org/10.1016/j.compedu.2014.12.020. [15]

de Vries, L., M. van de Weijer and M. Bartels (2022), *The human physiology of well-being: A systematic review on the association between neurotransmitters, hormones, inflammatory markers, the microbiome and well-being*, Elsevier Ltd, https://doi.org/10.1016/j.neubiorev.2022.104733. [165]

Degner, J. and D. Wentura (2008), "The extrinsic affective Simon task as an instrument for indirect assessment of prejudice", *European Journal of Social Psychology*, Vol. 38/6, pp. 1033-1043, https://doi.org/10.1002/ejsp.536. [36]

Denson, T. and E. Fabiansson Tan (2023), "Anger, hostility, and anger management", in *Encyclopedia of Mental Health*, Elsevier, https://doi.org/10.1016/b978-0-323-91497-0.00139-9. [219]

DeRosier, M., A. Craig and R. Sanchez (2012), "Zoo U: A Stealth Approach to Social Skills Assessment in Schools", *Advances in Human-Computer Interaction*, Vol. 2012, pp. 1-7, https://doi.org/10.1155/2012/654791. [128]

DeRosier, M. and J. Thomas (2018), "Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes", *Journal of Applied Developmental Psychology*, Vol. 55, pp. 52-61, https://doi.org/10.1016/j.appdev.2017.03.001. [130]

Dever, D. and R. Azevedo (2019), *Examining gaze behaviors and metacognitive judgments of informational text within game-based learning environments*, Springer Verlag, https://doi.org/10.1007/978-3-030-23204-7_11. [162]

Dever, D. et al. (2020), "The Impact of Autonomy and Types of Informational Text Presentations in Game-Based Environments on Learning: Converging Multi-Channel Processes Data and Learning Outcomes", *International Journal of Artificial Intelligence in Education*, Vol. 30/4, pp. 581-615, https://doi.org/10.1007/s40593-020-00215-1. [163]

Dicerbo, K. (2013), "Game-Based Assessment of Persistence", *Educational Technology & Society*, Vol. 17, pp. 17-28. [210]

Dirzyte, A. et al. (2021), "Self-and Other-Focused Emotional Intelligence, Situational Emotional Understanding, and Experience of Loss", *Journal of Intellectual Disability - Diagnosis and Treatment*, pp. 641-650, https://doi.org/10.6000/2292-2598.2021.09.06.7. [26]

Doesum, N., D. Van Lange and P. Van Lange (2013), "Social mindfulness: Skill and will to navigate the social world", *Journal of Personality and Social Psychology*, Vol. 105/1, pp. 86-103, https://doi.org/10.1037/a0032540. [138]

Dormal, V., N. Vermeulen and S. Mejias (2020), "Is heart rate variability biofeedback useful in children and adolescents? A systematic review", *Journal of Child Psychology and Psychiatry*, Vol. 62/12, pp. 1379-1390, https://doi.org/10.1111/jcpp.13463. [150]

Drimalla, H. et al. (2019), "From face to face: the contribution of facial mimicry to cognitive and emotional empathy", *Cognition and Emotion*, Vol. 33/8, pp. 1672-1686, https://doi.org/10.1080/02699931.2019.1596068. [67]

Duckworth, A. et al. (2007), "Grit: Perseverance and passion for long-term goals.", *Journal of Personality and Social Psychology*, Vol. 92/6, pp. 1087-1101, https://doi.org/10.1037/0022-3514.92.6.1087. [47]

Duckworth, A. and P. Quinn (2009), "Development and validation of the short Grit Scale (Grit-S)", *Journal of Personality Assessment*, Vol. 91/2, pp. 166-174, https://doi.org/10.1080/00223890802634290. [50]

Duckworth, A. and D. Yeager (2015), "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes", *Educational Researcher*, pp. 237-251, https://doi.org/10.3102/0013189X15584327. [38]

Durlak, J. et al. (2011), "The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions", *Child Development*, Vol. 82/1, pp. 405-432, https://doi.org/10.1111/j.1467-8624.2010.01564.x. [5]

Dweck, C. and E. Leggett (1988), "A social-cognitive approach to motivation and personality.", *Psychological Review*, Vol. 95/2, pp. 256-273, https://doi.org/10.1037/0033-295X.95.2.256. [55]

Dziobek, I. et al. (2008), "Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET)", *Journal of Autism and Developmental Disorders*, Vol. 38/3, pp. 464-473, https://doi.org/10.1007/s10803-007-0486-x. [83]

Elder, J., L. Wilson and J. Calanchini (2023), "Estimating the Reliability and Stability of Cognitive Processes Contributing to Responses on the Implicit Association Test", *Personality and Social Psychology Bulletin*, https://doi.org/10.1177/01461672231171256. [93]

Emerson, A. et al. (2020), "Multimodal learning analytics for game-based learning", *British Journal of Educational Technology*, Vol. 51/5, pp. 1505-1526, https://doi.org/10.1111/bjet.12992. [164]

Emihovich, B., L. Arrington and X. Xu (2019), "Press Play! How Immersive Environments Support Problem-Solving Skills and Productive Failure", https://doi.org/10.1007/978-3-030-15569-8_7. [8]

Evans, V. and C. Bond (2020), "Characteristics of effective small group social skill interventions in mainstream primary education: a systematic literature review", *Journal of Research in Special Educational Needs*, Vol. 20/4, pp. 331-342, https://doi.org/10.1111/1471-3802.12493. [86]

Fernández-Martín, F., J. Arco-Tirado and M. Hervás-Torres (2020), "Grit as a Predictor and Outcome of Educational, Professional, and Personal Success: A Systematic Review", *Psicología Educativa*, Vol. 26/2, pp. 163-173, https://doi.org/10.5093/PSED2020A11. [52]

Fletcher-Watson, S. and S. Hampton (2018), "The potential of eye-tracking as a sensitive measure of behavioural change in response to intervention", *Scientific Reports*, Vol. 8/1, https://doi.org/10.1038/s41598-018-32444-9. [153]

Fulya Eyupoglu, T. and J. Nietfeld (2019), "Intrinsic Motivation in Game-Based Learning Environments", https://doi.org/10.1007/978-3-030-15569-8_5. [9]

Galla, B. and A. Duckworth (2015), "More than resisting temptation: Beneficial habits mediate the relationship between self-control and positive life outcomes.", *Journal of Personality and Social Psychology*, Vol. 109/3, pp. 508-525, https://doi.org/10.1037/pspp0000026. [37]

Galla, B. et al. (2014), "The Academic Diligence Task (ADT): assessing individual differences in effort on tedious but important schoolwork", *Contemporary Educational Psychology*, Vol. 39/4, pp. 314-325, https://doi.org/10.1016/j.cedpsych.2014.08.001. [48]

Gascoine, L., S. Higgins and K. Wall (2017), "The assessment of metacognition in children aged 4–16 years: a systematic review", *Review of Education*, Vol. 5/1, pp. 3-57, https://doi.org/10.1002/rev3.3077. [106]

Geraets, C. et al. (2021), "Virtual reality facial emotion recognition in social environments: An eye-tracking study", *Internet Interventions*, Vol. 25, https://doi.org/10.1016/j.invent.2021.100432. [189]

Giglioli, I. et al. (2021), "An immersive serious game for the behavioral assessment of psychological needs", *Applied Sciences (Switzerland)*, Vol. 11/4, pp. 1-17, https://doi.org/10.3390/app11041971. [191]

Giglioli, I. et al. (2017), *Virtual stealth assessment: A new methodological approach for assessing psychological needs*, Springer Verlag, https://doi.org/10.1007/978-3-319-70111-0_1. [214]

Golub, Y. et al. (2019), "Salivary and hair cortisol as biomarkers of emotional and behavioral symptoms in 6–9 year old children", *Physiology and Behavior*, Vol. 209, https://doi.org/10.1016/j.physbeh.2019.112584. [170]

González Ramírez, M. et al. (2023), *Wearables for Stress Management: A Scoping Review*, Multidisciplinary Digital Publishing Institute (MDPI), https://doi.org/10.3390/healthcare11172369. [147]

Greenwald, A., D. McGhee and J. Schwartz (1998), "Measuring individual differences in implicit cognition: The implicit association test.", *Journal of Personality and Social Psychology*, Vol. 74/6, pp. 1464-1480, https://doi.org/10.1037/0022-3514.74.6.1464. [92]

Gross, N. (2023), "What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI", *Social Sciences*, Vol. 12/8, https://doi.org/10.3390/socsci12080435. [202]

Hamilton, D. et al. (2021), "Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design", *Journal of Computers in Education*, Vol. 8/1, pp. 1-32, https://doi.org/10.1007/s40692-020-00169-2. [178]

Hayward, E. and B. Homer (2017), "Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence", *British Journal of Developmental Psychology*, Vol. 35/3, pp. 454-462, https://doi.org/10.1111/bjdp.12186. [79]

Hellhammer, D., S. Wüst and B. Kudielka (2009), "Salivary cortisol as a biomarker in stress research", *Psychoneuroendocrinology*, Vol. 34/2, pp. 163-171, https://doi.org/10.1016/j.psyneuen.2008.10.026. [169]

Hennefield, L. and L. Markson (2022), "The development of optimistic expectations in young children", *Cognitive Development*, Vol. 63, https://doi.org/10.1016/j.cogdev.2022.101201. [64]

Hickey, B. et al. (2021), *Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review*, MDPI AG, https://doi.org/10.3390/s21103461. [148]

Higuera-Trujillo, J., J. López-Tarruella Maldonado and C. Llinares Millán (2017), "Psychological and physiological human responses to simulated and real environments: A comparison between Photographs, 360° Panoramas, and Virtual Reality", *Applied Ergonomics*, Vol. 65, pp. 398-409, https://doi.org/10.1016/j.apergo.2017.05.006. [185]

Howard, M. and M. Gutworth (2020), "A meta-analysis of virtual reality training programs for social skill development", *Computers and Education*, Vol. 144, https://doi.org/10.1016/j.compedu.2019.103707. [180]

James, K. et al. (2023), *Understanding the relationships between physiological and psychosocial stress, cortisol and cognition*, Frontiers Media S.A., https://doi.org/10.3389/fendo.2023.1085950. [166]

Jankowska, D. et al. (2018), "Exploring the creative process: Integrating psychometric and eye-tracking approaches", *Frontiers in Psychology*, Vol. 9/OCT, https://doi.org/10.3389/fpsyg.2018.01931. [159]

Jiménez-Soto, A. et al. (2022), "Beach balls: Assessing frustration tolerance in young children using a computerized task", *Acta Psychologica*, Vol. 224, p. 103528, https://doi.org/10.1016/j.actpsy.2022.103528. [62]

Johannesen, J. et al. (2013), "The Social Attribution Task-Multiple Choice (SAT-MC): A Psychometric and Equivalence Study of an Alternate Form", *ISRN Psychiatry*, Vol. 2013, pp. 1-9, https://doi.org/10.1155/2013/830825. [80]

Jones, S. and J. Kahn (2017), *The Evidence Base for How We Learn Supporting Students' Social, Emotional, and Academic Development Consensus Statements of Evidence From the Council of Distinguished Scientists National Commission on Social, Emotional, and Academic Development The Aspen Institute*, https://www.aspeninstitute.org/wp-content/uploads/2018/03/FINAL_CDS-Evidence-Base.pdf. [1]

Joseph, N., Y. Jiang and S. Zilioli (2021), "Momentary emotions and salivary cortisol: A systematic review and meta-analysis of ecological momentary assessment studies", *Neuroscience & Biobehavioral Reviews*, Vol. 125, pp. 365-379, https://doi.org/10.1016/j.neubiorev.2021.02.042. [172]

Kankaraš, M. and J. Suarez-Alvarez (2019), "Assessment framework of the OECD Study on Social and Emotional Skills", *OECD Education Working Papers*, No. 207, OECD Publishing, Paris, https://doi.org/10.1787/5007adef-en. [39]

Katz, D. et al. (2016), "Associations between the awakening responses of salivary α-amylase and cortisol with self-report indicators of health and wellbeing among educators", *Teaching and Teacher Education*, Vol. 54, pp. 98-106, https://doi.org/10.1016/j.tate.2015.11.012. [174]

Keil, J. et al. (2017), "The Pizzagame: A virtual public goods game to assess cooperative behavior in children and adolescents", *Behavior Research Methods*, Vol. 49/4, pp. 1432-1443, https://doi.org/10.3758/s13428-016-0799-9. [90]

Keil, M. (2012), "Salivary Cortisol: A Tool for Biobehavioral Research in Children", *Journal of Pediatric Nursing*, Vol. 27/3, pp. 287-289, https://doi.org/10.1016/j.pedn.2012.02.003. [171]

Ketonen, E. et al. (2023), "Can you feel the excitement? Physiological correlates of students' self-reported emotions", *British Journal of Educational Psychology*, Vol. 93/S1, pp. 113-129, https://doi.org/10.1111/bjep.12534. [212]

Khan, S. (2023), "The Ethical Imperative: Addressing Bias and Discrimination in AI-Driven Education", *Social Science Spectrum*, Vol. 2/1, https://sss.org.pk/index.php/sss/article/view/23. [203]

Khorrami, M., M. Khorrami and F. Farhangi (2022), "Evaluation of tree-based ensemble algorithms for predicting the big five personality traits based on social media photos: Evidence from an Iranian sample", *Personality and Individual Differences*, Vol. 188, p. 111479, https://doi.org/10.1016/j.paid.2021.111479. [194]

Kim, H. et al. (2018), "Social perspective-taking performance: Construct, measurement, and relations with academic performance and engagement", *Journal of Applied Developmental Psychology*, Vol. 57, pp. 24-41, https://doi.org/10.1016/j.appdev.2018.05.005. [76]

Kim, W. and J. Kim (2020), "Individualized AI Tutor Based on Developmental Learning Networks", *IEEE Access*, Vol. 8, pp. 27927-27937, https://doi.org/10.1109/ACCESS.2020.2972167. [198]

Kim, Y. and D. Ifenthaler (2019), "Game-Based Assessment: The Past Ten Years and Moving Forward", in *Game-Based Assessment Revisited*, https://doi.org/10.1007/978-3-030-15569-8_1. [122]

Kurdi, B. et al. (2019), "Relationship between the implicit association test and intergroup behavior: A meta-analysis", *American Psychologist*, Vol. 74/5, pp. 569-586, https://doi.org/10.1037/amp0000364. [91]

Kyllonen, P. and H. Kell (2018), "Ability tests measure personality, personality tests measure ability: Disentangling construct and method in evaluating the relationship between personality and ability", *Journal of Intelligence*, Vol. 6/3, pp. 1-26, https://doi.org/10.3390/jintelligence6030032. [16]

Lampropoulos, G. et al. (2022), "Augmented Reality and Gamification in Education: A Systematic Literature Review of Research, Applications, and Empirical Studies", *Applied Sciences*, Vol. 12/13, https://doi.org/10.3390/app12136809. [179]

Lane, R. and R. Smith (2021), "Levels of emotional awareness: Theory and measurement of a socio-emotional skill", *Journal of Intelligence*, Vol. 9/3, https://doi.org/10.3390/jintelligence9030042. [107]

Larradet, F. et al. (2020), "Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection", *Frontiers in Psychology*, Vol. 11, https://doi.org/10.3389/fpsyg.2020.01111. [145]

Lea, R. et al. (2023), "Do emotionally intelligent adolescents flourish or flounder under pressure? Linking emotional intelligence to stress regulation mechanisms", *Personality and Individual Differences*, Vol. 201, https://doi.org/10.1016/j.paid.2022.111943. [27]

Lester, J. et al. (2014), "Designing game-based learning environments for elementary science education: A narrative-centered learning perspective", *Information Sciences*, Vol. 264, pp. 4-18, https://doi.org/10.1016/j.ins.2013.09.005. [160]

Li, Y. et al. (2022), "Multitask learning for emotion and personality traits detection", *Neurocomputing*, Vol. 493, pp. 340-350, https://doi.org/10.1016/j.neucom.2022.04.049. [207]

Lobel, A. et al. (2016), *Designing and utilizing biofeedback games for emotion regulation: The case of Nevermind*, Association for Computing Machinery, https://doi.org/10.1145/2851581.2892521. [152]

Luhmann, M. et al. (2015), "Loneliness and social behaviours in a virtual social environment", *Cognition and Emotion*, Vol. 29/3, pp. 548-558, https://doi.org/10.1080/02699931.2014.922053. [143]

MacCann, C. and R. Roberts (2008), "Supplemental Material for New Paradigms for Assessing Emotional Intelligence: Theory and Data", *Emotion*, Vol. 8/4, pp. 540-551, https://doi.org/10.1037/a0012746.supp. [24]

Mandryk, R. and M. Birk (2019), "The Potential of Game-Based Digital Biomarkers for Modeling Mental Health", *JMIR Mental Health*, Vol. 6/4, p. e13485, https://doi.org/10.2196/13485. [209]

Marín-Morales, J. et al. (2018), "Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors", *Scientific Reports*, Vol. 8/1, https://doi.org/10.1038/s41598-018-32063-4. [186]

Marín-Morales, J. et al. (2020), "Emotion recognition in immersive virtual reality: From statistics to affective computing", *Sensors*, Vol. 20/18, pp. 1-26, https://doi.org/10.3390/s20185163. [184]

Mawalim, C. et al. (2023), "Personality trait estimation in group discussions using multimodal analysis and speaker embedding", *Journal on Multimodal User Interfaces*, Vol. 17/2, pp. 47-63, https://doi.org/10.1007/s12193-023-00401-0. [195]

McKown, C. (2019), "Reliability, Factor Structure, and Measurement Invariance of a Web-Based Assessment of Children's Social-Emotional Comprehension", *Journal of Psychoeducational Assessment*, Vol. 37/4, pp. 435-449, https://doi.org/10.1177/0734282917749682. [113]

McKown, C. et al. (2023), "Development and Validation of a Shortened Form of SELweb EE, a Web-Based Assessment of Children's Social and Emotional Competence", *Assessment*, Vol. 30/1, pp. 171-189, https://doi.org/10.1177/10731911211046044. [115]

McKown, C., N. Russo-Ponsaran and A. Karls (2023), "Web-based assessment of social and emotional competence in the late elementary grades", *Social Development*, Vol. 32/1, pp. 73-97, https://doi.org/10.1111/sode.12641. [114]

McRae, K. et al. (2012), "Individual differences in reappraisal ability: Links to reappraisal frequency, well-being, and cognitive control", *Journal of Research in Personality*, Vol. 46/1, pp. 2-7, https://doi.org/10.1016/j.jrp.2011.10.003. [58]

Mehlsen, M. et al. (2019), "Performance-based assessment of distraction in response to emotional stimuli: Toward a standardized procedure for assessing emotion regulation performance", *Personality and Individual Differences*, Vol. 150, https://doi.org/10.1016/j.paid.2019.06.026. [33]

Mehta, Y. et al. (2020), "Recent trends in deep learning based personality detection", *Artificial Intelligence Review*, Vol. 53/4, pp. 2313-2339, https://doi.org/10.1007/s10462-019-09770-z. [193]

Meindl, P. et al. (2019), "A brief behavioral measure of frustration tolerance predicts academic achievement immediately and two years later.", *Emotion*, Vol. 19/6, pp. 1081-1092, https://doi.org/10.1037/emo0000492. [45]

Musek, J. (2017), "Biological Aspects of General Factor of Personality", in *The General Factor of Personality*, Elsevier, https://doi.org/10.1016/b978-0-12-811209-0.00006-6. [41]

Nebel, S. and M. Ninaus (2019), "New Perspectives on Game-Based Assessment with Process Data and Physiological Signals", https://doi.org/10.1007/978-3-030-15569-8_8. [146]

Ng, Z. et al. (2022), "A Systematic Review of Emotion Regulation Assessments in US Schools: Bridging the Gap Between Researchers and Educators", *Educational Psychology Review*, Vol. 34/4, pp. 2825-2865, https://doi.org/10.1007/s10648-022-09691-4. [60]

OECD (2023), *OECD CENTRE ON PHILANTHROPY Data and analysis for development Philanthropy for Social and Emotional Learning*, OECD. [6]

OECD (2022), *THINKING OUTSIDE THE BOX The PISA 2022 Creative Thinking Assessment Thinking outside the box 2*, http://www.oecd.org/pisa/innovation. [32]

OECD (2021), *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*, OECD Publishing, Paris, https://doi.org/10.1787/92a11084-en. [66]

OECD (2020), *Curriculum Overload: A Way Forward*, OECD Publishing, Paris, https://doi.org/10.1787/3081ceca-en. [4]

OECD (2017), *How does PISA measure students' ability to collaborate?*. [31]

OECD (Forthcoming), *Education 2030 Conceptual Framework Development: Construct Analysis on Key Knowledge, Skills, Attitudes and Values for 2030*, OECD, Paris. [30]

Oranje, A. et al. (2019), "Summative Game-Based Assessment", in *Game-Based Assessment*, Springer, https://doi.org/10.1007/978-3-030-15569-8_3. [123]

Papanastasiou, G. et al. (2019), "Virtual and augmented reality effects on K-12, higher and tertiary education students' twenty-first century skills", *Virtual Reality*, Vol. 23/4, pp. 425-436, https://doi.org/10.1007/s10055-018-0363-2. [176]

Parra Vargas, E. et al. (2022), "Virtual reality stimulation and organizational neuroscience for the assessment of empathy", *Frontiers in Psychology*, Vol. 13, https://doi.org/10.3389/fpsyg.2022.993162. [190]

Parsons, T. (2015), "Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences", *Frontiers in Human Neuroscience*, Vol. 9/DEC, https://doi.org/10.3389/fnhum.2015.00660. [183]

Patterson, F. et al. (2012), "Evaluations of situational judgement tests to assess non-academic attributes in selection", *Medical Education*, Vol. 46/9, pp. 850-868, https://doi.org/10.1111/j.1365-2923.2012.04336.x. [21]

Perna, G. et al. (2020), "Heart rate variability: Can it serve as a marker of mental health resilience?: Special Section on "Translational and Neuroscience Studies in Affective Disorders" Section Editor, Maria Nobile MD, PhD", *Journal of Affective Disorders*, Vol. 263/15, pp. 754-761, https://doi.org/10.1016/j.jad.2019.10.017. [151]

Perry, A. and E. Karpova (2017), "Efficacy of teaching creative thinking skills: A comparison of multiple creativity assessments", *Thinking Skills and Creativity*, Vol. 24, pp. 118-126, https://doi.org/10.1016/j.tsc.2017.02.017. [97]

Persich, M., S. Krishnakumar and M. Robinson (2020), "Are You a Good Friend? Assessing Social Relationship Competence Using Situational Judgments", *Personality and Social Psychology Bulletin*, Vol. 46/6, pp. 913-926, https://doi.org/10.1177/0146167219880193. [28]

Peters, C., J. Kranzler and E. Rossen (2009), "Validity of the mayer - Salovey - Caruso emotional intelligence test: Youth version - Research edition", *Canadian Journal of School Psychology*, Vol. 24/1, pp. 76-81, https://doi.org/10.1177/0829573508329822. [120]

Polyak, S., A. von Davier and K. Peterschmidt (2017), "Computational psychometrics for the measurement of collaborative problem solving skills", *Frontiers in Psychology*, Vol. 8/NOV, https://doi.org/10.3389/fpsyg.2017.02029. [136]

Porter, T. et al. (2020), "Changing Learner Beliefs in South African Townships: An Evaluation of a Growth Mindset Intervention", *Social Psychological and Personality Science*, Vol. 11/7, https://doi.org/10.1177/1948550620909738. [57]

Porter, T. et al. (2020), "Measuring mastery behaviours at scale: The persistence, effort, resilience, and challengeseeking (PERC) task", *Journal of Learning Analytics*, Vol. 7/1, pp. 5-18, https://doi.org/10.18608/jla.2020.71.2. [54]

Porter, T. et al. (2020), "Intellectual humility predicts mastery behaviors when learning", *Learning and Individual Differences*, Vol. 80, p. 101888, https://doi.org/10.1016/J.LINDIF.2020.101888. [56]

Poustka, L. et al. (2008), "Dissociation of Cognitive and Emotional Empathy: The Multifaceted Empathy Test for Children and Adolescents: MET-J", *Journal of Autism and Developmental Disorders*, Vol. 38/3, pp. 464-473, https://doi.org/10.1007/s10803-007-0486-x. [84]

Primi, R. et al. (2020), "True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement", *International Journal of Testing*, Vol. 20/2, pp. 97-121, https://doi.org/10.1080/15305058.2019.1673398. [13]

Quinde-Zlibut, J. et al. (2021), "Multifaceted empathy differences in children and adults with autism", *Scientific Reports*, Vol. 11/1, https://doi.org/10.1038/s41598-021-98516-5. [69]

Rafner, J. et al. (2022), "Digital Games for Creativity Assessment: Strengths, Weaknesses and Opportunities", *Creativity Research Journal*, Vol. 34/1, pp. 28-54, https://doi.org/10.1080/10400419.2021.1971447. [126]

Rafner, J. et al. (2020), *Crea.blender: A Neural Network-Based Image Generation Game to Assess Creativity*. [105]

Rafner, J. et al. (2023), "Towards Game-Based Assessment of Creative Thinking", *Creativity Research Journal*, Vol. 35/4, pp. 763-782, https://doi.org/10.1080/10400419.2023.2198845. [99]

Rappa, N. et al. (2022), "The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: a scoping review", *Interactive Learning Environments*, Vol. 30/7, pp. 1338-1350, https://doi.org/10.1080/10494820.2019.1702560. [154]

Rasipuram, S. and D. Jayagopi (2020), "Automatic multimodal assessment of soft skills in social interactions: a review", *Multimedia Tools and Applications*, Vol. 79/19-20, pp. 13037-13060, https://doi.org/10.1007/s11042-019-08561-6. [199]

Ren, X. (2019), "Stealth Assessment Embedded in Game-Based Learning to Measure Soft Skills: A Critical Review", in *Game-Based Assessment Revisited*, https://doi.org/10.1007/978-3-030-15569-8_4. [10]

Richardson, D. et al. (2020), "Engagement in video and audio narratives: contrasting self-report and physiological measures", *Scientific Reports*, Vol. 10/1, https://doi.org/10.1038/s41598-020-68253-2. [213]

Rivers, S. et al. (2012), "Measuring Emotional Intelligence in Early Adolescence With the MSCEIT-YV: Psychometric Properties and Relationship With Academic Performance and Psychosocial Functioning", *Journal of Psychoeducational Assessment*, Vol. 30/4, pp. 344-366, https://doi.org/10.1177/0734282912449443. [116]

Robinson, M., M. Persich and R. Irvin (2022), "An ego effectiveness perspective of successful self-control: An individual difference and its links to social functioning and well-being", *Journal of Research in Personality*, Vol. 97, https://doi.org/10.1016/j.jrp.2022.104207. [29]

Rosso, A. and A. Riolfo (2020), "A Further Look at Reading the Mind in the Eyes-Child Version: Association With Fluid Intelligence, Receptive Language, and Intergenerational Transmission in Typically Developing School-Aged Children", *Frontiers in Psychology*, Vol. 11, https://doi.org/10.3389/fpsyg.2020.586065. [71]

Rousseau, D. et al. (1998), "Not so different after all: A cross-discipline view of trust", *Academy of Management Review*, Vol. 23/3, pp. 393-404, https://doi.org/10.5465/AMR.1998.926617. [85]

Russo-Ponsaran, N. et al. (2018), "Virtual Environment for Social Information Processing: Assessment of Children with and without Autism Spectrum Disorders", *Autism Research*, Vol. 11/2, pp. 305-317, https://doi.org/10.1002/aur.1889. [139]

Russo-Ponsaran, N. et al. (2021), "Psychometric properties of Virtual Environment for Social Information Processing, a social information processing simulation assessment for children", *Social Development*, Vol. 30/3, pp. 615-640, https://doi.org/10.1111/sode.12512. [140]

Salovey, P. and J. Mayer (1990), *EMOTIONAL INTELLIGENCE*. [119]

Scherer, K. and U. Scherer (2011), "Assessing the Ability to Recognize Facial and Vocal Expressions of Emotion: Construction and Validation of the Emotion Recognition Index", *Journal of Nonverbal Behavior*, Vol. 35/4, pp. 305-326, https://doi.org/10.1007/s10919-011-0115-4. [73]

Schlegel, K., D. Grandjean and K. Scherer (2014), "Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development.", *Psychological Assessment*, Vol. 26/2, pp. 666-672, https://doi.org/10.1037/a0035246. [74]

Schlegel, K. and K. Scherer (2016), "Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation", *Behavior Research Methods*, Vol. 48/4, pp. 1383-1392, https://doi.org/10.3758/s13428-015-0646-4. [75]

Schönbrodt, F. and J. Asendorpf (2012), "Attachment Dynamics in a Virtual World", *Journal of Personality*, Vol. 80/2, pp. 429-463, https://doi.org/10.1111/j.1467-6494.2011.00736.x. [142]

Schönbrodt, F. and J. Asendorpf (2011), "Virtual Social Environments as a Tool for Psychological Assessment: Dynamics of Interaction With a Virtual Spouse", *Psychological Assessment*, Vol. 23/1, pp. 7-17, https://doi.org/10.1037/a0021049. [141]

Schubring, D. et al. (2020), "Virtual reality potentiates emotion and task effects of alpha/beta brain oscillations", *Brain Sciences*, Vol. 10/8, pp. 1-19, https://doi.org/10.3390/brainsci10080537. [187]

Schutte, N. and J. Malouff (2020), "A Meta-Analysis of the Relationship between Curiosity and Creativity", *Journal of Creative Behavior*, Vol. 54/4, pp. 940-947, https://doi.org/10.1002/jocb.421. [95]

Shaw, A. (2022), "Creative Minecrafters: Cognitive and Personality Determinants of Creativity, Novelty, and Usefulness in Minecraft", *Psychology of Aesthetics, Creativity, and the Arts*, Vol. 17/1, pp. 106-117, https://doi.org/10.1037/aca0000456. [98]

Sher, K., M. Levi-Keren and G. Gordon (2019), "Priming, enabling and assessment of curiosity", *Educational Technology Research and Development*, Vol. 67/4, pp. 931-952, https://doi.org/10.1007/s11423-019-09665-4. [96]

Shute, V. and F. Ke (2012), "Games, learning, and assessment", in *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, Springer New York, https://doi.org/10.1007/978-1-4614-3546-4_4. [124]

Shute, V. et al. (2016), "Advances in the Science of Assessment", *Educational Assessment*, Vol. 21/1, pp. 34-59, https://doi.org/10.1080/10627197.2015.1127752. [211]

Shute, V. and S. Rahimi (2021), "Stealth assessment of creativity in a physics video game", *Computers in Human Behavior*, Vol. 116, https://doi.org/10.1016/j.chb.2020.106647. [132]

Shute, V., M. Ventura and Y. Kim (2013), "Assessment and learning of qualitative physics in Newton's playground", *Journal of Educational Research*, Vol. 106/6, pp. 423-430, https://doi.org/10.1080/00220671.2013.832970. [134]

Siefer, K., T. Leuders and A. Obersteiner (2021), "Which Task Characteristics Do Students Rely on When They Evaluate Their Abilities to Solve Linear Function Tasks? – A Task-Specific Assessment of Self-Efficacy", *Frontiers in Psychology*, Vol. 12, https://doi.org/10.3389/fpsyg.2021.596901. [111]

Siegling, A., D. Saklofske and K. Petrides (2015), "Measures of Ability and Trait Emotional Intelligence.", in *Measures of Personality and Social Psychological Constructs*, Elsevier Inc., https://doi.org/10.1016/B978-0-12-386915-9.00014-0. [109]

Snow, E. et al. (2016), "Taking Control: Stealth Assessment of Deterministic Behaviors Within a Game-Based System", *International Journal of Artificial Intelligence in Education*, Vol. 26/4, pp. 1011-1032, https://doi.org/10.1007/s40593-015-0085-5. [14]

Špiljak, B. et al. (2022), "A Review of Psychological Stress among Students and Its Assessment Using Salivary Biomarkers", *Behavioral Sciences*, Vol. 12/10, pp. 1-15, https://doi.org/10.3390/bs12100400. [168]

Steponavičius, M., C. Gress-Wright and A. Linzarini (2023), "Social and emotional skills: Latest evidence on teachability and impact on life outcomes", *OECD Education Working Papers*, No. 304, OECD Publishing, Paris, https://doi.org/10.1787/ba34f086-en. [2]

Stoeffler, K. et al. (2020), "Gamified performance assessment of collaborative problem solving skills", *Computers in Human Behavior*, Vol. 104, https://doi.org/10.1016/j.chb.2019.05.033. [135]

Suman, C. et al. (2022), "A multi-modal personality prediction system", *Knowledge-Based Systems*, Vol. 236, https://doi.org/10.1016/j.knosys.2021.107715. [205]

Sun, C. et al. (2022), "The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment", *Computers in Human Behavior*, Vol. 128, https://doi.org/10.1016/j.chb.2021.107120. [133]

Sutter, M., A. Untertrifaller and C. Zoller (2022), "Grit increases strongly in early childhood and is related to parental background", *Scientific Reports*, Vol. 12/1, https://doi.org/10.1038/s41598-022-07542-4. [51]

Taub, M. et al. (2017), "Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND", *Computers in Human Behavior*, Vol. 76, pp. 641-655, https://doi.org/10.1016/j.chb.2017.01.038. [161]

Terhürne, P. et al. (2022), "Validation and application of the Non-Verbal Behavior Analyzer: An automated tool to assess non-verbal emotional expressions in psychotherapy", *Frontiers in Psychiatry*, Vol. 13, https://doi.org/10.3389/fpsyt.2022.1026015. [197]

Thielmann, I. et al. (2021), "Economic Games: An Introduction and Guide for Research", *Collabra: Psychology*, Vol. 7/1, https://doi.org/10.1525/collabra.19004. [89]

Thielmann, I., B. Hilbig and I. Niedtfeld (2014), "Willing To Give But Not To Forgive: Borderline Personality Features And Cooperative Behavior", *Journal of Personality Disorders*, Vol. 28/6, pp. 778-795, https://doi.org/10.1521/pedi_2014_28_135. [88]

Thompson, N., C. van Reekum and B. Chakrabarti (2022), "Cognitive and Affective Empathy Relate Differentially to Emotion Regulation", *Affective Science*, Vol. 3/1, pp. 118-134, https://doi.org/10.1007/s42761-021-00062-w. [68]

Timmons, A. et al. (2023), "A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health", *Perspectives on Psychological Science*, Vol. 18/5, pp. 1062-1096, https://doi.org/10.1177/17456916221134490. [204]

Tor, N. and G. Gordon (2020), "Digital interactive quantitative curiosity assessment tool: Questions worlds", *International Journal of Information and Education Technology*, Vol. 10/8, pp. 614-621, https://doi.org/10.18178/ijiet.2020.10.8.1433. [137]

Torrance, E. (1966), *The Torrance Tests of Creative Thinking-Norms-Technical Manual Research Edition-Verbal Tests, Forms A and B- Figural Tests, Forms A and B.*, Personnel Press, Princeton. [102]

Tuomi, I. (2022), "Artificial intelligence, 21st century competences, and socio-emotional learning in education: More than high-risk?", *European Journal of Education*, Vol. 57/4, pp. 601-619, https://doi.org/10.1111/ejed.12531. [19]

Tyler, T. (2010), *Why people cooperate: The role of social motivations*, Princeton University Press. [87]

Vaida, S. and A. Opre (2014), *Emotional intelligence versus emotional competence*, https://www.researchgate.net/publication/287019335. [118]

Vaske, J., J. Beaman and C. Sponarski (2016), "Rethinking Internal Consistency in Cronbach's Alpha", *Leisure Sciences*, Vol. 39/2, pp. 163-173, https://doi.org/10.1080/01490400.2015.1127189. [218]

Ventura, M. and V. Shute (2013), "The validity of a game-based assessment of persistence", *Computers in Human Behavior*, Vol. 29/6, pp. 2568-2572, https://doi.org/10.1016/j.chb.2013.06.033. [131]

Ventura, S. et al. (2019), "Immersive Versus Non-immersive Experience: Exploring the Feasibility of Memory Assessment Through 360° Technology", *Frontiers in Psychology*, Vol. 10, https://doi.org/10.3389/fpsyg.2019.02509. [182]

Vogindroukas, I., E. Chelas and N. Petridis (2014), "Reading the Mind in the Eyes Test (Children's Version): A Comparison Study between Children with Typical Development, Children with High-Functioning Autism and Typically Developed Adults", *Folia Phoniatrica et Logopaedica*, Vol. 66/1-2, pp. 18-24, https://doi.org/10.1159/000363697. [72]

Walker, F. et al. (2017), "In the search for integrative biomarker of resilience to psychological stress", *Neuroscience & Biobehavioral Reviews*, Vol. 74, pp. 310-320, https://doi.org/10.1016/j.neubiorev.2016.05.003. [167]

Walton, K. et al. (2022), "A Big Five-Based Multimethod Social and Emotional Skills Assessment: The Mosaic™ by ACT® Social Emotional Learning Assessment", *Journal of Intelligence*, Vol. 10/4, https://doi.org/10.3390/jintelligence10040072. [112]

Watanabe, Y., Y. Motomura and E. Saeki (2022), "Development of emotional literacy and empathy among elementary-aged Japanese children", *International Journal of School and Educational Psychology*, Vol. 10/3, pp. 316-335, https://doi.org/10.1080/21683603.2020.1837699. [61]

Webb, T., E. Miles and P. Sheeran (2012), "Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation.", *Psychological Bulletin*, Vol. 138/4, pp. 775-808, https://doi.org/10.1037/a0027600. [59]

Weekley, J. et al. (2015), "Low-Fidelity Simulations", *Annual Review of Organizational Psychology and Organizational Behavior*, Vol. 2, pp. 295-322, https://doi.org/10.1146/annurev-orgpsych-032414-111304. [23]

Westera, W. et al. (2019), "Artificial intelligence moving serious gaming: Presenting reusable game AI components", *Education and Information Technologies*, Vol. 25/1, pp. 351-380, https://doi.org/10.1007/s10639-019-09968-2. [18]

West, M. et al. (2016), "Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling", *Educational Evaluation and Policy Analysis*, Vol. 38/1, pp. 148-170, https://doi.org/10.3102/0162373715597298. [12]

Wieckowski, A. and S. White (2017), "Eye-Gaze Analysis of Facial Emotion Recognition and Expression in Adolescents with ASD", *Journal of Clinical Child and Adolescent Psychology*, Vol. 46/1, pp. 110-124, https://doi.org/10.1080/15374416.2016.1204924. [157]

Wigelsworth, M. et al. (2010), "A review of key issues in the measurement of children's social and emotional skills", *Educational Psychology in Practice*, Vol. 26/2, pp. 173-186, https://doi.org/10.1080/02667361003768526. [11]

Wilkinson, M., S. Brantley and J. Feng (2021), *A Mini Review of Presence and Immersion in Virtual Reality*, SAGE Publications Inc., https://doi.org/10.1177/1071181321651148. [181]

Windingstad, S. et al. (2011), "Measures of Emotional Intelligence and Social Acceptability in Children: A Concurrent Validity Study", *Canadian Journal of School Psychology*, Vol. 26/2, pp. 107-126, https://doi.org/10.1177/0829573511406510. [121]

Wu, B., X. Yu and X. Gu (2020), *Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis*, Blackwell Publishing Ltd, https://doi.org/10.1111/bjet.13023. [177]

Xie, J. et al. (2023), "Brain Activation Differences of Six Basic Emotions Between 2D Screen and Virtual Reality Modalities", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 31, pp. 700-709, https://doi.org/10.1109/TNSRE.2022.3229389. [188]

Yoon, C. (2017), "A validation study of the Torrance Tests of Creative Thinking with a sample of Korean elementary school students", *Thinking Skills and Creativity*, Vol. 26, pp. 38-50, https://doi.org/10.1016/j.tsc.2017.05.004. [103]

Zamarro, G. et al. (2020), "Validation of survey effort measures of grit and self-control in a sample of high school students", *PLoS ONE*, Vol. 15/7, https://doi.org/10.1371/journal.pone.0235396. [46]

Zhang, Q. et al. (2018), "Correlations of hair level with salivary level in cortisol and cortisone", *Life Sciences*, Vol. 193, pp. 57-63, https://doi.org/10.1016/j.lfs.2017.11.037. [173]

Zhuang, X. et al. (2008), "Development and validity evidence supporting a teamwork and collaboration assessment for high school students", *ETS Research Report Series*, Vol. 2008/2, pp. i-51, https://doi.org/10.1002/j.2333-8504.2008.tb02136.x. [20]

Annex A: Search terms for SES

Annex Table 1. OECD SES Domains, key individual SES and their equivalent search terms

SES domains	SES	SES synonyms / associated keywords
Task performance (Conscientiousness)	Self-control	Self-discipline
	Responsibility	Trustworthiness
	Persistence	Perseverance
	Achievement motivation	Mastery skills
Emotion regulation (Emotional stability)	Stress Resistance	Resilience, Stress coping, Anxiety (antagonist term)
	Optimism	Positive emotion
	Emotional control	Emotional stability
Collaboration (Agreeableness)	Empathy	Compassion
	Trust	-
	Co-operation	Collaboration
Open-mindedness (Openness to experience)	Tolerance	Respect, Cultural flexibility
	Curiosity	-
	Creativity	Imagination, Flexible thinking
Engaging with others (Extraversion)	Sociability	-
	Assertiveness	Dominance, Leadership
	Energy	Enthusiasm
Supplementary skills	Critical thinking	Independence
	Metacognition	Self-awareness
	Self-efficacy	Self-esteem, Locus of control, Growth mindset
	Perspective-taking	Theory of Mind, Mentalising
	Social problem-solving	Conflict resolution
	Grit	(Direct links with Persistence and Achievement Motivation)

Note: Chernyshenko, O., M. Kankaraš and F. Drasgow (2018), *Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills*, OECD Education Working Papers, No. 173, OECD Publishing, Paris, https://doi.org/10.1787/db1d8e59-en

Conscientiousness *refers to, on the one side, the tendency of individuals for self-controlled, organised, and cautiously planned behaviour; and on the other, ambitious, persistent and dedicated effort in achieving personal goals.*
Extraversion *represents the tendency to seek the company of others, to initiate and maintain connections, and to feel comfortable in the presence of others. Extroverted individuals are also more likely to show assertiveness in social situations and provide leadership. They are often characterised by high levels of energy and zest for life. If extraversion partly refers to the quantity of interpersonal relations, agreeableness refers to their quality.*
Agreeableness *refers to the tendency to be more co-operative, maintain positive relations and minimise interpersonal conflict. These individuals are more likely to show active concern for the well-being of others and to hold positive beliefs about people in general.*
Emotional stability *represents the degree to which individuals are able to control their emotional responses and moods as well as the quality of their emotional states in general. Persons with high degrees of emotional stability will show more resilience in stressful situations, will be less likely to experience anger, irritation or sudden changes of mood, and will tend to have a better view of the world and outlook of the future.*
Openness to experience *is reflected in two main aspects. One involves the degree to which people are open to intellectual stimulation in general, as reflected in their intellectual curiosity, imagination, creativity, preference for novelty and variation. The other aspect is shown in the degree to which persons prefer experiential stimulation, as represented in their appreciation of art, aesthetic experiences, self-reflection and self-exploration.*

Annex B: Full list of identified behavioural assessment tools

Assessment tool	Source(s)
1) Academic Diligence Task	Galla et al. (2014): https://doi.org/10.1016/j.cedpsych.2014.08.001 Fuhrmann et al. (2019): https://doi.org/10.1080/17588928.2018.1504762 Zamarro et al. (2020): https://doi.org/10.1371/journal.pone.0235396 Wiese et al. (2018): https://doi.org/10.1111/jopy.12322
2) Assessment of Social Perspective-taking Performance (ASPP)	Kim et al. (2018): https://doi.org/10.1016/j.appdev.2018.05.005
3) Athenea	Giglioli et al. (2021): https://doi.org/10.3390/app11041971
4) Beach Balls Task	Jiminez-Soto et al. (2022): https://doi.org/10.1016/j.actpsy.2022.103528
5) Circuit Runner	Stoeffler et al. (2020): https://doi.org/10.1016/j.chb.2019.05.033 Polyak, von Davier & Peterschmidt (2020): https://doi.org/10.3389/fpsyg.2017.02029
6) Combined Stories Test (COST)	Achim et al. (2012): https://doi.org/10.1016/j.psychres.2011.10.011 Thibaudeau et al. (2018): https://doi.org/10.1016/j.psychres.2018.01.026
7) CREA	Rafner et al. (2023): https://doi.org/10.1080/10400419.2023.2198845
8) Crystal Island	Taub et al. (2017): https://doi.org/10.1016/j.chb.2017.01.038 Emerson et al. (2020): https://doi.org/10.1111/bjet.12992 Taub & Azevedo (2018): https://doi.org/10.5281/zenodo.3554711
9) Divergent Thinking Task	An, Song & Carr (2016): https://doi.org/10.1016/j.paid.2015.10.040
10) Emodiscovery	Pacella et al. (2018): https://doi.org/10.1016/j.compedu.2018.04.005 López-Pérez & Pacella (2021): https://doi.org/10.1037/emo0000690
11) EMOTICOM	Bland et al. (2016): https://doi.org/10.3389/fnbeh.2016.00025
12) Emotion Recognition Index	Scherer & Scherer (2011): https://link.springer.com/article/10.1007/s10919-011-0115-4 Schlegel et al. (2017): https://link.springer.com/article/10.1007/s11031-017-9631-9
13) Emotional Go/Nogo Task	Tottenham, Hare & Casey (2011): https://doi.org/10.3389/fpsyg.2011.00039 Thompson, van Reekum & Chakrabarti (2022): https://doi.org/10.1007/s42761-021-00062-w
14) Emotional Literacy Test with Hypothetical Scenarios	Watanabe, Motomura & Saeki (2020): https://doi.org/10.1080/21683603.2020.1837699
15) Emotional Stroop task	Zhang et al. (2023): https://psycnet.apa.org/doi/10.1037/bul0000389 Zinchenko et al. (2020): https://doi.org/10.1016/j.ijpsycho.2019.10.018
16) ENACT	Marocco et al. (2015): https://link.springer.com/chapter/10.1007/978-3-319-24258-3_37 Dell'Aquila et al. (2016): https://link.springer.com/chapter/10.1007/978-3-319-06311-9_5 Yaşar Akyar & Demirhan (2022): https://link.springer.com/article/10.1007/s10639-021-10823-6
17) Extrinsic Affective Simon Task	Degner & Wentura (2008): https://doi.org/10.1002/ejsp.536
18) Facial Emotion Recognition and Empathy Test (FERET)	Coskun (2019): https://doi.org/10.1017/cha.2018.51
19) Faculty Game	Sher, Levi-Keren & Gordon (2019): https://doi.org/10.1007/s11423-019-09665-4
20) Faux Pas Recognition Test	Baron-Cohen et al. (1999): https://link.springer.com/article/10.1023/A:1023035012436 Hayward & Homer (2017): https://doi.org/10.1111/bjdp.12186 Smogorzewska et al. (2022): https://doi.org/10.1016/j.lindif.2021.102111 Osterhaus & Bosacki (2022): https://doi.org/10.1016/j.dr.2022.101021
21) Future expectations task (FET)	Bamford & Lagattuta (2019): https://doi.org/10.1111/cdev.13293
22) Game for Social Anxiety	Dechant, Frommel & Mandryk (2021): https://doi.org/10.3389/fpsyg.2021.760850
23) Hall of Heroes	DeRosier & Thomas (2019): https://doi.org/10.1155/2019/6981698 Irava et al. (2019): https://doi.org/10.1177/0047239519854042
24) Hinting Task	Klein et al. (2020): https://doi.org/10.1002/mpr.1827 Nagendra et al. (2018): https://doi.org/10.1016/j.psychres.2017.09.074 Klein, Springfield & Pinkham (2022): https://doi.org/10.1080/23279095.2022.2082875 Pinkham et al. (2016): https://doi.org/10.1093/schbul/sbv056 Fiuza Cruz et al. (2022): https://pubmed.ncbi.nlm.nih.gov/36619846/

25) Intergroup Prisoner's dilemma	Thielmann et al. (2021): https://doi.org/10.1525/collabra.19004
26) Interpersonal Perception Task (IPT-15)	Iizuka, Patterson & Matchen (2022): https://link.springer.com/article/10.1023/A:1020761332372
27) Laboratory coping and emotion regulation task	Bettis et al. (2019): https://doi.org/10.1080/15374416.2018.1466306
28) Levels of Emotional Awareness Scale (LEAS) & Levels of Emotional Awareness Scale - Children (LEAS-C)	Lane & Smith (2021): https://doi.org/10.3390/jintelligence9030042 Siegling, Saklofske & Petrides (2015): https://doi.org/10.1016/B978-0-12-386915-9.00014-0 Bajgar et al. (2005): https://doi.org/10.1348/026151005X35417 Veirman, Fontaine & Van Ryckeghem (2016): https://doi.org/10.1037/pas0000261
29) Mayer–Salovey–Caruso Emotional Intelligence Test-Youth Version (MSCEIT-YV)	Peters, Kranzler & Rossen (2009): https://doi.org/10.1177/0829573508329822 Rivers et al. (2012): https://doi.org/10.1177/0734282912449443
30) Minecraft Task	Shaw (2023): https://psycnet.apa.org/doi/10.1037/aca0000456
31) Mirror Tracing Frustration Task (MTFT)	Meindl et al. (2019): https://psycnet.apa.org/doi/10.1037/emo0000492 Zamarro et al. (2020): https://doi.org/10.1371/journal.pone.0235396
32) Multifaceted Empathy Test	Drimalla et al. (2019): https://doi.org/10.1080/02699931.2019.1596068 Foell et al. (2018): https://psycnet.apa.org/doi/10.1007/s10862-018-9664-8 Grainger et al. (2023): https://doi.org/10.1177/10731911221127902
33) Noah Kingdom	Wang, Liu & Hau (2022): https://doi.org/10.1007/s10639-021-10777-9
34) Persistence, Effort, Resilience, and Challenge-Seeking Task (PERC Task)	Porter et al. (2020): http://dx.doi.org/10.18608/jla.2020.71.2 Porter et al. (2020b): https://doi.org/10.1177/1948550620909738 Selmeczy et al. (2021): https://doi.org/10.1016/j.cogdev.2021.101062
35) Physics Playground (Creativity)	Shute & Rahimi (2021): https://doi.org/10.1016/j.chb.2020.106647
36) Physics Playground (Persistence)	Ventura & Shute (2013): http://dx.doi.org/10.1016/j.chb.2013.06.033
37) Physics Playground (Co-operation)	Sun et al. (2022): https://doi.org/10.1016/j.chb.2021.107120 Andrews-Todd et al. (2023): https://doi.org/10.1016/j.compedu.2023.104928
38) Pizzagame	Keil et al. (2017): https://link.springer.com/article/10.3758/s13428-016-0799-9
39) Posterlet	Cutumisu, Chin & Schwartz (2019): https://doi.org/10.1111/bjet.12796 Cutumisu & Schwartz (2021): https://doi.org/10.1016/j.compedu.2021.104215
40) Prisoner's dilemma	Mengel (2017): https://doi.org/10.1111/ecoj.12548 Thielmann et al. (2021): https://doi.org/10.1525/collabra.19004
41) Public Goods Game	Thielmann et al. (2021): https://doi.org/10.1525/collabra.19004 van Dijk & De Dreu (2021): https://doi.org/10.1146/annurev-psych-081420-110718
42) Questions Worlds	Tor & Gordon (2020): https://doi.org/10.18178/ijiet.2020.10.8.1433
43) Reading the Mind in the Eyes Test - Child Version (RMET-C)	Rosso & Riolfo (2020): https://doi.org/10.3389/fpsyg.2020.586065 Hayward & Homer (2017): https://doi.org/10.1111/bjdp.12186
44) Rumble's Quest	Day et al. (2019): https://doi.org/10.1007/s10567-019-00282-4 Allen et al. (2023): https://doi.org/10.1002/ajs4.258 Freiberg et al. (2023): https://doi.org/10.1080/10888691.2023.2208753
45) Seaball	Song & Sparks (2017): https://doi.org/10.1177/0735633117740605
46) SELweb EE (early elementary)	McKown (2019): https://doi.org/10.1177/0734282917749682 McKown et al. (2023): https://doi.org/10.1177/10731911211046044
47) SELweb LE (late elementary)	McKown, Russo-Ponsaran & Karls (2022): https://doi.org/10.1111/sode.12641
48) Short version of the Geneva Emotion Recognition Test (GERT-S)	Schlegel & Scherer (2016): https://link.springer.com/article/10.3758/s13428-015-0646-4 Leutner et al. (2021): https://doi.org/10.1108/JMP-01-2020-0023
49) Simoland	Schönbrodt & Asendorpf (2011): https://psycnet.apa.org/record/2011-01588-001 Schönbrodt & Asendorpf (2012): https://doi.org/10.1111/j.1467-6494.2011.00736.x Luhmann et al. (2015): https://doi.org/10.1080/02699931.2014.922053
50) Simulation game	Kleitman et al. (2022): https://doi.org/10.3389/fpsyg.2021.717568
51) Social Attribution Task-Multiple Choice (SAT-MC)	Johannesen et al. (2013): http://dx.doi.org/10.1155/2013/830825 García-Guerrero et al. (2021): https://actaspsiquiatria.es/index.php/actas/article/view/166 Lee & Choi (2022): https://doi.org/10.3389/fpsyt.2022.883212
52) Torrance Test of Creative Thinking (TTCT)	Perry & Karpova (2017): http://dx.doi.org/10.1016/j.tsc.2017.02.017 Yoon (2017): http://dx.doi.org/10.1016/j.tsc.2017.05.004

53) Trust Game (also called Investment Game)	Thielmann et al. (2021): https://doi.org/10.1525/collabra.19004
54) Video-Social-Emotional Information Processing (V-SEIP)	Coccaro et al. (2017): http://dx.doi.org/10.1016/j.psychres.2016.11.004 Coccaro et al. (2021): https://doi.org/10.1016/j.jpsychires.2021.01.015
55) Virtual Environment for Social Information Processing (VESIP)	Russo-Ponsaran et al. (2018): https://doi.org/10.1002/aur.1889 Russo-Ponsaran et al. (2021): https://doi.org/10.1111/sode.12512
56) Vox Populi	Gjicali, Finn & Hebert (2020): https://doi.org/10.1016/j.compedu.2020.103959
57) Zoo U	DeRosier, Craig & Sanchez (2012): https://doi.org/10.1155/2012/654791 Craig, DeRosier & Watanabe (2015): https://doi.org/10.1089/g4h.2014.0075 DeRosier & Thomas (2018): http://dx.doi.org/10.1016/j.appdev.2017.03.001