frontiers in Education

REVIEW
published: 17 November 2021
doi: 10.3389/feduc.2021.687661

Formative Assessment of Social-Emotional Skills Using Rubrics: A Review of Knowns and Unknowns

Gina Pancorbo[1,2]*, *Ricardo Primi*[3,2], *Oliver P. John*[4,2], *Daniel Santos*[5,2] *and Filip De Fruyt*[1,2]

[1]*Department of Developmental, Personality and Social Psychology, Ghent University, Ghent, Belgium,* [2]*EduLab21, Instituto Ayrton Senna, Saõ Paulo, Brazil,* [3]*Post Graduate Program in Psychology, Universidade São Francisco, Campinas, Brazil,* [4]*Department of Psychology, University of California, Berkeley, Berkeley, CA, United States,* [5]*University of São Paulo, Ribeirão Preto, Brazil*

Educational practitioners have been increasingly interested in the use of formative assessment and rubrics to develop social-emotional skills in children and adolescents. Although social-emotional rubrics are nowadays commonly used, a thorough evaluation of their psychometric properties has not been conducted. In this scoping review, we examine the knowns and unknowns of the use of formative assessment approaches and rubrics in social-emotional learning. We first review initiatives of formative assessment and development of rubrics to assess social-emotional skills. Then, we discuss challenges associated with the development and use of rubrics to evaluate social-emotional skills in youth focusing on 1) assessment of single skills versus assessment of a comprehensive taxonomy of skills; 2) developing rubrics' performance level descriptions that accurately describe increasing mastery of skills; 3) obtaining adequate internal consistency and discriminant validity evidence; 4) self-reports versus observer reports of skills; and finally 5) how to account for adolescents' development in the construction of rubrics. This review outlines a research agenda for the psychometric study of rubrics to be used in social-emotional skill assessment.

Keywords: rubrics, social-emotional skills, formative assessment, scoping review, challenges

OPEN ACCESS

Edited by:
Anders Jönsson,
Kristianstad University, Sweden

Reviewed by:
Juan Fraile,
Universidad Francisco de Vitoria, Spain
Anastasiya A Lipnevich,
The City University of New York, United States

***Correspondence:**
Gina Pancorbo
Gina.Pancorbo@ugent.be

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *29 March 2021*
Accepted: *29 October 2021*
Published: *17 November 2021*

Citation:
Pancorbo G, Primi R, John OP, Santos D and De Fruyt F (2021) Formative Assessment of Social-Emotional Skills Using Rubrics: A Review of Knowns and Unknowns. Front. Educ. 6:687661. doi: 10.3389/feduc.2021.687661

INTRODUCTION

In the past decades, social-emotional skills (SEMS) have received increasing attention in educational settings because of their role in students' positive development (Kern et al., 2016; Taylor et al., 2017). Several studies have suggested that a *deeper* learning of SEMS involves the use of formative assessment approaches and self-assessment tools to enhance self-regulating capacities in students (Trilling and Fadel, 2009; Griffin et al., 2011; Pellegrino and Hilton, 2012; OECD, 2015). Rubrics are an attractive way to formatively assess SEMS because their concrete and behaviorally-oriented criteria may facilitate students' self-reflection in terms of where they situate themselves and what kind of social-emotional mastery level they want to achieve in the future (Panadero and Jönsson, 2013). However, despite their importance, very few attempts have been made to develop rubrics to assess social-emotional skills and evaluate their psychometric properties in youth. The objective of this scoping review is to present what we know and what we need to know on the use of formative assessment approaches and rubrics in SEMS learning, and to discuss the challenges associated with the development and use of rubrics to evaluate SEMS in youth.

TABLE 1 | Social-emotional skill domains and facets Primi et al. (2017).

Domain	Facet	Definition
Self-management	Determination	Setting goals and high standards, motivating oneself, working very hard, and applying oneself fully to the task, work, or project at hand
	Organization	Possessing organizational skills and meticulous attention to detail that are useful for planning and executing plans to reach longer-term goals
	Focus	Focusing attention and concentrating on the current task, and avoiding distractions
	Persistence	Overcoming obstacles in order to reach important goals
	Responsibility	Possessing time management skills, being punctual, and honoring commitments
Engaging with others	Social initiative	Approaching and connecting with others, both friends and strangers, initiating, maintaining, and enjoying social contact and connections
	Assertiveness	Speaking up, voicing opinions, needs, and feelings, and exerting social influence
	Enthusiasm	Showing passion and zest for life; approaching daily tasks with energy, excitement, and a positive attitude
Amity	Compassion	Using empathy and perspective-taking skills to understand the needs and feelings of others, acting on that understanding with kindness and consideration of others
	Respect	Treating others with respect and politeness
	Trust	Assuming that others generally have good intentions and forgiving those that have done wrong
Negative-emotion regulation	Stress modulation	Modulating anxiety and response to stress
	Self-confidence	Feeling satisfied with self and current life, having positive thoughts about self, and maintaining optimistic expectations
	Frustration tolerance	Regulating temper, anger, and irritation; maintaining tranquillity and equanimity in the face of frustrations
Open-mindedness	Intellectual curiosity	Mustering interest in ideas and a passion for learning, understanding, and intellectual exploration
	Creative imagination	Generating novel ways to think about or do things through experimenting, tinkering, learning from failure, insight, and vision
	Artistic interest	Valuing, appreciating, and enjoying design, art, and beauty, which may be experienced or expressed in writing, visual and performing arts, music, and other forms of self-actualization

Notes. *Extracted from Primi et al. (2017).*

Social-Emotional Skills

Social-emotional skills are defined as "individual characteristics that originate from biological predispositions and environmental factors, are manifested as consistent patterns of thoughts, feelings, and behaviors, continue to develop through formal and informal learning experiences, and that influence important socio-economic outcomes throughout the individual's life" (De Fruyt et al., 2015; p. 279). Cumulative evidence has shown the importance of SEMS on different spheres of individuals' life, ranging from educational attainment, job performance, employability, physical and mental health, and personal and societal well-being, among others (see Chernyshenko et al., 2018 for a review). Longitudinal research has provided evidence on the supportive and protective functions of SEMS; supportive, because they are associated with healthy development, and protective, as they prevent the exposure to or help to cope with risk factors across people's lives (Domitrovich et al., 2017).

Although there is considerable variability in the number and nature of skills included in different SEMS frameworks (see Abrahams et al., 2019, for a review), there is convincing evidence that these skills can be grouped under the umbrella of the Big Five personality factors of Conscientiousness, Neuroticism, Extraversion, Agreeableness and Openness to Experience (see Shiner et al., 2021 for a review of the Big Five)[1]. Providing additional support for this perspective, Primi et al. (2016) examined the overlap and commonalities across more than 200 items of eight scales that are frequently used to measure SEMS in children and youth (i.e., ages 10–17) with the idea of representing their common variance by a more manageable group of SEMS.

Based on the evidence above, Primi et al. (2017) proposed an integrative framework of five domains of SEMS with a set of more specific skills grouped under these five domains that cover the broad spectrum of social-emotional functioning in youth (see **Table 1** for further detail). The framework aims to capture skills that have predictive value and could serve as stand-alone skills or building blocks of more sophisticated "hybrid" constructs like citizenship, critical thinking, or entrepreneurship, among others. The framework is also useful to support policymakers, researchers, and educational specialists for policy decisions, for example to make decisions about the kind of skills that have to be included in educational curricula (Abrahams et al., 2019).

Formative Assessment and Rubrics

The assessment of SEMS is critical to elucidate students' social-emotional strengths and weaknesses, to provide useful information to guide social-emotional learning, and, ultimately, to contribute to students' short and long-term positive outcomes (Durlak et al., 2015). Durlak et al. (2015) proposed that in order to assess SEMS effectively, educational systems should consider clear standards (i.e., goals and benchmarks) to follow students' progress, and develop evidence-based curricula and instruction guidelines, as well as formative and summative approaches to stimulate, monitor and evaluate students' learning progress.

Contrary to summative assessment, where tests are used to evaluate students' learning at a given point of time, formative

[1]Where useful and necessary, we will refer to findings from personality research and literature to discuss parallels with SEMS.

assessment focuses on the use of tests to continuously improve students' performance *during* the learning process (Pellegrino and Hilton, 2012). Thus, formative assessment is the process where students actively and continuously engage in assessment activities such as self-, peer, and teacher feedback to achieve objectives and develop students' self-regulation and meta-cognitive capacities (Bolden et al., 2020). A growing number of studies support its pedagogical use (e.g., Andrade and Brookhart, 2020; P. P. Chen and Bonner, 2020; Durlak et al., 2015; Marshall and Drummond, 2006) and have provided evidence of its positive effects on students' achievement. A meta-analysis conducted by Hattie (2009) concluded that formative assessment was one of the most critical pedagogical strategies for improving students' learning. Likewise, Kingston and Nash (2011) found that formative assessment had a modest but significant effect on learning, while a meta-analysis by Graham et al. (2015) showed that the formative use of feedback by teachers yielded a larger effect size on students' writing achievement.

Rubrics are an attractive and innovative promising way to formatively assess SEMS because they have the potential to help students to reflect on their strengths and difficulties and guide their performance (Andrade, 2007; Panadero and Romero, 2014; Jönsson and Panadero, 2017). Moreover, the characteristics of rubrics' design may facilitate formative assessment processes. For doing so, rubrics should include explicit criteria that clearly explain what is assessed (Brookhart, 2018). In that sense, Brookhart (2018) stated that clear and quality criteria were crucial for students to conceptualize their learning goals and take the necessary steps to achieve them throughout the formative process. Rubrics should also include performance level descriptions that have descriptive rather than evaluative language, which can facilitate constructive feedback (Jönsson and Panadero, 2017). These characteristics are deemed to increase the transparency of the assessment and, consequently, promote self-reflection, self-regulated learning and feedback from peers and teachers (Jönsson and Panadero, 2017). However, very few studies have paid attention to clearly define and communicate how rubrics look like and how they can be used (Dawson, 2017). In other words, not enough information about the object of assessment, the scoring strategy, the evaluative criteria or the quality descriptions is provided in many studies, which might affect the transparency of rubrics' use. Additionally, rubrics have been mostly used to assess cognition-related competencies like writing, mathematics, or science (e.g., Lallmamode et al., 2016), and only a few attempts have been made to develop rubrics for assessing social-emotional skills in youth and evaluate their psychometric properties. Therefore, more steps are needed to maximize the use of rubrics for social-emotional skills assessment.

Psychometric Properties of Rubrics

Rubrics are defined as "a type of matrix that provides scaled levels of achievement or understanding for a set of criteria or dimensions of quality for a given type of performance" (Allen and Tanner, 2006, p. 197). They have been traditionally recognized as effective scoring guides because they enhance consistency in scores and support valid judgments of performance (Brookhart and Chen, 2015). Furthermore, research has suggested that rubrics can promote learning and support instruction because their defined skill levels create clear expectations of performance, making scoring more transparent and facilitating teachers' feedback on students' work (Jönsson and Svingby, 2007; Brookhart and Chen, 2015; Jönsson and Panadero, 2017). Likewise, other studies have claimed that rubrics' explicit assessment criteria may facilitate students' self-assessment in formative assessment settings, and help them to navigate in the learning progression of a specific competence or skill (Panadero and Jönsson, 2013). Hence, it is not surprising that rubrics have been widely used to assess individuals' academic achievements to evaluate educational programs and improve instruction across different education levels (Moskal and Leydens, 2000; Jönsson and Svingby, 2007; Reddy and Andrade, 2010).

Nevertheless, evidence on the contribution of rubrics to support students' learning is still inconclusive. On the one hand, Jönsson and Svingby (2007) review of 75 studies about rubrics could not draw a definite conclusion on the effect of rubrics on students' learning due to the diversity of methods and results they encountered. On the other hand, a more recent review by Brookhart and Chen (2015) of 63 studies showed that the use of rubrics contributed to students' achievement in different cognitive areas such as writing, general sciences, physics, and business education, among others. Additionally, other studies have shown that rubrics' use increases students' self-efficacy and self-regulation capacities in elementary and secondary school students (Andrade et al., 2009; Panadero et al., 2012).

Cumulative evidence has pointed out that rubrics' positive contributions may depend on a series of moderating factors (e.g., Wollenschläger et al., 2016). A review by Panadero and Jönsson (2013) on the formative use of rubrics concluded that rubrics might affect performance through one or more of the following processes: provide transparency to assessment, reduce students' anxiety, enable feedback, increase students' self-efficacy, or promote students' self-regulation. Likewise, several other studies have suggested that merely handing rubrics to students is not enough for improving performance (Panadero et al., 2012; Panadero and Jönsson, 2013; Wollenschläger et al., 2016). Instead, rubrics seem to be more conducive to learning when accompanied by constructive feedback with information about the task (i.e., "How am I going?") and the next steps to improve performance (i.e., "Where to go next?"). Moreover, rubrics' may promote students' positive SRL and prevent detrimental effects like stress when enough time is given to students to become familiar with the instrument through external guidance and practice (Panadero and Romero, 2014). As suggested by Brookhart and Chen (2015, p. 363), "scoring rubrics necessarily need to be part of instructional and formative assessment methods" to support students' learning.

In recent years, a growing number of studies have evaluated rubrics' psychometric properties. Two reviews by Jönsson and Svingby (2007) and Brookhart and Chen (2015) found that most studies report on the inter-rater reliability of rubrics (i.e., the

degree of consistency between different rater scores; Reddy and Andrade, 2010). In that sense, Brookhart and Chen (2015) suggested that rubrics yielded adequate inter-rater reliability levels and supported their use for decisions about individuals' performance, especially when their criteria are clear and raters receive training. By contrast, Jönsson and Svingby (2007) results showed that rubrics' inter-rater reliability estimates were relatively low for traditional testing, which led them to conclude that rubrics might not provide reliable scores for summative assessment purposes. However, they suggested that lower reliability levels could be considered acceptable for low-stakes assessments and when rubrics are used for classroom monitoring purposes (Jönsson and Svingby, 2007).

Similarly, a variety of methods have been used to collect evidence on different aspects of rubrics' validity (Jönsson and Svingby, 2007; Brookhart and Chen, 2015), including opinions of experts about rubrics' constructs (i.e., content-related validity), the correlation of rubrics' scores with external indicators (i.e., criterion-related validity), factor analyses to inspect the structural aspects of rubrics' scores (i.e., internal structure validity), as well as perceptions of teachers and students about the use of rubrics (i.e., consequential validity). Despite this great variety of information sources, Brookhart and Chen (2015) suggested that most studies used only one or two of these indicators for evaluating the validity of rubrics in mainly post-secondary school samples. Nevertheless, the authors concluded that "rubrics can produce valid and useful scores for grading or program evaluation in post-secondary education" (p. 362).

METHODS

The present study reviews the knowns and unknowns of the use of formative assessment and rubrics to evaluate social-emotional skills. We also discuss the challenges of the development and use of rubrics to evaluate SEMS in youth. We employed a scoping review method because of the novelty of the research area and because the objective was to map the key concepts and ideas as well as the main available evidence that supports the topic (Arksey and O'Malley, 2005). We followed the main guidelines of the five steps of the methodological framework proposed by Arksey and O'Malley (2005): 1) Identify the research question(s), 2) Identify relevant studies, 3) Select studies, 4) Chart the data, and 5) Collate, summarize, and report the results.

Identifying the Research Question

The research question of the present review is "What are the experiences of using formative assessment and rubrics to assess SEMS and what are the challenges to foster this field?"

Identifying Relevant Studies

To select articles, we searched for keywords in the databases Web of Science, ProQuest, and Google Scholar between January and December of 2020 limiting the start date to the year 2000. The search terms were extensive because of the nature of the topic. Examples of these terms were "assess*, AND social-emotional skills, AND rubrics", "assess*, AND social-emotional skills, AND formative assessment", "rubrics*, AND social-emotional skills, AND adolescence", etc. We included studies from peer-reviewed journals of the areas of education and psychology, as well as books, book chapters and reports from educational interventions. We also found valuable bibliography in the reference lists of the studies collected in the database searches. Additionally, we inquired five specialists in the area who suggested to review a small number studies (n = 4).

Study Selection: Inclusion and Exclusion Criteria

To be selected for inclusion in this review, we required that the papers were 1) published in English, Spanish or Portuguese; 1) report the analysis of quantitative and/or qualitative data or conduct a systematic or non-systematic review; 2) address the assessment of social-emotional skills; 3) address the use of rubrics to assess social-emotional skills; 4) address formative assessment of social-emotional skills; 5) address the psychometric properties of social-emotional skills' measures; 6) address the psychometric properties of rubrics; 7) address the development of social-emotional skills in children and adolescents. After familiarizing with the literature, new criteria were developed to guide decisions of inclusion and exclusion to the review (Arksey and O'Malley, 2005). For example, instead of "social-emotional skills", we used the terms "life skills", "soft skills" or specific nomenclature of the skills such as "creativity", "responsibility", etc. We followed the nomenclature of the integrative framework of Primi et al. (2017; see **Table 1**) to guide our search.

In the meanwhile, a study or article was excluded from this review if it was 1) published before year 2000; 2) published in a non-peer reviewed journal; 3) published in other language than English, Spanish or Portuguese; 4) published in journals that were not related to the education or psychology fields; 5) address the use of rubrics in summative assessment interventions.

Charting the Data

In this next step, we charted the data, which according to Arksey and O'Malley involves "sifting, charting and sorting" the collected material. After reading all the material and applying the inclusion and exclusion criteria, two researches sorted and classified the data in a "data charting form" by the authors' names, title, year, type research design (i.e., empirical, non-empirical, literature review, scoping review, etc.), objective, as well as age and origin of the sample used in the study. This classification helped us to select a total of 66 studies that dated from the year 2000–2020 that had a broad variety of research designs. We identified 41 empirical studies. From these, 30 had a cross-sectional design and 11 were longitudinal. Another group of 19 studies were reviews and from these, only three were literature or scoping reviews. The others were reviews that did not report following a systematic method to collect and analyze data. We also included one book and two book chapters. Three other documents were reports from international organizations. Around a third of the empirical studies involved the participation of adolescents (n = 22), eight involved the participation of children and 13 dealt with adults

(i.e., university students, teachers, pre-service teachers, parents or other caregivers, etc). Some studies involved the participation of more than one of these age groups. Most of the samples of the empirical studies were from the United States (n = 23), but other studies recruited participants from a wide variety of different countries such as Australia, Belgium, Brazil, Canada, Estonia, Spain, etc.

Collating, Summarizing, and Reporting the Results

The selected 66 studies were sorted again according to their thematic area. After reading all the material, two researches identified eight categories in which the studies could be classified: 1) studies that used formative assessment to evaluate one or more social-emotional skills; 2) studies that developed rubrics to assess one or more social-emotional skills; 3) studies that were about the implications of measuring social-emotional skills; 4) studies that were about the design and evaluation of performance level descriptions in rubrics; 5) studies that were about the psychometric properties of social-emotional skills' measures and/or rubrics; 6) studies that used rubrics as a self-assessment and/or observers' report method; 7) studies that involved self-reports and observers' reports of social-emotional skills; 8) studies that were about the development of social-emotional skills in youth.

In the following section we present the results of our findings giving a brief description of the reviewed studies and we also discuss the main implications of our results for researchers and practitioners.

RESULTS

Using Formative Assessment and Rubrics to Assess Social-Emotional Skills

A group of initiatives has already put into practice formative assessment strategies for developing specific social-emotional skills in classrooms with promising results (e.g., Brookhart, 2013; Andrade et al., 2014; Valle et al., 2016; Chen et al., 2017; Chen and Andrade, 2018; Bolden et al., 2020). Bolden et al. (2020) reviewed the effect of summative and formative assessment strategies on creativity learning in classrooms showing that interventions that used explicit and transparent criteria and which practices promoted students' self-assessment were effective in supporting creativity. This review further suggested that creativity assessment is more accurate and meaningful when teachers and students are provided with assessment criteria (i.e., the definition of creativity and a list of related behaviors) and have a clear conceptual understanding of what they are assessing. Concerning self-assessment practices, Bolden et al. (2020) showed that strategies of self-reflection (i.e., students using criteria to reflect on their or others' work) could enhance higher levels of creative and divergent thinking and verbal and figural creativity, among other processes.

Similarly, other studies have evaluated the effect of formative assessment on more specific social-emotional behaviors like arts' performance (e.g., Valle et al., 2016; Chen et al., 2017; Chen and Andrade, 2018; Fei). A group of these studies showed that formative assessment strategies where students used rubrics or checklists to self-assess their work or assess their peers' work had a significant positive effect of around 0.25 on arts achievement (Chen et al., 2017; Chen and Andrade, 2018). Chen et al. (2017) concluded that: "student learning in the arts is measurably deepened when students know what counts, receive feedback from their teachers, themselves, and each other, and have opportunities to revise" (p. 308).

Meanwhile, research on the use of rubrics to assess social-emotional skills is still limited. However, some studies have focused on developing rubrics to assess social-emotional related competencies such as theater arts' skills, creativity, music, and critical thinking, and evaluating how much they can contribute to assessing students' performance (e.g., F. Chen and Andrade, 2018; Lindström, 2006; Menéndez-Varela and Gregori-Giralt, 2018; Vincent-Lancrin et al., 2019). From these studies, very few have paid attention to assess the psychometric properties of rubrics. For example, a recent study of Susman-Stillman et al. (2018) constructed and tested a Preschool Theatre Arts Rubric including a group of scales (e.g., vocalization, focus/persistence/commitment to the play, and collaboration/awareness of others) for the observation of theater arts skills in preschool children. Their results showed adequate internal consistency and inter-rater reliability but weak evidence of convergent validity (i.e., the degree to which two measures that assess similar constructs are related) with measures of preschool-learning related behaviors and oral narrative skills.

Other studies have reported the educational benefits of using rubrics in areas like creativity and music learning. Brookhart (2013) created a rubric that measured creativity to help teachers and students to clarify the criteria and share with students "what they are aiming for, where they are now, and what they should do next" (Brookhart, 2013, p.29). The four-level rubric (i.e., very creative, creative, ordinary/routine, and imitative) assessed four different areas of creativity: the variety of ideas, the variety of sources, the novelty of idea combinations, and the novelty of communication. As Brookhart (2013) reported, the assessment of creativity using rubrics not only helped teachers to assess and give feedback to students but also helped students in the process of thinking creatively. Concerning music learning, a literature review conducted by Blom et al. (2015) concluded that the use of rubrics to assess music performance enhanced students' self-reflection and motivated them to be more sensitive and critical about their work. Additionally, the authors highlighted that rubrics are a valuable peer and self-assessment tool for music learning. Hence, despite the growing interest in rubrics' use to assess social-emotional behaviors, many questions on the topic remain unanswered. Thus, the following section raises a group of challenges related to the development and use of rubrics to assess children and adolescents' social-emotional skills.

Challenges for Developing and Using Rubrics to Assess Social-Emotional Skills in Youth

Single Social-Emotional Skills or Social-Emotional Skills Taxonomies

Research on the assessment of SEMS using rubrics has mainly focused on a small number of specific skills instead of

operationalizing a full taxonomy of social-emotional skills (e.g., like the Senna framework described in **Table 1**). Assessing only one or a few skills imposes fewer constraints, whereas representing a full taxonomy of skills raises questions on the assumed structure of skills and their convergent and divergent validities. There are also more basic concerns and choices to make.

First, several authors recognize that there is considerable variability among skills included in different SEMS models (Kyllonen et al., 2014; Primi et al., 2016). Many of these models include skills that have different names but describe the same underlying construct (i.e., "jangle fallacy"; Kyllonen, 2016; Voogt and Roblin, 2012). This lack of shared definitions and overlap among constructs may have several implications for measurement (Ziegler and Brunner, 2016). Hence, the advantage of constructing an instrument based on a SEMS taxonomy that includes well-defined constructs and specifies how these are related to each other has the potential to improve the accuracy of the assessment (Kyllonen et al., 2014; Ziegler and Brunner, 2016).

Second, SEMS can be assessed at different levels, i.e., at a higher-order or more abstract domain level, representing the common core of a group of lower-order or more specific facets. The broad domain of Self-management, for example, groups variance shared by the specific skills of Organization, Focus, Persistence, Responsibility, and Determination. This choice may have different implications for measurement. On the one hand, an instrument that measures higher-order domains will have the advantage of comprising many different behavioral manifestations and predicting a wide variety of outcomes (i.e., high bandwidth; Ozer and Benet-Martínez, 2006). On the other hand, an instrument that measures lower-order facets may have the benefit of describing more specific behaviors and predicting particular outcomes with greater accuracy (i.e., high fidelity; Paunonen and Ashton, 2001). Moreover, an instrument that assesses a specific facet (e.g., organization skills) may have higher internal consistency, stronger inter-item correlations, and a more simple factor structure because all its items measure similar patterns of behavior (Soto and John, 2017). By contrast, an instrument that measures broader domains (e.g., the five domains of the Senna taxonomy) may have lower internal consistency estimates and a more complex factor structure with higher chances of shared variances among domains (Soto and John, 2017).

Building Performance Level Descriptions for Rubrics

Traditionally, rubrics have been constructed by first identifying the criteria for good work on which the scoring will be based (e.g., the structural organization and clarity in a writing assignment) and then defining the content of the categories of quality of work based on examples of good and bad performance (e.g., examples of good or bad organization in writing assignments, Nitko and Brookhart, 2007).

By contrast, when constructing rubrics to assess SEMS, the definition of the categories of "quality of work" or "performance" level descriptions should be based on developmental theories that describe the increasing mastery levels of skills according to the age of the student. For example, a rubric that assesses emotion regulation in adolescents would include as the first and the latest performance levels descriptions that define, respectively, the least and the highest level of development that we can expect for emotion regulation at that age. The performance level descriptions in-between the first and the last would describe the intermediate steps adolescents could take on the hypothetical continuous path of emotion regulation performance. On top of these developmental processes and "defined" stages, there are also individual differences shaping and affecting such development.

Hence, designing rubrics for the assessment of SEMS in youth is not a simple task because social-emotional instruments have traditionally included descriptors of the absence (i.e., false-keyed items) and presence (i.e., true-keyed items) of the constructs (e.g., the BFI-2, Soto and John, 2017). However, much less is known about descriptors for the "middle steps" or "in-between levels" of the social-emotional construct continuum. Hence, one of the significant contributions of rubrics to social-emotional skills assessment is the inclusion of performance level descriptions that reflect the continuous and increasing mastery levels of the skills (Abrahams et al., 2019). Due to these performance level descriptions, students better understand the expectations associated with increasing skills' mastery and hence gain a clear picture of the learning objectives that they need to achieve (Rusman and Dirkx, 2017).

A major challenge when constructing rubrics is to statistically evaluate whether the performance level descriptions can be meaningfully differentiated by the rater and empirically related to the scores on the measured skills (Tierney and Simon, 2004; Brookhart, 2018; Panadero and Jönsson, 2020). In that sense, Brookhart (2018) emphasized that the number of performance level descriptions should correspond to "the number of reliable distinctions in student work that are possible and helpful" (p. 2). Humphry and Heldsinger (2014) stated that this is particularly important for construct validity as the scores of rubrics' performance level descriptions should have a continuous and smooth distribution that reflects the fine-grained variations in the levels of the construct being measured (Humphry and Heldsinger, 2014; Brookhart, 2018). By contrast, there is considerable uncertainty in the literature about the formulation and the methods to statistically evaluate the adequacy of rubrics' levels used to describe skill mastery (Rusman and Dirkx, 2017; Brookhart, 2018).

Item Response Theory (IRT) models could help to solve this challenge. IRT is a set of latent variable techniques designed to model the relationship between an individual's response to a particular item and that individual's position in probabilistic terms along the latent trait measured (Baker and Kim, 2017). As such, IRT would allow evaluating rubrics' rating scale structure by inspecting the Category Response Curves (CRCs) that represent the probabilities of endorsing rubrics' performance level descriptions displayed along the continuum of the underlying skill. Thus, the CRCs could reflect how well the

performance level descriptions represent the measured skill and help diagnose how well these level descriptions are used in the rating process (Linacre, 2002). However, it should be noted that an unbalanced distribution of responses in rubrics' performance level descriptions (i.e., low endorsement of rubrics' lowest level description) might affect the effectiveness of the rubrics' rating scale.

Additionally, IRT multidimensional models could contribute to evaluating whether rubrics' performance level descriptions are ordered as expected along the different dimensions of the measured skills (Bolt and Adams, 2017). Similarly, based on IRT modeling, rubrics could be graphically and empirically expressed in Wright maps or construct maps (Wilson, 2011), which indicate how well rubrics' performance level descriptions unfold with students' increasingly more elaborated responses. Construct maps could also contribute to represent rubrics' difficulty levels (i.e., from the easiest to the most difficult ones) and locate them together with students' observable scores on a single scale to provide insight into students' learning progression on a skill (Wilson, 2011).

Reaching an Optimal Internal Consistency and Discriminant Validity

Rubrics have been used to assess multidimensional constructs like theater arts' skills (Susman-Stillman et al., 2018), creative writing (Mozaffari, 2013), or pre-service teaching competencies (Bryant et al., 2016), among others. A typical multidimensional rubric includes several criteria, each of which has a scale of at least three performance level descriptions. For example, the rubric of creative writing developed by Mozaffari (2013) has four criteria—i.e., image, characterization, voice, and story- and four levels of achievement-i.e., excellent, good, fair, and poor. A few studies have evaluated whether rubrics' criteria that were not supposed to be related were actually related (i.e., discriminant validity or the correlation among scales that measure different traits) although with no promising results. A group of studies, for instance, found a high degree of collinearity ($r > 0.80$) among the dimensions of rubrics that assessed teaching competencies and dispositions when testing their structure through confirmatory factor analysis (Flowers, 2006; Bryant et al., 2016). In that sense, Flowers (2006) acknowledged that this result could be due to respondents' difficulties in distinguishing among the separate rating categories of the tested rubric. Similarly, Ciorba and Smith (2009) found high correlations (0.81–0.89) among the three scale dimensions of a rubric that aimed to assess music performance. As a result, the authors evaluated whether a unidimensional indicator could replace their rubric. Still, they argued that valuable information on students' strengths and weaknesses could be lost if the different rubric dimensions were not considered distinctively.

One plausible explanation for rubrics' discriminant validity challenge is that ratees and raters might not be able to account for differences among multiple skills when describing their own performance using rubrics (Sadler, 2009; Panadero and Jönsson, 2020). Sadler (2009) stated, for example, that raters might not be interested in evaluating individual criteria and, instead, prefer to assess performance as a whole. In response, Panadero and Jönsson (2020) stated that empirical evidence had suggested that judges can reliably assess multiple criteria using rubrics. The authors added that rubrics with multiple criteria of commonly known "tacit competencies" such as creativity had been well-differentiated and evaluated by teachers in other studies (e.g., Lindström, 2006). Still, Sandler's claims cast doubts on whether differentiating between criteria such as SEMS is possible for or even interest students when they self-assess their performance. Perhaps the idea that students need to have a clear compartmentalization of their skills may not be strictly necessary to promote the awareness of their own strengths and difficulties.

Meanwhile, research on social-emotional skills has paid greater attention to test the discriminant validity and internal consistency of their measurement tools. However, this effort has not been exempt from difficulties. On the one hand, research on Big Five personality instruments has shown adequate discrimination among the five domains and the facets within each domain (e.g., Soto and John, 2017). On the other hand, a large body of literature has claimed that it is unreasonable to assume that the Big Five measure completely independent clusters that could be tested via restricting approaches like confirmatory factor analysis (Marsh et al., 2010; Aichholzer, 2014). Additionally, evidence on the internal consistency and discrimination among the Big Five seems to be less strong in children or adolescents' self-ratings of personality (Soto et al., 2008). Allik et al. (2004) showed, for example, that the intercorrelations among the NEO-FFI scales gradually decreased from 0.24 at age 12 to 0.12 around age 18. Similarly, Soto et al. (2008) found that the between-domain differentiation (i.e., discriminant validity) and within-domain consistency (i.e., internal consistency) of various personality instruments increased in magnitude from late childhood until late adolescence.

Moreover, several studies have found that the internal consistency of short personality measures might be lower than the ones typically found in standard multi-item measures of the Big Five (Gosling et al., 2003; Woods and Hampson, 2005). Gosling et al. (2003) developed the Ten-Item Personality Inventory (TIPI), a 10-item measure of the Big Five. They found that although the instrument had adequate validity evidence, it showed lower reliability estimates than longer personality scales such as the Big Five Inventory. Likewise, other authors have argued that Cronbach's alpha internal consistency estimate is often misleading when used on short personality measures (Woods and Hampson, 2005). Woods and Hampson (2005), for instance, showed that Cronbach's alpha decreased from an average value of 0.85 for a personality measure of 100 items (i.e., Trait Descriptive Adjectives TDA) to a value of 0.50 for the TIPI.

Hence, as the previous evidence has shown, the evaluation of discriminant validity and internal consistency in traditional Likert-scale personality and skill measures is challenging. Thus, several questions remain concerning the evaluation of similar characteristics in non-traditional measures like rubrics when they include multidimensional criteria to assess SEMS in youth.

Accuracy in Reporting SEMS Using Rubrics

Youth's Self-Reports. Children and adolescents experience several developmental changes that may have considerable implications for their ability to think about themselves and use rubrics to report on their social-emotional skills. Children, compared to adolescents, have fewer capacities to think abstractly and logically about statements, as well as to ask questions about their identity (i.e: "Who I am?", "How I am different from others?"; Soto et al., 2008). However, this gradually changes as children approach adolescence. At this age, adolescents "are more likely to think about and describe themselves with abstract and psychological terms to differentiate among multiple aspects of the self, to recognize consistencies and inconsistencies among these self-aspects, and to organize them in a clear way" (Soto et al., 2008; p. 719).

Additionally, the abilities of verbal comprehension and information processing gradually grow from late childhood to adolescence. Youth uses more frequently new and more complex words to describe themselves in psychological terms (Soto et al., 2008), comprehend better what they read (Siegler, 1998), and process information faster and more fluently (Anderson, 2002). Similarly, longitudinal studies have shown that children's self-reflective skills involved in their metacognitive processing capacity gradually improve towards adolescence, although their growth might remain stable after age 15 (van der Stel and Veenman, 2010, van der Stel and Veenman, 2013).

The above-mentioned developmental changes may have consequences on how well children and adolescents can provide ratings of their own behavior and performance (Soto et al., 2008; Brown et al., 2015; Panadero et al., 2016). Compared to late adolescents or adults, younger children seem to be more optimistic and lenient, less coherent and rely more on others' opinions when self-reporting their performance and behavior (Soto et al., 2008; Brown et al., 2015). For example, some studies have shown that young students self-assess their performance influenced by their parents and teachers' academic standards and normative values but become more independent as they get older (Hogaboam-Gray, 2002; Kasanen and Räty, 2002). Meanwhile, a study of children's narrative abilities from 5 to 12 years indicated that students with less ability were more prone to overestimate their academic performance than those with better skills, but that this tendency decreased with age. This optimistic bias has also been found in self-reports of personality characteristics. Soto et al. (2008) showed, for example, that young children tend to systematically agree more with items (or disagree with negatively keyed items) in personality questionnaires (i.e., acquiescent responding) than adolescents or adults. In Soto et al. (2008) 's study, this response style affected the factor structure of personality self-reports at age 10 in such a way that the Big Five Factors could not be well recovered, although this improved in older ages (Soto et al., 2008).

Observers' Reports. Besides self-assessment, rubrics have also been used by teachers or other raters to evaluate students' performance in diverse cognitive abilities such as writing, math, or science. Overall, research has found that raters can provide reliable and consistent judgments of performance using rubrics, although several factors might influence their rating accuracy. Among others, the expertise and training of the raters, their attitudes towards students' ethnicity or the content, or the lack of clarity of the rubrics (Jönsson and Svingby, 2007; Reddy and Andrade, 2010; Brookhart and Chen, 2015) have been listed as factors that may affect the validity of ratings. Relative to the knowledge on self-assessments, much less is known, however, about how well raters report on students' SEMS using rubrics. Experiences in the fields of music, arts, and creativity, for example, have investigated the degree of agreement between raters (i.e., inter-rater reliability) when evaluating students' performance using rubrics (Lindström, 2006; Ciorba and Smith, 2009; Latimer et al., 2010; Susman-Stillman et al., 2018). Overall, most studies have found moderate to high levels of agreement among raters' judgments, which supports that rubrics can yield reliable results when raters have a good understanding of the criteria and are well-trained. Despite these promising results, many questions about rubrics observer reports remain unanswered. Several questions warrant further investigation, including: "What is the degree of consistency between raters' reports and students' self-reported SEMS scores?"; "Are teachers good raters of students' SEMS performance?"; and "How can teachers' personal characteristics affect their reports?".

On the other hand, evidence from research on SEMS assessment using Likert-type questionnaires has shown that teachers' reports are a valuable source of information of students' social-emotional characteristics across time (Measelle et al., 2005; Wienke Totura et al., 2009; Edmonds et al., 2013; Margherio et al., 2019). Results from a longitudinal study of Measelle et al. (2005), for example, indicated that teachers and parents' scores of children's Extraversion, Agreeableness, and Conscientiousness increasingly but moderately converged from ages 5 to 7. By contrast, another group of studies has suggested that teachers might not be good informants of students' social and emotional difficulties using Likert-based scales. Margherio et al. (2019) found, for example, that it was easier for teachers to recognize conduct problems than emotional problems in students, while Wienke Totura et al. (2009) reported a low agreement between teachers and students on experiences of bullying and victimization. Meanwhile, Kokkinos and Kargiotidis' (2016) study concluded that teachers' mental health characteristics (i.e., psychopathological symptoms) influenced their perceptions of students' emotional and behavioral problems. For example, teachers' interpersonal sensitivity symptoms (i.e., feelings of personal inferiority and inadequacy) predicted their ratings of students anxiety, affective, and somatic problems.

Accounting for Development of Social-Emotional Skills

A teacher that uses rubrics to formatively assess SEMS during the school year might expect that all his/her adolescent students would develop their SEMS in a similar way. However, a group of studies has shown that youth personality traits do not develop uniformly and increasingly. That is, youth's mean levels of

personality traits do not continuously increase with age (e.g., Soto et al., 2011), but may show temporary dips, which Soto and Tackett (2015) called the *disruption hypothesis* in personality development.

The disruption hypothesis proposes that "the biological, social, and psychological transitions from childhood to adolescence are accompanied by temporary dips in some aspects of personality maturity" (Soto and Tackett, 2015; p. 360). In that sense, evidence from self- and parents' reports have shown that mean levels of Agreeableness, Conscientiousness, and Openness to Experience tend to decrease from late childhood into early adolescence, and then increase from late adolescence into early adulthood (Soto et al., 2011; Denissen et al., 2013; Van den Akker et al., 2014; Soto, 2016). Similarly, Extraversion and Activity levels tend to considerably decrease from childhood to adolescence until they stabilize during adulthood. Special attention deserves Neuroticism as it develops differently in boys and girls. During childhood, girls and boys have similar levels of Neuroticism, but this pattern changes dramatically when they arrive in adolescence when girls increase their levels of negative affect while boys remain almost stable (Klimstra et al., 2009; Soto et al., 2011; Van den Akker et al., 2014; De Bolle et al., 2015; Soto, 2016). At this age, girls may experience more social and psychological difficulties than boys, such as negative stereotyping, gender-biased roles, body image concerns, and negative self-perceptions (Stice and Bearman, 2001). These gender differences persist until early adulthood when both girls and boys return to have more similar levels of Neuroticism (Klimstra et al., 2009; Soto et al., 2011; Van den Akker et al., 2014; De Bolle et al., 2015; Soto, 2016).

Besides mean level changes in personality, research has suggested that we should look at the way youth differ in how they express their personality characteristics and how these individual differences change across the lifespan. Several recent studies have shown that there are individual differences in personality trait development that tend to increase with age, from early childhood into early adolescence, and then remain relatively stable during adulthood (Damian et al., 2019; Mõttus et al., 2017, 2019). Mõttus et al. (2017) suggested that these increasing individual differences could be due to developmental mechanisms that manifest strongly during childhood and adolescence. Among several plausible factors, Mõttus et al. (2017) highlight socialization pressures on behavior, intrinsic personality maturation, or the expansion of the behavior repertory driven by the acquisition of new cognitive, self-regulatory, and emotional capacities, among others.

Altogether, the mentioned evidence highlights the importance of considering the particular characteristics of SEMS development in youth when drawing expectations about their learning progressions. In other words, educators should consider that not all youth will have similar learning trajectories or goals when using rubrics to develop their SEMS. For example, girls might not learn to regulate their emotions at the same "rhythm" as boys during adolescence; therefore, their goals should adapt to their different learning progression.

CONCLUSIONS AND FUTURE DIRECTIONS

In this review, we examined literature and empirical studies on the use of formative assessment and rubrics for SEMS learning and discussed some of the key challenges for the construction and use of rubrics to assess social-emotional skills. First, we identified an increasing number of initiatives that have implemented formative assessment strategies and constructed rubrics to assess social-emotional dimensions such as creativity, critical thinking, and arts learning (e.g., Vincent-Lancrin et al., 2019). However, these dimensions are only some of the many social-emotional skills included in SEMS taxonomies that are considered crucial for students' social-emotional learning. In that sense, more efforts are needed to expand the use of rubrics for the multidimensional skills proposed in the existing social-emotional taxonomies. Evidence-based guidelines and recommendations could be used to effectively design rubrics to measure SEMS (Pancorbo et al., 2020). In that sense, Dawson (2017) provided a framework of 14 elements (e.g., evaluative criteria, specificity, quality levels, etc.) that can be useful for researchers to make informed decisions on rubrics' design and use. Likewise, these endeavors should put emphasis on evaluating rubrics' psychometric properties with diverse methods as well as examining their power in predicting relevant outcomes.

Second, we highlighted the importance of testing the organization of rubrics' most basic components—performance level descriptions—to evaluate whether they capture the different dimensions of the construct they intend to measure. Sadly, very few attempts have been made so far to assess the organization of rubrics' descriptions using innovative statistical methods (e.g., Humphry and Heldsinger, 2014). This challenging task could be tackled using IRT models, which have the advantage of providing valuable information of the degree to which performance level descriptions contribute to rubrics' rating scale characteristics.

Third, we also raised some questions of what we could expect concerning the discriminant validity and internal consistency of SEMS rubrics, especially when they are administered to young respondents. As mentioned before, the discriminant properties of rubrics' scores in previous studies were overall weak. In the study of Pancorbo et al. (2020), for example, these properties were even weaker in the social-emotional rubrics' scores of young respondents with low language proficiency. To avoid these constraints as much as possible, future research initiatives could consider maximizing the differentiation of the content of rubrics that measure different SEMS. Building several rubrics per assessed domain could also improve the internal consistency of rubrics' scores.

Lastly, we emphasized that research should further explore how individual differences in youth's social-emotional development might affect the measurement of SEMS using rubrics. To our knowledge, no study has explored this topic, which points out its significance in designing a rubrics' research agenda. This could be ideally investigated in longitudinal studies that focus on exploring the developmental trajectories of rubrics' psychometric properties across adolescents' life span so that the interaction of age and cognitive abilities can be further understood.

The present review raised a number of critical concerns on social-emotional rubrics' conceptual and psychometric properties that should not discourage their use especially for formative

assessment purposes, where the objective is to support the student and his/her learning environment. We outlined a number of ways to examine and improve rubrics' properties and hence increase their impact and effectiveness in students' development. Given their current limitations, social-emotional rubrics should be used in a tentative way, and not considered as robust information or landmarks for at stake decision making or summative evaluation purposes. We hope that this review will contribute to research advancing the status of rubrics as a critical method to be used by students, teachers and educators, providing also actionable information for policy makers.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

Our work received financial support of EduLab21, Instituto Ayrton Senna, Sao Paulo, Brazil. RP also receives a scholarship from the National Council on Scientific and Technological Development (CNPq), funding from São Paulo Research Foundation (FAPESP) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) PVEX - 88881.337381/2019-01.

ACKNOWLEDGMENTS

We gratefully acknowledge the technical and financial support of EduLab21 and its team from the *Instituto Ayrton Senna* to conduct this study.

REFERENCES

Abrahams, L., Pancorbo, G., Primi, R., Santos, D., Kyllonen, P., John, O. P., et al. (2019). Social-emotional Skill Assessment in Children and Adolescents: Advances and Challenges in Personality, Clinical, and Educational Contexts. *Psychol. Assess.* 31 (4), 460–473. doi:10.1037/pas0000591

Aichholzer, J. (2014). Random Intercept EFA of Personality Scales. *J. Res. Pers.* 53, 1–4. doi:10.1016/j.jrp.2014.07.001

Allen, D., and Tanner, K. (2006). Rubrics: Tools for Making Learning Goals and Evaluation Criteria Explicit for Both Teachers and Learners. *CBE Life Sci. Educ.* 5, 197–203. doi:10.1187/cbe.06-06-0168

Allik, J., Laidra, K., Realo, A., and Pullmann, H. (2004). Personality Development from 12 to 18 Years of Age: Changes in Mean Levels and Structure of Traits. *Eur. J. Pers.* 18 (6), 445–462. doi:10.1002/per.524

Anderson, P. (2002). Assessment and Development of Executive Function (EF) during Childhood. *Child. Neuropsychol.* 8 (2), 71–82. doi:10.1076/chin.8.2.71.8724

Andrade, H., Hefferen, J., and Palma, M. (2014). Formative Assessment in the Visual Arts. *Art Educ.* 67 (1), 34–40. doi:10.1080/00043125.2014.11519256

Andrade, H. L., and Brookhart, S. M. (2020). Classroom Assessment as the Co-regulation of Learning. *Assess. Educ. Princ. Pol. Pract.* 27, 350–372. doi:10.1080/0969594x.2019.1571992

Andrade, H. L. (2007). Self-assessment through Rubrics. *Educ. Leadersh.* 65 (4), 60–63.

Andrade, H. L., Wang, X., Du, Y., and Akawi, R. L. (2009). Rubric-Referenced Self-Assessment and Self-Efficacy for Writing. *J. Educ. Res.* 102 (4), 287–302. doi:10.3200/joer.102.4.287-302

Arksey, H., and O'Malley, L. (2005). Scoping Studies: Towards a Methodological Framework. *Int. J. Soc. Res. Methodol.* 8 (1), 19–32. doi:10.1080/1364557032000119616

Baker, F. B., and Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Cham: Springer

Blom, D., Stevenson, I., and Encarnacao, J. (2015). "Assessing Music Performance Process and Outcome through a Rubric: Ways and Means," in *Assessment in Music Education: From Policy to Practice*. Editors D. Lebler, G. Carey, and S. D. Harrison (Cham: Springer International Publishing), 125–139. doi:10.1007/978-3-319-10274-0_9

Bolden, B., DeLuca, C., Kukkonen, T., Roy, S., and Wearing, J. (2020). Assessment of Creativity in K-12 Education: A Scoping Review. *Rev. Educ.* 8, 343–376. doi:10.1002/rev3.3188

Bolt, D. M., and Adams, D. J. (2017). Exploring Rubric-Related Multidimensionality in Polytomously Scored Test Items. *Appl. Psychol. Meas.* 41 (3), 163–177. doi:10.1177/0146621616677715

Brookhart, S. M. (2018). Appropriate Criteria: Key to Effective Rubrics. *Front. Educ.* 3, 22. doi:10.3389/feduc.2018.00022

Brookhart, S. M. (2013). Assessing Creativity. *Educ. Leadersh.* 70 (5), 28–34.

Brookhart, S. M., and Chen, F. (2015). The Quality and Effectiveness of Descriptive Rubrics. *Educ. Rev.* 67 (3), 343–368. doi:10.1080/00131911.2014.929565

Brown, G. T. L., Andrade, H. L., and Chen, F. (2015). Accuracy in Student Self-Assessment: Directions and Cautions for Research. *Assess. Educ. Princ. Pol. Pract.* 22 (4), 444–457. doi:10.1080/0969594x.2014.996523

Bryant, C. L., Maarouf, S., Burcham, J., and Greer, D. (2016). The Examination of a Teacher Candidate Assessment Rubric: A Confirmatory Factor Analysis. *Teach. Teach. Educ.* 57, 79–96. doi:10.1016/j.tate.2016.03.012

Chen, F., and Andrade, H. (2018). The Impact of Criteria-Referenced Formative Assessment on Fifth-Grade Students' Theater Arts Achievement. *J. Educ. Res.* 111 (3), 310–319. doi:10.1080/00220671.2016.1255870

Chen, F., Lui, A. M., Andrade, H., Valle, C., and Mir, H. (2017). Criteria-referenced Formative Assessment in the Arts. *Educ. Asse. Eval. Acc.* 29 (3), 297–314. doi:10.1007/s11092-017-9259-z

Chen, P. P., and Bonner, S. M. (2020). A Framework for Classroom Assessment, Learning, and Self-Regulation. *Assess. Educ. Princ. Pol. Pract.* 27, 373–393. doi:10.1080/0969594x.2019.1619515

Chernyshenko, O. S., Kankaraš, M., and Drasgow, F. (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. Paris: OECD Publishing. doi:10.1787/db1d8e59-en

Ciorba, C. R., and Smith, N. Y. (2009). Measurement of Instrumental and Vocal Undergraduate Performance Juries Using a Multidimensional Assessment Rubric. *J. Res. Music Educ.* 57 (1), 5–15. doi:10.1177/0022429409333405

Damian, R. I., Spengler, M., Sutu, A., and Roberts, B. W. (2019). Sixteen Going on Sixty-Six: A Longitudinal Study of Personality Stability and Change across 50 Years. *J. Pers. Soc. Psychol.* 117 (3), 674–695. doi:10.1037/pspp0000210

Dawson, P. (2017). Assessment Rubrics: towards Clearer and More Replicable Design, Research and Practice. *Assess. Eval. Higher Educ.* 42 (3), 347–360. doi:10.1080/02602938.2015.1111294

De Bolle, M., De Fruyt, F., McCrae, R. R., Löckenhoff, C. E., Costa, P. T., Aguilar-Vafaie, M. E., et al. (2015). The Emergence of Sex Differences in Personality Traits in Early Adolescence: A Cross-Sectional, Cross-Cultural Study. *J. Pers. Soc. Psychol.* 108 (1), 171–185. doi:10.1037/a0038497

De Fruyt, F., Wille, B., and John, O. P. (2015). Employability in the 21st century: Complex (Interactive) Problem Solving and Other Essential Skills. *Ind. Organ. Psychol.* 8 (2), 276–281. doi:10.1017/iop.2015.33

Denissen, J. J. A., van Aken, M. A. G., Penke, L., and Wood, D. (2013). Self-Regulation Underlies Temperament and Personality: An Integrative Developmental Framework. *Child. Dev. Perspect.* 7 (4), 255–260. doi:10.1111/cdep.12050

Domitrovich, C. E., Durlak, J. A., Staley, K. C., and Weissberg, R. P. (2017). Social-Emotional Competence: An Essential Factor for Promoting Positive Adjustment and Reducing Risk in School Children. *Child. Dev.* 88 (2), 408–416. doi:10.1111/cdev.12739

Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., and Gullotta, T. P. (2015). *Handbook of Social and Emotional Learning: Research and Practice*. New York, NY, US: The Guilford Press.

Edmonds, G. W., Goldberg, L. R., Hampson, S. E., and Barckley, M. (2013). Personality Stability from Childhood to Midlife: Relating Teachers' Assessments in Elementary School to Observer- and Self-Ratings 40 Years Later. *J. Res. Pers.* 47 (5), 505–513. doi:10.1016/j.jrp.2013.05.003

Flowers, C. (2006). Confirmatory Factor Analysis of Scores on the Clinical Experience Rubric. *Educ. Psychol. Meas.* 66 (3), 478–488. doi:10.1177/0013164405282458

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A Very Brief Measure of the Big-Five Personality Domains. *J. Res. Personal.* 37 (6), 504–528. doi:10.1016/s0092-6566(03)00046-1

Graham, S., Hebert, M., and Harris, K. R. (2015). Formative Assessment and Writing. *Elem. Sch. J.* 115 (4), 523–547. doi:10.1086/681947

Griffin, P., McGaw, B., and Care, E. (2011). *Assessment and Teaching of 21st Century Skills*. Springer Publishing Company.

Hattie, J. (2009). *A Synthesis of over 800 Meta-Analyses Relating to Achievement*. New York: Routledge.

Humphry, S. M., and Heldsinger, S. A. (2014). Common Structural Design Features of Rubrics May Represent a Threat to Validity. *Educ. Res.* 43 (5), 253–263. doi:10.3102/0013189x14542154

Jönsson, A., and Panadero, E. (2017). "The Use and Design of Rubrics to Support Assessment for Learning," in *Scaling up Assessment for Learning in Higher Education*. Editors D. Carless, S. Bridges, C. Chan, and R. Glofcheski (Singapore: Springer), 5, 99–111. doi:10.1007/978-981-10-3045-1_7

Jönsson, A., and Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educ. Res. Rev.* 2 (2), 130–144. doi:10.1016/j.edurev.2007.05.002

Kasanen, K., and Räty, H. (2002). "You Be Sure Now to Be Honest in Your Assessment": Teaching and Learning Self-Assessment. *Soc. Psychol. Educ. Int. J.* 5 (4), 313–328. doi:10.1023/A:1020993427849

Kern, M. L., Benson, L., Steinberg, E. A., and Steinberg, L. (2016). The EPOCH Measure of Adolescent Well-Being. *Psychol. Assess.* 28 (5), 586–597. doi:10.1037/pas0000201

Kingston, N., and Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educ. Meas. Issues Pract.* 30 (4), 28–37. doi:10.1111/j.1745-3992.2011.00220.x

Klimstra, T. A., Hale, W. W., Raaijmakers, Q. A., Branje, S. J., and Meeus, W. H. (2009). Maturation of Personality in Adolescence. *J. Pers. Soc. Psychol.* 96 (4), 898–912. doi:10.1037/a0014746

Kokkinos, C. M., and Kargiotidis, A. (2016). Rating Students' Problem Behaviour: the Role of Teachers' Individual Characteristics. *Educ. Psychol.* 36 (8), 1516–1532. doi:10.1080/01443410.2014.993929

Kyllonen, P. C. (2016). "Designing Tests to Measure Personal Attributes and Noncognitive Skills," in *Handbook of Test Development*. Editors S. Lane, M. R. Raymond, and T. M. Haladyna (New York: Routledge), 190–211.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., and Roberts, R. D. (2014). Personality, Motivation, and College Readiness: A Prospectus for Assessment and Development. ETS Research Report No. RR-14–06. Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12004

Lallmamode, S. P., Mat Daud, N., and Abu Kassim, N. L. (2016). Development and Initial Argument-Based Validation of a Scoring Rubric Used in the Assessment of L2 Writing Electronic Portfolios. *Assessing Writing* 30, 44–62. doi:10.1016/j.asw.2016.06.001

Latimer, M. E., Bergee, M. J., and Cohen, M. L. (2010). Reliability and Perceived Pedagogical Utility of a Weighted Music Performance Assessment Rubric. *J. Res. Music Educ.* 58 (2), 168–183. doi:10.1177/0022429410369836

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *J. Appl. Meas.* 3 (1), 85–106.

Lindström, L. (2006). Creativity:What Is it? Can You Assess it? Can it Be Taught? *Int. J. Art Des. Educ.* 25 (1), 53–66. doi:10.1111/j.1476-8070.2006.00468.x

Margherio, S. M., Evans, S. W., and Owens, J. S. (2019). Universal Screening in Middle and High Schools: Who Falls through the Cracks? *Sch. Psychol.* 34 (6), 591–602. doi:10.1037/spq0000337

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., et al. (2010). A New Look at the Big Five Factor Structure through Exploratory Structural Equation Modeling. *Psychol. Assess.* 22 (3), 471–491. doi:10.1037/a0019227

Marshall, B., and Jane Drummond, M. M. (2006). How Teachers Engage with Assessment for Learning: Lessons from the Classroom. *Res. Pap. Educ.* 21 (2), 133–149. doi:10.1080/02671520600615638

Measelle, J. R., John, O. P., Ablow, J. C., Cowan, P. A., and Cowan, C. P. (2005). Can Children Provide Coherent, Stable, and Valid Self-Reports on the Big Five Dimensions? A Longitudinal Study from Ages 5 to 7. *J. Pers Soc. Psychol.* 89 (1), 90–106. doi:10.1037/0022-3514.89.1.90

Menéndez-Varela, J.-L., and Gregori-Giralt, E. (2018). Rubrics for Developing Students' Professional Judgement: A Study of Sustainable Assessment in Arts Education. *Stud. Educ. Eval.* 58, 70–79. doi:10.1016/j.stueduc.2018.06.001

Moskal, B. M., and Leydens, J. A. (2000). Scoring Rubric Development: Validity and Reliability. *Pract. Assess. Res. Eval.* 7 (10), 1–6. doi:10.7275/q7rm-gg74

Mõttus, R., Briley, D. A., Zheng, A., Mann, F. D., Engelhardt, L. E., Tackett, J. L., et al. (2019). Kids Becoming Less Alike: A Behavioral Genetic Analysis of Developmental Increases in Personality Variance from Childhood to Adolescence. *J. Pers. Soc. Psychol.* 117 (3), 635–658. doi:10.1037/pspp0000194

Mõttus, R., Soto, C. J., Slobodskaya, H. R., and Back, M. (2017). Are All Kids Alike? the Magnitude of Individual Differences in Personality Characteristics Tends to Increase from Early Childhood to Early Adolescence. *Eur. J. Pers* 31 (4), 313–328. doi:10.1002/per.2107

Mozaffari, H. (2013). An Analytical Rubric for Assessing Creativity in Creative Writing. *TPLS* 3 (12), 2214–2219. doi:10.4304/tpls.3.12.2214-2219

Nitko, A. J., and Brookhart, S. M. (2007). *Educational Assessment of Students*. 5th ed. Upper Saddle River: Prentice Hall.

OECD (2015). *Skills for Social Progress*. Paris: OECD Publishing.

Ozer, D. J., and Benet-Martínez, V. (2006). Personality and the Prediction of Consequential Outcomes. *Annu. Rev. Psychol.* 57, 401–421. doi:10.1146/annurev.psych.57.102904.190127

Panadero, E., Brown, G. T. L., and Strijbos, J.-W. (2016). The Future of Student Self-Assessment: a Review of Known Unknowns and Potential Directions. *Educ. Psychol. Rev.* 28 (4), 803–830. doi:10.1007/s10648-015-9350-2

Panadero, E., and Jönsson, A. (2020). A Critical Review of the Arguments against the Use of Rubrics. *Educ. Res. Rev.* 30, 100329. doi:10.1016/j.edurev.2020.100329

Panadero, E., and Jönsson, A. (2013). The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review. *Educ. Res. Rev.* 9, 129–144. doi:10.1016/j.edurev.2013.01.002

Panadero, E., and Romero, M. (2014). To Rubric or Not to Rubric? the Effects of Self-Assessment on Self-Regulation, Performance and Self-Efficacy. *Assess. Educ. Princ. Pol. Pract.* 21 (2), 133–148. doi:10.1080/0969594X.2013.877872

Panadero, E., Tapia, J. A., and Huertas, J. A. (2012). Rubrics and Self-Assessment Scripts Effects on Self-Regulation, Learning and Self-Efficacy in Secondary Education. *Learn. Individual Differ.* 22 (6), 806–813. doi:10.1016/j.lindif.2012.04.007

Pancorbo, G., Primi, R., John, O. P., Santos, D., Abrahams, L., and De Fruyt, F. (2020). Development and Psychometric Properties of Rubrics for Assessing Social-Emotional Skills in Youth. *Stud. Educ. Eval.* 67, 100938. doi:10.1016/j.stueduc.2020.100938

Paunonen, S. V., and Ashton, M. C. (2001). Big Five Factors and Facets and the Prediction of Behavior. *J. Pers. Soc. Psychol.* 81, 524–539. doi:10.1037/0022-3514.81.3.524

Pellegrino, J. W., and Hilton, M. L. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, D.C: The National Academy Press.

Primi, R., John, O. P., Santos, D., and De Fruyt, F. (2017). *SENNA Inventory*. São Paulo, Brazil: Institute Ayrton Senna.

Primi, R., Santos, D., John, O. P., and Fruyt, F. D. (2016). Development of an Inventory Assessing Social and Emotional Skills in Brazilian Youth. *Eur. J. Psychol. Assess.* 32 (1), 5–16. doi:10.1027/1015-5759/a000343

Reddy, Y. M., and Andrade, H. (2010). A Review of Rubric Use in Higher Education. *Assess. Eval. Higher Educ.* 35 (4), 435–448. doi:10.1080/02602930902862859

Ross, J. A., Rolheiser, C., and Hogaboam-Gray, A. (2002). Influences on Student Cognitions about Evaluation. *Assess. Educ. Princ. Pol. Pract.* 9 (1), 81–95. doi:10.1080/09695940220119201

Rusman, E., and Dirkx, K. (2017). Developing Rubrics to Assess Complex (Generic) Skills in the Classroom: How to Distinguish Skills' Mastery Levels? *Pract. Assess. Res. Eval.* 22 (12), 1–9.

Sadler, D. R. (2009). Indeterminacy in the Use of Preset Criteria for Assessment and Grading. *Assess. Eval. Higher Educ.* 34 (2), 159–179. doi:10.1080/02602930801956059

Shiner, R. L., Soto, C. J., and De Fruyt, F. (2021). Personality Assessment of Children and Adolescents. *Annu. Rev. Dev. Psychol.* 3 (1). doi:10.1146/annurev-devpsych-050620-114343

Siegler, R. S. (1998). *Children's Thinking*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Soto, C. J., John, O. P., Gosling, S. D., and Potter, J. (2011). Age Differences in Personality Traits from 10 to 65: Big Five Domains and Facets in a Large Cross-Sectional Sample. *J. Pers. Soc. Psychol.* 100 (2), 330–348. doi:10.1037/a0021717

Soto, C. J., John, O. P., Gosling, S. D., and Potter, J. (2008). The Developmental Psychometrics of Big Five Self-Reports: Acquiescence, Factor Structure, Coherence, and Differentiation from Ages 10 to 20. *J. Pers. Soc. Psychol.* 94 (4), 718–737. doi:10.1037/0022-3514.94.4.718

Soto, C. J., and John, O. P. (2017). The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model with 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power. *J. Pers. Soc. Psychol.* 113 (1), 117–143. doi:10.1037/pspp0000096

Soto, C. J. (2016). The Little Six Personality Dimensions from Early Childhood to Early Adulthood: Mean-Level Age and Gender Differences in Parents' Reports. *J. Pers.* 84 (4), 409–422. doi:10.1111/jopy.12168

Soto, C. J., and Tackett, J. L. (2015). Personality Traits in Childhood and Adolescence. *Curr. Dir. Psychol. Sci.* 24 (5), 358–362. doi:10.1177/0963721415589345

Stice, E., and Bearman, S. K. (2001). Body-image and Eating Disturbances Prospectively Predict Increases in Depressive Symptoms in Adolescent Girls: a Growth Curve Analysis. *Dev. Psychol.* 37 (5), 597–607. doi:10.1037//0012-1649.37.5.597

Susman-Stillman, A., Englund, M., Webb, C., and Grenell, A. (2018). Reliability and Validity of a Measure of Preschool Children's Theatre Arts Skills: The Preschool Theatre Arts Rubric. *Early Child. Res. Q.* 45, 249–262. doi:10.1016/j.ecresq.2017.12.001

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

Tierney, R., and Simon, M. (2004). What's Still Wrong with Rubrics: Focusing on the Consistency of Performance Criteria across Scale Levels. *Pract. Assess. Res. Eval.* 9, 2. doi:10.7275/JTVT-WG68

Trilling, B., and Fadel, C. (2009). *21st Century Skills Learning for Life in Our Times*. San Francisco: Jossey-Bass.

Valle, C., Andrade, H., Palma, M., and Hefferen, J. (2016). Applications of Peer Assessment and Self-Assessment in Music. *Music Educ. J.* 102 (4), 41–49. doi:10.1177/0027432116644652

Van den Akker, A. L., Deković, M., Asscher, J., and Prinzie, P. (2014). Mean-level Personality Development across Childhood and Adolescence: a Temporary defiance of the Maturity Principle and Bidirectional Associations with Parenting. *J. Pers. Soc. Psychol.* 107 (4), 736–750. doi:10.1037/a0037248

van der Stel, M., and Veenman, M. V. J. (2010). Development of Metacognitive Skillfulness: A Longitudinal Study. *Learn. Individ. Differ.* 20 (3), 220–224. doi:10.1016/j.lindif.2009.11.005

van der Stel, M., and Veenman, M. V. J. (2013). Metacognitive Skills and Intellectual Ability of Young Adolescents: a Longitudinal Study from a Developmental Perspective. *Eur. J. Psychol. Educ.* 29 (1), 117–137. doi:10.1007/s10212-013-0190-5

Vincent-Lancrin, S., González-Sancho, C., Bouckaert, M., de Luca, F., Fernández-Barrerra, M., Jacotin, G., et al. (2019). *Fostering Students' Creativity and Critical Thinking: What it Means in School*. Paris: OECD Publishing. doi:10.1787/62212c37-en

Voogt, J., and Roblin, N. P. (2012). A Comparative Analysis of International Frameworks for 21stcentury Competences: Implications for National Curriculum Policies. *J. Curriculum Stud.* 44 (3), 299–321. doi:10.1080/00220272.2012.668938

Wienke Totura, C. M., Green, A. E., Karver, M. S., and Gesten, E. L. (2009). Multiple Informants in the Assessment of Psychological, Behavioral, and Academic Correlates of Bullying and Victimization in Middle School. *J. Adolesc.* 32 (2), 193–211. doi:10.1016/j.adolescence.2008.04.005

Wilson, M. (2011). Some Notes on the Term: "Wright Map", 25. Rasch Measurement Transactions, 3. Available at: https://www.rasch.org/rmt/rmt253b.htm#:~:text=The%20item%2Dperson%20map%2C%20often,axis%20marked%20with%20a%20scale (Accessed October 8, 2019).

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., and Harms, U. (2016). What Makes Rubrics Effective in Teacher-Feedback? Transparency of Learning Goals Is Not Enough. *Contemp. Educ. Psychol.* 44-45, 1–11. doi:10.1016/j.cedpsych.2015.11.003

Woods, S. A., and Hampson, S. E. (2005). Measuring the Big Five with Single Items Using a Bipolar Response Scale. *Eur. J. Pers.* 19 (5), 373–390. doi:10.1002/per.542

Ziegler, M., and Brunner, M. (2016). "Test Standards and Psychometric Modeling," in *Psychosocial Skills and School Systems in the 21st Century: Theory, Research, and Practice*. Editors A. A. Lipnevich, F. Preckel, and R. D. Roberts (Cham: Springer International Publishing), 29–55. doi:10.1007/978-3-319-28606-8_2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.