Validating self-assessment measures for quality of center-based childcare: a meta-analysis

Hay, S.M.; Vermeer, H.J.; Fukkink, R.G.; Nefs, H.T.

Early Education and Development

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/heed20

Validating Self-Assessment Measures for Quality of Center-Based Childcare: A Meta-Analysis

Simon M. Hay, Harriet J. Vermeer, Ruben G. Fukkink & Harold T. Nefs

To cite this article: Simon M. Hay, Harriet J. Vermeer, Ruben G. Fukkink & Harold T. Nefs (2025) Validating Self-Assessment Measures for Quality of Center-Based Childcare: A Meta-Analysis, Early Education and Development, 36:1, 165-185, DOI: 10.1080/10409289.2024.2360882

To link to this article: https://doi.org/10.1080/10409289.2024.2360882

EARLY EDUCATION AND DEVELOPMENT
2025, VOL. 36, NO. 1, 165–185
https://doi.org/10.1080/10409289.2024.2360882

Routledge
Taylor & Francis Group

Validating Self-Assessment Measures for Quality of Center-Based Childcare: A Meta-Analysis

Simon M. Hay [a], Harriet J. Vermeer [a], Ruben G. Fukkink [b,c], and Harold T. Nefs [a]

[a]Social and Behavioural Sciences, Institute of Education and Child Studies Leiden University; [b]Research Institute of Child Development and Education, University of Amsterdam; [c]Faculty of Social and Behavioural Sciences, Amsterdam University of Applied Sciences

ABSTRACT

Research Findings: A growing number of stakeholders in early childhood education (ECE) rely on self-assessment to assess and improve the quality of ECE. In this systematic review, we investigated the reliability and validity of self-assessment in ECE, summarizing findings from 27 publications. We meta-analytically synthesized findings from 25 publications for 1,882 groups and 79,163 children aged 0–72 months in center-based childcare. Most studies reported high internal consistency, but one study reported a lower consistency. Inter-rater reliability was generally high. A three-level meta-analysis ($k = 13$, $ES = 45$) revealed a positive association between self-assessment ratings and ratings with validated measures of ECE quality ($r = .38$), indicating a moderate convergent validity. Studies with lower methodological quality and published "peer reviewed" studies reported somewhat higher correlations between self-assessment ratings and ratings with validated measures. The meta-analytic correlation remained significant after removal of studies with lower methodological quality ($r = .33$) or studies from the "grey" literature ($r = .44$). A second meta-analysis ($k = 16$, $ES = 71$) with a focus on the predictive validity of self-assessment ratings showed a small significant association between self-assessed ECE quality and child outcomes ($r = .09$); there were no significant moderators. *Practice or Policy:* Despite empirical evidence for the validity of self-assessment, further studies are needed to investigate potential bias in self-assessment. Future studies should further explore the validity and reliability of self-assessment measures in ECE, including countries outside the United States.

Introduction

The quality of early childhood education (ECE) is an important factor for child development (Campbell et al., 2002; Deming, 2009; Farrell et al., 2019; Garca et al., 2020; Halpern, 2000; National Institute of Child Health and Human Development Early Child Care Research Network, 2000; National Association for the Education of Young Children, 2022; Peisner-Feinberg & Burchinal, 1999). A growing number of stakeholders rely on self-assessment to assess, measure, and improve ECE quality. Following the literature on ECE quality assessment, we define self-assessment as a systematic self-directed accountability method of monitoring and evaluating service provision in ECE settings to enhance quality, with the ultimate goal to improve children's development (Center for Early Childhood Professional Development, 2020; Munton et al., 1997; Norris et al., 2003). Self-assessment may not only involve internal ratings conducted by the organization itself but also external ratings by trained observers (Center for Early Childhood Professional Development, 2020; National Association for the Education of

CONTACT Harriet J. Vermeer vermeer@fsw.leidenuniv.nl Social and Behavioural Sciences, Institute of Education and Child Studies Leiden University, Wassenaarseweg 52, Leiden 2333 AK, The Netherlands

Young Children, 2011; Quality Compendium, 2022b). Self-assessment may include a variety of instruments to monitor quality including observation, document review, and self-report (Tout et al., 2009). Several self-assessment initiatives have been taken (Elicker & Ruprecht, 2019) and, therefore, the need for high-quality self-assessment in ECE with valid measures has become more urgent.

The most common goal of self-assessment is supporting continuous ECE quality improvement by improving the structural and/or process quality of ECE (Elicker et al., 2007; Markowitz et al., 2020; Organisation for Economic Co-operation and Development, 2013, 2021; Picchio et al., 2012; Quality Compendium, 2022a). Self-assessment may stimulate critical reflection on current practice by recognizing strengths and identifying opportunities for improvement in the own organization (Australian Children's Education & Care Quality Authority, 2019; National Association for the Education of Young Children, 2010; Quality Compendium, 2022a). It enables professionals to become aware of their own strengths and weaknesses, and to identify their needs for improvement (Isoré, 2009). Self-assessment may also stimulate an ongoing dialog on ECE quality with childcare providers, researchers, policymakers and ECE inspection.

In the late nineties, Munton and colleagues developed the first ECE self-assessment tools in the United Kingdom (Munton & Mooney, 1999; Munton et al., 1997). In the United States Quality Improvement Systems (QISs, also known as Quality Rating Systems, QRSs and Quality Rating and Improvement Systems, QRISs), were developed and have evolved over the past two decades from a rare approach to inform parents about childcare quality to a standardized tool for ECE programs in more than half of all states in 2014 (Elicker & Ruprecht, 2019; Goffin & Barnett, 2015; Tout et al., 2017). The first QRIS was implemented in 1997. As of fall 2021, 42 states had at least one fully developed QRIS, and 3 states are currently developing a QRIS (Quality Compendium, 2022a). Although there is variety in QRIS design, there is broad agreement regarding its primary features. Most of the QRISs are completely voluntary for ECE programs (Herbst, 2018; Swenson-Klatt & Tout, 2011), while sometimes basic licensure was incorporated into the QRIS by assigning all licensed ECE programs to the lowest rating; programs can then decide whether or not they want to participate in further QRIS procedures to attain higher ratings (Swenson-Klatt & Tout, 2011). Also in other countries accountability initiatives have been taken, including Australia (NSW Department of Education, 2020), Belgium (Janssen et al., 2016), Canada (Perlman et al., 2017), Greece (Rentzou, 2010), China (Hu & Li, 2012; Hu et al., 2015; Li et al., 2016; Pan et al., 2010), Hong Kong (Wong & Li, 2010), and the Netherlands (Boogaard et al., 2011; Gevers Deynoot-Schaub et al., 2009).

Quality Indicators in ECE Self-Assessment

Self-assessment often focuses both on structural quality and process quality. Process quality refers to the child's day-to-day experiences in ECE and includes the social, emotional, physical, and instructional activities and interactions with caregivers, peers, and the physical environment (Howes et al., 2008; Pianta et al., 2005; Thomason & La Paro, 2009). Structural quality characteristics include, among other things, group size, children-to-caregiver ratio, and educators' qualifications, and is considered an important precondition of proximal process quality (Cryer et al., 1999; Phillips et al., 2000; Phillipsen et al., 1997; Pianta et al., 2005; Vandell, 2004). Self-assessment in ECE often includes multiple components to assess both structural and process quality of ECE and to make these transparent for staff and parents (Zellman et al., 2008). Self-assessment, specifically QRISs, often includes the evaluation of quality of ECE at classroom level. The most common type of QRIS indicators for center-based programs were related to the physical environment (86% of 44 QRISs). In addition, 82% of QRISs included indicators regarding interactions between staff and children. In 2021, 43 QRISs used an observational tool, often the Environment Rating Scales (ERS) (Cryer et al., 2003, 2004), and/or the Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008). QRIS may use the ERS and/or the CLASS in self-assessment (Garcia-Sanchez et al., 2015), and also may involve self-developed tools (Quality Compendium, 2022b). About half of QRISs (44%) also include

the support of children's healthy development and school readiness (Organisation for Economic Co-operation and Development, 2021; Quality Compendium, 2022b).

Validation of Self-Assessment in ECE

Measurement and the psychometric quality of the various measures involved play a pivotal role in self-assessment. Seen from the perspective of professionals, who are responsible for quality management, self-assessment measures should measure ECE quality in a reliable and valid way in order to give ECE staff and other stakeholders meaningful insights. Validation is important to ensure that the use of a self-assessment measure produces accurate ratings that capture meaningful differences in program quality (Karoly, 2014; Tout et al., 2017). It is, however, not clear whether outcomes with self-assessments are systematically related to evaluation with validated measures of ECE quality, and whether quality ratings from self-assessment predict children's development (Language ENvironment Analysis, 2022). Examples of validated ECE measures most often used in publications included in this study are the ERS (Cryer et al., 2003, 2004), and the CLASS (Pianta et al., 2008), used by external raters with an academic affiliation. In this systematic review, we focus on the reliability and validity of self-assessment measures in ECE.

Almost all validation studies on ECE self-assessment measures were conducted in the United States. In these studies, self-assessed ECE quality is compared to observation data collected by researchers (Elicker & Ruprecht, 2019; Karoly, 2014; Organisation for Economic Co-operation and Development, 2013; Schilder, 2013; Zellman & Fiene, 2012). The findings from the first validation studies in the field of self-assessment in an ECE context have generally been discouraging, showing weak and inconsistent associations (i.e., convergent validity), as measured with validated measures (ERS and/or CLASS) (seefor example, Hestenes et al., 2015; Hu et al., 2015; Jeon et al., 2014; Lahti et al., 2013, 2015; Lipscomb et al., 2017; Magnuson & Lin, 2015; Pan et al., 2010; Perlman et al., 2017; Rentzou, 2010; Soliday Hong et al., 2015; Wong & Li, 2010). A gradually growing number of studies explored the predictive value of self-assessment quality ratings for children's development (measured with validated instruments), reporting small associations (Hestenes et al., 2015; Jeon & Buettner, 2015; Lahti et al., 2013, 2015; Le et al., 2015; Magnuson & Lin, 2016; Sabol & Pianta, 2015; Sabol et al., 2013; Soliday Hong et al., 2015; Yazejian et al., 2017). These findings indicate that self-assessed ECE quality was moderately represented in the data collected with validated measures, acknowledging that predictive validity for validated measures is also modest. The review of published validation studies by Cannon et al. (2017) suggests some progress toward the development of valid rating systems, but the evidence is still limited and often contradictory, which prevents drawing firm conclusions about the validity of current QRIS ratings. Regarding convergent validity, most QRIS validation studies have shown that programs with higher ratings had higher scores on the validated Environment Rating Scales (Karoly, 2014).

Considering that the ERSs are often embedded in the QRISs, a significant positive relationship between the ERS and QRIS ratings may be expected (Karoly, 2014), acknowledging the differences between self-assessment and external evaluation. However, independent measures of ECE quality have not always shown the expected positive relationship with QRIS quality ratings (Karoly, 2014). In addition, QRIS ratings have not always shown robust relationships with children's development (Elicker & Ruprecht, 2019). This raises the question whether current QRISs accurately capture differences in program quality that predict child development (Cannon et al., 2017; Elicker & Ruprecht, 2019; Sabol & Pianta, 2015; Tout, 2013). The lack of predictive validity has led to calls for new ways of measuring quality in childcare settings (Cannon et al., 2017; Elicker & Ruprecht, 2019; Perlman et al., 2017; Rentzou, 2010; Thornburg et al., 2009; Wong & Li, 2010). This prompted the emergence of second-generation QRISs. Second-generation QRISs focus on direct quality improvement support activities such as use of evidence-based approaches to quality improvement and evaluating the effectiveness of promising new approaches to improve childcare quality (Cannon et al., 2017; Elicker et al., 2022; Markowitz et al., 2020; Schmitt et al., 2023; Sussman et al., 2022;

Vitiello et al., 2018). In sum, many challenges and important questions remain concerning self-assessment of ECE quality, including its reliability and validity (Elicker & Ruprecht, 2019).

The Present Study

The need for validation of self-assessment in ECE has become more urgent after the growth of self-assessment in several countries (Karoly, 2014; Markowitz et al., 2020), and because we know relatively little about the reliability and validity of self-assessment measures in this context (Goffin & Barnett, 2015). In the current study, we focus on the reliability and validity of self-assessment measures in the context of center-based ECE. We conducted a meta-analysis to integrate and analyze the results of different validation studies. This meta-analysis is to our knowledge the first review which synthesizes findings from the literature on the validation of self-assessment instruments regarding ECE quality. We focus on the reliability, convergent validity, and predictive validity of self-assessment measures in ECE by addressing the following research questions.

Q1: What is the reliability of self-assessment in ECE?

Q2: Do the outcomes of self-assessed ECE quality correspond with ECE quality ratings measured with validated measures (convergent validity)?

Q3: Which variables at the self-assessment method level and study level moderate this association (see Q2)?

Q4: Do the outcomes of self-assessed ECE quality predict children's development measured with validated measures (predictive validity)?

Q5: Which variables at the self-assessment level, study level and developmental outcome level moderate this association (see Q4)?

Method

Search Strategy

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). The electronic databases PsycINFO, ERIC and Web of Science were searched in July 2020, and was updated in January 2024. In addition, we conducted a search in Google Scholar in January 2024 in order to find unpublished studies. We used the following query: (self-assess* OR self-evaluat* OR rating) AND (quality* OR improve* OR accountab*) AND (childcare OR "child care" OR daycare OR "day care" OR kindergarten OR "kinder garten" OR preschool OR pre-school OR nurser* OR ECE OR "early childhood education" OR ECEC OR "early childhood education and care" OR "after school care") AND ("educational quality" OR "child care quality" OR "quality rating" OR "quality improvement" OR "quality rating system*" OR QRS* OR "quality rating and improvement system*" OR QRIS* OR "program quality" OR QIS* OR "quality improvement system*" OR TQRIS* OR "tiered quality rating and improvement system*"). In the initial pilot, we found that relevant publications used different terms referring to validation studies on self-assessment in ECE and these key words were combined in an extensive query in order to identify all relevant publications meeting the inclusion criteria. All authors discussed draft versions of the search protocol during multiple online meetings to reach agreement on the final version. The timeframe was set from 1990 to 2023, as the first self-assessment tools for ECE were developed and implemented in the late nineties.

Selection of Publications

We included publications that (1) focused on quality of center-based ECE for children from 0 to 72 months through self-assessment; (2), reported empirical quantitative data concerning reliability and/or validity of the self-assessment instrument; (3) and used validated measures for validation of self-assessment. We included both published and unpublished studies in our sample. All publications since 1990, which met the inclusion criteria, were included in our study. We attempted to include studies related to after school care, but the search did not yield any publications that met our inclusion criteria. Regarding research design, we included correlational validation studies comparing ECE quality self-assessment ratings with ratings conducted with validated measures. We considered a rating as self-assessment if the authors of the study labeled the reported rating as self-assessment. We screened studies based on title, abstract, and full text. The first author screened the reference lists of all 187 identified publications at title and abstract level to identify additional publications. When an additional publication was identified, the reference list of this publication was also screened by the first author. This amounted to a total of 105 additional publications. The abstracts were screened to determine whether all inclusion criteria were met. A flow chart of the literature search is shown in Figure 1.

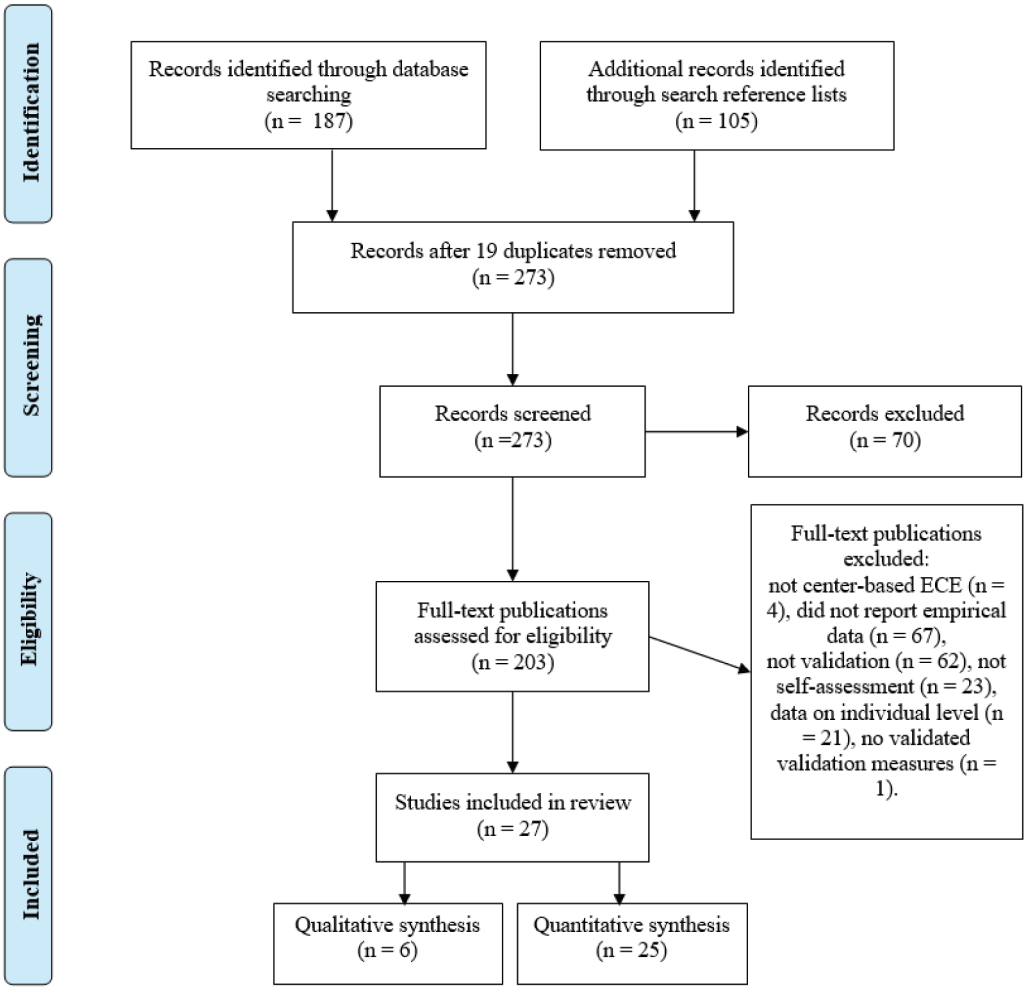

Figure 1. PRISMA flowchart.

Coding

Each publication was coded independently by two authors with a coding scheme that included general study characteristics (year of study, country, study design, sample size, age range, and quality rating of included studies; see Table 1). The quality of each publication was assessed by applying an overall rating of "good," "fair" or "poor" using the "Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies" (National Heart Lung and Blood Institute, 2013). The quality of a publication was rated based on items concerning selection bias, blinding of assessors, data collection methods, and quality of the analyses. In line with the guidelines of the Quality Assessment Tool, we considered the overall rating good when the study design was not flawed and the potential risk of bias was low. If a publication showed some risk of bias the publication was rated "fair," and if a study design was flawed and/or the risk of bias was high the publication was rated "poor." The quality of all publications was coded by two independent coders. The percentage of agreement between the first and the second author was 88% with a corresponding Cohen's kappa of .82, whereas agreement between the first and the third author was 87% with a corresponding Cohen's kappa of .80. In case of disagreement (i.e., 4 publications) with respect to the rating of a publication, researchers achieved consensus by discussion. Furthermore, we coded the type of self-assessment measure (QRIS or another measure), the quality domains of the self-assessment measure (general process quality or quality of interaction caregiver-children), whether self-assessment focused on social or cognitive developmental outcome, whether self-assessment was conducted by internal or external raters, whether the self-assessment instrument was first or second generation, and the publication status of the study. Finally, we coded reliability (i.e., Cronbach's α or the related KR20) and validity coefficients (i.e., Pearson correlation between self-assessment ratings and ratings of validated measures or between categorical quality ratings and children's outcomes). The coding of the full-text publications resulted in an additional exclusion of publications. In total, we included 27 publications in our study (see Figure 1).

Data Analysis

We conducted two meta-analyses: a meta-analysis of the validity coefficients of the studies with regard to convergent validity (1) and with regard to predictive validity (2) of self-assessment measures. Specifically, for QRIS publications we used the observational data from the ERS (Cryer et al., 2003, 2004) or CLASS (Pianta et al., 2008). The effect size for both meta-analyses is Pearson's correlation *r*. Findings within one publication were included as separate effect sizes (Pearson's correlation *r*) when they pertained to different outcomes, subsamples or measurement points. Estimates were extracted from the research report wherever possible; only in a few cases a (standardized) beta coefficient was converted to *r*, which was used in our analyses. The correlation coefficient *r* and the beta coefficients from multiple regression models are found to be strongly correlated ($r = .84$ in the review study by Peterson & Brown, 2005). Following the recommendation of Bowman (2012), *r* values were converted to Fisher's *z* for the meta-analytic aggregation of outcomes. In addition, some studies reported means, standard deviations and sample size for the different quality levels (categories). Based on this data we computed Cohen's *d*, which then was converted to *r* and Fisher's *z* for the meta-analytic aggregation of outcomes. Results were finally converted back to *r* for interpretation of the outcomes. We did not perform a meta-analysis if there were less than three studies.

We conducted three-level multilevel meta-analyses with sampling variance of the extracted effect sizes at level 1, variance between effect sizes extracted from the same study at level 2, and variance between studies at level 3, using R Statistical Software and the Metafor package. This model allows effect sizes to vary between participants (level 1), outcomes (level 2), and studies (level 3) (Viechtbauer, 2010). We explored possible moderator effects for the convergent validity and predictive validity. We attempted to test the same moderators for both convergent and predictive validity, but this was not always possible as in some cases $k < 3$. The moderators concerning convergent validity were the

Table 1. Descriptives of the included studies.

Study Characteristics							Self-Assessment			Reliability and Validity	
Study	Country	Study design	Sample (groups)	Age range children in months	Published (peer reviewed) or not published (PU, NP)	Quality rating of included studies	Self-assessment instrument	Quality domains included (CI, GP, QI)	Internal/ external rating	Reliability (IC, IR)	Validity (CT, CV, KG, PV)
Elicker et al. (2022)	United States	Longitudinal	75	12–36	PU	Good	QRIS	QI	External	NA	PV
Hestenes et al. (2015)	United States	Cross-sectional	97	32–69	PU	Good	QRIS	GP, QI	External	NA	CV, PV
Hu et al. (2015)	China	Cross-sectional	177	36–72	PU	Fair	KQRS	GP, QI	NI	NA	CV
Jeon et al. (2014)	United States	Cross-sectional	71	NI	PU	Fair	QRIS	QI	Internal	NA	CV, KG
Jeon & Buettner (2015)	United States	Cross-sectional	36	NI	PU	Fair	QRIS	QI	External	NA	PV
Lahti et al. (2013)	United States	Cross-sectional	196	NI	NP	Good	QRIS	GP, QI	External	NA	CV
Lahti et al. (2015) (study Indiana)	United States	Cross-sectional	149	NI	PU	Fair	QRIS	GP, QI	Internal	NA	CT, CV
Lahti et al. (2015) (study Maine)	United States	Cross-sectional	194	NI	PU	Fair	QRIS	GP	Internal	NA	CT, CV
Le et al. (2015)	United States	Longitudinal	49	36–60	PU	Good	QRIS	GP	External	NA	PV
Lipscomb et al. (2017)	United States	Cross-sectional	304	15–60	NP	Fair	QRIS	GP, QI	External	NA	CV
Magnuson and Lin (2015)	United States	Cross-sectional	109	36–60	NP	Fair	QRIS	GP	External	NA	CV
Magnuson and Lin (2016)	United States	Longitudinal	507	36–60	NP	Fair	QRIS	GP	External	NA	PV
Markowitz et al. (2020)	United States	Longitudinal	76	48–60	PU	Good	QRIS	QI	External	NA	PV
Pan et al. (2010)	China	Cross-sectional	18	48–72	PU	Fair	BKQRS	GP	Internal	NA	CV
Perlman et al. (2017)	Canada	Cross-sectional	93	0–30	PU	Fair	AQI	GP	External	IC, IR	CT, CV
Rentzou (2010)	Greece	Cross-sectional	27	53	PU	Poor	ACEI GGA	GP	Internal	NA	CV
Sabol et al. (2013)	United States	Cross-sectional	703	NI	PU	Good	QRIS	GP, QI	External	IC	PV

(*Continued*)

Table 1. (Continued).

Study Characteristics							Self-Assessment			Reliability and Validity	
Study	Country	Study design	Sample (groups)	Age range children in months	Published (peer reviewed) or not published (PU, NP)	Quality rating of included studies	Self-assessment instrument	Quality domains included (CI, GP, QI)	Internal/ external rating	Reliability (IC, IR)	Validity (CT, CV, KG, PV)
Sabol & Pianta (2015)	United States	Longitudinal	71	24–60	PU	Good	QRIS	GP, QI	External	NA	PV
Schmitt et al. (2023)	United States	Longitudinal	NI	NI	PU	Fair	QRIS	GP	External	NA	PV
Soliday Hong et al. (2015)	United States	Cross-sectional	101	36–48	PU	Poor	QRIS	GP, QI	External	IC	CV
Sussman et al. (2022)	United States	Longitudinal	1725	54–66	NP	Fair	QRIS	NI	External	NA	PV
Thornburg et al. (2009)	United States	Cross-sectional	66	36–60	NP	Fair	QRIS	NI	External	NA	PV
Vitiello et al. (2018)	United States	Cross-sectional	85	48–60	PU	Fair	QRIS	QI	External	IC, IR	PV
Wong & Li (2010)	China	Cross-sectional	80	NI	PU	Poor	QAI	CI	Internal	NA	CV
Yazejian et al. (2017)	United States	Longitudinal	22	36–60	NP	Fair	QRIS	GP	External	IC	CV, PV

Note. NA = not applicable, NI = not indicated; Published: PU = published, NP = not published; Quality domains: CI = Child indicators, GP = General process quality, QI = Quality of interaction caregiver-children; Reliability: IC = internal consistency, IR = inter rater reliability; Validity: CV = convergent validity, CT = content validity, KG = known groups validity, PV = predictive validity.

domains of ECE quality (general process quality vs. quality of interaction between caregiver and children), type of rater (internal vs. external raters), studies that we rated as "poor" compared to studies rated as "fair" or "good," study origin (from the Unites States vs. other), and published studies versus not published literature. Regarding predictive validity we explored if the association between self-assessed ECE quality and children's development was moderated by the domain of children's development (social vs. cognitive development). In addition, we tested whether the association between self-assessed quality ratings and child development was moderated by the version of self-assessment instrument (first or second generation) or by the publication status of a study (published vs. unpublished). We tested for publication bias using a multilevel meta-analyses (MLMA) implementation of the Egger test using effect size variance as a moderator (Rodgers & Pustejovsky, 2021). Heterogeneity of outcomes was tested using the *Q* test (Huedo-Medina et al., 2006). Following the guidelines from (Cooper et al., 2019), moderators were tested individually in the moderator analysis.

Results

Twenty-seven publications were included in our review, of which 25 in the meta-analysis. Most publications ($k = 18$; 72%) were from the United States, but we also included studies from Canada, China, Greece and Hong Kong. The self-assessment measures were administered on groups with children in the age range of 0–72 months. Six publications examined the reliability of self-assessment instruments. In our review, we focus on validity of self-assessment, but because reliability is a prerequisite for validity, we briefly report reliability figures from the included studies in a narrative review. Due to the insufficient number of eligible publications (which was set at $k \geq 3$), we were not able to meta-analytically examine the data on known groups validity.

Narrative Review

Six publications reported internal consistency of the self-assessment instrument (Perlman et al., 2017; Quick et al., 2016; Sabol et al., 2013; Vitiello et al., 2018; Yazejian et al., 2017; Zellman et al., 2008) and three also reported interrater reliability (Perlman et al., 2017; Sabol et al., 2013; Vitiello et al., 2018). In these studies, research into reliability concerned validated instruments in the context of self-assessment like ERS or CLASS, applied by trained local raters. Perlman et al. (2017) reported a Cronbach's alpha of .84 for the infant version and .86 for the toddler version. Sabol et al. (2013) reported a Cronbach's alpha of .88 related to the CLASS overall score. Yazejian et al. (2017) reported a value of .82 (KR20). Vitiello et al. (2018) reported Cronbach's alpha values ranging from .77 to .96. Zellman et al. (2008) found a moderate relation ($r = .25$) between the components of the self-assessment instrument. Only Quick et al. (2016) reported a relatively low internal consistency ($\alpha = .54$).

Inter-rater reliability was generally high. Perlman et al. (2017) reported interrater reliability ranging from 96% to 100% concerning the infant instrument, and 88% to 100% for the toddler instrument. Sabol et al. (2013) reported a weighted kappa of .69 for the ERS, and .66 for CLASS. Finally, Vitiello et al. (2018) reported a minimum score of 80% agreement with the master codes; no exact estimate or reliability range was reported.

The included studies evaluated different types of validity. Content validity supported the quality components included in a QRIS (e.g., staff qualifications, learning environment, family engagement), which fitted in with how various stakeholders essentially define childcare quality in the concerning states (Lahti et al., 2015). Two publications (Jeon et al., 2014; Squibb, 2013) investigated the known groups validity. Known groups validity is a form of construct validity that refers to the extent to which an assessment can differentiate between groups that are expected to differ based on certain characteristics (Hattie & Cooksey, 1984). Jeon et al. (2014) reported that programs which participated in QRIS, had significantly better scores than nonparticipating programs on the ECERS-R (Cryer et al., 2003), Emotional Support and Instructional Support domains from the CLASS (Pianta et al., 2008), and on

physical literacy environment as measured with the Early Language & Literacy Classroom Observation (ELLCO) (Smith et al., 2008). Squibb (2013) reported mixed empirical evidence regarding known groups validity. Associations between QRIS ratings and quality profiles in Virginia's QRIS showed that each of the four quality profiles described a different type of quality regarding four quality components (i.e., teacher education, interactions, ratio and group size, and environment). Strong, significant positive relationships were found between certain quality profiles and QRIS ratings, whereas for another quality profile a negative, strong relationship was found with the QRIS Star Rating.

Meta-Analyses of Convergent and Predictive Validity

In total, 25 publications examined convergent and/or predictive validity of self-assessment in ECE and were included in our meta-analyses. The meta-analysis with regard to the convergent validity of self-assessment synthesized findings for a total of 1,882 groups in the age range from 0 to 72 months. The meta-analysis related to the predictive validity included a total of 79,163 children in the age range from 12 to 72 months.

Convergent Validity of Self-Assessment in ECE

The meta-analysis of validity coefficients with regard to convergent validity ($k = 13$, ES = 45) demonstrated a significant positive relationship between self-assessment ECE quality measures and validated quality ECE measures, $r = .38$, 95% CI [.29–.46], $p < .001$ (see Figure 2). This corresponds to a medium effect size (Cohen, 1988). The funnel plot in Figure 3 of the effect sizes showed that studies are symmetrically distributed, and publication bias seems absent, which was confirmed with the MLMA Egger test ($\beta = 0.44$, $z = 0.59$; $p = .56$).

The Q-tests of homogeneity indicated significant variation among studies, $Q(44) = 172.26$, $p < .0001$. We explored possible moderator effects for the convergent validity outcomes (see Table 2). The domains of ECE quality (general process quality vs. quality of interaction between caregiver and children) did not significantly moderate the relationship between self-assessment ratings and ratings with validated measures of ECE quality, $r = -.03$, 95% CI [−0.14–0.08], $p = .56$. We found no significant moderation effect of internal vs. external raters, $r = .09$, 95% CI [−0.12–0.30], $p = .40$. In addition, studies from the United States vs. other countries did not significantly moderate the relationship between self-assessment ratings and ratings with validated measures of ECE quality, $r = -.17$, 95% CI [−0.37–0.04], $p = .11$. We did find a moderator effect of study quality, showing that studies with weaker designs reported higher correlations between self-assessment ratings and ratings with validated quality measures vs. studies with stronger designs, $r = .26$, 95% CI [0.10–0.40], $p = .0018$. After removal of lower quality studies, an additional analysis ($k = 10$) resulted in a significant medium positive relationship between childcare self-assessment quality measures and validated childcare quality measures, $r = .33$, 95% CI [.27–.39], $p < .001$. Finally, published studies reported higher correlations between self-assessment ratings and ratings with validated quality measures than unpublished studies, $r = .25$, 95% CI [.09–.41], $p = .0028$. After removal of unpublished studies, an additional analysis ($k = 9$) resulted in a significant medium positive relationship between childcare self-assessment quality measures and validated childcare quality measures, $r = .44$, 95% CI [.36–.52], $p < .001$.

Predictive Validity of Self-Assessment in ECE

In the meta-analysis ($k = 16$, ES = 71) of predictive validity findings, children's outcome measures were coded as either "positive" (ES = 50; 70,4%) or "negative" (ES = 21; 29.6%) (see Figure 4). Effect sizes for negative outcomes were converted by changing the sign, and, hence, outcomes from our analysis thus reflect positive relationships between ECE quality and children's favorable development. The meta-analysis of predictive validity coefficients resulted in a significant positive

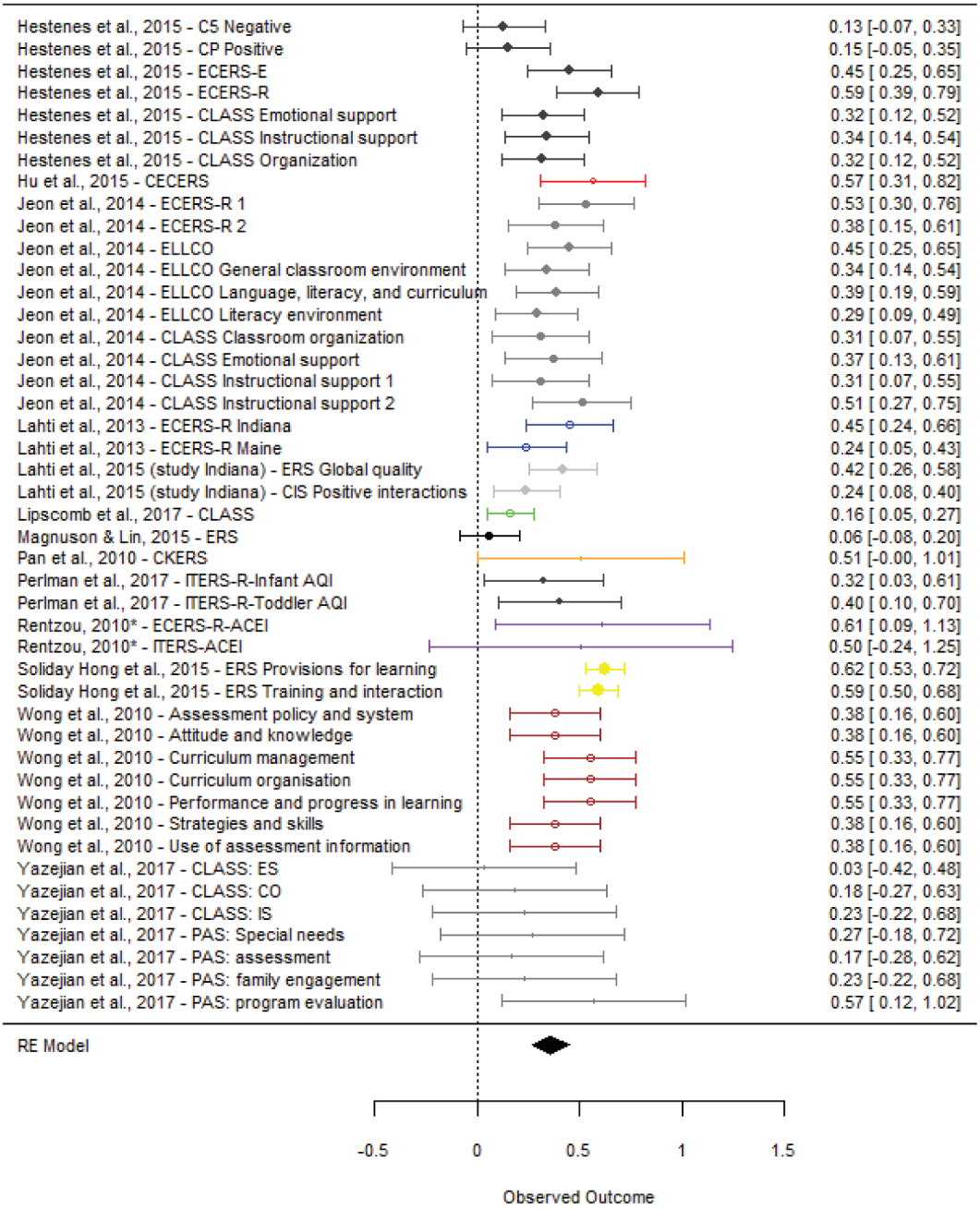

Figure 2. Outcomes meta-analysis: convergent validity of self-assessment instruments in ECE.

relationship between self-assessed childcare quality and child developmental outcome ($r = .09$, 95% CI [.04–.14], $p = .0005$), i.e., a small effect size (Cohen, 1988). The funnel plot in Figure 5 shows little asymmetry, which was confirmed by the MLMA Egger test ($\beta = -.63$, $z = -.62$, $p = .53$).

The *Q*-tests of homogeneity indicated significant variation among studies, $Q(43) = 169.92$, $p < .0001$. There was no moderator effect found for predictive validity regarding the comparison between self-assessment and validated ratings regarding social vs. cognitive developmental outcomes, $r = -.03$, 95% CI [−.12–.06], $p = .51$ (see Table 2). In addition, the association

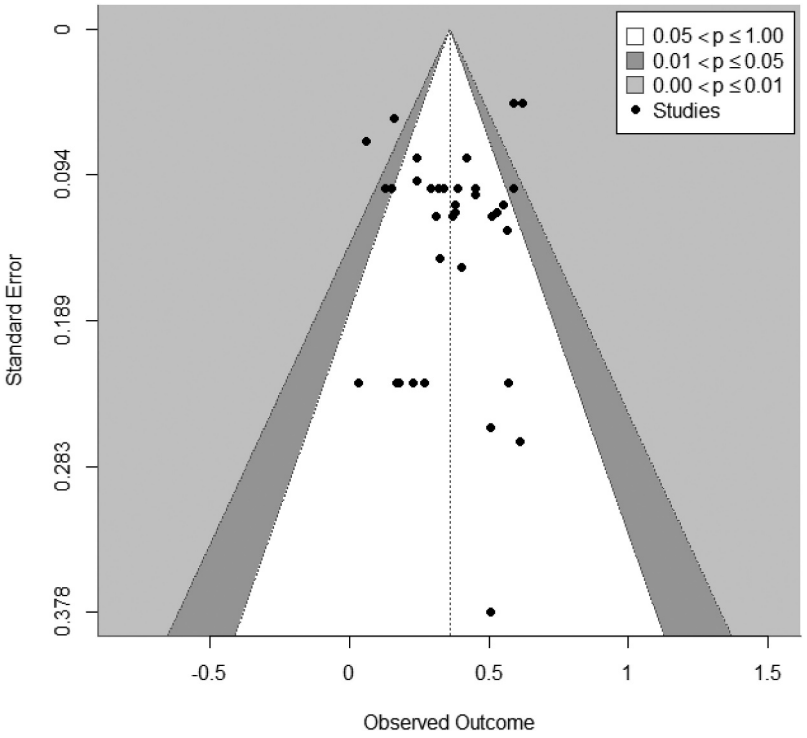

Figure 3. Funnel plot for publications related to convergent validity of self-assessment instruments in ECE.

Table 2. Overview of results of the moderator analyses.

Type of Validity	Moderator	Category	ES (studies)[1]	*N*	Effect Size (*r*)
Convergent validity	Quality domain(s)	General process quality	36 (13)	1,739	.38
		Quality of interaction caregiver-children	9 (4)	743	.35
	Rater	Internal	23 (5)	345	.43
		External	22 (8)	1,393	.35
	Quality study***	Fair-good	34 (10)	1,202	.32
		Poor	11 (3)	536	.53
	Origin study	US	32 (8)	1,458	.33
		Not US	13 (5)	280	.48
	Publication status**	Published (peer reviewed)	34 (9)	1,026	.44
		Not published	11 (4)	712	.21
Predictive validity	Developmental outcomes	Social development	17 (9)	74,234	.07
		Cognitive development	54 (14)	78,813	.10
	Version self-assessment	First generation	61 (12)	78,248	.10
		Second generation	10 (3)	915	.05
	Publication status	Published (peer reviewed)	47 (10)	70,905	.11
		Not published	24 (5)	9,237	.07

Note: [1]The number of studies and/or effect sizes for convergent and predictive validity do not always add up because not every study contained data on every moderator of our study.
** = $p < .01$, *** = $p < .001$; For convergent validity, *N* is reported at group level. For predictive validity *N* is reported at child level.

between self-assessed ECE quality and ratings with validated measures concerning children's development outcome was not moderated by the application of a second vs. first-generation self-assessment instrument either ($r = -.05$, 95% CI [−.20–.10], $p = .52$), or by publication status ($r = .04$, 95% CI [−.08–.15], $p = .50$).

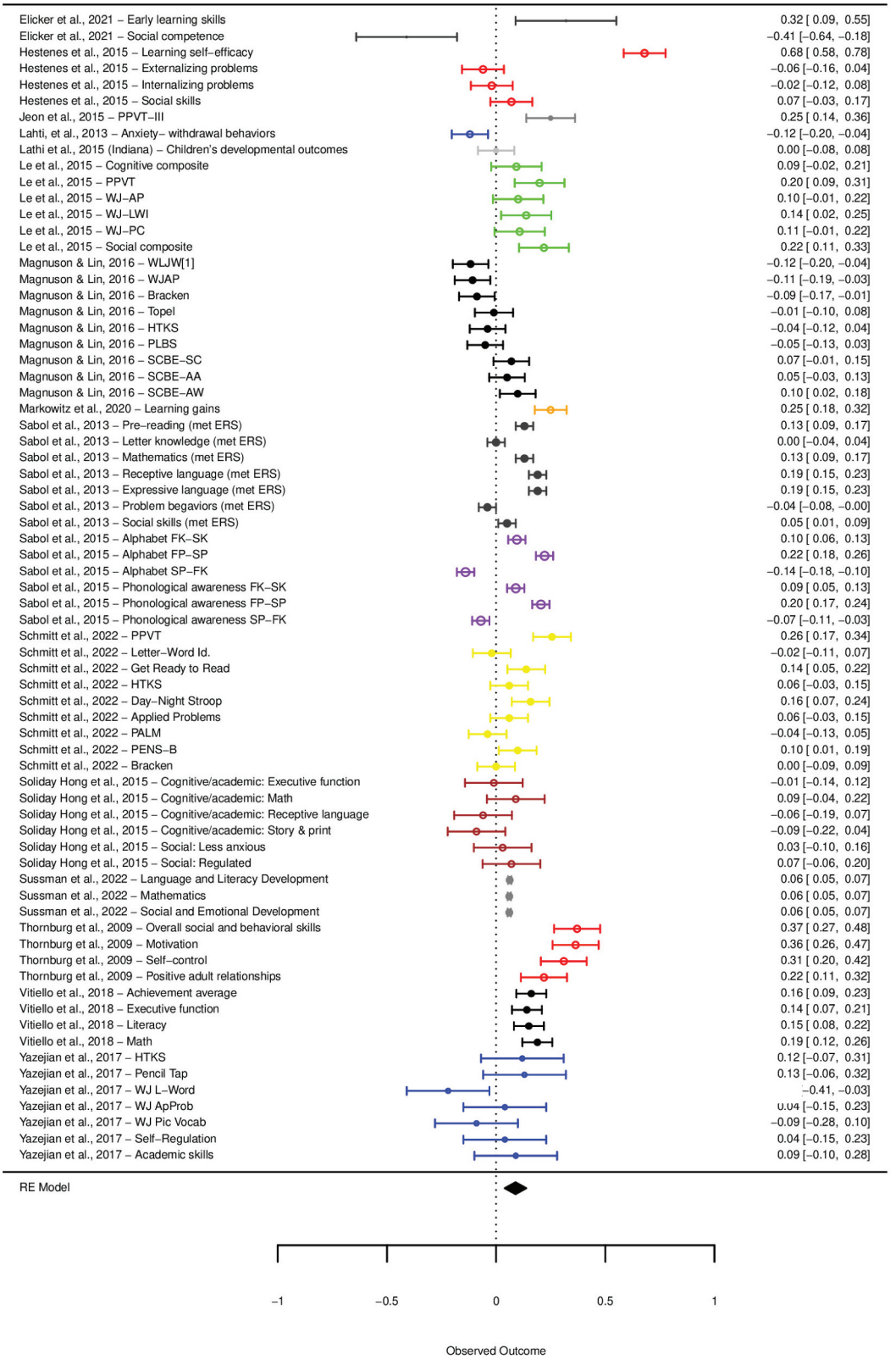

Figure 4. Outcomes meta-analysis: predictive validity of self-assessment instruments in ECE.

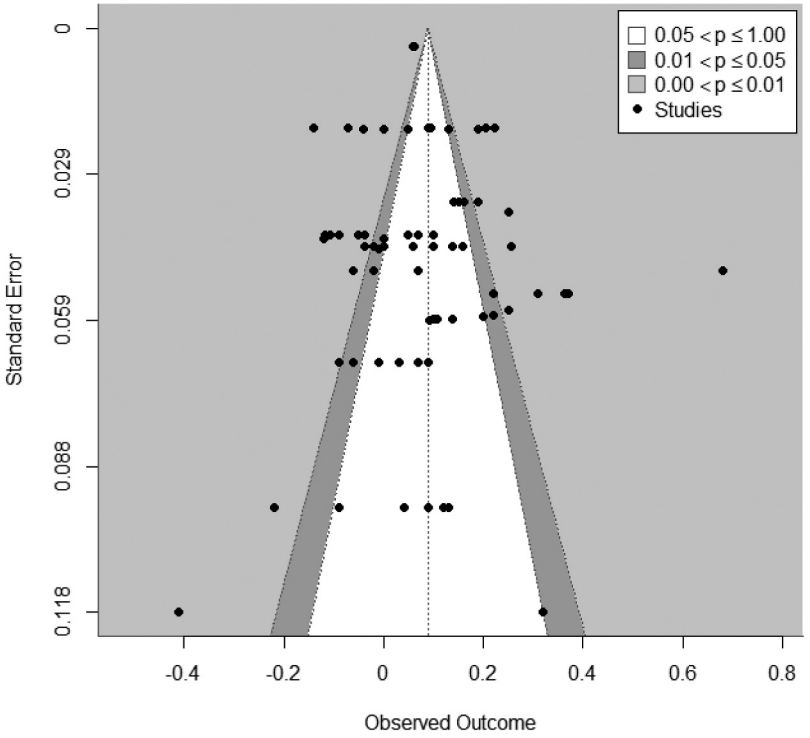

Figure 5. Funnel Plot for Publications related to Predictive Validity of Self-Assessment Instruments in ECE.

Discussion

The findings from our review provide empirical support for the construct validity and reliability of self-assessment measures as currently used in ECE, which means that self-assessment can, to a certain extent, capture ECE quality.

The meta-analysis concerning convergent validity showed a medium significant correlation between self-assessment measures and validated measures of the quality of ECE. Our meta-analysis shows that self-assessment ratings, which include general process quality and/or measures for caregiver–child interaction, converge with external ratings with validated scales. An explanation for the relatively modest convergence is that self-assessment measures assess similar but not identical constructs compared to validated measures. Each self-assessment measure has slightly different items and procedures. QRIS includes items related to the quality of ECE programs, but also kindergarten readiness and parent education (Elicker & Ruprecht, 2019; Heckman, n.d.; National Association for the Education of Young Children, 2022). These dimensions may not (all) be represented in the validated measures for the external evaluation of ECE quality. Second, self-assessment was often completed prior to the measurements with validated instruments and the time between the waves of data collection sometimes was several weeks, which may have lowered the correlation between self-assessment ratings and ratings of validated measures of ECE quality. The results of the current review complement earlier research in the United States that reported higher self-assessment ratings for centers with higher scores on validated measures of ECE quality, as measured with the ERS (Karoly, 2014).

The domains of ECE quality, type of rater, and origin of the study did not significantly moderate the relationship between self-assessment ratings and ratings with validated measures of ECE quality, suggesting that these variables do not influence the overall effect. We did find a moderator effect for study quality, indicating higher effect sizes for studies with weaker designs. Further, peer-reviewed

publications generally reported higher correlations, possibly due to publication bias (i.e., studies with higher correlations are more likely to be accepted for publication (Cooper et al., 2019). However, an additional analysis after removal of the weaker studies or studies from the gray literature resulted in similar outcomes, and our meta-analytic findings seem therefore robust.

A second meta-analysis which focused on the predictive validity of self-assessment ratings, revealed that the association between self-assessed quality and child development was modest. Our finding that self-assessment measures in ECE has limited predictive validity is in line with the predictive value of validated and widely used measures in ECE research. The meta-analysis of Perlman et al. (2016) regarding the predictive validity of the CLASS in ECE has revealed comparably small associations between the CLASS and children's outcomes. Also the meta-analysis of longitudinal relations between ECE quality, as measured with the CLASS, and academic outcomes reported positive but small association (Ulferts et al., 2019). Our results are slightly more positive than previous studies in the United States that reported mixed results (Elicker & Ruprecht, 2019). Whereas some studies found positive effects of high-quality ECE on early literacy skills and social-emotional development (Thornburg et al., 2009; Zellman et al., 2008), other researchers did not find a significant relationship between QRIS-ratings and child development (Sabol et al., 2013). Our findings provide some support, at a meta-analytic level, for the assumed positive relationship between quality ratings from self-assessment in ECE and children's social and cognitive development in early childhood education and the first years of elementary school.

A possible explanation for the small positive association between self-assessment and child developmental outcomes, in line with Burchinal et al. (2010), is that the ratings of self-assessed ECE quality (just like many validated measures, like ERS or CLASS) are not conducted at child level but at group level, which may not provide an accurate assessment of input at the level of individual children (Howard et al., 2024; Meek et al., 2020; Policy Equity Group, 2022; Swartz, 2022;). Second, there may be erratic developmental patterns for children at a young age. Children differ in developmental trajectories for the physical, cognitive, social, and emotional growth and in the way they interact with their environment as well as in play, affection, and other factors (Ruffin, 2009), and, hence, strong predictive relationships cannot be expected. Finally, various variables each influence child development in the period of preschool (e.g., the quality of instruction, transition from preschool to school, parent support of children's learning). The self-assessments of quality show similarly quite small associations with children's development as is the case for observational ratings of quality. To conclude, self-assessment has some predictive validity for the socio-emotional, cognitive and academic domain, acknowledging the weak association between ECE quality and various outcomes at child level during their preschool years.

We found no moderators for our meta-analytic results for predictive validity of self-assessment. The small but significant association between self-assessed ECE quality and children's development outcomes appeared similar for different developmental domains, for first and second-generation self-assessment instruments, and for published and unpublished studies.

Limitations of This Study

This systematic review of self-assessment measures in ECE has several limitations. First, we narrowed our scope to center-based childcare and excluded home-based childcare. Therefore, we cannot generalize our findings to self-assessment in home-based care. We also found relatively few publications meeting our inclusion criteria regarding the reliability of self-assessment instruments in ECE. Third, the statistical power is limited due to the small number of studies, which affected the moderator analysis of our review (Huedo-Medina et al., 2006). The body of research concerning validity of self-assessment measures in ECE is still relatively small despite the recent growth of publications and more studies would allow a more in-depth analysis for the exploration of variation among studies. Fourth, we noticed that often ECE quality was not self-assessed at the same time as the measurement with validated measures, which may suppress convergence of ratings. In order to create a representative

comparison between self-assessment and ratings with validated measures the procedures during the day (e.g., number of children and caregivers present, time of day, etc.) should be as similar as possible. In addition, the validity of self-assessment measures in ECE has mainly been examined in the United States, and specifically predictive validity has only been examined in the United States. Hence, we cannot generalize the findings related to predictive validity to other countries. Further, an important limitation is that there is potential bias of self-assessment, despite the reliability and validity of the measures. Particularly in a high stakes' context, in which incentives are involved, a personal interest of raters could affect their objectivity. We therefore recommend that external ratings should be used in high-stakes self-assessment. Self-reported ECE quality may over-report quality, but the reverse may also be true. The fact is that providers may under-report quality, when they are unmotivated to participate in a QRIS and when some items require a lot of effort from the staff, as Esplin and colleagues (2019) showed. Finally, self-assessment measures assess similar but not identical constructs compared to validated measures and there is thus room for improvement.

Implications for Research and Practice

Results of this systematic review support, at meta-analytic level, the construct validity of self-assessment in ECE. Although there is empirical evidence for the validity of self-assessment, there is a need for further study. In addition, more research on the potential bias of self-assessments is needed, distinguishing between high-stake and regular contexts. A common view on self-assessment is that it is not meant to "collect data related to quality" but rather to enable ECE programs to be aware of their own strengths and weaknesses, and to identify their needs for improvement (Organisation for Economic Co-operation and Development, 2013, 2021; Picchio et al., 2012; Quality Compendium, 2022a). Validation of self-assessment instruments potentially adds to the use of self-assessment for ECE quality measurement. In addition, more research is needed regarding convergent and predictive validity of self-assessment measures in ECE, specifically in countries other than the United States. However, considering the modest predictive validity of validated quality measures (Perlman et al., 2016; Ulferts et al., 2019), it is important to have realistic expectations of the predictive validity of self-assessment instruments. Furthermore, we encourage more research concerning the reliability of self-assessment instruments in ECE.

Self-assessment measures in ECE have been described as elaborate, complex, and labor-intensive standards. It is therefore conceivable that systems for monitoring quality may not be sustainable for childcare providers (Elicker & Ruprecht, 2019). Esplin et al. (2019), for example, found that ECE providers who use self-assessment may restrict their focus on quality characteristics that require relatively little documentation, which may suppress variation in ECE ratings. Other studies suggest to adapt QRISs by focusing on teacher–child interactions and the promotion of equity (Language ENvironment Analysis, 2022; Meek et al., 2020; Policy Equity Group, 2022; Swartz, 2022). This is in line with the emergence of second-generation QRIS, which focuses explicitly on quality improvement (Cannon et al., 2017). Therefore, we recommend more research on the practical feasibility of self-assessment measures for ECE providers, and adjustment of the content of self-assessment instruments in line with research findings, in order to ensure that these measures remain of practical value. Specifically, a relevant question for the near future is whether self-assessment measures with a focused approach may offer a logistic advantage without affecting their reliability and validity.

In addition, more research is needed to reflect on the content and construct validity of current self-assessment measures in ECE. The recent history of quality rating in ECE reveals different concepts of ECE quality and the focus on quality has shifted somewhat over time. Whereas the initial emphasis was on process quality, we see that in recent years the focus shifts toward individual child experiences. Self-assessment, in this case QRIS, needs to focus on teacher–child interactions (Language ENvironment Analysis, 2022; Meek et al., 2020; Policy Equity Group, 2022; Swartz, 2022). Although ECE quality has been defined extensively by various stakeholders, we recommend a broad, constructive discussion between researchers, practitioners, and policy makers

regarding the content of ECE quality to evaluate the definition of ECE quality and to explore possibilities to arrive at new, shared definitions of ECE quality and their operationalization in practice. Also ECE providers can initiate this dialog on the content of self-assessment measures with researchers, parents, policymakers, and ECE inspection. Finally, research on self-assessment in the context of family childcare and afterschool care is needed in order to know whether our findings also apply to these contexts.

Conclusion

This systematic review provides empirical evidence for the validity of current self-assessment measures for ECE quality. Validated self-assessment measurements can be a valuable addition to other quality ratings conducted with validated measures, and they could help breach the gap between research and practice. Providing professionals with valid self-assessment quality rating may increase awareness of their strengths and weaknesses, which is a first step in identifying specific areas of improvement of ECE quality (Isoré, 2009). By providing ECE practice with validated self-assessment instruments, it becomes possible for childcare providers to bring their own perspective into the ongoing discussion on ECE quality. This perspective may stimulate childcare providers to take more responsibility for the quality of the services they provide, and may facilitate the ongoing dialog about ECE quality with researchers, parents, policymakers and ECE inspection.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Simon M. Hay http://orcid.org/0000-0003-3013-1741
Harriet J. Vermeer http://orcid.org/0000-0001-7363-1792
Ruben G. Fukkink http://orcid.org/0000-0001-6212-9553
Harold T. Nefs http://orcid.org/0000-0002-4943-7938

References

References marked with an * were included in the meta-analyses

Australian Children's Education & Care Quality Authority. (2019). *Self-assessment tool. The most effective improvements to service delivery are initiated from within the service*. https://www.acecqa.gov.au/assessment/materials

Boogaard, M., Van Daalen-Kapteins, M., & Gevers Deynoot-Schaub, M. (2011). *De Kwaliteitsmonitor BSO* [The BSOQuality monitor]. SWP.

Bowman, N. A. (2012). Effect sizes and statistical methods for meta-analysis in higher education. *Research in Higher Education*, *53*(3), 375–382. https://doi.org/10.1007/s11162-011-9232-5

Burchinal, M., Tarullo, L., & Zaslow, M. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quaterly*, *25*(2), 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the abecedarian project. *Applied Developmental Science*, *6*(1), 42–57. https://doi.org/10.1207/S1532480XADS0601_05

Cannon, J. S., Zellman, G. L., Karoly, L. A., & Schwartz, H. L. (2017). *Quality rating and improvement systems for early care and education programs: Making the second generation better*. RAND Corporation. https://www.rand.org/pubs/perspectives/PE235.html

Center for Early Childhood Professional Development. (2020). *Self-assessment overview*. CECPD. https://cecpd.org/Portals/1400/Assets/Documents/Self-Assessment%20Overview%203-8-22.pdf?ver=-Rt6eyeE9t0JfOCgzG9pVQ%3D%3D

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Lawrence Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds). (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation.

Cryer, D., Harms, T., & Riley, C. (2003). *All about the ECERS-R; a detailed guide in words & pictures to be used with the ECERS-R*. PACT House Publishing.
Cryer, D., Harms, T., & Riley, C. (2004). *All about the ITERS-R; a detailed guide in words and pictures to be used with the ITERS-R*. PACT House Publishing.
Cryer, D., Tietze, W., Burchinal, M., Leal, T., & Palacios, J. (1999). Predicting process quality from structural quality in preschool programs: A cross-country comparison. *Early Childhood Research Quarterly*, *14*(3), 339–361. https://doi.org/10.1016/S0885-2006(99)00017-4
Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics*, *1*(3), 111–134. https://doi.org/10.1257/app.1.3.111
*Elicker, J., Gold, Z. S., Mishra, A. A., & Lane, S. F. (2022). Toddlers' developmental trajectories as a function of QRIS rated child care quality. *Child and Youth Care Forum*, *51*, 633–660. https://doi.org/10.1007/s10566-021-09643-z
Elicker, J., Langill, C. C., Ruprecht, K., & Kwon, K.-A. (2007). *Paths to Quality: A child care quality rating system for Indiana: What is its scientific basis?* (Report #1). Center for Families Publications. https://doi.org/10.13140/2.1.4310.1765
Elicker, J., & Ruprecht, K. (2019). Child care quality rating and improvement systems (QRIS): A national experiment in improving quality. In C. P. Brown, M. B. McMullen, & N. K. File (Eds.), *Handbook of early childhood care and education* (pp. 515–536). Wiley.
Esplin, J. A., Neilson, B. G., Berghout Austin, A. M., & Fronk, A. (2019). Self-report QRIS: Challenges with validation. *Early Education and Development*, *30*(6), 817–834. https://doi.org/10.1080/10409289.2019.1591045
Farrell, A. F., Collier-Meek, M. A., & Furman, M. J. (2019). Supporting out-of-school time staff in low resource communities: A professional development approach. *American Journal of Community Psychology*, *63*(3–4), 378–390. https://doi.org/10.1002/ajcp.12330
Garca, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2020). Quantifying the life-cycle benefits of an influential early-childhood program. *Journal of Political Economy*, *128*(7), 2502–2540. https://doi.org/10.1086/705718
Garcia-Sanchez, C., Guzman, M., Cordero, M., Smith-Sanchez, S., & Woods, T. (2015). *Self-assessment 2015 guidance & checklist for governing boards & program administrators*. NYC Administration for Children's Services EarlyLearnNYC. https://www.nyc.gov/assets/acs/pdf/self_assessment/SELF_ASSESSMENT_GUIDANCE_AND_CHECKLIST_2015_FINAL_clean.pdf
Gevers Deynoot-Schaub, M., Fukkink, R., Riksen-Walraven, M., De Kruijf, R., Helmerhorst, K., & Tavecchio, L. (2009). *De NCKO Kwaliteitsmonitor* [The NCKOQualitymonitor]. SWP.
Goffin, S. G., & Barnett, W. S. (2015). Assessing QRIS as a change agent. *Early Childhood Research Quarterly*, *30*, 179–182. https://doi.org/10.1016/j.ecresq.2014.08.005
Halpern, R. (2000). The promise of after-school programs for low-income children. *Early Childhood Research Quarterly*, *15*(2), 185–214. https://doi.org/10.1016/S0885-2006(00)00056-9
Hattie, J., & Cooksey, R. (1984). Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement*, *8*(3), 295–305. https://doi.org/10.1177/014662168400800306
Heckman, J. (n.d.). *ABC/CARE: Elements of quality early childhood programs that produce quality outcomes*. Retrieved August 21, 2023, from https://heckmanequation.org/wp-content/uploads/2018/06/F_Heckman_ABC_CARE_Quality_110117.pdf
Herbst, C. (2018). The impact of quality rating and improvement systems on families' child care choices and the supply of child care labor. *Labour Economics*, *54*, 172–190. https://doi.org/10.1016/j.labeco.2018.08.007
*Hestenes, L. L., Kintner-Duffy, V., Wang, Y. C., La Paro, K., Mims, S. U., Crosby, D., Scott-Little, C., & Cassidy, D. J. (2015). Comparisons among quality measures in child care settings: Understanding the use of multiple measures in North Carolina's QRIS and their links to social-emotional development in preschool children. *Early Childhood Research Quarterly*, *30*, 199–214. https://doi.org/10.1016/j.ecresq.2014.06.003
Howard, S. J., Lewis, K. L., Walter, E., Verenikina, I., & Kervin, L. K. (2024). Measuring the quality of adult–child interactions in the context of ECEC: A systematic review on the relationship with developmental and educational outcomes. *Educational Psychology Review*, *36*(6), 1–50. https://doi.org/10.1007/s10648-023-09832-3
Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*(1), 27–50. https://doi.org/10.1016/j.ecresq.2007.05.002
Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 Index? *Psychological Methods*, *11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193
Hu, B. Y., & Li, K. (2012). The quality rating system of Chinese preschool education: Prospects and challenges. *Childhood Education*, *88*(1), 14–22. https://doi.org/10.1080/00094056.2012.643711
*Hu, B. Y., Vong, K. I., & Chi Kuan Mak, M. (2015). An evaluation of China's kindergarten quality rating system through the Chinese early childhood environment rating scale – the Zhejiang case. *Early Years*, *35*(1), 50–66. https://doi.org/10.1080/09575146.2014.978266
Isoré, M. (2009). *Teacher evaluation: Current practices in OECD countries and a literature review. OECD education working Paper No. 23*. https://www.lena.org/qris-and-interaction/

Janssen, J., Declercq, B., Daems, M., Hulpia, H., Van Cleynenbreugel, C., Laevers, F., & Vandenbroeck, M. (2016). *MeMoQ Sub-report 6. Methodologie van het zelfevaluatie-instrument* [Methodology of the self-evaluation instrument]. Kind & Gezin. https://biblio.ugent.be/publication/8552332

*Jeon, L., & Buettner, C. K. (2015). Quality rating and improvement systems and children's cognitive development. *Child and Youth Care Forum*, *44*(2), 191–207. https://doi.org/10.1007/s10566-014-9277-7

*Jeon, L., Buettner, C. K., & Hur, E. (2014). Examining pre-school classroom quality in a statewide quality rating and improvement system. *Child and Youth Care Forum*, *43*(4), 469–487. https://doi.org/10.1007/s10566-014-9248-z

Karoly, L. A. (2014). *Validation studies for early learning and care quality rating and improvement systems: A review of the literature*. RAND Corporation. https://www.rand.org/pubs/working_papers/WR1051.html

*Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly*, *30*, 280–290. https://doi.org/10.1016/j.ecresq.2014.04.005

*Lahti, M., Sabol, T., Starr, R., Langill, C., & Tout, K. (2013). *Validation of quality rating and improvement systems (QRIS): Examples from four States. Research-to-policy, research-to-practice brief*. Office of Planning, Research and Evaluation. https://www.acf.hhs.gov/sites/default/files/documents/opre/qris_brief_validation_in_4_states.pdf

Language ENvironment Analysis. (2022, August 10). *QRIS and interaction: Do quality rating systems reflect individual experience?* https://www.lena.org/qris-and-interaction/

*Le, V. N., Schaack, D. D., & Setodji, C. M. (2015). Identifying baseline and ceiling thresholds within the qualistar early learning quality rating and improvement system. *Early Childhood Research Quarterly*, *30*, 215–226. https://doi.org/10.1016/j.ecresq.2014.03.003

Li, K., Pan, Y., Hu, B., Burchinal, M., De Marco, A., Fan, X., & Qin, J. (2016). Early childhood education quality and child outcomes in China: Evidence from Zhejiang province. *Early Childhood Research Quarterly*, *36*, 427–438. https://doi.org/10.1016/j.ecresq.2016.01.009

*Lipscomb, S. T., Weber, R. B., Green, B. L., & Patterson, L. B. (2017). *Oregon's quality rating improvement system (QRIS) validation study one: Associations with observed program quality*. Oregon State University. https://health.oregonstate.edu/sites/health.oregonstate.edu/files/occrp/pdf/oregons_qris_validation_study1report_withappendices_jan2017.pdf

*Magnuson, K., & Lin, Y. (2015). *Validation of the QRIS YoungStar rating scale. Report 1: Wisconsin early child care study findings on the validity of youngstar rating for observed classroom quality*. MSW UW–Madison. https://dcf.wisconsin.gov/files/youngstar/pdf/validationreport1.pdf

*Magnuson, K., & Lin, Y. (2016). *Validation of the QRIS YoungStar rating scale. Report 2: Wisconsin early child care study findings on the validity of youngstar rating for children's school readiness*. MSW UW–Madison. https://dcf.wisconsin.gov/files/youngstar/pdf/validationreport2.pdf

*Markowitz, A. J., Bassok, D., & Player, D. (2020). Simplifying quality rating systems in early childhood education. *Children and Youth Services Review*, *112*, 1–14. https://doi.org/10.1016/j.childyouth.2020.104947

Meek, S., Iruka, I. U., Allen, R., Yazzie, D., Fernandez, V., & Catherine, E., McIntosh, K., Gordon, L., Gilliam, W., Hemmeter, M. L., & Powell, T. (2020). Fourteen priorities to dismantle systematic racism in early care and education. *The Children's Equity Project*. https://childandfamilysuccess.asu.edu/cep

Munton, A. G., & Mooney, A. (1999). Improving the quality of childcare provision through self-assessment: Organizational characteristics of nurseries as predictors of the ability to plan and implement change. *Early Child Development and Care*, *148*(1), 21–34. https://doi.org/10.1080/0300443991480103

Munton, A. G., Mooney, A., & Rowland, L. (1997). Quality in group and family day care provision: Evaluating self-assessment as an agent of change. *European Early Childhood Education Research Journal*, *5*(1), 59–75. https://doi.org/10.1080/13502939785208001

National Association for the Education of Young Children. (2010). *QRIS and NAEYC accreditation fact sheet*. NAEYC. https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/our-work/public-policy-advocacy/QRSFactSheet.pdf

National Association for the Education of Young Children. (2011). *NAEYC position on quality rating and improvement systems*. NAEYC. https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/our-work/public-policy-advocacy/2011_QRIS_Statement_0.pdf

National Association for the Education of Young Children. (2022). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8* (4th ed.). NAEYC.

National Heart Lung and Blood Institute. (2013). *Quality assessment tool for observational cohort and cross-sectional studies*. https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools

National Institute of Child Health and Human Development Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development*, *71*(4), 960–980. https://doi.org/10.1111/1467-8624.00202

Norris, D. J., Dunn, L., & Eckert, L. (2003). *"Reaching for the stars" center validation study final report*. Early Childhood Collaborative of Oklahoma. https://oklahoma.gov/content/dam/ok/en/okdhs/documents/okdhs-report-library/reachingforthestarscentervalidationstudyfinalreport_dcc_05212007.pdf

NSW Department of Education. (2020). *Self-assessment - working document*. https://education.nsw.gov.au/content/dam/main-education/early-childhood-education/operating-an-early-childhood-education-service/media/documents/regulation-assessment-rating/self-assessment-working-document.pdf

Organisation for Economic Co-operation and Development. (2013). *Literature review on monitoring quality in early childhood education and care (ECEC)*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2021). *Starting strong VI: Supporting meaningful interactions in early childhood education and care, starting strong*. OECD Publishing.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., & Moher, D. (2021). The PRISMA 2020 checklist. *BMJ*, *71*, 372.

*Pan, Y. J., Liu, Y., & Lau, E. Y. H. (2010). Evaluation of the kindergarten quality rating system in Beijing. *Early Education and Development*, *21*(2), 186–204. https://doi.org/10.1080/10409281003631134

Peisner-Feinberg, E. S., & Burchinal, M. R. (1999). *The Children of the cost, quality, and outcomes study go to school. Technical report* (Vol. 449). University of North Carolina. https://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/NCEDL_CQO_technical_report.pdf

*Perlman, M., Brunsek, A., Hepditch, A., Gray, K., & Falenchuck, O. (2017). Instrument development and validation of the infant and toddler assessment for quality improvement. *Early Education and Development*, *28*(1), 115–133. https://doi.org/10.1080/10409289.2016.1186468

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., Shah, P. S., & van Wouwe, J. P. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the classroom assessment scoring system) in early childhood education and care settings and child outcomes. *PLOS ONE*, *11*(12), e0167660. https://doi.org/10.1371/journal.pone.0167660

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175–181. https://doi.org/10.1037/0021-9010.90.1.175

Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly*, *12*(3), 281–303. https://doi.org/10.1016/S0885-2006(97)90004-1

Phillips, D., Mekos, D., Scarr, S., McCartney, K., & Abbott-Shim, M. (2000). Within and beyond the classroom door: Assessing quality in child care centers. *Early Childhood Research Quarterly*, *15*(4), 475–496. https://doi.org/10.1016/S0885-2006(01)00077-1

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science*, *9*(3), 144–159. https://doi.org/10.1207/s1532480xads0903_2

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431–451. https://doi.org/10.1016/j.ecresq.2008.02.001

Picchio, M., Giovannini, D., Mayer, S., & Musatti, T. (2012). Documentation and analysis of children's experience: An ongoing collegial activity for early childhood professionals. *Early Years*, *32*(2), 159–170. https://doi.org/10.1080/09575146.2011.651444

Policy Equity Group. (2022). *QRIS: Knocking it down & rebuilding early childhood: Are you ready?* https://policyequity.com/qris-building-the-case-for-knocking-it-down/

Quality Compendium. (2022a). *A catalog and comparison of quality improvement systems (QIS)*. https://qualitycompendium.org/

Quality Compendium. (2022b). *Profile report keystone stars*. https://www.pakeys.org/keystone-stars/

Quick, H. E., Hawkinson, L. E., Holod, A., Anthony, J., Muenchow, S., & Parrish, D., Martin, A., Weinberg, E., Hoon Lee, D., Cannon, J., & Lee, D. H. (2016). *Independent evaluation of California's race to the top-early learning challenge quality rating and improvement system: Cumulative technical report*. RAND Corporation. https://www.air.org/sites/default/files/downloads/report/RTT-ELC%20QRIS%20Cumulative%20Technical%20Report%20-%20FINAL.pdf

*Rentzou, K. (2010). Using the ACEI global guidelines assessment to evaluate the quality of early child care in Greek settings. *Early Childhood Education Journal*, *38*(1), 75–80. https://doi.org/10.1007/s10643-010-0382-4

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160. https://doi.org/10.1037/met0000300

Ruffin, N. J. (2009). *Understanding growth and development patterns of infants*. Virginia Tech. https://www.pubs.ext.vt.edu/content/dam/pubs_ext_vt_edu/350/350-055/350-055(FCS-132P).pdf

*Sabol, T. J., & Pianta, R. C. (2015). Validating virginia's quality rating and improvement system among state-funded pre-kindergarten programs. *Early Childhood Research Quarterly*, *30*, 183–198. https://doi.org/10.1016/j.ecresq.2014.03.004

*Sabol, T. J., Soliday Hong, S. L., Pianta, R., & Burchinal, M. (2013). Can rating of pre-K programs predict children's learning? *Science*, *341*(6148), 845–846. https://doi.org/10.1126/science.1233517

Schilder, D. (2013). *Quality rating and improvement system (QRIS) validation study designs*. Center on Enhancing Early Learning Outcomes. https://researchconnections.org/childcare/resources/26184

*Schmitt, S. A., Elicker, J. A., Purpura, D. J., Duncan, R. J., Schmerold, K. L., Budrevich, A., Bryant, L. M., & Finders, J. K. (2023). The effects of a high quality state-run preschool program as rated by a quality rating and improvement system on children's school readiness. *Early Childhood Research Quarterly*, *62*, 89–101. https://doi.org/10.1016/j.ecresq.2022.07.013

Smith, M. W., Brady, J. P., & Anastasopoulos, L. M. P. P. (2008). *User's guide to the early language & literacy classroom observation pre-K tool, pre-K (ELLCO pre-K)*. Paul H. Brookes Publishing Co.

*Soliday Hong, S. L., Howes, C., Marcella, J., Zucker, E., & Huang, Y. (2015). Quality rating and improvement systems: Validation of a local implementation in LA county and children's school-readiness. *Early Childhood Research Quarterly*, *30*(PB), 227–240. https://doi.org/10.1016/j.ecresq.2014.05.001

Squibb, K. M. (2013). *Quality profiles in early childhood: An example from virginia's quality rating improvement system*. Old Dominion University. https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1011&context=earlychildhood_etds

*Sussman, J., Melnick, H., Newton, E., Kriener-Althen, K., Draney, K., Mangione, P., & Gochyyev, P. (2022). *Preschool quality and child development: How are learning gains related to program ratings?* WestEd & BEAR Center. https://learningpolicyinstitute.org/media/3742/download?inline&file=ECE_DRDP_Quality_REPORT.pdf

Swartz, M. (2022, June 3). *The quality trap: The movement to reconsider QRIS*. https://earlylearningnation.com/2022/06/the-quality-trap-the-movement-to-reconsider-qris/

Swenson-Klatt, D., & Tout, K. (2011). Measuring and rating quality. In M. Zaslow, I. Martinex-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 347–362). Paul H. Brookes Publishing Co.

Thomason, A. C., & La Paro, K. M. (2009). Measuring the quality of teacher-child interactions in toddler child care. *Early Education and Development*, *20*(2), 285–304. https://doi.org/10.1080/10409280902773351

*Thornburg, K. R., Mayfield, W. A., Hawks, J. S., & Fuger, K. L. (2009). *The missouri quality rating system school readiness study*. Center for Family Policy & Research University of Missouri. https://www.academia.edu/71221146/The_Missouri_Quality_Rating_System_School_Readiness_Study

Tout, K. (2013). Look to the stars: Future directions for the evaluation of quality rating and improvement systems. *Early Education and Development*, *24*(1), 71–78. https://doi.org/10.1080/10409289.2013.741912

Tout, K., Magnuson, K., Limpscomb, S., Karoly, L., Starr, R., Quick, H., Early, D., Epstein, D., Joseph, G., Maxwell, K., & Wenner, J. (2017). *Validation of the quality ratings used iN quality rating and improvement systems (QRIS): A synthesis of state studies. OPRE Report #2017-92*. Office of Planning, Research and Evaluation.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of quality rating and improvement systems. Issue brief #2009-14*. Office of Planning, Research and Evaluation.

Ulferts, H., Wolf, K. M., & Anders, Y. (2019). Impact of process quality in early childhood education and care on academic outcomes: Longitudinal meta-analysis. *Child Development*, *90*(5), 1–17. https://doi.org/10.1111/cdev.13296

Vandell, D. L. (2004). Early child care: The known and the unknown. *Merrill-Palmer Quarterly*, *50*(3), 387–414. https://doi.org/10.1353/mpq.2004.0027

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

*Vitiello, V. E., Bassok, D., Hamre, B. K., Player, D., & Williford, A. P. (2018). Measuring the quality of teacher–child interactions at scale: Comparing research-based and state observation approaches. *Early Childhood Research Quarterly*, *44*, 161–169. https://doi.org/10.1016/j.ecresq.2018.03.003

*Wong, M. N. C., & Li, H. (2010). From external inspection to self-evaluation: A study of quality assurance in Hong Kong kindergartens. *Early Education and Development*, *21*(2), 205–233. https://doi.org/10.1080/10409281003638725

*Yazejian, N., Burchinal, M., Soliday Hong, S., & Howard, E. (2017). *ExceleRate Illinois validation study report*. Frank Porter Graham Child Development Institute. https://buildinitiative.org/wp-content/uploads/2021/06/ExceleRate-Illinois_Validation-Study-and-Child-Outcomes-Report_12-31-17.pdf

Zellman, G. L., & Fiene, R. (2012). *Validation of quality rating and improvement systems for early care and education and school-age care, research-to-policy, research-to-practice brief OPRE 2012-29*. Office of Planning, Research and Evaluation. https://files.eric.ed.gov/fulltext/ED534457.pdf

Zellman, G. L., Perlman, M., Le, V.-N., & Setodji, C. M. (2008). *Assessing the validity of the qualistar early learning quality rating and improvement system as a tool for improving child-care quality*. RAND Corporation. https://www.rand.org/pubs/monographs/MG650.html