

Quality of Child Care Using the Environment Rating Scales: A Meta-Analysis of International Studies

Harriet J. Vermeer¹ · Marinus H. van IJzendoorn¹ ·
Rodrigo A. Cárcamo^{1,2} · Linda J. Harrison³

Published online: 8 January 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract The current study provides a systematic examination of child care quality around the globe, using the Environment Rating Scales (ERS). Additional goals of this study are to examine associations between ERS process quality and structural features (group size, caregiver–child ratio) that underpin quality and between ERS and more proximal aspects of child care quality (caregiver sensitivity). Furthermore, we consider possible differences in ERS associations arising from scale characteristics (infant vs early childhood version, original version vs revised scale, full version vs shortened version). The reported meta-analysis combines results of ERS child care quality reported in 72 studies from 23 countries across five international geographic regions. Group center care appeared to be of average quality with higher quality levels in Australia/New Zealand and North America. Our results suggest that: (1) ERS characteristics are not associated with differences in ERS scores and (2) ERS scores are related to indicators of proximal quality of care (caregiver sensitivity) and, to a lesser degree, structural quality of care (caregiver–child ratio). The meta-analysis provided cross-cultural comparisons on child care quality on a common instrument as a means to advance discussion on child care quality internationally.

✉ Harriet J. Vermeer
vermeer@fsw.leidenuniv.nl; vermeer@leidenuniv.nl

Marinus H. van IJzendoorn
vanijzen@fsw.leidenuniv.nl

Rodrigo A. Cárcamo
rodrigo.carcamo@umag.cl

Linda J. Harrison
lharrison@csu.edu.au

¹ Centre for Child and Family Studies, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands

² Department of Psychology, University of Magallanes, Avenida Bulnes 01855, P.O. Box 113-D, Punta Arenas, Chile

³ School of Teacher Education, Charles Sturt University, Bathurst, NSW 2795, Australia

Keywords Child care quality · Center care · International · Environment Rating Scales

Résumé Cette étude fournit un examen systématique de la qualité des services de garde à l'enfance à travers le monde, en utilisant les échelles d'évaluation de l'environnement (ERS). D'autres objectifs de cette étude sont d'examiner les associations entre le processus de la qualité et les caractéristiques structurelles des ERS (taille du groupe, ratio éducateur-enfant) qui étayent la qualité, et entre les ERS et les aspects plus proximaux de la qualité de la garde des enfants (sensibilité de l'éducateur). De plus, nous regardons les différences possibles dans les associations des ERS découlant des caractéristiques des échelles (la version nourrissons et tout-petits contre la version préscolaire, la version originale contre la version révisée, la version complète contre la version abrégée). La méta analyse rapportée combine les résultats de qualité de la garde d'enfants aux ERS provenant de 72 études dans 23 pays de cinq régions géographiques internationales. La garde de groupe en centre s'est révélée de qualité moyenne avec des niveaux de qualité plus élevés en Australie/Nouvelle-Zélande et Amérique du Nord. Nos résultats suggèrent (1) que les caractéristiques d'échelle des ERS ne sont pas associées aux différences de scores aux ERS, et (2) que les scores aux ERS sont liés aux indicateurs de qualité proximale (sensibilité de l'éducateur) et, à un degré moindre, à la qualité structurelle (ratio éducateur-enfant). La méta analyse donne des comparaisons interculturelles de la qualité des services de garde à un instrument commun comme moyen de faire avancer la discussion sur la qualité des services de garde au niveau international.

Resumen El presente estudio provee una revisión sistemática de la calidad del cuidado infantil en guarderías alrededor del mundo, usando las Escalas de Calificación del Ambiente (ERS). Los objetivos adicionales de este estudio son examinar las asociaciones entre la calidad de procesos y características estructurales (tamaño del grupo, proporción cuidadores - niños) que sustentan la calidad, y entre ERS y aspectos más próximos de la calidad del cuidado infantil (sensibilidad del cuidador). Además consideramos posibles diferencias en la asociación en ERS surgidas desde las características de las escalas (infantes versus niñez temprana, versión original versus escala revisada, versión completa versus versión abreviada). El meta-análisis reportado combina resultados de la calidad del cuidado de guarderías ERS de 72 estudios reportados en 23 países de cinco regiones geográficas internacionales. El cuidado grupal en estos centros apareció estar en calidad promedio, con altos niveles de calidad en Australia/Nueva Zelanda y Norte América. Nuestros resultados sugieren que: (1) Las características de la escala ERS no están asociadas con diferencias en los puntajes ERS y (2) Los puntajes de ERS están asociados a indicadores proximales de la calidad del cuidado (sensibilidad del cuidador) y, en menor grado, a la calidad estructural del cuidado (proporción niños-cuidadoras). El meta-análisis provee comparaciones transculturales sobre la calidad del cuidado en guarderías con un instrumento común, como una manera de avanzar en la discusión sobre la calidad del cuidado de niños a nivel internacional.

Introduction

Worldwide, children spend a large part of their early childhoods in some form of out-of-home child care (UNICEF 2008). A large body of research has shown that high-quality child care has advantages for child development (e.g., Belsky et al. 2007). Child care quality is operationalized in many ways, but generally refers to the broad range of environmental features and interactions in non-parental care and education settings that have been linked to children's development (Zaslow et al. 2011). Internationally, there is substantial agreement about what is considered essential in providing for children's developmental needs in center-based child care (e.g., Lamb and Ahnert 2006). Core elements that are universally recognized as required for children's positive development are: safe and healthy care settings, developmentally appropriate stimulation and opportunities for learning, positive interactions with adults, and the promotion of individual emotional growth and positive relationships with other children (Cryer et al. 2002). Despite this consensus regarding the components that contribute to child care quality, little is known about levels of child care quality across different countries. By covering a wide range of regulatory systems, a cross-country comparison of quality may provide more insight into possible determinants of high-quality care. The current meta-analysis focuses on the most widely used instruments to measure quality: the Early Childhood Environment Rating Scale (ECERS; Harms et al. 1980), the Infant/Toddler Environment Rating Scale (ITERS; Harms et al. 1990), and its revisions (ECERS-R; Harms et al. 1998 and ITERS-R; Harms et al. 2003). Whilst these measures have been used extensively to compare levels of center-based child care quality, or to assess change over time within countries, this study is the first systematic examination of quality assessments around the globe.

The Environment Rating Scales (ERS) were developed to evaluate process quality in early child care settings. Process quality refers to the experiences of children within the child care environment including their interactions with others, materials, and activities (Phillipsen et al. 1997). Process quality is assessed primarily through observation and has been found to be more predictive of child outcomes than structural indicators of quality such as staff-to-child ratio, group size, cost of care, and type of care (Whitebook et al. 1989).

The ERS have a long history of use worldwide; the first use of the original ECERS in the USA dates back more than 30 years ago (Harms and Clifford 1983). Outside the USA, the first studies using the ERS were published almost 20 years ago (e.g., Kärrby and Giota 1994). The long-term international use of these scales, which in itself is suggestive of cross-cultural validity or at least feasibility, provides an excellent opportunity for an international comparison of child care quality. Despite the extensive use of the ERS, however, psychometric research is restricted. Gordon et al. (2013, 2015) recently tested the scales' criterion validity in a US sample and found little evidence with respect to child developmental outcomes and moderate evidence of validity with alternative observational measures of quality. In the present paper, we add to this work by examining associations between ERS process quality and observed proximal aspects of child care quality (caregiver sensitivity) in a wide range of international studies. We also consider possible

differences in ERS associations arising from scale characteristics (infant vs early childhood version, original vs revised scale, full vs shortened version). A further goal is to examine associations between ERS and structural features (group size, caregiver–child ratio). In the following sections, these goals are further explained.

Process Quality and Structural Features

Government regulation and other forms of quality assurance within countries impact a range of structural features that have been shown to affect process measures of quality (NICHD ECCRN 2002). There are major differences across countries in policy focus, regulations, and subsidies with respect to child care, which are likely to have an effect on levels of ERS quality. The diversity is evident in the UNICEF Innocenti Research Center (2008) *Report Card 8* that compared 25 OECD countries on ten benchmarks, representing minimum standards for protecting the rights of children (e.g., regulation and accreditation of early childhood services, training of early childhood staff, minimum staff-to-child ratios). This comparison, for example, showed that in 2004 not even half of the countries met the minimum staff-to-children ratio of 1:15 for preschool education.

Two structural features, group size and child–caregiver ratio, are easily quantifiable and are operationalized in the same way across studies, thereby allowing comparisons across countries. In within-country studies, the range of these structural indicators may be limited because of government legislations, and therefore across-country studies are expected to include a wider range of group size and child–caregiver ratios. Previous cross-country studies that encompassed a wider range of quality than typically seen in one country (Love et al. 2003) have demonstrated that this approach can uncover stronger associations than is possible in within-country studies. Drawing on earlier work (Phillipsen et al. 1997), our hypothesis is that process quality will be negatively associated with group size and child–caregiver ratio; that is, it will be higher when group sizes are smaller and/or when fewer children are cared for by one caregiver.

Process Quality and Caregiver Sensitivity

The ERS are usually considered instruments for measuring process quality, targeting children’s interactions with caregivers and peers and their participation in different activities (Vandell and Wolfe 2000). In the ERS, features of the physical environment (personal care, space, furniture, and physical safety) have been emphasized, leading to criticism because of a perceived underrepresentation of caregiver–child interactions (e.g., Cassidy et al. 2005a). Because of this, researchers have often used additional instruments beyond the ERS for capturing caregiver–child interactions, such as the Arnett Caregiver Interaction Scale (CIS; Arnett 1989) or, more recently, the Classroom Assessment Scoring System Pre-K (CLASS pre-K; Pianta et al. 2008). The inclusion of the CIS measure in this meta-analysis not only provides a cross-cultural comparison of caregiver sensitivity in group settings around the world, but also contributes to the evidence for the validity of the ERS as

a measure of process quality. We hypothesize that ERS scores will be positively associated with caregiver sensitivity as assessed by alternative measures.

Process Quality and Scale Characteristics

Because different variants of the ERS have been widely used in studies with both comparative and longitudinal designs, we investigate whether different ERS yield different results. For instance, in longitudinal studies where children make the transition from infant to toddler to preschool groups, different types of ERS must be used; or, after the publication of revised editions, researchers need to decide whether they will stick to the original version or change to the new version of the scale. In this study, we distinguish between type of scale (ITERS vs ECERS), original versus revised scales (ITERS vs ITERS-R; ECERS vs ECERS-R), and full versus shortened versions of the ERS.

Although the revised versions of the ERS share the same rationale and underlying constructs as the original scales, they slightly differ from the original versions. Items have been added (e.g., cultural diversity, inclusion of children with disabilities), the scoring of the items is somewhat different than in the original scales, and more information with respect to scoring has become available (i.e., via a Web site and through handbooks). Although a previous study by Sakai et al. (2003) showed that the ECERS scores and ECERS-R scores were comparable for the same sample, researchers in Germany (Tietze et al. 2001) have reported half a scale point difference in quality estimates in favor of the original ECERS. Because these findings diverge and are based on a limited number of studies, it is currently unclear whether observed differences in quality, as measured with the original and the revised versions, represent measurement differences or real differences. The current meta-analytic study examines possible differences in mean scores of the ERS in its original versus revised version. In addition, we test whether quality ratings have decreased or increased across time.

Shortened versions of the ERS are used for different reasons. First, many researchers do not include the scores of the subscale Parents and Staff (or “Adult needs” dependent on the type of scale) when calculating a mean score, because the items from this subscale do not reflect the children’s everyday experiences and thus do not cover process quality. Second, in some countries, concerns have been raised about the applicability of particular items of the scales, and one response of researchers is to delete inappropriate items. In Sweden for instance, some items are excluded because they do not reflect the Swedish preschool care practice (Kärrby and Giota 1994). Third, researchers may decide to limit the number of items to reduce the time-consuming training and administration of the full ERS. Previous analyses of US studies (e.g., Cassidy et al. 2005b; Perlman et al. 2004; Scarr et al. 1994) have shown that a shorter version of the ECERS-R is a good proxy for scores on the full scale. The current study examines whether shortened versions of the ERS yield the same results as the full scale in a larger international sample of studies.

Research Aims

In summary, this study serves three goals. First, we provide an international perspective on child care quality by reviewing a wide range of studies in which the ERS were used. Geographic region was included as a moderator in the analyses because it was assumed that there would be some cultural similarities by regional area. Second, we examine whether assessments with the ERS are influenced by scale characteristics. Third, we provide insight into the associations between ERS process quality, caregiver sensitivity, and structural features of child care (group size, caregiver–child ratio). Research questions are: (1) Does process quality of child care differ across geographic regions? (2) Does process quality of child care depend on ERS characteristics (type of scale, original versus revised versions, full versus shortened versions)? and (3) How are structural components of child care quality and caregiver sensitivity related to ERS ratings of process quality?

Method

Selection Procedure

We systematically searched the electronic databases ERIC, Current Contents, PsychInfo, and PubMed using single and combined search terms as follows: Early Childhood Environment Rating Scale*, Infant/Toddler Environment Rating Scale*, ECERS*, ITERS*, child care, day care, center/centre care, and preschool. Please note that the ECERS Extension (ECERS-E; Sylva et al. 2003) that was developed to supplement the ECERS-R in terms of curricular aspects of quality was not included in the meta-analysis. Second, the references of the collected papers were searched for relevant studies. Studies were included if the following criteria were met: Studies (1) were carried out within child care centers for children up to 5 years, (2) provided descriptive statistics for the ERS, and (3) achieved pre-established levels of inter-rater reliability for these measures.

A further criterion for inclusion in this meta-analysis was the representativeness of the centers in each study. Two steps were taken. Studies that selected only high-quality centers (e.g., Perlman et al. 2004) and studies that targeted specific populations, such as in Head Start settings (e.g., Ontai et al. 2002), were not included. Of the remaining studies, we examined to what extent the authors acknowledged that the centers were an adequate representation of that particular country, and based on the information provided, we distinguished between three levels of representativeness (high, moderate, and low). We considered studies to be *highly* representative if stratified random sampling was used at the country or state (US) level; *moderately* representative if stratified random sampling was used but only for part of the country or state (US), for example, metropolitan areas; and *low* either if no random sampling was reported or if random sampling was reported but only for part of the country or state (US) and without considering strata in the selection. If authors did not explicitly mention the term “stratified,” but described random sampling maximizing diversity with respect to relevant features of a

particular country, such as geography, program settings, and family SES, we considered this as stratified random sampling.

The number of groups or classrooms observed for each study was noted, but we did not establish a minimum number of centers within a study as a criterion for inclusion.

Further, if a study reported on the results of a quality improvement program or an intervention, only pretest scores for that sample were used. If more than one publication was found for the same study or dataset, only the results from the most recent publication were used in the meta-analysis, unless an earlier publication provided more relevant information than the latest publication. If a publication reported on both the ITERS(-R) and ECERS(-R) in different subsamples, these subsamples were considered as separate studies. Similarly, if a publication reported quality scores for more than one country, each country sample was considered as a separate study.

We finished the search in summer 2012. This procedure yielded 72 publications, published from 1989 to 2012, covering a total of 7737 child care groups or classrooms.

Quality Measures

Environment Rating Scales

The ITERS and its revision the ITERS-R have been developed for groups with children under 2.5 years of age, whereas the ECERS and its revision the ECERS-R are intended for groups with children between 2.5 and 5 years of age. Each of these scales comprises seven subscales, that is: (a) Space and Furnishings (e.g., indoor space, room arrangement for play, child-related display), (b) Personal Care Routines (e.g., greeting/departing, nap/rest, health practices), (c) Language–Reasoning (e.g., books/pictures, informal use of language), (d) Activities (e.g., fine motor, dramatic play), (e) Interaction (e.g., supervision of children, staff–child interactions, interactions between children), (f) Program Structure (e.g., free play, group time), and (g) Parents and Staff (e.g., provisions for parents, staff interaction). The items (ranging from 35 for the original ITERS to 43 for the ECERS-R) are presented on a seven-point scale with detailed criteria for 1 (*inadequate*), 3 (*minimal*), 5 (*good*), and 7 (*excellent*). For each item, a score is given from 1 to 7, generating an average score for process quality across all items. Each item equally contributes to the average process quality score. Scoring is based on observation (minimal 2 h) as well as caregiver responses to questions about aspects of the program that are not directly observable.

Arnett Caregiver Interaction Scale

The Arnett Caregiver Interaction Scale (CIS; Arnett 1989) consists of 26 descriptions of caregiver behavior that are scored on a four-point scale, with scores ranging from 1 (*not at all*) to 4 (*very much*). Ratings are based on how often (after a few hours of observation) a caregiver was observed to perform the behavior

described in the item. Arnett (1989) originally distinguished four subscales: (a) Positive Interaction, (b) Punitiveness, (c) Permissiveness, and (d) Detachment; however, subsequent analysis by Whitebook et al. (2004) found a three-factor solution that included the factors Sensitivity, Harshness, and Detachment. This was replicated by other researchers (e.g., Tietze et al. 1996). A more recent validation study (Colwell et al. 2013) showed that the Arnett scale measures one substantive dimension (sensitive caregiver interaction) rather than four subscales. For the purposes of this meta-analysis, therefore, we used the reported score for caregiver sensitivity (although sometimes labeled differently), which is defined as caregiver behavior that is warm, attentive, and engaged.

Data Extraction

Sample characteristics comprised publication year, the geographical area from which the sample originated (country, geographic region), sample size (number of groups or classrooms observed), and the assigned level of representativeness (high, moderate, low). The coded outcomes were the mean scores on process quality as measured with the ERS, mean scores on Arnett caregiver sensitivity, mean reported group size, child–caregiver ratio, and reported Pearson’s correlations (*rs*) between process quality and caregiver sensitivity, group size, and child–caregiver ratio. Group size was defined as the mean number of children present during the observation. Because studies varied in how ratio was reported, that is, as child–caregiver ratios and caregiver–child ratios, we extracted a figure for mean child–caregiver ratio (number of children present divided by number of caregivers present; e.g., 4:1). If only caregiver–child ratio was reported (i.e., 1: 4), we calculated its inverse (child–caregiver ratio = 1/caregiver–child ratio). Studies that only provided correlations between caregiver–child ratio and process quality could not be included in the meta-analysis, because it is not possible to convert these correlations to extract child–caregiver ratio. In order to examine the reliability of the ERS, the reported internal consistencies (Cronbach’s alpha) of the scales were also coded.

Scale moderators included type of scale (ITERS[-R] vs ECERS[-R]), original versus revised versions, and full versus shortened versions. As for the latter, we coded whether subscales or individual items were excluded from the scales before analysis. If so, the scale was coded as shortened version.

To assess inter-coder reliability, 21 publications (29 %) were coded by two coders. Agreement between the coders for outcome variables and moderators was satisfactory (mean kappa for categorical variables .77; percentage agreement between 76 and 100 %; mean intra-class correlations for continuous variables .99).

Meta-analytic Procedures

The meta-analysis was performed using the Comprehensive Meta-Analysis Program (CMA; Borenstein et al. 2009). Tests for significance and moderator analyses were performed through random-effects models (Borenstein et al. 2007). A random-effects model allows for the possibility that there are random differences between studies that are associated with variations in procedures, measures, settings, that go

beyond subject-level sampling error and thus point to different study populations. Q -statistics (Borenstein et al. 2009) were computed to test the homogeneity of the overall set and specific sets of effect sizes. Contrasts were only tested if a subset consisted of at least four studies ($k \geq 4$) (Bakermans-Kranenburg et al. 2003).

To address possible publication bias, we (1) used the “trim and fill” method (Duvall and Tweedie 2002a, b) to calculate the effect of potential publication bias on the outcomes of the meta-analysis and (2) computed the fail-safe N according to the method proposed by Orwin (1983), referring to the number of studies necessary to bring the effect size down to trivial levels (e.g., $r < .10$). Using the “trim and fill” method, a funnel plot is constructed of each study’s effect size against the sample size or the standard error (usually plotted as $1/SE$ or precision). These plots should be shaped like a tunnel if no publication bias is present. However, since smaller studies and statistically nonsignificant studies are less likely to be published, studies in the bottom left hand corner of the plot are often omitted (Duvall and Tweedie 2002a, b). With the “trim and fill” procedure, the k rightmost studies considered to be symmetrically unmatched are trimmed and their missing counterparts are imputed or “filled” as mirror images of the trimmed outcomes. This leads to an adjusted estimate of the combined effect size taking into account potential publication bias.

Results

These results provide a summary on the features of the studies identified in the systematic search of the literature. The analyses then report findings on the research questions which explore how process quality of child care, measured by mean ERS scores, varies for moderator variables of geographic region, sampling characteristics, and scale characteristics. Then, we examine how structural components of child care quality (group size, child–caregiver ratio) and associations between structural components and process quality vary for moderator variables of geographic region and scale characteristics. Finally, we examine how caregiver sensitivity and associations between caregiver sensitivity and process quality vary for moderator variables of geographic region and scale characteristics.

We found 72 studies (56 publications) with a total of 7737 child care groups (infant groups, toddler groups, preschool groups, mixed age groups) in 23 countries covering five geographic regions (Asia, Australia, Europe, North America, South America). Table 1 provides an overview of study characteristics and moderators. About one-third of the studies ($k = 25$) were conducted in the USA. Reported Cronbach’s alpha ($k = 34$) of the full ERS ranged from .66 to .97 with a mean alpha of .90.

Mean ERS score for the combined set of studies ($k = 72$, $N = 7737$) was 3.96 (CI 3.79–4.12), which is just below the midpoint of the seven-point scale (see Table 2). Mean scores ranged from 2.4 to 5.98. Duval and Tweedie’s (2002a, b) trim and fill approach revealed no asymmetry in the funnel plots (see Fig. 1); the absence of unmatched studies on the left side suggests that publication bias is unlikely. A cumulative meta-analysis confirmed the absence of an association

Table 1 Study characteristics (country, authors, year of publication, sample size, additional quality measures) and moderators (geographic region, representativeness, type of scale, full scale)

Geographic region	Country	Authors (year of publication)	Representativeness	Sample size	Type of scale ^a	Full scale	Additional quality measures
Asia	Bangladesh	Aboud (2006)	Moderate	22	ECERS-R	No	
	China	Ying Hu and Szentze (2009)	Low	40	ECERS-R	No	
	Jordan	Al-Hassan et al. (2010)	High	107	ECERS-R	Yes	
	Singapore	Kwan et al. (1998)	High	16	ECERS-R	Yes	
	South Korea	Sheridan et al. (2009)	Moderate	24	ECERS	No	
Australia	Australia	Fenech et al. (2010)	Moderate	69	ECERS-R	No	
		Fenech et al. (2010)	Moderate	48	ITERS	No	
		Ishimine et al. (2010)	Moderate	35	ECERS-R	Yes	
		Skouteris et al. (2007)	Low	70	ITERS	No	R
		Farquhar (1989)	Moderate	8	ECERS	Yes	
Europe	New Zealand	Tietze et al. (1996)	High	37	ECERS	No	S, rS
	Austria	Anders et al. (2012)	Moderate	97	ECERS-R	Yes	G, R
	Germany	Cryer et al. (1999)	High	82	ECERS	No	G, rG, R, S, rS
		Sheridan and Schuster (2001)	Moderate	10	ECERS	Yes	
		Tietze and Cryer (2004)	High	75	ITERS	No	
Greece	Greece	Petrogannis (2002)	Low	25	ITERS	Yes	S
		Rentzou (2010)	Low	17	ECERS-R	Yes	
		Rentzou (2010)	Low	10	ITERS	Yes	
Ireland	Ireland	Hennessy and Delaney (1999)	Low	16	ECERS	No	
		Melhuish et al. (2002)	High	78	ECERS-R	Yes	
Italy	Italy	Varin et al. (1996)	Low	6	ITERS	Yes	

Table 1 continued

Geographic region	Country	Authors (year of publication)	Representativeness	Sample size	Type of scale ^a	Full scale	Additional quality measures
North America	Netherlands	Gevers Deynoot-Schaub and Riksen-Walraven (2005)	Moderate	51	ITERS	No	G, rG, R, rR
		Vermeer et al. (2008)	High	27	ECERS-R	No	G, rG, R, rR, S, rS
		Vermeer et al. (2008)	High	15	ITERS-R	No	G, rG, R, rR, S, rS
		Vermeer et al. (2010)	Moderate	20	ECERS-R	No	G, rG, R
	Portugal	Van IJzendoorn et al. (1998)	High	43	Combined	No	G, rG, R, rR, S, rS
		Abreu-Lima et al. (2013)	Moderate	60	ECERS-R	Yes	
		Barros and Aguiar (2010)	Moderate	160	ITERS-R	No	G, rG, R, rR
	Spain	Cryer et al. (1999)	High	80	ECERS	No	G, rG, R, S, rS
		Pessanha et al. (2007)	Low	30	ITERS	Yes	rG, rR
		Cryer et al. (1999)	High	55	ECERS	No	G, rG, R, S, rS
Lera et al. (1996)		Low	59	ECERS	Yes		
Sp. Basque Country	Vermeer et al. (2010)	Moderate	26	ECERS-R	No	G, rG, R	
Sweden	Kärby et al. (1994)	Moderate	40	ECERS	Yes		
	Sheridan and Schuster (2001)	Moderate	10	ECERS	Yes		
UK	Sheridan et al. (2009)	Moderate	27	ECERS	No		
	Sylva et al. (2006)	High	141	ECERS-R	Yes		
	Goelman et al. (2006)	High	210	ECERS-R	Yes	S, rS	
	Goelman et al. (2006)	High	115	ITERS	Yes	S	
	Japel et al. (2005)	Moderate	1024	ECERS-R	Yes		
	Schliecker et al. (1991)	Low	10	ECERS	Yes		
Neth. Antilles	Meerdink and Schonenburg (2010)	Low	45	ECERS-R	No	G, rG, R, rR	

Table 1 continued

Geographic region	Country	Authors (year of publication)	Representativeness	Sample size	Type of scale ^a	Full scale	Additional quality measures
USA		Bisceglia et al. (2009)	Low	153	ITERS-R	No	R, rR
		Boller et al. (2010)	Low	25	ITERS-R	Yes	S
		Bryant et al. (1999)	Low	180	ECERS	Yes	
		Burchinal et al. (2002)	Moderate	418	ECERS	Yes	R, S
		Burchinal et al. (2002)	Moderate	135	ITERS	Yes	R, S
		Cryer et al. (1999)	High	288	ECERS	No	G, rG, R, S, rS
		De Kruif et al. (2000)	High	63	Combined	Yes	G, R, S
		Dunn (1993)	Low	30	ECERS	Yes	G, rG, R
		Fiene et al. (2002)	High	111	ECERS-R	Yes	
		Hestenes et al. (1993)	Low	30	ECERS	Yes	
		Hestenes et al. (2007)	Low	400	ITERS-R	No	rG, rR
		Hooks et al. (2006)	Low	68	ECERS-R	Yes	
		Howes and Smith (1995)	High	99	ECERS	Yes	G, R
		Howes and Smith (1995)	High	75	ITERS	Yes	G, R
		Jones-Branch et al. (2004)	Low	34	ECERS-R	Yes	
		LoCasale-Crouch et al. (2007)	High	676	ECERS-R	Yes	G, R
		Peisner-Feinberg et al. (2008)	Moderate	177	ECERS-R	Yes	G, S
		Powell et al. (2008)	Low	12	ECERS-R	Yes	
		Sakai et al. (2003)	Moderate	68	ECERS-R	Yes	rS
		Scarr et al. (1994)	High	120	ECERS	Yes	G, rG
	Scarr et al. (1994)	High	238	ITERS	Yes	G, rG	

Table 1 continued

Geographic region	Country	Authors (year of publication)	Representativeness	Sample size	Type of scale ^a	Full scale	Additional quality measures
		Tietze and Cryer (2004)	High	219	ITERS	No	
		Tonyan and Howes (2003)	High	192	Combined	Yes	
		Torquati et al. (2011)	High	192	ITERS	Yes	S
		Torquati et al. (2011)	High	114	ECERS-R	Yes	S
South America	Brazil	Campos et al. (2011)	Moderate	138	ECERS-R	Yes	
	Chile	Domínguez et al. (2008)	Moderate	39	ITERS-R	Yes	rR, S, rS
	Chile	Herrera et al. (2005)	Moderate	120	ECERS	Yes	rS
	Chile	Herrera et al. (2005)	Moderate	63	ITERS	Yes	rS
	Chile	Villalón et al. (2002)	Moderate	120	ECERS	Yes	

^a R, revised scale; G, group size (M and SD) reported; rG, correlation between group size and ERS reported; R, child-caregiver ratio (M and SD) reported; rR, correlation between child-caregiver ratio and ERS reported; S, sensitivity (M and SD) reported; rS, correlation between sensitivity and ERS reported

Table 2 Results of moderator analyses: number of studies and classrooms and combined mean quality scores including 95 % confidence intervals (CI)

	<i>K</i>	<i>N</i>	<i>M</i>	SE	95 % CI	<i>Q</i> heterogeneity
Overall	72	7737	3.96	.08	3.79–4.12	5457.73**
<i>Continent</i>						
Europe	27	1297	3.82 ^c	.11	3.59–4.04	1624.65**
North America ^d	30	5521	4.09 ^b	.11	3.87–4.30	1672.33**
South America	5	480	3.58 ^c	.26	3.07–4.09	30.16**
Australia ^e	5	230	4.97 ^a	.27	4.44–5.51	256.64**
Asia	5	209	3.29 ^c	.26	2.78–3.81	87.25**
<i>Sample representativeness</i>						
High	26	3468	4.00	.14	3.73–4.27	1121.47**
Moderate	26	3009	3.94	.14	3.66–4.22	2015.40**
Low	20	1260	3.92	.16	3.60–4.24	2063.41**
<i>Scale characteristics</i>						
Infant–toddler ERS ^f	21	2144	3.85	.16	3.54–4.16	3216.51**
Early childhood ERS ^f	48	5295	3.98	.10	3.78–4.19	2036.40**
Original scale	40	3466	4.03	.11	3.80–4.25	2176.34**
Revised scale	32	4271	3.87	.12	3.62–4.12	3216.27**
Full scale	47	5645	3.91	.11	3.71–4.12	1591.31**
Shortened scale	25	2092	4.03	.14	3.75–4.31	3835.63**

* $p < .05$; ** $p < .001$ ^{a,b,c} Means with different letters differ significantly^d Including Netherlands Antilles^e Including New Zealand^f Including revision

between year of publication and mean ERS score: A trend toward higher or lower scores across time was not present.

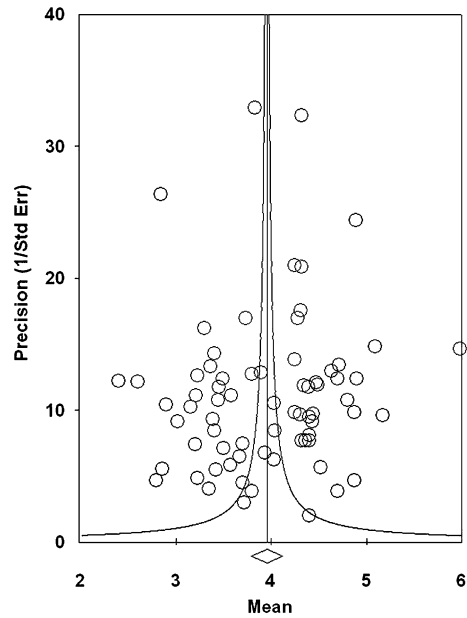
At the country level, average child care quality was lowest (mean scores < 3) in Bangladesh (Aboud 2006), the Netherlands Antilles (Meerdink and Schonenburg 2010), and South Korea (Sheridan et al. 2009) and highest (mean scores > 5) in Australia (Fenech et al. 2010; Skouteris et al. 2007).

How Does Process Quality of Child Care Vary by Geographic Region?

We conducted a moderator analysis contrasting mean ERS scores across five geographic regions, which gave statistically significant results, $Q(4) = 24.75$, $p < .001$ (see Table 2).

Pairwise post hoc contrasts indicated that child care quality in Australia and New Zealand was significantly higher than in all other geographic regions (see Fig. 1). Child care quality in North America (including the Netherlands Antilles) was

Fig. 1 Funnel plots showing each study's effect size against the standard error (plotted as precision). Mean ERS scores are plotted on the X-axis. Each dot represents a study



significantly higher than in Europe, South America, and Asia (see Fig. 2). Other contrasts across geographic regions were not statistically significant.

In addition to *mean scores*, we examined whether *variances* in ERS scores differed across geographic regions. For this purpose, an ANOVA (using SPSS 19) was applied to the data, with mean SDs on the ERS as dependent variables and geographic region as the independent variable. Results showed a main effect for geographic region ($F[4, 67] = 8.02$; $p < .001$). Post hoc tests (Bonferroni) demonstrated that SDs were significantly lower in the European and Asian samples ($d = 1.35$ and $d = .97$ respectively). This implies that, in general, there is more variety in ERS scores within North American countries than within Asian and European countries. Because variances in scores could be confounded with scale characteristics, we added type of scale, original versus revised scale, and full versus shortened scale, as covariates in three subsequent analyses. After including these covariates, the effects remained unchanged.

How Does Process Quality of Child Care Vary by Sampling Characteristics?

As for representativeness, 26 studies were designated as highly representative, 26 studies were moderately representative, and 20 studies were low representative of a particular country. We conducted a moderator analysis contrasting mean ERS scores across the three levels, which was not statistically significant (see Table 2), indicating that level of study representativeness for a particular country did not

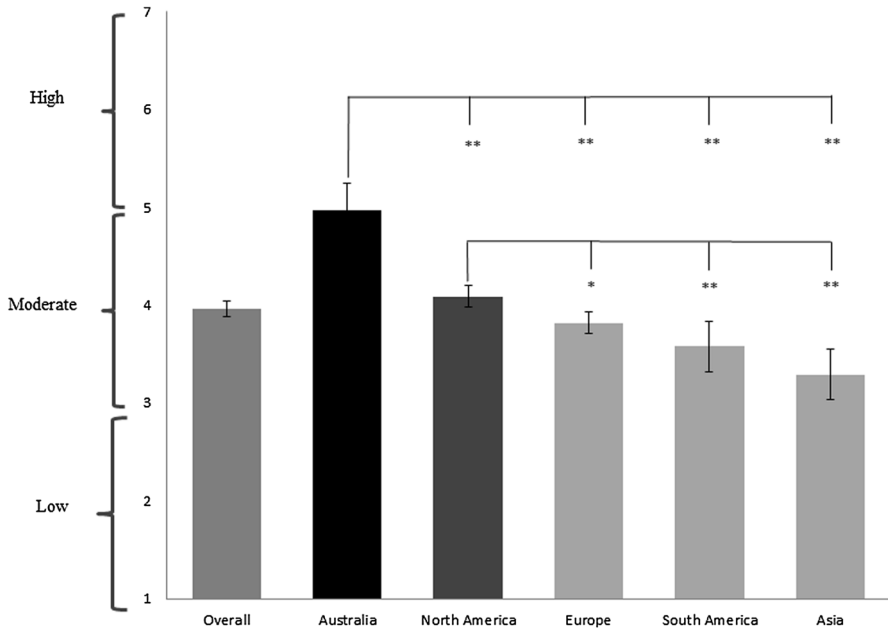


Fig. 2 Mean process quality as observed with the ERS in five geographic regions. Mean ERS scores are plotted on the Y-axis and classified as low (mean score < 3), moderate ($3 \leq$ mean score < 5), and high (mean score \geq 5). * $p < .05$; ** $p < .01$

affect quality levels. We conducted a further test for differences in ERS scores across geographic regions by including only the 52 studies that were highly or moderately representative. Results were significant, $Q(4) = 22.88$, $p < .001$. Pairwise post hoc contrasts confirmed our findings from the whole dataset: Child care quality in Australia and New Zealand was significantly higher than in all other geographic regions. Child care quality in North America was significantly higher than in South America and Asia. The differences in quality between North America and Europe disappeared, although a trend toward higher scores in North America was still visible.

How Does Process Quality of Child Care Vary by Scale Characteristics?

Type of Scale

Twenty-one studies reported on the ITERS or its revision (ITERS-R); 48 studies used the ECERS or ECERS-R, and three studies reported on combined outcomes for ITERS and ECERS (these studies were excluded from the moderator analysis). A moderator analysis contrasting studies reporting on the scale for infants and toddlers (ITERS[-R]) versus studies reporting on the scale for older children (ECERS[-R]) did not yield a statistically significant result.

Original Versus Revised Scale

The original ITERS or ECERS was used in 40 studies, whereas the revised version was used in 32 studies. A moderator analysis contrasting scores on the original scales versus the revised scales did not yield a statistically significant result: Scores on process quality did not depend on the version of the scale.

Full Scale Versus Shortened Scale

In 25 studies, a shortened version of the ERS was used, most often (74 %) excluding the subscale Parents and Staff or Adult Needs. A moderator analysis contrasting scores on the full scale versus shortened scale was not statistically significant.

Additionally, we examined whether interactions between the moderators geographic region and scale characteristics yielded different scores on process quality. We tested whether geographic region in interaction with respectively type of scale, original scale versus revision, and full scale versus shortened scale yielded different results for mean ERS ratings. We found only one interaction effect: Within Europe, a moderator analysis contrasting studies reporting on the scale for infants and toddlers (ITERS[-R]) versus studies reporting on the scale that was developed for older children (ECERS[-R]), yielded a statistically significant result, $Q(1) = 7.70, p < .01$, with higher scores on the early childhood version compared with the infant–toddler version.

How do Structural Features and Associations Between Structural Features and Process Quality Vary by Geographic Region and Scale Characteristics?

Group Size

Mean group sizes and standard deviations were reported in 21 studies, from which 11 were European and 10 were from North America. Mean group size for the combined set of studies ($k = 21, N = 2467$) was around 15 ($M = 15.19$; CI 13.25–17.13), with a range from 9.1 to 30.0. A moderator analysis showed that group size in Europe was not statistically significantly different from group size in North America. Moderator analyses on group size including other geographic regions were not possible, because less than four studies ($k < 4$) were involved.

A moderator analysis with scale characteristics showed that group size differed dependent on type of scale. As was expected, group sizes were smaller in groups in which the ITERS(-R) was used than in groups in which the ECERS(-R) was used, $Q(1) = 15.44, p < .001$. No differences in group sizes were found when comparing groups in which the original versus revised scales were used and groups in which the full scale versus shortened scales were used.

Pearson's correlation coefficients (r_s) between mean ERS scores and group size were reported in 17 studies (see Table 3) and ranged from $-.40$ to $.29$. Mean r for the combined set of studies ($k = 17, N = 1710$) was $-.03$ (CI $-.11-.06$; $p = .53$), demonstrating no overall statistically significant association between process quality

Table 3 Pearson's correlations between ERS scores, group size, ratio, and caregiver sensitivity: number of studies and classrooms, and combined r 's including 95 % confidence intervals (CI)

	K	N	r	95 % CI	Q heterogeneity
Group size	17	1710	-.03	-.11-.06	37.23**
Child-caregiver ratio	10	963	-.17**	-.27 to -.07	16.82*
Caregiver sensitivity	13	1127	.62**	.56-.67	22.44*

* $p < .05$; ** $p < .01$

and group size. Moderator analyses showed no effect of continent or scale characteristics.

Child-Caregiver Ratio

In 21 studies (11 in Europe and 10 in North America), mean child-caregiver ratio and standard deviations were reported. Mean child-caregiver ratio for the combined set of studies ($k = 21$, $N = 2638$) was 8.60 (CI 7.21–9.51), with a range from 3.1 to 25. Across all studies, on average eight to nine children were cared for by one caregiver. A moderator analysis showed that child-caregiver ratio in Europe did not statistically significantly differ from child-caregiver ratio in North America. Moderator analyses on ratio including other geographic regions were not possible, because less than four studies ($k < 4$) were involved.

Further analyses showed that ratio differed dependent on type of scale. Expectedly, ratios in groups in which the ITERS(-R) were used were lower than in groups in which the ECERS(-R) was used, $Q(1) = 33.74$, $p < .001$. No differences were found for original scale versus revision and full scale versus shortened scale.

Pearson's correlation coefficients (r s) between mean ERS scores and child-caregiver ratio were reported by 10 studies (6 in Europe, 3 in North America, 1 in South America) and ranged from $-.33$ to $.22$ (see Table 3). Mean r for the combined set of studies ($k = 10$, $N = 963$) was $-.17$ (CI $-.27$ to $-.07$; $p = .001$), indicating that process quality was significantly associated with child-caregiver ratio: Process quality was higher when fewer children were under the care of a caregiver. Moderator analyses either were not possible because of too small k (geographic region, type of scale) or did not yield any statistically significant results (original vs revised scale, full vs shortened scale).

How do Caregiver Sensitivity and Associations Between Caregiver Sensitivity and Process Quality Vary by Geographic Region and Scale Characteristics?

Caregiver Sensitivity

In 19 studies (8 in Europe and 11 in North America), caregiver sensitivity, as measured with the Arnett scale, was reported. Mean caregiver sensitivity for the

combined set of studies ($k = 19$, $N = 2212$) was 3.06 (CI 2.98–3.15). A moderator analysis showed a significant difference between caregiver sensitivity in Europe and North America, $Q(1) = 5.85$, $p = .02$. On average, caregivers in North America received higher ratings for sensitive behavior than caregivers in Europe. Moderator analyses on caregiver sensitivity including other geographic regions were not possible, because less than four studies ($k < 4$) were involved.

Pearson's correlation coefficients (r_s) between mean ERS scores and caregiver sensitivity were reported in 13 studies (see Table 3). Mean r for the combined set of studies ($k = 13$, $N = 1127$) was .62 (CI .56–.67; $p < .01$), demonstrating a statistically significant positive association between scores on the ERS and caregiver sensitivity. After correcting for attenuation based on the reliabilities of both scales, the correlation was .73, showing strong convergence. Effects were not moderated by scale characteristics. Moderator effects for geographic region and type of scale could not be tested because $k < 4$. As a further test, Duval and Tweedie's (2002a, b) trim and fill approach revealed no asymmetry in the funnel plots; the absence of unmatched studies on the left side suggests that publication bias is unlikely. Orwin's fail-safe method showed that, using a "trivial" correlation of .10, it would need another 86 unpublished studies with null effects for the association between mean ERS scores and caregiver sensitivity to bring the effect size under $r = .10$.

Discussion

Across five geographic regions and 23 countries, average center child care quality as measured by the ERS was moderate: nearly 4 on a seven-point rating scale. The lowest quality levels (mean scores < 3) were reported in Bangladesh (Aboud 2006), the Netherlands Antilles (Meerdink and Schonenburg 2010), and South Korea (Sheridan et al. 2009). Good quality care (mean scores > 5) was reported in Australia (Fenech et al. 2010; Skouteris et al. 2007). Our results, based on over 7700 observations, showed that, on average, levels of child care quality were higher in Australia than in all other geographic regions and that child care quality in North America was higher than in Europe, South America, and Asia. These findings were largely confirmed in a further analysis excluding studies that were not representative of child care centers for a particular country.

Differences in Process Quality Across Geographic Regions

The results of this meta-analysis provide the first opportunity for researchers and policymakers to take a global view of center-based quality for children under 5 years of age. Whilst the results for ERS scores are compelling, accounting for the observed differences in quality between geographic regions and the variations in quality within geographic regions and countries, requires consideration of multiple factors that could not be addressed within our analyses. Because the design of the current meta-analysis does not allow a cause-effect analysis, we can only speculate on possible influences on quality.

There are major differences across the countries included in this meta-analysis in policy focus, regulations, and subsidies with respect to child care. As noted in the UNICEF (UNICEF Innocenti Research Center 2008) benchmarks for early childhood education and care, government regulations and accreditation systems are key factors influencing quality. These vary from one country to another, as well as between states or provinces, and even municipalities or organizations, within a country. We note that 14 of the countries included in the current meta-analysis were also reported on in the UNICEF (2008) comparison of 25 OECD countries. Surprisingly, the countries achieving the highest scores for child care quality in our meta-analysis did not achieve a high ranking in the UNICEF list: The USA met three out of ten benchmarks; Australia achieved only two benchmarks; and Canada achieved one benchmark. In contrast, European countries met more benchmarks, but Europe had a lower average ERS score in our meta-analysis. It seems unlikely, therefore, that the higher scores on the ERS that we found in the USA and Australia can be explained in full by broad government policy initiatives (at least those identified by UNICEF).

Four additional issues are of importance when seeking to explain the observed differences in quality across countries. First, the ERS place importance on features of the physical environment (individual care, hygiene, space, furniture, equipment, and physical safety) as indicators of basic quality of care. This may be of greater importance in the USA and Australia/New Zealand than in other countries. Alternately, it may be more difficult to meet these standards in less prosperous countries or in countries in which national legislation, regulation, or policies do not stress the importance of these indicators. Consequently, the hierarchically ordering of items within the ERS, which require meeting basal aspects of hygiene and physical environment before other higher-order indicators can be assessed, may have resulted in lower scores. For example, ratings for some items will not exceed a score of 1 (*inadequate*) because requirements for specific hygienic procedures or furnishings have not been met, even though higher scale indicators that might be assessed positively are observed.

Second, child care centers in the USA and Australia/New Zealand may be expected to provide environments for children's educational outcomes to a greater extent than in other countries. Lower scores on some items or subscales (e.g., Activities) may be obtained because the specified amount or variety of equipment for stimulating children's development (e.g., blocks, fine motor equipment, supplies for pretend play, sand and water) are not evident in daycare centers in Europe or other parts of the world (e.g., Vermeer et al. 2008).

Third, in some parts of the USA, ERS have been seen as important tools for the evaluation and improvement of quality of care (e.g., Park et al. 2012). Although studies that specifically targeted centers involved in quality improvement programs were not included in the meta-analysis, it is conceivable that the expectations for child care organizations to meet the ERS criteria are more common in the USA than in other countries. It may also be the case that the attention given to quality improvement in general in the USA, through state-based Quality Improvement and Rating Systems, may have raised awareness of the importance of caregiver sensitivity in interactions with children. In turn, this may account for the observed

differences in caregiver sensitive behavior that, on average, was higher in North America than in Europe, as measured with the Arnett Caregiver Interaction Scale.

Fourthly, studies examining the internal and cross-cultural validity of the ERS are limited. Gordon et al. (2013, 2015) showed that, among other things, the category ordering of the ECERS-R (response process validity) assumed by the structure of the scale is not consistently evident; that is, indicators attached to higher rating categories do not necessarily reflect higher quality. Furthermore, few studies have assessed how well the items of the ERS represent indicators of child care quality in countries outside the USA. An exception is the detailed consideration of linguistic, functional, cultural, and metric equivalence of the ERS in Chile and Bangladesh (Limlingan 2011).

Differences in Process Quality Within Geographic Regions

Although average ERS scores for child care quality in North America exceeded that in three other geographic regions, our analyses revealed higher variance in scores in North America than in Europe and Asia. Thus, in North America scores on the ERS were less clustered around the mean than in Europe and Asia. This higher variance in scores may be partly explained by a greater diversity in regulatory standards for child care across states within the USA. For example, in states with the most stringent child care center standards, specific training is required for caregivers and child-caregiver ratios and group sizes are very small. In other states, there are no educational requirements, one adult may supervise as many as ten or 12 toddlers, and group sizes are unregulated (Clarke-Stewart and Allhusen 2005).

Australia, on the other hand, has a universal quality assurance mechanism for center-based child care that requires centers to meet standards for accreditation and quality improvement in order to qualify for parent subsidies for the cost of care (National Child care Accreditation Council [NCAC], 1993). Harrison (2010) has argued that, as well as providing greater consistency in regulatory standards, NCAC processes have resulted in a higher minimum standard in Australia, evidenced by ERS scores that rarely fall below 3.

On the other hand, cultural and country-specific interpretations are less likely to explain the observed differences in caregiver sensitive behavior that, on average, was higher in North America than in Europe, as measured with the Arnett scale. Evidence that beliefs about ideal maternal sensitivity do not differ across the globe (Mesman et al. 2015) and that contingent responsiveness, an important aspect of sensitivity, is a universal component of parenting (Kärtner et al. 2010) suggests that these differences may be genuine rather than cultural. It should be noted, however, that specific expressions of sensitive responsiveness (such as face-to-face interactions) may differ across cultures (Kärtner et al. 2010) and that the Arnett scale is a global scale that is not well suited to detect these specific expressions. In the current study, differences in mean caregiver sensitivity were small and caregiver sensitivity was reported in 19 studies (26 %) only, of which eight were from the USA. However, the design of our study does not allow conclusions about the origin of these differences, and we can therefore only speculate about possible causes. It may also be the case that the attention given to quality improvement in the USA may

have raised awareness of the importance of caregiver sensitivity in interactions with children.

ERS Process Quality, Scale Characteristics, and Structural Features

The results of our meta-analyses suggest that scores on ERS process quality were not significantly associated with ERS characteristics. This is useful psychometric information for researchers, providing evidence that different versions of the scales (type of scale, original vs revised, full vs shortened) could be reliably used in within-study and across-study comparisons without strongly affecting or distorting the outcomes.

Results of our hypothesis testing showed that child care quality was not associated with group size, but was negatively associated with child–caregiver ratio: Process quality was somewhat higher when fewer children were under the care of one caregiver. Thus, the role of structural characteristics of the care setting (in this study: child–caregiver ratio and group size) is mixed when explaining quality differences: Ratio matters more than group size. It is important to note, however, that the findings for group size were inconsistent across the combined set of studies; for example, the overall low correlation ($r = -.06$) between ERS process quality and group size was due to the fact that the reported outcomes covered both negative correlations (in the expected direction) and positive correlations (unexpected).

ERS Process Quality and Caregiver Sensitivity

Our meta-analysis showed that process quality was positively associated with caregiver sensitivity, as measured by the Arnett Caregiver Interaction Scale ($r = .62$). After correcting for attenuation, the estimated r even reached .73, explaining about half of the variance. This is an important finding, given that the ERS have often been criticized in relation to a perceived omission of process items and/or indicators. For example, Cassidy et al. (2005a) performed a content analysis of the ECERS-R at the indicator level, concluding that over half of the indicators of the ECERS-R measured structural quality rather than process quality. A specific feature of the ERS, however, is that process and structural features are intertwined within the items and indicators. Thus, the ERS items do not allow a disentanglement of process quality and structural quality.

Although the Arnett scale has been widely used in studies to measure caregiver–child interactions, validation studies are scarce. Colwell et al. (2013) showed that the Arnett scale is not well suited to distinguish between caregivers who are “highly” versus “moderately” positive in their interactions with children and that the scale measures one substantive dimension (sensitive caregiver interaction) rather than four subscales. For the purpose of our meta-analysis—to examine whether the ERS are associated with more “proximal” measures of child care quality—the Arnett scale is an appropriate measure. It should be noted, however, that in some studies, data collection for the ERS and the Arnett scale may have been done by the same observers. It is possible that this has led to an overestimation of the correlation between process quality and caregiver sensitivity. Of key importance is that our

meta-analysis showed that the ERS cover a core element that is universally recognized as required for children's positive development, that is, sensitive and responsive care by one or more adults.

Limitations

A limitation of the current meta-analysis concerns the variation of studies in terms of representativeness for a particular country. We have approached this problem in two ways. First, we excluded studies screening for only high-quality centers and studies that targeted specific populations. Second, we controlled for representativeness in additional moderator analyses that largely confirmed our results, showing that level of representativeness did not or only marginally affect quality ratings.

Another limitation is that we only included group size and child–caregiver ratio as structural indicators and that the number of studies able to be included in our hypothesis testing was relatively small (e.g., 10 for ratio). These two indicators were selected because of their importance, but also because they are objectively measured, which allowed a reliable comparison across countries. Whilst other indicators, such as caregiver qualifications, education, and experience, have often been cited as key structural features influencing process quality (see Burchinal et al. 2002), these differ worldwide and are not measured consistently. Therefore, it was not possible to include them in our meta-analysis.

The same holds for structural features at the country level: The countries included in the current meta-analyses vary widely in terms of beliefs (e.g., child care as a universal right or a social welfare provision) and policies, which in turn may influence the quality of care that is offered. Also, the distinction between public and private centers could not be included as a moderator, because this distinction does not exist in every country, and if it does, definitions concerning public and private vary across countries and may change over time.

Conclusion

Taken together, this international meta-analysis has shown that group center care, as measured by the ERS, is of average quality, with higher quality levels in Australia/New Zealand and North America. Our results emphasize the on-going need for continuing efforts by government policy makers, early childhood service providers, and educators to enhance, or maintain high levels of, childcare quality. This is especially salient in Asia, South America, and Europe. Our results suggest that scale characteristics are not responsible for differences in scores and that the ERS are related to indicators of proximal quality of care (caregiver sensitivity) and, to a lesser degree, structural quality of care (child–caregiver ratio).

This paper not only presents a state-of-the-art analysis of the ERS around the world, but also shows the need for further research focusing on psychometrics and cross-cultural issues. We believe that this work aligns with the benefits, identified by Limlingan (2011, p. 45) of making “cross-cultural comparisons using a common

instrument” which when “composed and utilized in the right way, provides a good method to facilitate discussions which allow us to learn from one another.”

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix: Key Terms

Cumulative meta-analysis: a meta-analysis that is performed first with one study, then with two studies, and so on, until all relevant studies have been included in the analysis.

Funnel plot: a graphical method to display possible publication bias. If there is no publication bias, a typical symmetric funnel shape can be observed.

Meta-analysis refers to the statistical synthesis of results from a series of studies.

Publication bias: Bias is produced when the published studies do not represent adequately all the studies carried out on a specific topic.

Q statistic: a test to assess heterogeneity (variability) across the studies included in a metaanalysis, in which the effect size of study is compared with a pooled estimate across studies for the overall effect.

Trim and fill method: a non-parametric technique to correct for publication bias in meta-analysis. The method can be used to estimate the number of missing studies and to adjust the estimate of the “treatment” effect for potential publication bias in a meta-analysis.

References

The asterisked references in this list indicate that these are the research studies included in the meta-analysis.

- *Aboud, F. E. (2006). Evaluation of an early childhood preschool program in rural Bangladesh. *Early Childhood Research Quarterly*, 21, 46–60.
- *Abreu-Lima, I. M., Leal, T. B., Cadima, J., & Gamelas, A. M. (2013). Predicting child outcomes from preschool quality in Portugal. *European Journal of Psychology of Education*, 28(2), 399–420.
- *Al-Hassan, S. M., Obeidat, O. M., & Lansford, J. E. (2010). Education reform and the quality of kindergartens in Jordan. *Early Child Development & Care*, 180(9), 1203–1213. doi:10.1080/03004430902907640.
- *Anders, Y., Rossbach, H. G., Weinert, S., Ebert, S., Kuger, S., Lehl, S., & Von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly*, 27, 231–244.
- Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*, 10, 541–552.
- Bakermans-Kranenburg, M. J., Van IJzendoorn, M. H., & Juffer, F. (2003). Less is more: Meta-analyses of sensitivity and attachment interventions in early childhood. *Psychological Bulletin*, 129(2), 195.
- *Barros, S., & Aguiar, C. (2010). Assessing the quality of Portuguese child care programs for toddlers. *Early Childhood Research Quarterly*, 25, 527–535.

- Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., & Owen, M. T. (2007). Are there long-term effects of early child care? *Child Development*, 78(2), 681–701.
- *Bisceglia, R., Perlman, M., Schaack, D., & Jenkins, J. (2009). Examining the psychometric properties of the Infant–Toddler Environment Rating Scale-Revised edition in a high stakes context. *Early Childhood Research Quarterly*, 24, 121–132.
- *Boller, K., Del Grosso, P., Blair, R., Jolly, Y., Fortson, K., Paulsell, D., et al. (2010). *The seeds to success modified field test: Findings from the impact and implementation studies*. Princeton, NJ: Mathematic Policy Research Inc.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.
- Borenstein, M., Hedges, L., & Rothstein, H. (2007). Meta-analysis: Fixed effect vs. random effects. www.Meta-Analysis.com.
- *Bryant, D. M., Maxwell, K. L., & Burchinal, M. (1999). Effects of a community initiative on the quality of child care. *Early Childhood Research Quarterly*, 14, 449–464.
- *Burchinal, M. R., Cryer, D., & Clifford, R. M. (2002). Caregiver training and classroom quality in child care centers. *Applied Developmental Science*, 6, 2–11.
- *Campos, M. M., Bhering, E. B., Esposito, Y., Gimenes, N., Abuchaim, B., Valle, R., & Unbehau, S. (2011). The contribution of quality early childhood education and its impacts on the beginning of fundamental education. *Educação e Pesquisa*, 37, 15–33.
- Cassidy, D. J., Hestenes, L. L., Hansen, J. K., Hegde, A., Shim, J., & Hestenes, S. (2005a). Revisiting the two faces of child care quality: Structure and process. *Early Education and Development*, 16, 505–520.
- Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005b). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the Early Childhood Environment Rating Scale-Revised. *Early Childhood Research Quarterly*, 20, 345–360.
- Clarke-Stewart, A., & Allhusen, V. D. (2005). *What we know about child care*. Cambridge: Harvard University Press.
- Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the Early Childhood Longitudinal Study-Birth Cohort. *Early Childhood Research Quarterly*, 28, 218–233.
- *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, 339–361.
- Cryer, D., Tietze, W., & Wessels, H. (2002). Parents' perceptions of their children's child care: A cross-national comparison. *Early Childhood Research Quarterly*, 17, 259–277.
- *De Kruijff, R. E. L., McWilliam, R. A., Ridley, S. M., & Wakely, M. B. (2000). Classification of teachers' interaction behaviors in early childhood classrooms. *Early Childhood Research Quarterly*, 15, 247–268.
- *Domínguez, P., Moreno, L., Narvaez, L., Herrera, M. O., & Mathiesen, M. E. (2008). *Prácticas Pedagógicas de Calidad. Informe Final [Teaching Practices of Quality. Final Report]*. Ministerio de Educación de Chile MINEDUC. <http://bibliotecacentrodeestudios.mineduc.cl/CEUsuarioANONIMO/>.
- *Dunn, L. (1993). Proximal and distal features of day care quality and children's development. *Early Childhood Research Quarterly*, 8, 167–192.
- Duvall, S., & Tweedie, R. (2002a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95, 89–98.
- Duvall, S., & Tweedie, R. (2002b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463.
- *Farquhar, S. (1989). Assessing New Zealand child day care quality using the Early Childhood Environment Rating Scale (1). *Early Child Development and Care*, 47, 93–105.
- *Fenech, M., Sweller, N., & Harrison, L. J. (2010). Identifying high-quality centre-based child care using quantitative data sets: What the numbers do and don't tell us. *International Journal of Early Years Education*, 18(4), 283–296.
- *Fiene, R., Greenberg, M., Bergsten, M., Fegley, C., Carl, B., & Gibbons, E. (2002). *The Pennsylvania early childhood quality settings study: Report prepared for the governor's task force on early childhood care and education*. Pennsylvania: Harrisburg.
- *Gevers Deynoot-Schaub, M. J. J. M., & Riksen-Walraven, J. (2005). Child care under pressure: The quality of Dutch centers in 1995 and in 2001. *The Journal of Genetic Psychology*, 166, 280–296.

- *Goelman, H., Forer, B., Kershaw, P., Doherty, G., Lero, D., & LaGrange, A. (2006). Towards a predictive model of quality in Canadian child care centers. *Early Childhood Research Quarterly*, *21*, 280–295.
- Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology*, *49*(1), 146–160.
- Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N., Kaestner, R., & Korenman, S. (2015). Identifying high-quality preschool programs: New evidence on the validity of the Early Childhood Environment Rating Scale-Revised (ECERS-R) in relation to school readiness goals. *Early Education and Development*, *26*(8), 1086–1110.
- Harms, T., & Clifford, R. (1983). Assessing preschool environments with the Early Childhood Environment Rating Scale. *Studies in Educational Evaluation*, *8*, 261–269.
- Harms, T., Clifford, R. M., & Cryer, D. (1980). *Early Childhood Environment Rating Scale*. New York: Teachers College Press.
- Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale—Revised*. New York: Teachers College Press.
- Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York: Teachers College Press.
- Harms, T., Cryer, D., & Clifford, R. M. (2003). *Infant/Toddler Environment Rating Scale—Revised*. New York: Teachers College Press.
- Harrison, L. J. (2010). *International comparisons of quality using a methodology of objective criteria for centre-based programs for children under five*. Paper presented at the Australian Association for Research in Education (AARE) national conference.
- *Hennessy, E., & Delaney, P. (1999). Using the Early Childhood Environment Rating Scale in Ireland: Do parents and service providers share its value? *Early Years: An International Journal of Research and Development*, *19*, 12–25.
- *Herrera, M. O., Mathiesen, M. E., Merino, J. M., & Recart, I. (2005). Learning contexts for young children in Chile: Process quality assessment in preschool centres. *International Journal of Early Years Education*, *13*, 13–27.
- *Hestenes, L. L., Cassidy, D. J., Hegde, A. V., & Lower, J. K. (2007). Quality in inclusive and noninclusive infant toddler classrooms. *Journal of Research in Childhood Education*, *22*, 69–84.
- *Hestenes, L. L., Kontos, S., & Bryan, Y. (1993). Children's emotional expression in child care centers varying in quality. *Early Childhood Research Quarterly*, *8*, 295–307.
- *Hooks, L. M., Scott-Little, C., Marshall, B. J., & Brown, G. (2006). Accountability for quality: One state's experience in improving practice. *Early Childhood Education Journal*, *33*, 399–403.
- *Howes, C., & Smith, E. W. (1995). Relations among child care quality, teacher behavior, children's play, emotional security, and cognitive activity in child care. *Early Childhood Research Quarterly*, *10*, 381–404.
- *Ishimine, K., Wilson, R., & Evans, D. (2010). Quality of Australian child care and children's social skills. *International Journal of Early Years Education*, *18*, 159–175.
- *Japel, C., Tremblay, R. E., & Côté, S. (2005). Quality counts! Assessing the quality of daycare services based on the Quebec Longitudinal Study of Child Development. *IRPP Choices*, *11*, 1–42.
- *Jones-Branch, J. A., Torquati, J. C., Raikes, H., & Edwards, C. P. (2004). Child care subsidy and quality. *Early Education and Development*, *15*, 327–342.
- *Kärby, G., & Giota, J. (1994). Dimensions of quality in Swedish day care centers—an analysis of the Early Childhood Environment Rating Scale. *Early Child Development and Care*, *104*, 1–22.
- Kärtner, J., Keller, H., & Yovsi, R. D. (2010). Mother-infant interaction during the first 3 months: The emergence of culture-specific contingency patterns. *Child Development*, *81*, 540–554.
- *Kwan, C., Sylva, K., & Reeves, B. (1998). Day care quality and child development in Singapore. *Early Child Development and Care*, *144*, 69–77.
- Lamb, M.E., & Ahnert, L. (2006). Nonparental child care: Context, concepts, correlates and consequences. In W. Damon & R.M. Lerner (Eds.) & K.A. Renninger & I.E. Siegel (Vol. Eds.), *Handbook of child psychology*, Vol. 4, *Child psychology in practice* (6th ed., pp. 950-1-16). Hoboken, NJ: Wiley.
- *Lera, M., Owen, C., & Moss, P. (1996). Quality of educational settings for four-year-old children in England. *European Early Childhood Education Research Journal*, *4*(2), 21–32.
- Limlingan, M. C. (2011). On the right track. Measuring early childhood development program quality internationally. *Current Issues in Comparative Education*, *14*, 38–47.

- *LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., et al. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly*, 22, 3–17.
- Love, J. M., Harrison, L., Sagi-Schwartz, A., Van IJzendoorn, M. H., Ross, C., Ungerer, J. A., et al. (2003). Child care quality matters: How conclusions may vary with context. *Child Development*, 74(4), 1021–1033.
- *Meerdink, M. A., & Schonenburg, L. (2010). *Pedagogical quality of daycare centers on the Netherlands Antilles: Current quality and comparison with quality measures of 2007*. Utrecht: Department of Social Sciences, Utrecht University.
- *Melhuus, E., Quinn, L., McSherry, K., Sylva, K., Sammons, P., Siraj-Blatchford, I., et al. (2002). *Characteristics of pre-school environments in Northern-Ireland: an analysis of observational data*. Belfast: The Stranmillis Press.
- Mesman, J., Van IJzendoorn, M. H., Behrens, K., Carbonell, O. A., Cárcamo, R. A., Cohen-Paraira, I., et al. (2015). Is the ideal mother a sensitive mother? Beliefs about early childhood parenting in mothers across the globe. *International Journal of Behavioral Development*,. doi:10.1177/0165025415594030.
- National Childcare Accreditation Council [NCAC]. (1993). *Putting children first. Quality Improvement and Accreditation System handbook*. Sydney: National Childcare Accreditation Council.
- National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network. (2002). Child-care structure → process → outcome: Direct and indirect effects of child-care quality on children's development. *Psychological Science*, 13(3), 199–206.
- Ontai, L. L., Hinrichs, S., Beard, M., & Wilcox, B. L. (2002). Improving child care quality in Early Head Start programs: A partnership model. *Infant Mental Health Journal*, 23(1–2), 48–61.
- Orwin, R.G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8(2), 157–159.
- Park, S., Ferreti, Ch., & Ames, G. (2012). Assessing quality of early care and education: Lessons learned from San Francisco's 'Gateway to quality'. *International Journal of Child Care and Education Policy*, 6(2), 55–77.
- *Peisner-Feinberg, E., Bryant, D., Fischer, R., & Grayson, L. (2008). *Cuyahoga County child care quality study*. Cleveland: Case Western Reserve University.
- Pearlman, M., Zellman, G. L., & Le, V. N. (2004). Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly*, 19, 398–412.
- *Pessanha, M., Aguiar, C., & Bairrão, J. (2007). Influence of structural features on Portuguese toddler child care quality. *Early Childhood Research Quarterly*, 22, 204–214.
- *Petrogannis, K. (2002). Greek day care centres' quality, caregivers' behaviour and children's development. *International Journal of Early Years Education*, 10, 138–148.
- 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, 281–303.
- Pianta, R. C., La Paro, K. M., & Hamre, B. C. (2008). *The Classroom Assessment Scoring System*. Baltimore, MD: Brooks.
- *Powell, D. R., Burchinal, M. R., File, N., & Kontos, S. (2008). An eco-behavioral analysis of children's engagement in urban public school classrooms. *Early Childhood Research Quarterly*, 23, 108–123.
- *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.
- *Sakai, L. M., Whitebook, M., Wishard, A., & Howes, C. (2003). Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing differences between the first and revised edition. *Early Childhood Research Quarterly*, 18(4), 427–445.
- *Scarr, S., Eisenberg, M., & Deater-Deckard, K. (1994). Measurement of quality in child care centers. *Early Childhood Research Quarterly*, 9, 131–151.
- *Schliecker, E., White, D. R., & Jacobs, E. (1991). The role of day care quality in the prediction of children's vocabulary. *Canadian Journal of Behavioural Science*, 23(1), 12–24.
- *Sheridan, S., & Schuster, K. M. (2001). Evaluation of pedagogical quality in early childhood education: A cross-national perspective. *Journal of Research in Childhood Education*, 16(1), 109–124.
- *Sheridan, S., Giota, J., Han, Y., & Kwon, J. (2009). A cross-cultural study of preschool quality in South Korea and Sweden: ECERS evaluations. *Early Childhood Research Quarterly*, 24(2), 142–156.

- *Skouteris, H., McNaught, S., & Dissanayake, C. (2007). Mothers' transition back to work and infants' transition to child care: Does work-based child care make a difference? *Child Care in Practice*, 13(1), 33–47.
- Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2003). *Assessing quality in the early years: Early Childhood Environmental Rating Scale Extension (ECERS-E) four curricular subscales*. London: Trentham Books.
- *Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., & Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*, 21, 76–92.
- Tietze, W., Bethke, C., Braukhane, K., Roßbach, H., Schlecht, D., & Schmidt, N. (2001). *Analyse und Sicherung pädagogischer Qualität in den Tageseinrichtungen der Bremischen Evangelischen Kirche [Analysis and guarantee of pedagogical quality in daycare centers of the Bremish Evangelical Church]*. Berlin: Freie Universität.
- *Tietze, W., & Cryer, D. (2004). Comparisons of observed process quality in German and American infant/toddler programs. *International Journal of Early Years Education*, 12, 43–62.
- *Tietze, W., Cryer, D., Bairrão, J., Palacios, J., & Wetzel, G. (1996). Comparisons of observed process quality in early child care and education programs in five countries. *Early Childhood Research Quarterly*, 11, 447–475.
- *Tonyan, H. A., & Howes, C. (2003). Exploring patterns in time children spend in a variety of child care activities: Associations with environmental quality, ethnicity, and gender. *Early Childhood Research Quarterly*, 18, 121–142.
- *Torquati, J. C., Raikes, H. H., Huddleston-Casas, C. A., Bovaird, J. A., & Harris, B. A. (2011). Family income, parent education, and perceived constraints as predictors of observed program quality and parent rated program quality. *Early Childhood Research Quarterly*, 26, 453–464.
- UNICEF Innocenti Research Center (2008). *Report Card 8: The child care transition*. Florence: UNICEF Innocenti Research Center.
- *Van IJzendoorn, M. H., Tavecchio, L. W. C., Stams, G. J. J. M., & Reiling, E. J. (1998). Quality of center day care and attunement between parents and caregivers: center day care in cross-national perspective. *The Journal of Genetic Psychology*, 159, 437–454.
- Vandell, D. L., & Wolfe, B. (2000). *Child care quality: Does it matter and does it need to be improved?*. Madison: Institute for Research on Poverty.
- *Varin, D., Crugnola, C. R., Molina, P., & Ripamonti, C. (1996). Sensitive periods in the development of attachment and the age of entry into day care. *European Journal of Psychology of Education*, 11(2), 215–229.
- *Vermeer, H. J., Groeneveld, M. G., Larrea, I., Van IJzendoorn, M. H., Barandiaran, A., & Linting, M. (2010). Child care quality and children's cortisol in Basque Country and the Netherlands. *Journal of Applied Developmental Psychology*, 31(3), 339–347.
- *Vermeer, H. J., Van IJzendoorn, M. H., De Krijf, R. E., Funnink, R. G., Tavecchio, L. W. C., Riksen-Walraven, J. M., & Van Zeijl, J. (2008). Child care in The Netherlands: Trends in quality over the years 1995–2005. *The Journal of Genetic Psychology*, 169, 360–385.
- *Villalón, M., Suzuki, E., Herrera, M. O., & Mathiesen, M. E. (2002). Quality of Chilean early childhood education from an international perspective. *International Journal of Early Years Education*, 10(1), 49–59.
- Whitebook, M., Howes, C., & Phillips, D. (1989). *Who cares? Child care teachers and the quality of care in America: Final report: National Child Care Staffing Study*. Berkeley, CA: Child Care Employee Project.
- Whitebook, M., Sakai, L. M., & Howes, C. (2004). Improving and sustaining center quality: The role of NAEYC accreditation and staff stability. *Early Education and Development*, 15(3), 305–326.
- *Ying Hu, B., & Szentze, J. (2009). Exploring the quality of early childhood education in China: Implications for early childhood teacher education. *Journal of Early Childhood Teacher Education*, 30, 247–262.
- Zaslow, M., Tout, K., & Halle, T. (2011). Differing purposes for measuring quality in early childhood settings. Aligning purpose with procedures. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 389–410). Baltimore: Paul Brookes Publishing.