The when and what of measuring ECE quality: Analysis of variation in the Classroom Assessment Scoring System (CLASS) across the ECE day

Abstract

Observational assessment has become an integral component of the quality improvement agenda in Early Care and Education (ECE) and has been significant in directing attention to quality of interactions within the ECE setting. Understanding the functioning of observational assessment is increasingly important as the stakes are high. *Assessment outcomes* influence program funding. *Assessment criteria* direct educators to preference particular types of experiences for children, grounded in the assumptions that these criteria are those that best predict children's' school readiness, ongoing achievements and wellbeing. We examine two contextual factors that might affect assessment functioning: (1) *When*- the timing of assessment in the ECE day and (2) *What*- the content and format of the activities observed. We provide the example of the Classroom Assessment Scoring System (CLASS), analysing 11,341 observations cycles undertaken in a representative sample of 2,306 Australian pre-K (age 3-4 years) through Year 2 (age 7-8 years) classrooms. Our data show a generalised decline in instructional, organisational and emotional support across the ECE day (8am to 4pm) with recovery in emotional support at the end of the day. Within-classroom analyses demonstrate that whole group and small group formats and science, math, and social science content inflate, while meal times, physical activity, and transitions constrain CLASS scores. Separate analyses for pre-K classrooms showed similar patterns. We discuss the findings in terms of the purpose of assessment and suggest that particular times and events in the ECE day might serve as *barometers of quality*.

Keywords: Early childhood education; Quality assessment; CLASS; Measure reliability; Longitudinal multilevel modelling

Highlights

- Time matters- CLASS scores declined across the ECE day, emotional support recovered at end of day
- Format matters- whole and small group formats upward-bias Instructional Support scores
- Content matters- Educational content upward-bias CLASS scores, whilst care and health content downward- bias CLASS scores
- Specific times, formats and contents may act as *barometers of quality*

Introduction

Internationally, the strong body of evidence identifying the potential for Early Childhood Education (ECE) experiences to promote children's school readiness and ongoing educational achievements has driven a policy agenda focused on quality improvement (Organisation for Economic Co-operation and Development [OECD], 2017). Quality assessment of programs, has become an integral component of this agenda (Cannon, Zellman, Karoly, & Schwartz, 2017; Cohen & Goldhaber, 2016; Grant, Comber, Danby, Theobald, & Thorpe, 2018; Lingard, Martino, & Rezai-Rashti, 2013; Roberts-Holmes, 2015). For policy-makers, assessment is designed to provide assurance that the large public investment in ECE is being well directed. For provider organisations and services, assessment outcomes are used to identify areas of strength within a program and to direct attention to those components of the program that are targets for quality improvement. Positive assessments and associated ratings may also serve a marketing function especially in the competitive market of programs offered in the prior-to-school sector (e.g., Australian Children's Education & Care Quality Authority, Australia; Quality Rating and Improvement Systems, USA). For families and communities, publication of ratings is intended as a source of data to inform choice or identify inequities. The assessment process and attendant actions, therefore, are high stakes.

Funding decisions for ECE, whether directly or through the mechanism of parent choice, are increasingly driven by assessment outcomes (Cannon et al., 2017; Cleveland, Susman-Stillman, & Halle, 2013; Degotardi, Sweller, Fenech, & Beath, 2018; Mashburn, 2017; Yamamoto & Li, 2012). Content and nature of educator-child interactions in ECE programs may also be targeted to meet assessment criteria, and therefore impact children's experiences (Harms, Clifford, & Cryer, 2014; Pianta, Mashburn, Downer, Hamre, & Justice, 2008). Yet the significant faith placed in quality assessment tools is increasingly being

questioned (Mashburn, 2017; Pianta, Downer, & Hamre, 2016; Setodji, Schaack, & Le, 2018).

In this paper we seek to contribute to understanding of the functioning of quality assessment tools by examining contextual factors that may influence their outcome. We take the example of the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) and analyze data from a large representative sample of Australian classrooms, pre-K (age 3–4 years) through Year 2 (age 7–8 years) (N = 2,306 classrooms). There is a particularly strong policy focus on measuring quality in the years prior to school entry. However, methods of assessing interactional quality, and particularly the use of the CLASS observation examined in this paper, extend into the early school years. Our analyses therefore utilise all data available across ages 3–8 years to provide a comprehensive analysis of the quality assessment process. However, in response to the policy imperative to understand provision in the preschool years, we also conduct sensitivity analyses that evaluate the pre-K context separately.

We attend to two aspects of context that might influence reliability and attendant outcomes:

1. *When* observations are made: We ask if time of day systematically biases CLASS emotional, organisational and instructional scores.
2. *What* formats and content are observed: We ask if instructional formats and content areas systematically bias CLASS emotional, organisational and instructional scores.

Our findings have implication for classroom quality assessment practices, the use of assessment outcomes to inform policy and practice. Importantly, given that criteria used to assess quality of classroom practice can direct experiences for young children, our findings also have implications for the diversity and intensity of experiences provided for children in their ECE program.

Defining and Measuring ECE Quality

Two distinct bodies of research evidence have highlighted the potential of ECE programs to promote children's development and address social and educational inequalities. First, research from neuroscience has identified the mechanisms connecting early experience and neural development. Studies identify rich early interactional experiences as setting-down the neural architecture required for positive trajectories of learning and behaviour (Luby, Belden, Harms, Tillman, & Barch, 2016; Nelson, Zeanah, & Fox, 2019). Second, longitudinal and experimental studies have shown that, at least in the case of children from disadvantaged backgrounds, ECE programs *can* have positive effects on child outcomes (Melhuish et al., 2015). However, effectiveness is predicated on the quality of experiences provided, and not all ECE programs deliver positive child outcomes (Melhuish et al., 2015). Significant effort has been directed to understanding ECE quality and specifically the components of ECE programs that support positive development (Burchinal, Magnuson, Powell, & Hong, 2015). To date while our understanding of what distinguishes higher ECE experience has advanced considerably, what constitutes a *rich* or *high quality* ECE experience is at best partial, and effects, though consistently found, are low (Soliday Hong et al., 2019).

Theoretical perspectives on quality have long distinguished between structural features of an ECE program (staff ratios, group size, teacher qualification, hours of provision), and process features pertaining to interactions and relationships (Slot, Leseman, Verhagen, & Mulder, 2015; Thorpe, Cloney, & Tayler, 2010). Structural features are objective and therefore readily and reliably measured. However, relational environments have been found to better predict child outcomes in studies of ECE (Mashburn et al., 2008; Slot et al., 2015). Compared to the simple measurement of structural features, measuring interactional quality is a less certain science, however. Such measurement requires specification of behavioural criteria that index the quality of a child's experience in their ECE program. To be

generalizable to a population level, these criteria must be specified such that they can be consistently and reliably assessed across the diversity of ECE contexts and experiences. If they are to inform policy and practice, the criteria must also be measurable at scale requiring training of a large body of assessors who can both maintain fidelity to the observation criteria and provide feedback that serves to deliver practice change and effective outcomes for children.

A range of theoretically-driven observational measures has been developed for use at scale (Arnett, 1989; Harms et al., 2014; Pianta, La Paro, et al., 2008b). They have taken professional opinion and a diverse body of research findings to specify behavioural indices of ECE quality. Broadly, the measures draw from the parenting literature and incorporate constructs from attachment theory and neuroscience studies (e.g., responsiveness, warmth, consistency) and from learning and education theory (e.g., scaffolding, sustained shared thinking, organisation). Most measures adopt a global rating using either Likert scaling (e.g., Arnett, 1989; Colwell, Gordon, Fujimoto, Kaestner, & Korenman, 2013) or a hierarchical scoring process based on the presence or absence of behavioural indicators and structural criteria that are graded to represent degrees of quality on a broad diversity of criteria (e.g., Early Childhood Environment Rating Scale [ECERS]; Harms et al., 2014). While extremely detailed these measures to do not provide a time signal in the rating procedure.

The Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) is unique in providing a time signal that links quality scores to time of day, activity and pedagogical formats. CLASS requires a minimum of four sequential 30-minute cycles of observation (20 minutes) and scoring (10 minutes) to generate average quality scores that index the quality of the ECE day. Scores are collected for multiple cycles with simultaneous records of classroom activities. A common feature of the most frequently used large scale measures (CLASS [Pianta et al., 2008] and ECERS [Harms et al., 2014]) includes training in observation and

testing of fieldworker reliability through comparison with gold standard video-recorded interactions or gold standard assessors in the field. Similarly, both measures yield summary scores that generalise the quality of educator behaviour and classroom level interactions from the observed period.

High Stakes-Assessing ECE Quality

Across time, the dominant research tools used to measure ECE quality have been reified. That is, they have become synonymous with accepted understanding of quality (Hunkin, 2018; Rentzou, 2017). ECE quality assessment measures have been used in large scale, publicly funded research studies designed to inform policy decisions and practice interventions. These include decisions about program funding and investments in professional development. While the distinction has been made between made policy actions as 'high stakes" (e.g., Mashburn, 2017) while research is *low stakes* (e.g., Cohen & Goldhaber, 2016) we assert all of this work is high stakes. Research decisions about what constitutes ECE quality and the confidence with which this is asserted have underpinned uptake into policy and practice. Observation methods and attendant assessment scores used in quality rating improvement systems to categorise services as meeting or failing to meet a criterion are entirely founded in research. Their use reflects the considerable faith placed in quality assessment measurement tools both to capture the program elements that achieve positive child outcomes and to discriminate between services in the provision of these elements (Cannon et al., 2017). The consequences of assessment escalate the stakes from high (being defined as adequate or not) to higher when attendant funding of a service is terminated or considerable financial investments made to change practices such that they conform to the measure criteria (Mashburn, 2017; Pianta, Mashburn, et al., 2008). Highest of all stakes is the effect on children's experiences and ongoing development and learning outcomes. A key test of the value of quality assessment tools is the extent to which

interventions in response to an assessment outcome achieve improvement in children's experiences and developmental attainments.

To date the evidence is that the standard observation assessments in current use are poor predictors of child outcomes (Perlman et al., 2016). While longitudinal population cohort studies show statistically significant associations between quality assessments and child outcomes, the size of the effect is small. Emerging data from meta-analyses of intervention studies (a stronger test) identify important, but weak combined effect size of 0.14 or non-significant effects of improvement in standard quality assessments scores on child outcomes (Egert, Fukkink, & Eckhardt, 2018; Fukkink, Jilink, & Oostdam, 2017). Indeed, in a review published in the *Future of Children*, Pianta and colleagues (2016) describe the achievements in improving child outcomes at scale as 'ineffectual' (p.119). The disappointingly low impact of large-scale quality improvement interventions has initiated a range of challenging questions about the assumptions that underpin current assessments of ECE quality.

Examining the Assumptions of Quality Assessment: A Focus on When and What

Recent critiques of quality assessment in ECE have highlighted the need to establish if existing measures can safely guide policy and practice decisions (Mashburn, 2017; Setodji et al., 2018). A well-functioning assessment measure of ECE quality must meet the two fundamental psychometric assumptions of validity and reliability. Validity refers to the efficacy of measurement in truly capturing a construct, in this case ECE quality, and the extent to which the measure is effective in predicting intended child outcomes (see Reichardt, 2011). In lay terms, these might be thought as achieving adequate conceptualisation and precision in measuring ECE quality (*Coverage*) and in verifying this against predicted outcomes (*Consequence).* Reliability refers to the accuracy of measurement (*Credibility)* and its stability across time and context (*Consistency*). In the context of ECE assessments this

means a measure of ECE quality should function in the same way regardless of the individual using it or the time and context in which it is applied.

In establishing the value of any ECE quality assessment measure, attention to reliability and validity is requisite. To date, evidence of poor prediction of child outcomes (Egert et al., 2018; Fukkink et al., 2017; Pianta et al., 2016) and emerging studies identifying limitations of coverage (Farran, Meador, Christopher, Nesbitt, & Bilbrey, 2017; Pattinson, Staton, Smith, Sinclair, & Thorpe, 2014; Thorpe, Pattinson, Smith, & Staton, 2018) raise concerns about validity. They also suggest problems of reliability because if the content of a measure is theoretically sound, the problem may lie with the functioning of the test or testing procedure.

Exacerbating concerns about reliability and validity is the practice of modifying ECE quality assessment tools for application in policy and practice decisions (Buell, Han, & Vukelich, 2017; Mashburn, 2017; Setodji et al., 2018). Mashburn (2017), for example, examined the assumptions underpinning the use of CLASS in determining funding renewal for Head Start grantees. He reported limited evidence regarding the generalizability, and predictive validity of CLASS as used for this purpose. One particular concern raised was the divergence from standard CLASS protocols to reduce the duration of observation and exchange sequential observation for random time-sampling. In this example the observation duration is considerably reduced from the standard minimum of four observation cycles each of 20 minutes duration (2 hours) to two cycles each of at least 10 minutes duration (20 minutes) (Early Childhood Learning & Knowledge Center, n.d.). In random sampling of observation periods, the time of observation across samples and the format and content within the observation are not controlled. The underpinning assumption, therefore, is that quality assessments do not vary across time, format or content. These assumptions are tested in the current study in which we take a large representative population sample of classrooms,

assessed using CLASS (versions pre-K and K as appropriate). We test three hypotheses, relating to the when and what of ECE quality.

***When* assessments are made?**

H1: Time of observation systematically biases CLASS emotional, organisational and instructional scores. Our hypothesis is based on the extant evidence, comprising two studies, suggesting that time of assessment matters. Curby and colleagues (2011), based on a study assessing 693 publicly funded pre-K classrooms, identified relative stability in CLASS assessments across the first two hours of the day. They reported correlation coefficients for CLASS domains in the moderate to strong range (.52 - .77). While statistically strong, these associations relate to only two hours in the day and fall far short of unity. When a full day is studied, greater variability is reported. Available data are from 1,600 Year 3 and 5 classrooms (Curby et al., 2011) in the NICHD Study of Early Childcare and Youth Development. In this study the Classroom Observation System (COS), a measure utilising the same time-linked scoring system later used in CLASS, was the quality assessment measure. Results showed only moderate stability .38 - .68, In which start of day and transition times evidenced the lowest assessment scores (Curby et al., 2011). The current study builds on this evidence by examining the functioning of CLASS (Pianta, La Paro, et al., 2008b) in a large (N = 2,603 classrooms), representative sample of pre-K to Year 2 classrooms observed across the ECE day.

***What* formats and contents are observed?**

H2a There will be systematic bias in CLASS scores associated with instructional formats.

H2b There will be systematic bias in CLASS scores associated with instructional content.

Two studies are available that provide evidence on the association of CLASS assessment with teaching format and content. With regard to teaching formats, Cabell, DeCoster, LoCasale-Crouch, Hamre and Pianta (2013) conducted a study of teachers working in very low-income preschool classrooms (N = 314) and reported higher mean CLASS scores in large group formats and lower mean scores associated with free choice, meals, and routines. Curby and colleagues (2011) in their study of Year 3 and 5 classrooms report a similar finding in which group instruction yielded higher mean scores. With regard to content, both studies found that academic content was positively associated with higher mean scores. Cabell and colleagues' (2013) study of preschool classrooms specifically identified science content in large group or free choice format and literacy content in large group as yielding the highest mean global interaction scores. Our data of more than 11,000 CLASS observation cycles, using the standard format (20 minutes observation, 10 minutes recording) undertaken in a representative sample of 2,306 classrooms, advances understanding of context effects. Unique to our study is analysis of within-classroom variation of scores. Such analysis elucidates how changes in instructional format and content alter CLASS scores within classrooms. Further, we provide fine-scale coding of field notes to provide a fuller understanding of the content of the ECE day beyond the broad categories provided in the CLASS proforma.

Method and Analysis

Data

The data derive from the *Effective Early Education for Children (E4Kids)*, a study of N = 2,600 children recruited from ECE services when they were aged 3. The study protocol has been previously published (Tayler et al., 2016; Tayler, Ishimine, Cloney, Cleveland, & Thorpe, 2013). The sample, recruited in urban (Melbourne and Brisbane), regional (Shepparton, Victoria) and remote (Mt Isa, Queensland) locations using a social stratification

procedure is representative of Australian children attending ECE services at age 3, presenting an over-representation of parent employment but representation of the Australian population in all other demographic characteristics (Tayler, 2016). For all classrooms with a study child, observations using the Classroom Assessment Scoring System (CLASS; [Pianta, La Paro, & Hamre, 2008a; Pianta, La Paro, et al., 2008b]) were made longitudinally each year from 2010 – 2013 capturing the pre-K (including home-based day care, centre-based day care and stand-alone education programs) through Year 2 classroom experiences. Table 1 provides definitions of these programs for international comparison. Observations were undertaken by 92 research staff who were trained and certified as reliable both at initial training and through annual re-certification. Following CLASS protocol, certified reliability was expressed as being within one rating of the gold standard coder, with at least 80% agreement across all observations. In-field assessment of fieldworkers against a gold standard CLASS-coder was conducted in 2011 and agreement within one rating was high (96.4%; Cloney et al. 2017). Consistent with the CLASS manual, using Pre-K and K versions of CLASS as appropriate, observers completed ratings of Emotional, Instructional and Classroom Organisation supports.

During observations, records were made of instructional format (whole group, small groups, individual time, free choice, meals and snacks and routine) and content type (art, science, math, social studies, other and literature/language arts) across observation cycles. Field notes also provided greater detail of format and content. Consistent with the CLASS manual specification, sleep-rest times were not rated except for 130 rooms where a specific study was undertaken (Pattinson et al., 2014).

Sample selection.

For this study, we focused on a sample of 2,306 class groups. Following the CLASS protocol, each of the 2,306 class groups were observed on multiple occasions (CLASS

observation cycles) on a single day, yielding a total of 11,341 observations. All observations occurred in the first half of the Australian school year. While 4-6 CLASS observation cycles (2 - 3 hours) were available for most classrooms, interruptions in the scheduled day meant that for 197 (8.5%) classes only 3 valid cycles (1.5 hours) were available. These classes were retained in the analysis. Further details of the sampling process and summary statistics for variables used in the analyses are provided in the online supplementary material.

Coding format and content types. Observed content types and formats were not mutually exclusive with overlap evident for both content (25%) and format (52%) and were recoded to generate mutually exclusive categories to facilitate comparative analyses. The process and consequent changes in proportions are presented in the online supplementary material. In brief, there were three stages in the management of content coding. First, because content type had an 'other category', whenever *other* occurred with only one other content type, only the specified content type was included. Second, when *other* was an exclusive category (n = 3,555) new codes were derived with reference to the field note activities. The work was undertaken by an early years pedagogy specialist, with reliability checks undertaken independently. The 'other' category was thus recoded to include transitions, play, digital technology, gross motor, food, rest and relaxation, learning centres and music. Finally, for all remaining cases when there was more than one content type concurrently listed, this was coded as 'multiple content'.

For format, recoding was focused on the 'whole group' category as this was the most common category with additional format types indicated. A coding of 'whole group' was frequently used to indicate that all children were engaging in a format; thus, this code was removed when it occurred with one additional format type. Further, as coding of both whole- and small- group format types concurrently represented a large proportion of format coding,

an additional whole/small group coding category was created. All remaining format combinations were collapsed into an 'other format' category.

Analysis

Measurement properties. The reliability of the CLASS scores over time was assessed using single factor, congeneric Confirmatory Factor Analyses (CFA) in R (version 3.4.2; R Core Team, 2017) using the lavaan (version 0.6-2; Rosseel, 2012) and semTools (version 0.5-1; Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018) packages. These CFA's were tested for longitudinal measurement invariance (Liu et al., 2017), within each classroom in each assessment year to assure that CLASS had been adequately measured. That is, longitudinal measurement invariance indicates if the factor structure of the measure is the same for each observation. Longitudinal measurement invariance would indicate, for example, that the measures of CLASS in the morning and afternoon have the same factor structure. If the measures have the same factor structure they can then be reliably compared. Single factor structure was relevant because later analyses use CLASS factors independently. The longitudinal invariance testing was carried out for rooms with three (all), four (91%), five (72%), and six (28%) observations. In cases where an observation cycle did not have the range of other cycles for an indicator (e.g., 2-7 vs. 1-7) the lower scores were increased by one until all observation cycles had the same range (e.g., all 2-7).

Variables were specified as ordinal and the Robust Weighted Least Squares (WLSMV) estimator (Flora & Curran, 2004) was used and fit statistics were scaled using the second order corrected χ_2 (Asparouhov & Muthen, 2010; Muthén, du Toit, & Spisic, 1997; Satorra, 2000). Adequate fit was achieved when CFI $\geq$ .95, RMSEA $\leq$ .06 and SRMR $\leq$ .08 (Schreiber, Nora, Stage, Barlow, & King, 2006). Changes in CFI < .01, RMSEA < .015 and SRMR < .015 between nested models were taken to indicate invariance (Chen, 2007; Putnick & Bornstein, 2016), though there is debate on the best criteria (Putnick & Bornstein, 2016).

Following Putnick & Bornstein (2016), the invariance models tested were configural (same factor loading patterns – and thresholds due to ordinal data), metric (equivalent loadings; weak), scalar (equivalent item intercepts; strong) and residual (equivalent item residuals; strict).

When. Multilevel random effects linear models were fitted in R using maximum likelihood estimation via the lme4 (version 1.1-17; Bates, Mächler, Bolker, & Walker, 2015), lmerTest (version 3.0-1; Kuznetsova, Brockhoff, & Christensen, 2017), and merTools (version 0.4.1; Knowles, Frederick, & Whitworth, 2018) packages to identify particular times of day that distinguished emotional, organisational and instructional quality. Sequential model fitting to determine what pattern best explained CLASS scores proceeded as follows:

1) Models predicted the three CLASS subscale scores with random intercepts for each classroom year grouping (e.g., room 1 year 2010, room 1 year 2011, room 2 year 2010) compared to a model that had only an overall intercept. This was to establish if there is a differences in the average CLASS score between classrooms and between the same classrooms in different years. We used classroom year grouping because a different teacher could be in each room in each year.
2) Models predicted the three CLASS subscale scores with random intercepts for each classroom year grouping and each fieldworker year grouping. This was to establish if some fieldworkers were systematically associated with higher or lower CLASS scores. This would suggest, and partially account for, possible observer bias.
3) Models predicted the three CLASS subscale scores with random intercepts for each classroom year grouping, each fieldworker year grouping and fixed slope for mean-centred time of day in hours. This was to determine if CLASS scores changed across the day. Time was mean-centred (subtract the grand mean of time

of day from each observed time of day) to assist in the interpretation of coefficients and improve polynomial term calculation (referenced below).

4) Models included random slopes for each classroom year grouping for mean-centred time of day. This was to establish if classrooms had unique trajectories of CLASS across the day (e.g., some may increase in quality and some may decrease).

5) Models added quadratic and cubic (if quadratic was significant) fixed slopes for mean-centred time of day. The quadratic and cubic terms were included to evaluate if CLASS scores changed in a non-linear manner across the day (e.g., started high, dropped at midday, increased in the afternoon).

6) Models included random slopes for quadratic and cubic (if quadratic terms varied sufficiently) mean-centred time of day if they were significant ($p < .05$) in step five. This was to establish if classrooms had unique non-linear trajectories across the day.

Time of day was centred to reduce correlations between polynomial terms whilst improving interpretability of linear growth (Schielzeth, 2010). Fixed terms (time) were kept if they were statistically significant from zero ($p < .05$). Log-likelihood ratio tests, Akaike Information Criterion (AIC; Akaike, 1973) and Bayesian Information Criterion (BIC; Schwarz, 1978) comparisons and caterpillar plots (Knowles et al., 2018) were used to evaluate random effect variance components. The log-likelihood ratio test compares the difference in model fit of a model with and without a random effect variance component. If there is a significant improvement ($p < .05$) this suggests including the random effect variance may be useful. Likewise, a relatively meaningful reduction (e.g., > 10; Burnham & Anderson, 2003; Raftery, 1995) in AIC or BIC between models suggests improved model fit.

AIC compares parameters independent of sample size, and is more useful when we are concerned with avoiding false negatives (type II error). BIC penalizes models with larger sample sizes, and is more useful when we want to avoid false positives (type I error). *Caterpillar*, or *forest*, plots place simulated empirical Bayes estimates of individual random effects coefficients on the *y-axis* ordered from lowest to highest along the *x-axis* (1 to *n*) with 95% confidence intervals. If the vast majority of 95% confidence intervals overlap a line centred on zero on the *y-axis* it gives an indication that the estimated random effects may be statistically unreliable and adding variation to the coefficient may not be a relevant improvement to the model. For example, if only one random slope for time can be reliably estimated as statistically different from zero for 2,306 classrooms, adding random slopes for all classrooms may add unnecessary complexity.

What. Hybrid multilevel random effects linear models (Allison, 2009) were fitted in R using the aforementioned packages to determine if either instructional content or format distinguished emotional, organisational and instructional quality. In these models, time varying covariates (format and content) are included as within-classroom averages and within-classroom deviations from the average. Thus, the average terms capture differences between classrooms that on average had more of a particular content or format, and as such is also susceptible to bias from unobserved characteristics, whilst the deviation or change terms captures what happens to the CLASS score within a classroom when the content or format changes, making the change term more reliable. In these models, we estimated the effects of each content type and format separately. As the ratio of children to adults may affect classroom management, we controlled for the number of children and adults present during observation cycles. In addition, as lesson formats across grades may differ, and therefore influence scores on CLASS domains, we also controlled for school grade level. A random intercept and random slopes for mean centred time of day for each classroom year grouping

was included, as well as a random intercept for each fieldworker year grouping and a fixed coefficient for quadratic mean centred time of day (and cubic for classroom organisation) based on the findings from analysis part 2. These models were also run separately for children not yet in formal school (including centre-based and home-based childcare and education programs).

Model diagnostics: When and what. Diagnostic plots were examined in R using the *qqmath* function of the lattice package (version 0.20-35; Sarkar, 2008) to evaluate normality of the residuals and validate assumptions of multilevel regression models (Pinhero & Bates, 2000).

Results

Measurement properties

Tests of longitudinal measurement invariance across the ECE day and CFA results are presented in Table 2 for the first three observation cycles. These show that the CLASS measurements adhered to the factor structure and this generally did not change with observation cycle. Specifically, it can be concluded that Emotional Support (ES), Instructional Support (IS) and Classroom Organisation (CO) have strict measurement invariance between observations based on changes in the CFI (< .010), RMSEA (< .015) and SRMR (< .015). The one exception was that for IS the SRMR metric invariance test failed marginally with a change of .017 when a change less than .015 was required. Considering the CFI changed < .010 and RMSEA changed < .015, this result was taken as acceptable. All other observation cycles (four, five, and six) comprehensively passed tests of invariance, presented in the online supplementary material. The average Cronbach α was; ES = 0.69; IS = 0.83; CO = 0.72.

When

The estimated trajectories best describing CLASS across the day were from models with a random intercept for classroom year group (e.g., room 1 2010; room 1 2011; room 2 2010), random intercept for fieldworker year group and random slope for time in each classroom year group. The overall trajectory of scores across time followed a quadratic trend for IS ($\beta = -0.102$, $p < .001$, $\beta_2 = 0.01$ $p = .02$) and ES ($\beta = -0.079$, $p < .001$, $\beta_2 = 0.019$, $p < .001$), whilst CO followed a similar cubic trend ($\beta = -0.070$, $p < .001$; $\beta_2 = 0.023$, $p < .001$; $\beta_3 = -0.005$, $p = .002$). The rules concerning these model results are mentioned in the analytical strategy. Specifically, the best models for ES, IS, and CO were found at step 5. These trends saw the highest CLASS scores in the morning, which dipped in the middle of the day and returned for ES (Figure 1A), stabilised for IS (Figure 1B) and stabilised and then fell for CO (Figure 1C). Adding random variation in the quadratic trend for each classroom (step 6) improved model fit in terms of AIC and χ_2 difference test for ES, CO and IS, and BIC for ES, but the amount of additional variance explained was small and caterpillar plots revealed few reliable estimates, so this additional complexity was removed. Summary data tables for these models are provided in the online supplementary material.

What

Changes in teaching format affected the CLASS scores. These effects are summarised in Figure 2 and Table 3 and presented in the online supplementary material. Results suggest that meals and snacks, individual time, routine and other formats downward bias CLASS scores. Conversely, whole group, free choice centres, small group and small and combined groups bias CLASS scores upward. These effects are interpreted in Table 3.

Effects of the average format on CLASS, compared to change terms, were generally in the same direction or unreliably estimated ($p > .05$, based on overlapping 95% confidence intervals). Exceptions included (1) meals and snacks which were associated with lower effects on ES ($\beta = -0.465$); (2) the ‘other’ category which was associated with higher effects

on CO (β = 0.203) and IS (β = 0.057); (3) free choice centres which were associated with lower effects on IS (β = -0.243); and (4) whole group formats which were associated with lower effects on IS (β = 0.039).

Changes in content type influenced CLASS scores and these effects are summarised in Figure 3 and Table 4 and the online supplementary material. These effects indicate food, transitions, gross motor, rest and relaxation, digital technology and the remaining multiple content category downward bias most domains of CLASS scores, whilst science, social studies, math, art and literature/language art upward bias most CLASS domains. Play and music upward bias some aspects of CLASS and downward bias others. Directional bias of learning centres was less apparent. These effects are presented in Table 4.

Effects of average content type on CLASS, compared to change terms, were overwhelmingly in the same direction or unreliably estimated ($p > .05$, based on overlapping 95% confidence intervals). The exception is that the multiple content type was associated with higher effects on CO (β = 0.095).

Model diagnostics: When and what

Residual diagnostics plots revealed no major departures from the assumptions of linear models, however, they indicated lower CLASS scores tended to be overestimated and higher scores underestimated. Because IS is skewed to 1 and CO is skewed to 7 (skewed in opposite directions) they had similar trends in that higher IS had the most pronounced underestimation and lower CO had the most pronounced overestimation. On the other hand, ES tended to have a split of higher underestimation and lower overestimation.

One way to reduce the above heterogeneity of residuals is to model the square of each outcome (e.g., ES^2; Fox 2015). Therefore, to determine if an improved calibration to linear model assumptions was inconsequential, models were refit with ES and CO squared and IS reverse scored and squared. Additionally, the random effect of time was removed as high and

low slopes also tended to be under and overestimated, respectively. These changes substantially improved how well the models met the linear model assumptions. Further, there were no changes in our substantive inferences. Two minor changes in statistical significance emerged: ES had a statistically significant cubic function in the growth model component ($p = .044$); the effect of social studies on CO was no longer statistically significant (changing from $p = .042$ to $p > .05$). There were minor variations in comparative differences of average and change terms. Given that interpreting coefficients for squared outcomes is difficult, initial analyses that uses outcomes as scored are reported. Analysis of squared outcomes can be requested from the authors.

General covariates

Number of children in a room

An increase (change) in the number of children within an individual classroom was associated with a significant lowering of IS (β range -0.014 to -0.009, max $p < .001$), CO (β range -0.011 to -0.008, max $p < .001$) and ES (β range -0.006 to -0.004, max $p < .001$) in all estimated models. A higher average (over the day) number of children across classrooms was also associated with lower IS (β range -0.009 to -0.007, max $p = .027$), whilst effects on CO and ES were unreliably estimated ($p > .05$). The direction of these findings were replicated using squared outcomes.

Number of Adults

Increases in the number of adults, in all models, was associated with significantly higher CO (β range 0.033 to 0.041, max $p < .001$) and ES (β range 0.019 to 0.022, max $p = .005$), whilst effects on IS were unreliably estimated ($p > .05$). A higher average number of adults was also associated with higher ES (β range 0.047 to 0.052, max $p = .01$), whilst effects on CO and IS were unreliably estimated ($p > .05$). The same directions for these findings emerged using squared outcomes.

Service Type

Analyses of variation by service type are presented in Table 5. These findings are summarised across the models used to estimate effects of format and content and show two important trends. First, within the pre-K year (age 3-4 years) stand-alone education programs had higher CLASS scores than day care provision in the home or at a centre. Second, across year grade, reflecting increasingly formal approaches to learning, there was an increase in IS scores, but a commensurate decline in ES.

Sensitivity of results to using only classrooms younger than school (pre-K)

In general, results were very similar using the subset of classrooms before children entered formal schooling. Variations are outlined in the online supplementary material as they do not influence overall conclusions.

Sensitivity of results analyzing prior-to-school classrooms only

Results for comparisons of full-sample data and that for the prior-to-school sample only (Pre-K) are presented in the online supplementary material. In summary, there were few substantive differences. There were no significant differences in format effects and only one substantive finding (once frequency was accounted for) in content in which Art content increased IS scores in Pre-K settings, but not in school settings.

Discussion

ECE programs offer significant potential to support children's futures, but realisation of positive effects is dependent on the emotional *richness* and instructional qualities of the experiences they provide. For this reason assessment of ECE quality across the entire range of measure application is high stakes. Confidence in measures of ECE quality and the research findings they have generated have been central to policy and practice decisions that impact funding, educator training and the experiences provided for children. Yet a growing body of studies identify concerns about the reliability and validity of ECE quality assessment

measures. Despite strong theoretical underpinnings, meta-analysis of study outcomes across a range of correlational and intervention designs at scale show negligible predictive validity (Egert et al., 2018; Fukkink et al., 2017). One explanation for these findings is that the constructs assessed are not valid measures of quality. Alternatively, given the strong theoretical foundations of the measures, explanations for these disappointing findings may relate to problems with reliability of the measures and the circumstances of their application. Modifications to standard measure protocols (Mashburn, 2017) or measure usage (Setodji et al., 2018) in policy and practice applications have exacerbated concern.

In this study we examined functioning of one extensively used quality assessment measure, CLASS (Pianta, La Paro, et al., 2008b), and focused on reliability. We assessed the potential biasing effects of time of assessment and the content and formats of observation. Our results indicate that time of day matters. Ratings of instructional support and classroom organisation decline across the day, while emotional support first declines then recovers at end of day. Our results also show that format and content matter. Variations in content and format systematically biased CLASS scores.

While our example focused on CLASS, our results present a number of challenges that extend beyond this specific measure to the use of quality assessment measures in policy and practice interventions more broadly. Our findings also present directions for new approaches in measurement applications and present possibilities for interventions that ensure each child experiences *rich* interactions across the diversity of activities in the ECE day.

Reliability of quality assessment measures and application in policy and practice

Assessments outcomes based on standard measures such as ECERS and CLASS have been used to determine whether a service meets a determined minimum standard and/or to grade services using a rating system (e.g., Australia; Australian Children's Education & Care Quality Authority [ACECQA], 2018; Cannon et al., 2017). Failure to meet the specified

standard, or achieving a low grade, incurs reputational and financial costs whether through funding termination, reduced attendances or investment in program remediation (Cannon et al., 2017; Cleveland et al., 2013; Degotardi et al., 2018; Mashburn, 2017; Yamamoto & Li, 2012). High assessments outcomes are used for marketing (e.g., ACECQA, n.d.; Cannon et al., 2017). In the Head Start grant renewal designation, random time sampling applying CLASS was used to assess service quality (Mashburn, 2017).

The findings of the current study raise concerns about the practical application of the assessment measures. Our findings, like those of two prior studies (Curby et al., 2011), show that time matters. Assessments undertaken at different times of day therefore, are not equally comparable. Our evidence suggests assessments undertaken later in the day present risk of underestimating quality with attendant adverse policy implications. There are three key candidate hypotheses we propose to explain this. First, the observed decline in ratings of instructional support and classroom organisation may reflect general patterns of educators and/or children fatigue across the day. Second, the diurnal decline in scores may reflect deliberate decisions regarding the programing of content and format in response to child fatigue. The marginal increase in emotional climate at the end of the day could also reflect programming choices. Our data shows that less cognitively challenging contents and formats are often accompanied by a rise in emotional support. Third, the observed average trend could reflect sampling variability as fewer classrooms were observed earlier and later in the day, few rooms were observed across the entire day, and format and content fluctuate with time of day. Further, the best trajectory to describe individual classrooms varied from the proposed average trend line (e.g., random slopes were present). This highlights the need to specifically test timing effects in isolation of other explanations, for example, by standardising for format and content, observing rooms for the entire day, and observing the same room across multiple days.

A key message for our findings is that different formats and contents are not comparable. Random time-sampling is a particular concern as this may bias quality assessment dependent on the content or format sampled. Sampling of an educationally focused format and content such as a whole group science lesson, for example, will favour the assessment outcome. Whereas sampling health and care content and formats such as outdoor physical activity and mealtimes will present a less-favourable quality assessment. These outcomes are not benign and can cause harm. Termination of a program's funding affects the continuity of children's care and the lives of those invested in providing an ECE program. Furthermore, workforce morale, stress and loss to the sector are significant problems (Corr, Cook, LaMontagne, Davis, & Waters, 2017; Grant et al., 2018; Grant, Danby, Thorpe, & Theobald, 2016; Roberts-Holmes, 2015).

ECE quality assessments undertaken across a full day should be noted. While safer then random sampling and more extensive than a 2-hour sequential time sampling, they are subject to the assumption that the contents of each ECE day are standard and, therefore, that all ECE days are comparable. Yet a study across multiple days in a single classroom applying four different observation measures showed substantial daily variation that related to the composition of the class group and variation in activities (Rentzou, 2017). Further, as observations are time consuming, labour intensive and expensive they may not be feasible at scale (Rentzou, 2017). For this reason, short-duration time sampling (particularly in large scale assessments) may be necessary. Our data suggest that standardisation or adjustments for time of day, format and content of an observation likely present a more reliable and fair method of assessment. Such an approach raises a further question: What formats and contents should be selected to assess quality?

Identifying 'Barometer events' in assessment of ECE quality

Systematic variation in CLASS scores associated with the format and content of an observation cycle has been previously reported (Cabell et al., 2013; Curby et al., 2011). In this study, like its predecessors, whole group educationally focused content was found to positively bias CLASS scores while health and care content such as outdoor play and mealtimes presented negative bias. A literal interpretation of our finding would be an intervention to improve assessed quality in which educational content and large group formats are increased in place of health promoting and care content. In reality, the body of evidence from both neuroscience and education-focused studies infers that the ECE day should be rich and diverse and afford each child the opportunity to choose a range of activities. The ECE day necessarily includes care activities (meals, naps and toileting) and transitions. We therefore propose that our results should direct attention to increasing the *precision of quality assessment* and attendant interventions. Specifically, the timing of observation and the format and content of activities observed should consider the purpose of the quality assessment; including both making comparison across ECE settings and identification targets for improvement within an ECE setting.

Based on our findings that identify some contents as potential *barometers of quality*, two strategies are suggested. First, when the focus problem is comparison of classrooms or against a benchmark (e.g., quality rating improvement system) a standardised approach might be used. Here observations would focus on specific activities of interest including both care and educational activities. This approach reduces some biases and enhances comparability. However, use of a barometer events approach in context, requires careful consideration of the selection of activity types with regards to specific domains of interest. Second, and particularly pertinent in informing intervention, observations might focus on those parts of the ECE day that present the greatest challenge to educators in attaining higher quality assessments within specific domains of interest. Again, while such an approach may help to

guide selection of observational context for assessment, and enhance comparability across settings, careful consideration of the types of activities selected in relation to the domains/s of interest, goal of assessment and program types (e.g., activities across prior to school versus school based settings) is required.

Evidence is emerging that suggests targeting *barometer events* may be an effective method of both assessing and improving quality. This evidence comes from bottom-up studies in which variation in practices were mapped to quality assessment scores (Pattinson et al., 2014) and child outcomes (Thorpe et al., 2018). For example, Pattinson and colleagues (2014) found that CLASS emotional support scores across a naptime observation predicted emotional support scores for the rest of the assessed ECE day. Classrooms in which non-sleepers were required to lie down for long periods without alternative activity (> 60-minutes) were also those that had low emotional support scores outside the naptime. Those classrooms with more flexible practices were also those achieving higher emotional support scores outside of naptime, with a linear association between degree of flexibility in naptime and emotional support measured across the ECE day. Connecting these findings to child outcomes a further study showed effects of naptime practices on child stress indexed using diurnal cortisol patterns (Thorpe et al., 2018). Extending this finding to Instructional support, studies of sleep-rest times found that children who do not sleep during rest time spend an average of 60 minutes and up to 190 minutes lying down without instructional activity (Staton, Smith, Hurst, Pattinson, & Thorpe, 2017) or opportunity for positive learning experience (Gehret, Cooke, Staton, Irvine, & Thorpe, 2019). More powerful are intervention studies showing a pathway from change in educator behaviour to residual gains in child outcome, with focus on effect size that demonstrate substantive outcome (e.g., effect size > 0.2; Farran et al., 2017). Farran and colleagues (2017), in a three-year partnership with pre-K educators took this approach. Through observation, eight key areas of challenge for

educators, the *Magic 8*, were identified. The *Magic 8*, were then targeted for quality improvement through coaching support and the effect of intervention on child academic and social emotional outcomes was assessed. Consistent with the current study and prior studies (Curby et al., 2011), the *Magic 8* included transitions. Observation of transitions identified practices in which children spent unnecessary time without activity. Identifying this problem instigated actions to reduce the duration of transitions and was among the changes associated with improvement in children's outcomes with moderate to high effect. Such emerging examples suggest a new direction not only for quality assessment but for improving the richness of child experiences.

Providing rich interactional experiences for children attending ECE programs

Quality assessment measures are used to assess teacher behaviour, providing a generalised summary of the classroom environment across an observation period generally limited to a single day. The limited capacity of these measures to substantively predict child outcomes (Egert et al., 2018; Fukkink et al., 2017; Pianta et al., 2016) and the identification of systematic bias within the measurement, while disappointing, present an opportunity to consider new approaches to quality assessment. Neuroscience research has identified responsive and caring relationships, not content alone, as the mechanism supporting neural development (Luby et al., 2016; Nelson et al., 2019).

Teaching is profoundly relational. Content and formats are not only determined by the educator, but also occur in response to children. In this way, the individual needs, abilities and interests of children within a class play a role in defining the work of the educator (Buell et al., 2017; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014). Thus a growing body of work documents that both size and composition of a class group constrain the educator's possibility of attaining high assessment scores (Buell et al., 2017; Cohen & Goldhaber, 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014). Proportion of children

in a classroom with a disability (Soukakou, 2012), behavioural problems (Skalická, Belsky, Stenseng, & Wichstrøm, 2015), from racial minorities (LoCasale-Crouch et al., 2007), from low socioeconomic backgrounds (Reid & Ready, 2013), who are boys (Buell et al., 2017), or of varied age (Moller, Forbes-Jones, & Hightower, 2008) negatively affect quality assessment ratings. The educator's task, and attainments of quality assessment scores, varies across years, and is dependent on the children in the classroom (Buell et al., 2017; Cohen & Goldhaber, 2016).

The current study demonstrates that the content and formats of ECE influence quality assessment. However, to date, the input of children in quality assessment has been largely overlooked. Quality may well be best seen in conditions of challenge that relate to equitable delivery of rich environments across the diversity of children attending ECE programs. To date only one standard measure, ECERS-E (Sylva, Siraj-Blatchford, & Taggart, 2010), attends directly to diversity. Our data also direct attention to a focus on care and health as barometer events in which the quality of interactions could be significantly improved.

Strengths and Limitations

In this study we have analysed a large and representative sample of Australian ECE programs across Pre-K to Year 2, with detailed attention to limiting potential sources of coding and statistical error. While the generalizability of these data beyond the Australian context is a potential limitation, available evidence suggests our sample yields comparable scores on ECERS and CLASS assessment to those in UK and USA samples (Tayler et al., 2013). Our study addressed the policy imperative for quality improvement in preschool settings and the use of the CLASS measure. In addressing this question our strategy was to use all CLASS data with sensitivity analyses to examine variation for the subset of Pre-K observations. That findings were substantively similar indicates common measurement

functioning across the early years of education. We acknowledge, however, that the quality improvement policy focus may be different in the school sector.

Our study is subject to a range of limitations. First, while adhering to the standard CLASS protocol of multiple observations, our study was limited to observations within a single day in each classroom. Other studies indicate variations in quality assessment across days and times of year (Buell et al., 2017; Plank & Condliffe, 2013). Second, while fieldworkers met the CLASS reliability criteria, consistent with prior reports (Cohen & Goldhaber, 2016; Mashburn, 2017), analyses in the current study showed evidence of fieldworker bias. Fieldworker effects were statistically controlled within this study; however, understanding the extent and cause of fieldworker bias warrants further detailed exploration as these affect assessment outcomes. Third, in categorising the content and format of CLASS, fieldworkers endorsed the categories provided in the record form. However, manual coding of fieldwork records identified a range of contents that fell outside those provided. We responded by using field notes to reclassify. Our experience identifies the need for a more comprehensive set of content categories and clear definitions of each of these. Fourth, adherence to the standard protocol means that some aspects of the ECE day, most notably naptime, were not assessed, yet may well be important indicators of quality (Pattinson et al., 2014) and child outcomes (Thorpe et al., 2018). Finally, further exploration of content validity is a focus for future research. Predictive validity remains a key issue. Our data provide longitudinal standard assessments and data linkage to ongoing educational attainments and behavioural indices (school behaviour, behavioural sanctions, and crime) that will provide opportunity to explore the predictive capacity of barometer events.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Allison, P. D. (2009). *Fixed effects regression models*. Sage Publication. https://doi.org/10.4135/9781412993869

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*, *10*(4), 541–552. https://doi.org/10.1016/0193-3973(89)90026-9

Asparouhov, T., & Muthen, B. (2010). *Simple second order chi-square correction*. Retrieved from http://statmodel2.com/download/WLSMV_new_chi21.pdf

Australian Children's Education & Care Quality Authority. (n.d.). Promote your rating. Retrieved from https://www.acecqa.gov.au/assessment/promote-your-rating

Australian Children's Education & Care Quality Authority. (2018). *Guide to the national quality framework*. Retrieved from https://www.acecqa.gov.au/sites/default/files/2018-11/Guide-to-the-NQF_0.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Buell, M., Han, M., & Vukelich, C. (2017). Factors affecting variance in Classroom Assessment Scoring System scores: Season, context, and classroom composition. *Early Child Development and Care*, *187*(11), 1635–1648. https://doi.org/10.1080/03004430.2016.1178245

Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (2015). Early childcare and education. In R. Lerner (Ed.), *Handbook of Child Psychology and Developmental Science* (7th editio, pp. 1–45). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118963418.childpsy406

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference : A practical information-theoretic approach*. New York, NY: Springer New York.

Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly*, *28*(4), 820–830. https://doi.org/10.1016/J.ECRESQ.2013.07.007

Cannon, J., Zellman, G., Karoly, L., & Schwartz, H. (2017). Quality rating and improvement systems for Early Care and Education programs: Making the second generation better. RAND Corporation. https://doi.org/10.7249/PE235

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cleveland, J., Susman-Stillman, A., & Halle, T. (2013). *Parental perceptions of quality in early care and education*. Bethesda, MD: Child Trends Publication.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

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*(2), 218–233. https://doi.org/10.1016/j.ecresq.2012.12.004

Corr, L., Cook, K., LaMontagne, A. D., Davis, E., & Waters, E. (2017). Early childhood educator mental health: Performing the "National Quality Standard." *Australasian Journal of Early Childhood*, *42*(4), 97–105. Retrieved from https://search.informit.com.au/documentSummary;dn=355099924274238;res=IELHSS

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., … Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *The Elementary School Journal*, *112*(1), 16–37. https://doi.org/10.1086/660682

Degotardi, S., Sweller, N., Fenech, M., & Beath, A. (2018). Influences on parents' child care choices: A comparative analysis of preschool and long day care users. *Child and Youth Care Forum*, *47*(5), 683–700. https://doi.org/10.1007/s10566-018-9452-3

Early Childhood Learning & Knowledge Center. (n.d.). 1304.16 Use of CLASS: Pre-K instrument in the Designation Renewal System. Retrieved February 8, 2019, from https://eclkc.ohs.acf.hhs.gov/policy/45-cfr-chap-xiii/1304-16-use-class-pre-k-instrument-designation-renewal-system

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research*, *88*(3), 401–433. https://doi.org/10.3102/0034654317751918

Farran, D. C., Meador, D., Christopher, C., Nesbitt, K. T., & Bilbrey, L. E. (2017). Data-driven improvement in prekindergarten classrooms: Report from a partnership in an urban district. *Child Development*, *88*(5), 1466–1479. https://doi.org/10.1111/cdev.12906

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466.supp

Fukkink, R., Jilink, L., & Oostdam, R. (2017). A meta-analysis of the impact of early childhood interventions on the development of children in the Netherlands: An inconvenient truth? *European Early Childhood Education Research Journal*, *25*(5), 656–666. https://doi.org/10.1080/1350293X.2017.1356579

Gehret, H., Cooke, E., Staton, S., Irvine, S., & Thorpe, K. (2019). Three things I learn at sleep-time: children's accounts of sleep and rest in their early childhood education programs. *Early Years*, 1–18. https://doi.org/10.1080/09575146.2019.1634010

Grant, S., Comber, B., Danby, S., Theobald, M., & Thorpe, K. J. (2018). The quality agenda: Governance and regulation of preschool teachers' work. *Cambridge Journal of Education*, *48*(4), 515–532. https://doi.org/10.1080/0305764X.2017.1364699

Grant, S., Danby, S., Thorpe, K., & Theobald, M. (2016). Early childhood teachers' work in a time of change. *Australasian Journal of Early Childhood*, *41*(3), 38–45. Retrieved from https://search.informit.com.au/fullText;dn=461462610406178;res=IELHSS

Harms, T., Clifford, R. M., & Cryer, D. (2014). *Early Childhood Environment Rating Scale* (3rd ed.). Teachers College Press. Retrieved from https://www.tcpress.com/ecers-3-early-childhood-environment-rating-scale-9780807755709

Hunkin, E. (2018). Whose quality? The (mis)uses of quality reform in early childhood and education policy. *Journal of Education Policy*, *33*(4), 443–456. https://doi.org/10.1080/02680939.2017.1352032

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A., & Rosseel, Y. (2018). Package "semTools": Useful tools for Structural Equation Modeling. Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/web/packages/semTools/semTools.pdf

Knowles, J. E., Frederick, C., & Whitworth, A. (2018). MerTools: Tools for analyzing mixed effect regression models [R package merTools version 0.4.1]. Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/package=merTools

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lingard, B., Martino, W., & Rezai-Rashti, G. (2013). Testing regimes, accountabilities and education policy: Commensurate global and national developments. *Journal of Education Policy*, *28*(5), 539–556. https://doi.org/10.1080/02680939.2013.820042

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. https://doi.org/10.1037/met0000075

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., … Barbarin, O. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly*, *22*(1), 3–17. https://doi.org/10.1016/J.ECRESQ.2006.05.001

Luby, J. L., Belden, A., Harms, M. P., Tillman, R., & Barch, D. M. (2016). Preschool is a sensitive period for the influence of maternal support on the trajectory of hippocampal development. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(20), 5742–5747. https://doi.org/10.1073/pnas.1601443113

Mashburn, A. J. (2017). Evaluating the validity of classroom observations in the Head Start Designation Renewal System. *Educational Psychologist*, *52*(1), 38–49. https://doi.org/10.1080/00461520.2016.1207539

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., … Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749. https://doi.org/10.1111/j.1467-8624.2008.01154.x

Melhuish, E. C., Ereky-Stevens, K., Petrogiannis, K., Ariescu, A., Penderi, E., Rentzou, K., … Leseman, P. (2015). *A review of research on the effects of Early Childhood*

Education and Care (ECEC) upon child development. CARE project; Curriculum Quality Analysis and Impact Review of European Early Childhood Education and Care (ECEC).

Moller, A. C., Forbes-Jones, E., & Hightower, A. D. (2008). Classroom age composition and developmental change in 70 urban preschool classrooms. *Journal of Educational Psychology*, *100*(4), 741–753. https://doi.org/10.1037/a0013099

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes (technical report). Los Angeles, CA: University of California. Retrieved from http://www.scielo.br/scielo.php?script=sci_nlinks&ref=000173&pid=S0102-79722013000200003000039&lng=en

Nelson, C. A., Zeanah, C. H., & Fox, N. A. (2019). How early experience shapes human development: The case of psychosocial deprivation. *Neural Plasticity*, *2019*, 1–12. https://doi.org/10.1155/2019/1676285

OECD. (2017). *Starting Strong 2017: Key OECD indicators on early childhood education and care*. Paris: OECD Publishing. https://doi.org/10.1787/9789264276116-en

Pattinson, C. L., Staton, S. L., Smith, S. S., Sinclair, D. M., & Thorpe, K. J. (2014). Emotional climate and behavioral management during sleep time in early childhood education settings. *Early Childhood Research Quarterly*, *29*(4), 660–668. https://doi.org/10.1016/J.ECRESQ.2014.07.009

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (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

Pianta, R. C., Downer, J., & Hamre, B. (2016). Quality in early education classrooms: Definitions, gaps, and systems. *The Future of Children*, *26*(2), 119–137. Retrieved from www.futureofchildren.org

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008a). *Classroom Assessment Scoring System: Manual K-3*. Baltimore, MD: Paul H Brookes Publishing.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008b). *Classroom Assessment Scoring System (CLASS): Manual Pre-K*. Baltimore, MD: Brookes Publishing Co. Retrieved from https://www.bookdepository.com/Classroom-Assessment-Scoring-System-CLASS-Manual-Pre-K-Bridget-K-Hamre/9781557669414

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

Pinhero, J. C., & Bates, D. M. (2000). Linear mixed-effects models: Basic concepts and examples. In *Mixed-Effects Models in S and S-PLUS* (pp. 3–56). New York: Springer-Verlag. https://doi.org/10.1007/0-387-22747-4_1

Plank, S. B., & Condliffe, B. F. (2013). Pressures of the season: An examination of classroom

quality and high-stakes accountability. *American Educational Research Journal*, *50*(5), 1152–1182. https://doi.org/10.3102/0002831213500691

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/J.DR.2016.06.004

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Raftery, A. E. (1995). *Bayesian model selection in social research*. *Sociological Methodology* (Vol. 25). Retrieved from http://www.vcharite.univ-mrs.fr/PP/lubrano/BMC/Raftery-Socmeth1995.pdf

Reid, J. L., & Ready, D. D. (2013). High-quality preschool: The socioeconomic composition of preschool classrooms and children's learning. *Early Education & Development*, *24*(8), 1082–1111. https://doi.org/10.1080/10409289.2012.757519

Rentzou, K. (2017). Using rating scales to evaluate quality early childhood education and care: Reliability issues. *European Early Childhood Education Research Journal*, *25*(5), 667–681. https://doi.org/10.1080/1350293X.2017.1356599

Roberts-Holmes, G. (2015). The 'datafication' of early years pedagogy: 'If the teaching is good, the data should be good and if there's bad teaching, there is bad data.' *Journal of Education Policy*, *30*(3), 302–315. https://doi.org/10.1080/02680939.2014.924561

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-75969-2_1

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In *Innovations in Multivariate Statistical Analysis* (pp. 233–247). Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-4603-0_17

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Setodji, C. M., Schaack, D., & Le, V. N. (2018). Using the early childhood environment rating scale-Revised in high stakes contexts: Does evidence warrant the practice? *Early Childhood Research Quarterly*, *42*, 158–169. https://doi.org/10.1016/j.ecresq.2017.10.001

Skalická, V., Belsky, J., Stenseng, F., & Wichstrøm, L. (2015). Preschool-age problem behavior and teacher-child conflict in school: Direct and moderation effects by preschool organization. *Child Development*, *86*(3), 955–964.

https://doi.org/10.1111/cdev.12350

Slot, P. L., Leseman, P. P. M., Verhagen, J., & Mulder, H. (2015). Associations between structural quality aspects and process quality in Dutch early childhood education and care settings. *Early Childhood Research Quarterly*, *33*, 64–76. https://doi.org/10.1016/J.ECRESQ.2015.06.001

Soliday Hong, S. L., Sabol, T. J., Burchinal, M. R., Tarullo, L., Zaslow, M., & Peisner-Feinberg, E. S. (2019). ECE quality indicators and child outcomes: Analyses of six large child care studies. *Early Childhood Research Quarterly*, *49*, 202–217. https://doi.org/10.1016/j.ecresq.2019.06.009

Soukakou, E. P. (2012). Measuring quality in inclusive preschool classrooms: Development and validation of the Inclusive Classroom Profile (ICP). *Early Childhood Research Quarterly*, *27*(3), 478–488. https://doi.org/10.1016/J.ECRESQ.2011.12.003

Staton, S. L., Smith, S. S., Hurst, C., Pattinson, C. L., & Thorpe, K. J. (2017). Mandatory nap times and group napping patterns in child care: An observational study. *Behavioral Sleep Medicine*, *15*(2), 129–143. https://doi.org/10.1080/15402002.2015.1120199

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. https://doi.org/10.3102/0162373715616249

Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2010). *The Early Childhood Environment Rating Scale Curricular Extension*. Trentham Books Ltd.

Tayler, C. (2016). *The E4Kids study: Assessing the effectiveness of Australian early childhood education and care programs.* Brisbane. Retrieved from https://education.unimelb.edu.au/__data/assets/pdf_file/0006/2929452/E4Kids-Report-3.0_WEB.pdf

Tayler, C., Cloney, D., Adams, R., Ishimine, K., Thorpe, K., & Nguyen, T. K. C. (2016). Assessing the effectiveness of Australian early childhood education and care experiences: study protocol. *BMC Public Health*, *16*(1), 352. https://doi.org/10.1186/s12889-016-2985-1

Tayler, C., Ishimine, K., Cloney, D., Cleveland, G., & Thorpe, K. (2013). The quality of early childhood education and care services in Australia. *Australasian Journal of Early Childhood*, *38*(2), 13–21. Retrieved from https://search.informit.com.au/documentSummary;dn=449075796308754;res=IELHSS

Thorpe, K. J., Cloney, D., & Tayler, C. (2010). Rethinking early childhood education and care: Implications for research and evaluation. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 144–150). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-044894-7.01201-X

Thorpe, K. J., Pattinson, C. L., Smith, S. S., & Staton, S. L. (2018). Mandatory naptimes in childcare do not reduce children's cortisol levels. *Scientific Reports*, *8*(1), 4545. https://doi.org/10.1038/s41598-018-22555-8

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Retrieved from https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-

Classroom-Observations.pdf

Yamamoto, Y., & Li, J. (2012). What makes a high-quality preschool? Similarities and differences between Chinese immigrant and European American parents' views. *Early Childhood Research Quarterly, 27*(2), 306–315. https://doi.org/10.1016/J.ECRESQ.2011.09.005

Table 1
Definition of ECE programs included within the study.

	Age of children	Staff Ratio	Description	Number of classrooms (range number of study children in room)
Home-based Day Care	Multi-age	Staff Ratio: 1:7 Maximum of 4 children of prior-to-school age (0-5 years)	Licensed and assessed ECE provided in a family home usually by a single educator.	53 (1 - 5)
Centre-based Day Care	36 – 54 months (3 – 4.5 years)	1:11	Licensed and assessed ECE provided in a childcare centre, can be either for-profit or not-for profit. May include an embedded Kindergarten (preschool) program, provided by a 4- year college qualified teacher.	164 (1 - 32)
Pre-K	36 - 54 months (3 – 4.5 years)	1:11	Part-time program, provided by a 4- year qualified teacher as a stand-alone education program.	342 (1 - 20)
Kindergarten	4.5 – 5.5 years	1:24	First full-time year of education provided in a school context.	1031 (1 - 16)
Year 1	5.5 – 6.5 years	1:24	Second full-time year of education provided in a school context.	479 (1 - 14)
Year 2	6.5 – 7.5 years	1:24	Third full-time year of education provided in a school context.	193 (1 - 5)
Combined school				44 (1 - 12)

Table 2

Tests of Measurement Invariance for the First Three Observation Periods

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	$\Delta\chi^2$ (Δ*df*)	ΔCFI	ΔRMSEA	ΔSRMR	Decision
ES									
Configural	217 (39)***	.993	.045 (.039 - .050)	.029					Accept
Metric	253 (75)***	.993	.032 (.028 - .037)	.030	53.6 (36)*	.000	.012	-.001	Accept
Scalar	294 (81) ***	.991	.034 (.030 - .038)	.029	40.5 (6)***	.001	-.002	.000	Accept
Residual	338 (89) ***	.990	.035 (.031 - .039)	.031	46.6 (8)***	.001	-.001	-.002	Accept
IS									
Configural	59 (15) ***	.998	.036 (.026 - .046)	.014					Accept
Metric	78 (43) ***	.998	.019 (.012 - .025)	.014	25.6 (28)	.000	.017	.000	Marginal accept
Scalar	90 (47) ***	.998	.02 (.014 - .026)	.014	12.2 (4)*	.000	-.001	.000	Accept
Residual	102 (53) ***	.998	.02 (.014 - .026)	.015	12.4 (6)	.000	.000	-.001	Accept
CO									
Configural	63 (15) ***	.997	.037 (.028 - .047)	.015					Accept
Metric	90 (43) ***	.997	.022 (.015 - .028)	.016	32.5 (28)	.000	.016	-.001	Accept
Scalar	125 (47) ***	.995	.027 (.021 - .033)	.016	31.2 (4)***	.002	-.005	.000	Accept
Residual	147 (53) ***	.994	.028 (.022 - .033)	.017	23.6 (6)***	.001	-.001	-.002	Accept

Note. N = 6,918; groups n = 2,306. ES = emotional support; IS = instructional support; CO = classroom organisation; CFI = Confirmatory Factor Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardised Root Mean Residual. Change in metrics from previous level indicated in parentheses. χ^2 adjusted using second-order method (Asparouhov & Muthen, 2010) and CFI and RMSEA statistics scaled to this adjusted χ^2, SRMR uses the Bentler no mean method, whilst change in χ^2 uses (Satorra, 2000) adjustment.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3

Summary of the Effects (Unstandardized Coefficients) of Changes in Format on CLASS Scores

Format	β_{IS}	β_{ES}	β_{CO}	What happens when a classroom changes to this format or content?
Meals and snacks	-0.348***	-0.16***	-0.364***	Lower scores in all CLASS domains
Individual time	-0.209***	-0.03	-0.085**	Lower instructional support and classroom organisation
Routines	0.001	-0.085**	-0.109**	Lower emotional support and classroom organisation
Other	-0.239***	0.019	-0.095***	Lower instructional support and classroom organisation
Small and whole group	0.092*	-0.01	0.057	Higher instructional support
Small group	0.063	0.044*	0.134***	Higher emotional support and classroom organisation
Free choice	0.043	0.309***	0.221***	Higher emotional support and classroom organisation
Whole group	0.225***	-0.041**	0.064***	Higher instructional support and classroom organisation, lower emotional support
Average (SD)	2.67 (1.16)	5.41 (0.85)	5.26 (0.99)	Average and standard deviation of scores.

Note. * p < .05. ** p < .01. *** p < .001.

Table 4

Summary of the Effects (Unstandardized Coefficients) of Changes in Content on CLASS Scores

Content	β_{IS}	β_{ES}	β_{CO}	What happens when a classroom changes to this content?
Food	-0.489 (0.065)***	-0.262***	-0.57***	Lower scores in all three CLASS domains
Transitions	-0.519***	-0.181**	-0.483***	Lower scores in all three CLASS domains
Gross motor	-0.67***	-0.037	-0.006	Lower instructional support
Rest and relaxation	-0.481**	-0.248*	-0.38**	Lower scores in all three CLASS domains
Digital technology	-0.294*	-0.135	-0.312**	Lower instructional support and classroom organisation
Multiple content	-0.052*	-0.029	-0.086***	Lower instructional support and classroom organisation
Play	-0.201***	0.317***	0.174***	Lower instructional support, higher emotional support and classroom organisation
Music	-0.31***	-0.08	0.157**	Lower instructional support, higher classroom organisation
Rotations	-0.148*	-0.016	0.11	Lower instructional support
Literature/ language arts	0.151***	-0.079***	-0.029	Higher instructional support, lower emotional support
Art	-0.066	0.184***	0.148***	Higher emotional support and classroom organisation
Math	0.18***	-0.045*	0.139***	Higher instructional support and classroom organisation, lower emotional support
Social studies	0.507***	0.099	0.134*	Higher instructional support and classroom organisation
Science	0.677***	0.147**	0.196***	Higher instructional support and classroom organisation
Average (SD)	2.67 (1.16)	5.41 (0.85)	5.26 (0.99)	Average and standard deviation of scores.

Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 5

Summary of the Effects (Unstandardized Coefficients) of school grade on CLASS Scores. Coefficients are low to high point estimate ranges. See online supplementary Tables 6.1 to 6.6 for more detail.

School grade	β_{IS} Coefficients	β_{IS} p Values	β_{ES} Coefficient	β_{ES} p Values	β_{CO} Coefficient	β_{CO} p Values	Average difference in CLASS scores compared to pre-K
Pre-K program							Reference Group
Home-based day care	-0.275 to -0.22	.032 to .087 13 of 22 <.05	-0.009 to 0.02	.856 to > .999	-0.087 to -0.054	.463 to .651	Lower IS
Centre based day care	-0.527 to -0.486	< .001	-0.613 to -0.588	< .001	-0.569 to -0.545	< .001	Lower IS, ES, and CO
Kindergarten	0.12 to 0.276	<.001 to .061 21 of 22 < .05	-0.233 to -0.166	<.001 to .001	0.118 to 0.166	.002 to .023	Higher IS and CO Lower ES
Year 1	0.068 to 0.234	.001 to .399 20 of 22 < .05	-0.423 to -0.344	< .001	-0.073 to -0.017	.278 to .807	Higher IS Lower ES
Year 2	-0.022 to 0.146	.369 to .896	-0.272 to -0.181	.019 to .13 17 of 22 < .05	0.203 to 0.271	.057 to .153	Lower ES
Combined school	0.103 to 0.25	.045 to .421 3 of 22 < .05	-0.177 to -0.099	.1 to .366	.021 to .076	.516 to .856	Sporadic indication of higher IS

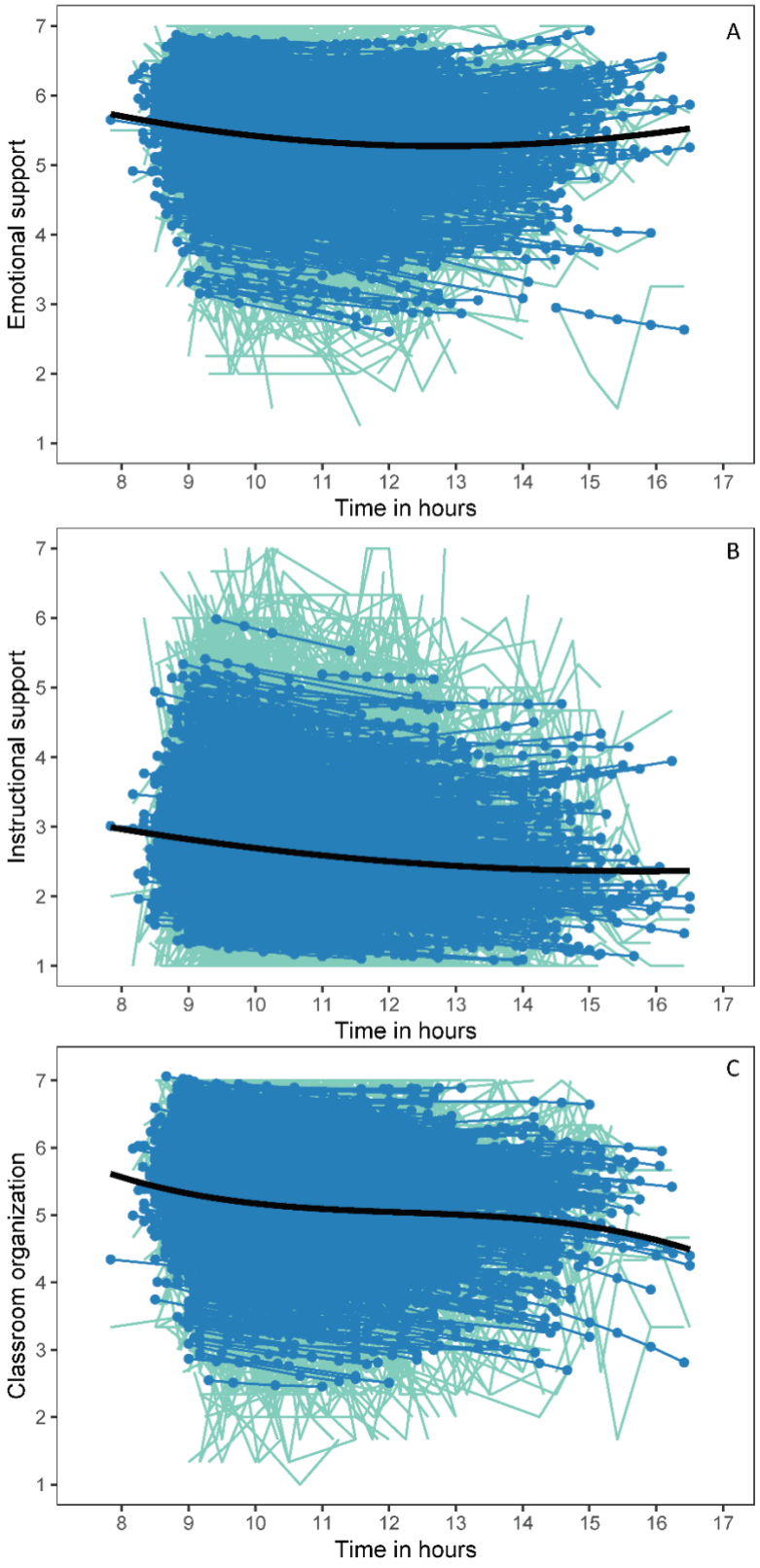

Figure 1. Visual summary of CLASS scores across the day. Large black line is the average conditional estimate (without confidence intervals as the software does not yet have an efficient way to incorporate uncertainty into variance parameters), small blue lines and dots are the predicted trajectories of classrooms and small light green lines are the observed data. A) ES = emotional support, B) IS = instructional support, C) CO = classroom organisation.

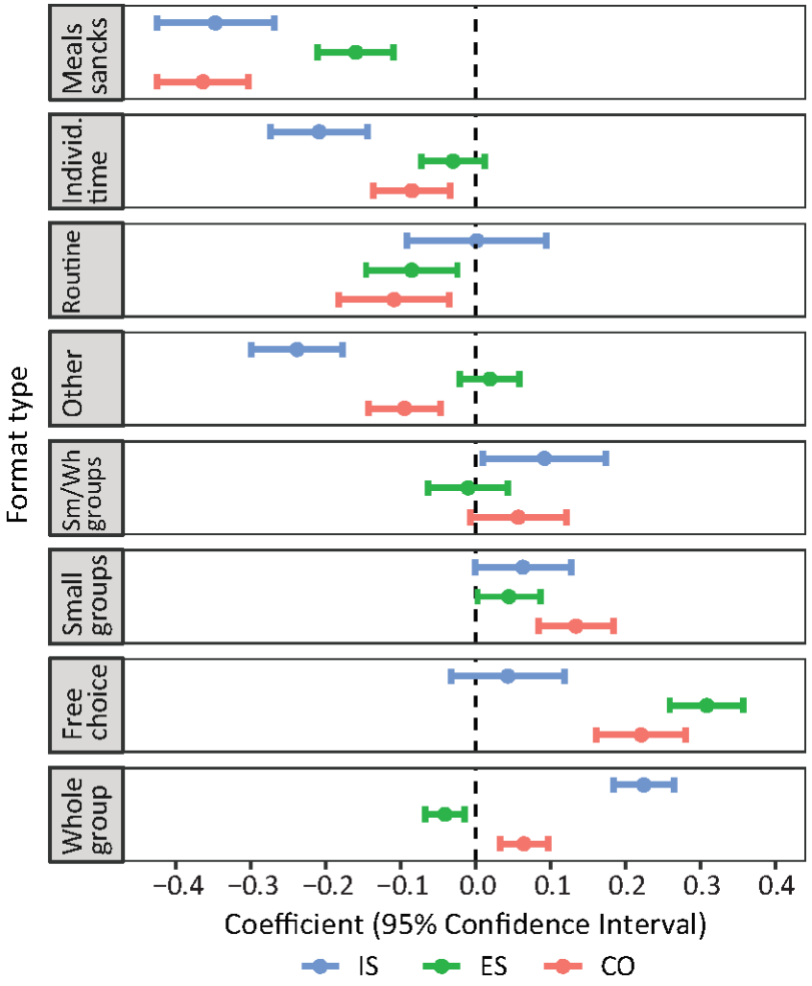

Figure 2. Effects of changes in format on CLASS scores. Coefficient estimates are unstandardized. IS = instructional support, ES = emotional support and CO = classroom organisation.

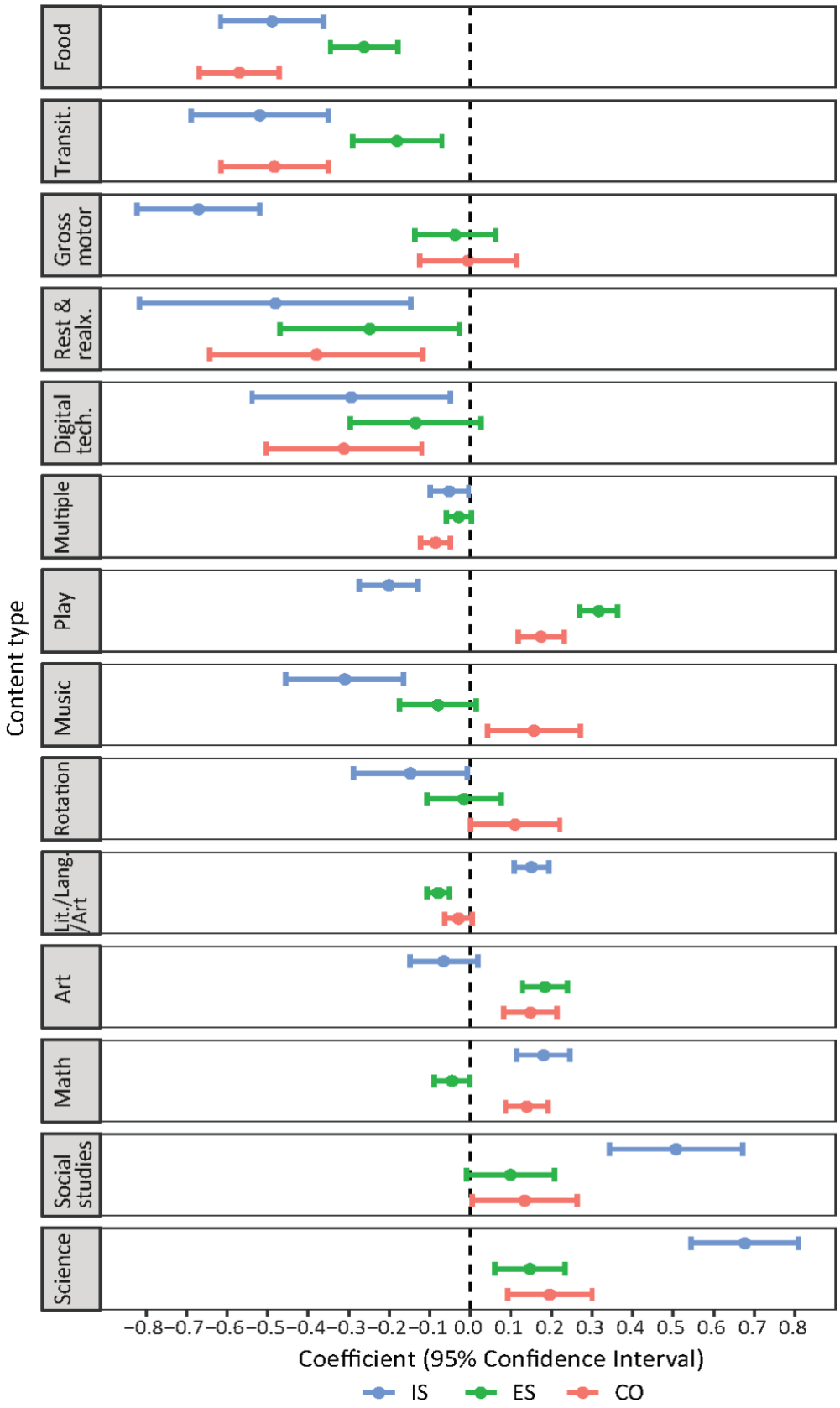

Figure 3. Effects of changes in content on CLASS scores. Coefficient estimates are unstandardized.