Organisation for Economic Co-operation and Development

EDU/WKP(2020)5

Unclassified **English text only**

11 March 2020

DIRECTORATE FOR EDUCATION AND SKILLS

Measuring process quality in early childhood education and care through Situational Judgement Questions: Findings from TALIS Starting Strong 2018 Field Trial

OECD Education Working Paper No. 217

Trude Nilsen (University of Oslo), Pauline Slot (Utrecht University), Hynek Cigler (Muni Masaryk University) and Minge Chen (IEA Hamburg)

Trude Nilsen, University of Oslo (trude.nilsen@ils.uio.no) ;
Pauline Slot, Utrecht University (p.l.slot@uu.nl);
Hynek Cigler, Muni Masaryk University (hynek.cigler@mail.muni.cz); and
Minge Chen, IEA Hamburg (minge.chen@iea-hamburg.de)

JT03458684

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

OECD EDUCATION WORKING PAPERS SERIES

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed herein are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the Directorate for Education and Skills, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

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

Acknowledgements

This paper was prepared by Trude Nilsen (University of Oslo), Pauline Slot (Utrecht University), Hynek Cigler (Muni Masaryk University) and Minge Chen (IEA Hamburg). The authors would like to thank the many people involved in the review process, especially Clara Barata and Arno Engel of the OECD Secretariat. In addition, the TALIS Starting Strong 2018 Consortium, Technical Advisory Group (TAG) and Questionnaire Expert Group (QEG) members provided suggestions and feedback both during the process of developing the Situational Judgement Questions (SJQs) and during the process of writing the paper.

The authors would also like to thank the TALIS Starting Strong 2018 National Project Managers (NPMs), the national experts in Early Childhood Education and Care consulted by NPMs, the QEG and the TAG for providing the expert scoring.

Théo Reybard, Mernie Graziotin, Leslie Greenhow and Deborah Fernandez of the OECD Secretariat provided editorial and production assistance.

Abstract

Situational Judgement Questions (SJQs) measuring process quality were included in the OECD Starting Strong Teaching and Learning International Survey 2018 (TALIS Starting Strong 2018) to address concerns of self-report bias in large-scale international surveys. These SJQs provide the staff in early childhood education and care with situations taken from a real-life professional context and offer several options on how to address these given situations.

Using TALIS Starting Strong 2018 field trial data, this paper evaluates the reliability and validity of the SJQs as measures of process quality in a large-scale international survey. The results showed that the SJQs were reliable, valid and worked well in evaluating process quality. High process quality was characterised by: 1) supporting child-directed play; 2) managing conflicts through behavioural management; and 3) supporting pro-social behaviour by encouraging sharing and collaboration among children. Staff self-efficacy and formal education were positively related to these practices. The paper further makes recommendations regarding the formats, type of analysis and interpretation of the SJQs in the Main Survey.

Résumé

Des questions de jugement situationnel (TJQ) évaluant la qualité des processus ont été utilisés dans l'Enquête internationale sur l'enseignement et l'apprentissage (TALIS) de l'OCDE sur la petite enfance 2018 (TALIS Petite enfance 2018) pour régler la question du biais d'autodéclaration dans les enquêtes internationales à grande échelle. Les TJQ proposent aux agents de l'éducation et de l'accueil des jeunes enfants diverses situations inspirées du monde professionnel réel, avec, au choix, plusieurs réactions possibles.

À l'aide des données des essais sur le terrain menés dans le cadre de l'Enquête Petite enfance, le présent document évalue la fiabilité et la validité des TJQ en tant qu'indicateurs de la qualité des processus dans une enquête internationale à grande échelle. Les résultats montrent que les TJQ sont fiables, valides et efficaces pour évaluer la qualité des processus. Les processus de qualité élevée se caractérisent par les pratiques suivantes : 1) encourager chez les enfants le jeu autodirigé ; 2) régler les conflits par le biais de la gestion du comportement ; et 3) favoriser les comportements prosociaux en encourageant les enfants à partager et à collaborer. La formation scolaire des agents ainsi que la perception qu'ils ont de leur propre efficacité sont positivement corrélées à ces pratiques. Le document formule en outre des recommandations à propos de la structure, du type d'analyse et de l'interprétation des TJQ dans l'enquête principale.

Table of contents

Tables

Figures

Boxes

List of abbreviations and acronyms

ECEC	Early Childhood Education and Care
IRT	Item Response Theory
NPM	National Project Manager
NRM	Nominal Response Modelling
QEG	Questionnaire Expert Group
SJQ	Situational Judgement Question
TAG	Technical Advisory Group

1. Introduction

High-quality early childhood education and care (ECEC) is vital for children's development, learning and well-being [e.g. Vermeer et al. (2016[1])]. ECEC quality is characterised by structural indicators such as staff-to-child ratio, environmental features such as space, furnishing and materials, staff formal education, and process quality (Vermeer et al., 2016[1]). Among these aspects, process quality has been shown to have the strongest impact on children's development, learning and well-being [e.g. Slot (2018[2])]. Process quality refers to children's daily experiences and interactions with staff and other children (Bronfenbrenner and Morris, 2006[3]), and thus, staff practices play an important role in these experiences and interactions.

It is important that research generates advice on process quality in particular with the increase of resources invested in ECEC (Bertram and Pascal, 2016[4]; OECD, 2018[5]; Slot, 2018[2]). Stakeholders need to know 1) what characterises high process quality. 2) how to promote high process quality, and 3) the associations of high process quality with children's learning and well-being (Sim et al., 2019[6]; Slot, 2018[2]).

As part of its data development strategy, the OECD ECEC Network and the OECD Secretariat agreed to implement an International Large-scale Survey that focussed on ECEC staff and centre leaders. The data generated served as a proxy for the quality of learning and well-being environments and the organisation of ECEC. Furthermore, the survey is a useful tool to investigate process quality at scale within feasible financial means, as well as providing comparisons at the international level (Sim et al., 2019[6]).

The Starting Strong Teaching and Learning International Survey 2018 (TALIS Starting Strong) is the first International Large-scale Survey of ECEC staff and centre leaders. TALIS Starting Strong builds on the established Teaching and Learning International Survey (TALIS) of school teachers and principals and seeks to identify strengths of, and improvement opportunities for, early childhood learning and well-being environments across and within different countries. TALIS Starting Strong also aims to provide valuable knowledge to inform policy makers of pedagogical and professional practices, staff working conditions and the overall quality of the ECEC. It is also the first International Large-scale Survey measuring process quality.

Despite an increasing body of research on the importance of process quality, there are still considerable challenges to be considered in the assessment of process quality. First, the way process quality is measured differs across studies and countries, making inferences and comparisons at the international level difficult [e.g. Slot (2018[2])]. Second, the most reliable and valid assessment of process quality has been conducted using class or playroom observations (e.g. live or through video), which is a technically challenging, expensive and a time-consuming approach for many countries (Vermeer et al., 2016[1]). Moreover, most of these studies used small samples, which limits the potential for policy makers and practitioners to draw conclusions at the country level.

International Large-scale Surveys of process quality address some of these challenges. Such surveys: 1) propose a common instrument shared and validated across countries; 2) are less expensive and time-consuming than observations; and 3) have sampling and analysis principles that enable them to generalise inferences at the country level.

Despite these advantages, such surveys may also be problematic in the comparative assessment of process quality because respondents often report what they believe is expected from them (referred to as social desirability), rather than their actual practices (Muijs, 2006[7]). Moreover, in comparative studies, constructs such as process quality may be prone to cultural bias, as respondents from one culture may understand and address the questions differently and have different response patterns (e.g. have an overall more negative or positive way of rating a scale) than respondents from other cultures.

Situational Judgement Questions (SJQs) can lower the social desirability and cultural bias related to self-reports by providing staff with real-life professional context and assessing how staff address these situations (Kyllonen and Bertling, 2013[8]; Lievens, Peeters and Schollaert, 2008[9]). SJQs include a context describing concrete examples reflecting real-life professional contexts and provide the respondent with several options on how to address the given situations.

To address the issues of social desirability and cultural bias in the assessment of process quality in TALIS Starting Strong 2018, process quality was measured in the survey's field trial using SJQs in addition to the more traditional formats with rating scales. Because SJQs had never been used to assess process quality, it is of the utmost importance to assess the psychometric properties (i.e. reliability and validity) and the overall interpretability of the items. The aim of this paper is to describe the properties of SJQs, evaluate how well they work in evaluating staff process quality in a large-scale international survey, and make recommendations for their use and interpretation in the Main Survey of TALIS Starting Strong 2018.

1.1. Theory on process quality

Process quality is a key driver of children's development and learning in Early Childhood Education and Care (ECEC) settings (Bronfenbrenner and Morris, 2006[3]; Melhuish et al., 2015[10]; OECD, 2018[5]). More specifically, the social, emotional, physical and instructional aspects of children's everyday interactions with staff and other children while being engaged in play, activities or routines lie at the heart of this concept (Barros et al., 2016[11]; Ghazvini and Mullis, 2002[12]; Howes et al., 2008[13]; Pianta et al., 2005[14]; Slot, 2018[2]). Staff-child interactions commonly include the following elements (OECD, 2018[15]; Slot et al., 2017[16])

- emotional climate, which includes the physical and emotional climate and sense of security children experience in interactions with staff;
- instructional and pedagogical quality, which involves the strategies, activities, and practices staff use to engage and support children in learning and development;
- organisation of routines and management to guide children's behaviour (Hamre et al., 2014[17]).

There is strong evidence supporting the importance of these three aspects (e.g. Hamre et al (2013[18]), OECD (2018[5])). First, the positive and affectionate relationships between staff and children provide children with a sense of security, which supports their well-being and allows them to explore their environment while developing autonomy and competence

(Ainsworth et al., 1978[19]; Bronson, 2000[20]; Sroufe, 2000[21]). Thus, an emotionally supportive environment provides the foundation for children to develop and learn.

Second, there is good evidence illustrating the importance of staff abilities to support children's development and learning by facilitating exploration, using scaffolding, and providing children with language-rich interactions during play and activities (e.g. Hamre et al., (2013[18]), OECD (2018[5])). Staff interactions to support children's development and learning can occur during organised, staff-initiated activities, such as shared book reading or circle time, or during child-initiated situations, such as play.

Third, staff ability to support children in showing appropriate behaviour is especially relevant in the early years of development as children gradually learn to regulate their emotions and interact with peers (Bronson, 2000[20]; Kopp, 1982[22]). Research has shown that staff use of effective classroom routines and behaviour management strategies is important for children's self-regulation skills and engagement in play and activities (e.g. Blair (2003[23]), Bronson (2000[20]), Emmer and Stough (2001[24]), Raver (2004[25])).

ECEC staff seem to favour providing support for children's development and well-being, while striking a good balance between social-emotional and more pre-academic domains of development, at least in Europe (Broekhuizen et al., 2015[26]). For children under three years old, a recent survey of 2 700 ECEC staff in nine European countries highlighted that social-emotional skills were considered more important than pre-academic skills (Broekhuizen et al., 2015[26]).

Some subdomains of process quality have recently been given more attention. Within behaviour management, the promotion of social and pro-social skills within peers is important. Adults play an important role in supervising and guiding children's social encounters with peers (Howes, 2011[27]; Howes et al., 2008[13]; van Schaik, Leseman and Huijbregts, 2014[28]), particularly during peer play, sharing and collaboration (Bhavnagri and Parke, 1991[29]; Ladd and Hart, 1992[30]). A Dutch study showed that ECEC staff scored the lowest on fostering peer interactions compared to interaction skills aimed at general emotional support and developmentally stimulating interactions (Helmerhorst et al., 2014[31]).

There are several strategies staff can use to facilitate peer interactions. Work by Williams et al. (2010[32]) distinguished three main categories: 1) adult-centred (e.g. distracting a child when conflicts occur); 2) child-centred (e.g. communicating about peers' presence, activity, objects or feelings and intentions); and 3) group interactions (e.g. staff interacts with child and peers). Staff use of child-centred scaffolding techniques appeared positively related to children's peer sociability, whereas adult-centred techniques showed negative associations (Williams, Mastergeorge and Ontai, 2010[32]). Including these aspects in the assessment of process quality offers a more comprehensive view of staff child-interactions in the ECEC context, and a better description of what can be considered high process quality.

1.2. Measurement of process quality

Observation measures are the most common form of process quality assessment. Researchers either make direct observations or through videos, and then rate process quality using a standardised rating scale (Slot, 2018[2]) characterising low, medium and high levels of process quality.

There are observational measures focusing on global quality, such as the Environmental Rating Scales (ERS), which include the assessment of a variety of aspects, such as furnishing, materials, activities, and interactions. Others observation measures focus more on the quality of interactions, such as the Classroom Assessment Scoring System (CLASS) (for more information about observational instruments, please see Vermeer et al. (2016[1]) or Slot (2018[2])). Even though ERS are broadly defined measures, neither these, nor other measures capture *all* aspects of process quality as described in the previous theory section, namely aspects of peer interactions.

Although observational measures are the most common way to assess process quality, there are some pending methodological issues in addition to the issues aforementioned (Slot et al., 2015[33]). First, observational measures have been criticised for not being able to measure process quality adequately as several studies have revealed problems with the scaling of items and their validity (e.g. Burchinal, Kainz and Cai (2011[34]), Cassidy et al. (2005[35]), Colwell et al. (2013[36]), Cryer et al. (1999[37]), de Kruif et al. (2000[38]), Gordon et al. (2013[39]), Layzer and Goodson (2006[40]), Perlman, Zellman and Le (2004[41]), Slot et al. (2017[16]) Zaslow et al (2010[42]) (2013[36])). For instance, several studies have used the Early Childhood Environmental Rating Scale, an ERS measure of process quality for settings serving children 3-5 year-olds, and found differences in the factor structure, that is, the studies were not able to identify the same factor structure including the same number of dimensions and underlying concepts (Cassidy et al., 2005; de Kruif et al., 2000; Gordon et al., 2013; Perlman et al., 2004). These differences point to challenges with the operationalisation of the underlying concepts, rendering comparisons between studies and between countries very difficult.

In addition, there is some evidence suggesting that items in the Early Childhood Environmental Rating Scale, as well as other observational measures (such as the Caregiver Interaction Scale) do not differentiate well between different levels of process quality, particularly in the mid- to high range (Colwell et al., 2013; Gordon et al., 2013), meaning that the measures are not able to distinguish between ECEC environments of medium and high process quality. However, it is important to distinguish between these different levels, as research has shown that a minimum level of process quality is required to positively affect children's well-being, development and learning (Burchinal, Kainz and Cai, 2011[34]; Burchinal et al., 2010[43]; Hatfield et al., 2016[44]; Ruzek et al., 2014[45]; Weiland et al., 2013[46]) .

In addition, the reliability and stability of classroom observations using these standardised observation measures have been questioned (Pianta and Hamre, 2009[47]). The variance in scoring from 4 to 14% can be explained by the fact that the same observer rated the quality. To increase reliability, studies need to have several observations per classroom within a certain time frame, or multiple raters conducting classroom observations simultaneously. Both solutions are costly and do not address the issue of lack of stability of quality over time, as research has shown that process quality varies across the year (Pianta and Hamre, 2009[47]). Global ratings of quality appeared to be more stable across the day than more detailed, time-sampling methods (Pianta and Hamre, 2009[47]).

1.3. Previous research on situational judgement questions

1.3.1. Challenges pertaining to self-reports

Staff self-report measures have been used to assess process quality as well (Charlesworth et al., 1993[48]; Walston and West, 2004[49]; Xue and Meisels, 2004[50]), sometimes in combination with observational measures (Kuger and Kluczniok, 2008[51]; Slot et al., 2015[33]). Staff self-reports yield to more stable results on process quality over time (Pianta and Hamre, 2009), and are also more cost-effective.

However, self-report measures, such as surveys, have are also known to be prone to specific challenges such as social desirability and cultural bias (van de Vijver and He, 2016[52]). Social desirability refers to the tendency of respondents to answer in a manner that will be viewed favourably by others. For instance, when teachers self-report their instructional practices, several studies have found that teachers tend to report practices they believe are high quality, rather than what they actually do in the classroom (Muijs, 2006[7]). For instance, teachers would probably not admit that they spend most of the lesson lecturing, as most teachers would know these are not ideal practices. Observational studies have, in fact, detected that these teachers lecture most of the lesson (ibid). This has also been observed in the Field Trial of TALIS Starting Strong 2018 where staff consistently rated themselves very high on self-reported measures of instructional support (scores around 3.5 out of 4 on the majority of items) whereas observational measures of instructional support (such as the CLASS) indicate an average score of 2 out of 7 (OECD, 2018[15]).

Cultural bias is another challenge and refers to interpreting and judging phenomena by standards inherent to one's own culture (He and Van de Vijver, 2013[53]). For instance, in self-reports of student motivation, students in Asian countries and Finland tend to respond more negatively, perhaps due to more humble and introvert ways of expression, while Western countries tend to respond more positively (He and Van de Vijver, 2016[54]). This is problematic because average levels of student motivation in the first set of countries is not comparable to the levels of student motivation in the second set of countries.

Response style is one of the challenges resulting from cultural bias. When respondents rate a question using Likert scales (i.e. respondents indicate the extent to which they agree with a statement), respondents in some cultures tend to use the extremities of the scales (e.g. I agree very much), while respondents in other cultures tend to choose the response in the middle (e.g. I agree). Such as in the example of student motivation, this is problematic when attempting to compare the same construct, across countries because self-reported levels of construct rating are not comparable.

The two types of response scales most commonly used in self-report are Likert scales and frequencies, and they pose different challenges to the interpretation of self-reported data. The challenge with Likert scales is that agreement is a subjective and affective measure and is therefore jeopardised by cultural bias. Moreover, if the respondent knows what is expected of them, they tend to choose what they assume is the “correct” response making Likert scales also prone to social desirability (van de Vijver and He, 2016[52]).

Frequency scales tend to be less prone to social desirability and cultural bias than Likert scales due to the lack of affective motions (agreement) involved; the respondents simply respond to how often they implement certain practices. However, there are also challenges related to the use of frequency scales for self-reported measures. The assumption behind applying a frequency scale to the measurement of teaching practices for example, is that the more often the practices are reported, the better the outcome; in other words, “more is better”. This may be problematic though, as performing the same type of practice all the time may not lead to better outcomes (Creemers and Kyriakides, 2006[55]; Teig, Scherer and Nilsen, 2018[56]). For instance, teaching children numeracy the whole day, every day, may not lead to improving children’s well-being, development or learning.

1.3.2. What are SJQs and how may they diminish the challenges with self-reports?

When measuring process quality in terms of staff interactions, neither Likert scales nor frequency scales capture what staff would *do*, but rather how often they exhibit certain behaviour or to what extent they believe that certain practices are of high quality and viewed favourably by others.

One innovative methodological approach to address these issues in international large-scale surveys is the use of Situational Judgment Questions (SJQs) (Bolt, Lu and Kim, 2014[57]). The idea behind SJQs is to provide a closer link than self-reports between the surveyed items and the respondents' actual daily work-life situations. SJQs present respondents with real-life professional contexts and possible ways to address these situations. Respondents then have to indicate which way they would act in that situation. SJQs are hence meant to measure how respondents would act in a given situation by asking them what they would actually *do*, rather than asking about their *agreement* with a particular way of acting (i.e. Likert scales).

SJQs have been around for about 50 years and were first employed in work settings and personnel selection (e.g. Motowidlo, Dunnette and Carter (1990[58])). Today they are also used in different research fields, such as the fields of education and psychology. In the context of international large-scale assessments of education, they were used in Programme for International Student Assessment (PISA) 2012 to measure students' problem-solving skills (Kyllonen and Bertling, 2013[8]). The items focussed on a person's initial response to a problem when facing the three following scenarios, 1) a problem with a text message on a mobile phone; 2) a route selection for getting to a zoo; and 3) a malfunctioning ticket vending machine. These SJQs were found to measure three distinct aspects or dimensions of problem-solving, namely 1) systematic problem-solving behaviour (i.e. behaviours resulting from an analysis of the situation); 2) unsystematic problem-solving behaviours (i.e. impulsive behavioural tendencies); and 3) help seeking behaviours (i.e. relying on others' knowledge and expertise). However, the reliabilities of the three dimensions were lower than "the standards for internal consistencies of indices used in PISA" (i.e. the Cronbach's α, a measure of the reliability of a psychometric test, did not reach the minimum threshold of 0.70), which was considered as "typical finding for situational judgement tests" [OECD (2014[59]), p. 369]. In other words, the items included in the situational judgement questions in PISA 2012 did not sufficiently capture the targeted underlying traits or concept.

Recently TALIS included SJQs in the Field Trial for the 2018 cycle of the survey. These SJQs attempted to measure instructional quality through four dimensions: *Classroom Management* (e.g. Quiet students down and then remind them of the classroom rules), *Clarity of Instruction* (e.g. Ask questions to check if the students have understood what I taught about topic), *Supportive climate* (e.g. Ask students who have finished task to help studs who are struggling) and *Cognitive Activation* (e.g. Give tasks that require students to apply what they have learned to new contexts). These items were based on the theoretical framework of Klieme and collegues (2009[60]), which has been shown to work well in international large-scale assessments such as *Trends In Mathematics and Science Study* (TIMSS) (Nilsen and Gustafsson, 2016[61]). Field Trial analysis demonstrated that while some SJQs worked well, others showed large variation across countries. Hence, it was decided not to include them in the Main Survey. One reason why the SJQs in TALIS 2018 may not have worked well could be because the items were originally meant for mathematics teachers, while TALIS includes teachers in all subjects.

1.3.3. The properties and formats of SJQs

Construct versus theory driven SJQs

The process of making the SJQs can rely heavily on psychometrics, so-called construct-driven SJQs, or rely more on experts and theory, so-called theory-driven SJQs (Lievens, 2017[62]). Both approaches have advantages and disadvantages. According to Lievens (2017[62]), theory-driven SJQs tend to be multi-dimensional (i.e. measuring several traits rather than one) with low reliability. This is due to the fact that SJQs were traditionally developed to capture a variety of work-related knowledge, skills and abilities, such as teamwork knowledge (Morgeson, Reider and Campion, 2005[63]) or academic performance (Oswald et al., 2004[64]).

Concurrently, a strong theoretical foundation is needed for high validity. In a pure construct-driven approach, a number of very similar items would be included to capture the common trait; however, this would jeopardise the conceptual broadness of the SJQs. In other words, only a fraction of the underlying trait would be measured resulting in low validity. However, some studies managed to build SJQs with high convergent validity such as Mussel, Gatzka, and Hewig (2016[65]) who developed SJQs for assessing compliance, gregariousness and self-discipline and reached Cronbach's α coefficient of 0.59. Becker (2005) also build a uni-dimensional construct using 30 items intending to measures several dimensions of integrity, such as respect, trust, career progress and good working relationships with colleagues.

An iterative process where subject matter experts and psychometricians collaborate would provide better measures (as was the case in TALIS Starting Strong 2018, see section 2.1.2). In the context of making SJQs on process quality in international large-scale surveys, this would mean that experts on process quality and ECEC collaborate closely with psychometricians (such as the Technical Advisory Group to TALIS Starting Strong) to enable better validity and increase reliability of the SJQs.

How reliable and valid are the SJQs?

Another challenge of SJQs is poor reliability, meaning that the items measuring the construct do not sufficiently capture the underlying trait. Meta-studies have found Cronbach's α coefficients of 0.4-0.5 (Catano, Brochu and Lamerson, 2012[66]). Yet, a review of SJQs found that SJQs have higher criterion-related validity (i.e. the SJQs are related to other constructs one would theoretically expect them to be related to) and incremental validity (i.e. increased predictive ability beyond that provided by an existing method of assessment.) than more traditional formats used in cognitive ability and personality tests (Lievens, Peeters and Schollaert, 2008[9]). Moreover, the authors found that SJQs were less prone to cultural bias than other measures and especially did not bias against minorities. Respondents also tend to respond more positively to SJQs as they tap into situations more relevant to them. However, even though SJQs are known to be less prone to social desirability, it is still possible to attempt to respond to these items in a manner that will be viewed favourably by others (Lievens, Peeters and Schollaert, 2008[9]).

Scoring of SJQs

A number of studies have examined how to score SJQs to foster the most reliable and valid results (e.g. Guo, Zu and Kyllonen (2016[67]); Kyllonen and Bertling (2013[8]); Whelpley (2014[68])). The response to items may be ranked from low to high based on theory. In order to validate these scores, it is advisable to also have a number of experts score the SJQs (i.e. collecting the so-called expert opinions). If the experts are unsure how to rank the items (i.e. which response options should be ranked high, and which should be ranked low), a popularity score, in which the items are scored according to the modal responses of the respondents, may also be used.

In addition to these scoring methods, previous research indicates that the Item Response Theory (IRT) model called Nominal Response Model (NRM) seems to be the most reliable and valid way to score SJQs (e.g. Guo, Zu and Kyllonen (2016[67]); Guo et al (2016[69])). Guo, Zu, and Kyllonen (2016[67]) for example, used SJQs to measure students' ability to manage emotions in a longitudinal study. This study showed that NRM scoring was more reliable than expert scoring. These results are in line with another study which found that NRM provided higher reliabilities than other types of scoring and higher reliabilities than the Generalized Partial Credit Model (Guo et al., 2016[69]).

SJQ formats

SJQs may have different formats depending on the scaling of response options. Two common formats are Likert-scale format and the forced choice format. For the Likert-scale format, the respondents have to rate each of the items in the task according to an agreement scale. Also other rating scales may be used, for instance rating items extending from "I would definitely not do this" to "I would definitely do this". In the forced choice format, the respondents have to choose between items in the situational judgement task instead of rating them. For instance, the respondents are only allowed to choose one of the items as a first choice, and one of the items as a second choice.

Which of these two formats has the highest reliability and validity is a difficult question as it may depend on the concept measured. Some research indicates that the Forced-choice format induces less social desirability and cultural bias than single items and Likert scale formats [e.g. Burus, Naemi and Kyllonen (2011[70])]. However, most of these studies measure personality traits, and the difference in reliability and validity of the two formats for process quality has never been studied.

1.3.4. How would SJQs help in the assessment of process quality, and how should they be interpreted?

SJQs have the potential to complement other measures of process quality. Because they are meant to measure how respondents would act in a given situation, if found to be valid and reliable, they may offer a more accurate picture of what takes place inside the playroom or classroom than other self-reports.

However, in order to provide also a valid assessment of process quality, responses to SJQs need to:

- assess the domains proposed by the previous literature and instruments, while addressing some of their limitations (namely in subdomains, such as promotion of social and pro-social skills with peers)
- correlate with other instruments, namely observations
- discriminate between high, medium and low process quality.

The best strategy to design SJQs to measure process quality is to map the SJQ context onto proposed domains from previous observational measures, and develop studies where both SJQs and observational measures are used for the same samples. This would mean having at least one SJQ describing a respondents' actual daily work-life situations per domain of process quality, and comparing scores of the SJQ to the observation scores.

If not feasible, studies with only SJQs need to search for agreement between participants within and across countries, and between experts, staff and theory to be considered valid measures of process quality. In these situations, SJQs can then be said to measure "best practices" in the domains of emotional climate, instructional quality or behavioural management to the same degree, or better, than previous measures.

Much care needs also to be put in the design of items. In order to avoid social desirability and other biases, effort needs to be made to create items that are not perceived as obvious "wrong choices". However, it is essential that items can still be ranked to reflect lower to higher levels of process quality. More specifically, most participants within and across countries need to agree that a particular item (i.e. directing the children's attention to the classroom rules) is the most appropriate practice given a particular context (i.e. situation of peer conflict).

1.4. Research questions

Given that Situational Judgment Questions (SJQs) are still relatively new measures in the assessment of ECEC process quality, the current paper takes an explorative approach to the use and data analysis of SJQs in the TALIS Starting Strong 2018 Field Trial. It describes iterative process of developing and adapting the SJQs to the assessment of process quality in TALIS Starting Strong 2018; conducting the pilot; and finally using the field trial data of TALIS Starting Strong 2018 to make well-informed decisions about the inclusion of SJQs in the Main Survey. It includes recommendations for the formats, type of analyses and interpretation of process quality SJQs in the Main Survey. The paper also aims to make recommendations for future use of this methodological approach in other international large-scale surveys. In short, the following questions will be addressed:

1. How does process quality measured using SJQs vary within and across countries participating in TALIS Starting Strong 2018?
2. What type of scoring provides the highest reliability?
3. What is the reliability of the SJQs?
4. To what degree are the SJQs comparable across the two TALIS Starting Strong 2018 target populations: staff working in centres for children under the age of three years and staff working in centres belonging to pre-primary education (ISCED Level 02)? What degree are they comparable across countries?
5. How valid are the SJQs in the assessment of ECEC process quality?

2. Methodology

The methodology section first describes the sample and data, including the process of developing the Situational Judgement Question (SJQ). The proposed methodological approach involves different types of analyses depending on the research question and the nature of the data, and these analyses are presented in the same order as the research questions.

This section and the following ones discusses the SJQs that were included in the TALIS Starting 2018 Field Trial. The format and number of SJQ have been selected for the Main Survey on the basis of this analysis.

2.1. Sample and data

2.1.1. The sample and design

Although the field trial data was used for the analyses in the present study, data from the pilot study provided information during the process of creating the SJQs. The pilot was conducted in October 2016 where National Project Managers (NPMs) from each of the participating countries conducted guided focus groups to test all items. Participants consisted of staff from two target populations, namely staff working in centres for children under the age of three years and staff working in centres belonging to ISCED Level 02.

In the Field Trial and the Main Survey, nine countries participated at the ISCED Level 02, and four of these countries also participated with staff working in centres for children under the age of three. All countries went through a process of translation verification, to ensure the high quality of the data. A description of the sample sizes for the Field Trial is provided in Table 2.1.

In the field trial, a rotated questionnaire design was implemented due to the large number of field trial survey questions. This design led to a requirement of a minimum sample size of 30 centres per country and 240 early childhood education and care (ECEC) staff members per country (Sim et al., 2019[6]).

Members of Questionnaire Expert Group (QEG), Technical Advisory Group (TAG), National project managers (NPMs) in collaboration with experts in early childhood education and care from each country also scored the SJQs for both target populations (36 scorers in total). Experts were identified by the participating countries' NPMs. In addition to scoring the SJQs, NPMs, members of QEG, TAG, and experts also provided information about their background in ECEC, and what they thought of the proposed SJQs (i.e. strengths and weaknesses). The background of the experts varied: some of the experts worked as teacher trainers in ECEC, some worked as researchers in ECEC, some worked in early childhood education and care and had studied ECEC, and some were PhDs within the field of ECEC. Many of the researchers in ECEC had previously worked on ECEC programmes.

Table 2.1. Number of staff for each participating country in the Field Trial, per target population

Country	Number of staff working in centres belonging to ISCED Level 02	Number of staff working in centres for children under the age of three years
1	121	157
2	277	Did not participate
3	152	131
4	421	Did not participate
5	363	280
6	222	Did not participate
7	220	Did not participate
8	149	152
9	307	Did not participate

Note: The countries are anonymous because the data are from the Field Trial.

2.1.2. The development of the SJQs

Data for this paper are from the TALIS Starting Strong 2018 Field Trial. However, information from the pilot and the process undergone to develop the SJQs from the very start of the project was also included.

In TALIS Starting Strong 2018 an iterative method of item development was used where the consultation of National Project Managers (NPMs), the Technical Advisory Group (TAG), the Questionnaire Expert Group (QEG), and experts in ECEC played an important role. Feedback was provided from these parties before and after both the Pilot and the Field Trial. This was done through webinars, meetings and by email. The nature of this iterative method was important for the quality of the items and the success of the item development in TALIS Starting Strong 2018.

The following box provides a brief overview of the iterative process of developing the SJQs.

Box 2.1. Developing the Situational Judgement Questions (SJQ)

The following describes the iterative process of developing the Situational Judgement Questions (SJQs)

1. In early 2016, six items were initially created based on both theory and previous research on SJQs capturing the relevant aspects of process quality and distinguished between three domains of quality: emotional support, behavioural management and support for learning and development. In addition, three other SJQs were developed to capture other, less frequently, studied aspects of process quality (support for child-directed play, support for pro-social behaviour, and support for diversity).

   - The first SJQ concerned emotional support and aimed to measure how staff would respond in case a child would call out for help. Because this item had a high risk of social desirability, response categories that distinguished between responding immediately or letting the child wait for a moment (capturing the timely response) were developed. Moreover, this SJQ also tried to distinguish between levels of sensitivity in staff reactions, namely whether the staff approached the child or responded from a distance.
   - The second SJQ concerned behavioural management and aimed to measure staff response to a quarrel between two children.
   - The third SJQ concerned support for learning and development and aimed to measure how staff would implement instructional practices by asking how staff would respond to a child's query about seasons.
   - The fourth SJQ concerned support for child-directed play and aimed to measure the staff role in facilitating and enriching children's free play and using this as an opportunity for development and learning. This SJQ was proposed to include measures of both the pre-academic and play orientation in early childhood education and care (ECEC). It suggested several responses to children's play ranging from non-involvement or encouraging peer play to playing along while following the children's lead and enriching the children's play.
   - The fifth SJQ was aimed at capturing staff support of children's pro-social behaviour during play. Although peer relations are highly valued by professionals in ECEC, little research has addressed this particular aspect in the context of process quality.
   - Lastly, in view of increasing globalisation and cultural diversity in ECEC, the sixth SJQ concerned support for diversity and aimed to measure staff responses on dealing with cultural diversity and inclusiveness in organising classroom practices and activities.

2. In September 2016, these SJQ proposals received feedback from members of the Technical Advisory Group (TAG), Questionnaire Expert Group (QEG), the OECD Secretariat as well as other experts on SJQs from education and human resources research and altered based on feedback. The SJQ on emotional support

was excluded due to concerns about social desirability. The remaining five items were reviewed and piloted by countries.

3. In November 2016, the Technical Advisory Group (TAG) and QEG provided feedback based on data from the pilot. Also, the countries' National Project Manager (NPMs) provided feedback on the items. The aim was to make SJQs that could be used in the questionnaires for both staff working in centres for children under the age of three years and to staff working in centres belonging to ISCED Level 02 It was difficult to make the SJQ on support for learning and development (learning about seasons) general enough to be applicable across both these age populations thus, it was excluded after the pilot. The reason was that measuring staff support of children's learning and development across such large age gaps was challenging. The SJQs on diversity and inclusiveness and emotional support were also excluded after the pilot, as they were sensitive to cultural differences. Additional feedback was used to further improve the SJQs for the Field Trial, which included the following SJQs:
    - Context for the SJQ on Encouraging pro-social behaviourSJQ: children sharing
    - Context for the SJQ on Behavioural management: handling children quarrelling
    - Context for the SJQ on Support for child-directed play: following the child's lead
4. In May 2017, it was agreed to include Likert and Forced-choice response scales for the remaining three SJQs in the Field Trial. In order to provide the respondent with only one correct item (reflecting high-process quality) in the forced-choice format, the number of items had to be reduced compared to the Likert format and some items had to be altered. This means that the items are not always the same across the two response scales, and there are often more items in the Likert than the forced choice response scales.
5. Field Trial data was collected in the nine participating countries in May-June 2017.
6. NPMs in collaboration with experts from the participating countries, and members of QEG and TAG scored the items in the Short Questionnaire on Situational Judgement Questions between July-August 2017.
7. Later in 2017, the results from the expert scoring, and the results from the field trial were discussed with TAG and QEG and further feedback was provided.
8. Based on the field trial analysis, which will be presented in the paper, two SJQs were retained for the Main Survey questionnaire administered in 2018.

A similar process was undertaken in TALIS 2018 to measure instructional quality (which is similar to process quality, but in a school context). The results from TALIS Starting Strong 2018 are discussed against those in TALIS in the discussion section.

2.1.3. The data

In the field trial of TALIS Starting Strong 2018 two sets of response options for SJQs were used: a forced choice option and a Likert scale. To make the SJQs applicable for both populations, the age of the children mentioned in the question stem was three years old.

As an example, one of the SJQs – Question 30 – is presented below in Box 2.2, whereas the rest are listed in Annex A.

Question 30 is about two children quarrelling and taps into the aspect of process quality referred to as behavioural management. The Likert format of this question includes a 4-point Likert scale (I would definitely do this - I would probably do this - I would probably not do this - I would definitely not do this). The question includes seven different items. The SJQs were scored in several ways that will be described in the next section (2.2.2).

Box 2.2. Question 30-Likert

Aspect of process quality: Behavioural management. Context: quarrelling among children.

Suppose that you see two children of the same age (three years) and size quarrelling, and one child (child A) hits the other (child B). Child B is crying. Child A has previously regularly hit other children. Several other children are watching. What would you do?

a) I would speak firmly to child A while the other children are listening.

b) I would ask child A and B what happened.

c) I would warn child A that if he/she hits again he/she will face negative consequences.

d) I would tell child A what he/she should have done differently in this situation.

e) I would ask child A to apologise to child B.

f) I would remind child A of our rules, that hitting is not allowed.

g) I would include the other children in the discussion after the conflict.

Question 30 with a forced choice format has the same stem (context), but the respondents are asked which of the items would be their first and second choice (see Box 2.3). The items in the forced choice question are different from the Likert scale format, because only one item could be the most correct. This was not necessary in the Likert scale format, as respondents could rate each item. Hence, the Likert format contains more items.

Box 2.3. Question 30 - Forced choice

Aspect of process quality: Behavioural management. Context: quarrelling among children

Suppose that you see two children of the same age (three years) and size quarrelling, and one child (child A) hits the other (child B). Child B is crying. Child A has previously regularly hit other children. Several other children are watching.

What would you do?

a) I would speak firmly to child A while the other children are listening.

b) I would focus on child B and comfort him/her.

c) I would tell the children who was wrong and who was right.

d) I would resolve the conflict together with child A and B.

2.2. Method of analysis

The following sub-sections describe the method of analysis according to the order of the research questions.

2.2.1. How does process quality measured using SJQs vary within and across countries participating in TALIS Starting Strong Survey 2018?

To examine within and between country variations in SJQs, the relative frequencies of each item within each SJQ were calculated, taking the different sample sizes into account, the modal responses, and the means for each country and for both formats (Likert scale and Forced-choice). This was done for both target populations and for expert scores.

Diagrams reflecting the profiles for each country were made and inspected to look for patterns of agreement. This enabled to examine the degree to which the respondents within and/or across countries agreed (i.e. whether they chose the same response option).

Descriptive statistics were computed using the SPSS software.

2.2.2. What type of scoring provides the highest reliability?

Ranking

The items were first scored according to ranking in line with the theoretical framework (representing an ordinal scale). For instance, Question 30a ("*I would speak firmly to child A while the other children are listening*") is not considered a high-quality behavioural management practice, as staff would embarrass the child in front of others. Hence, the response *I would definitely do this*, would be assigned a 0, the lowest score, while *I would probably do this* =1, *I would probably not do this* = 2, and *I would definitely not do this* (correct answer) would be scored 3, the highest score.

Popularity

Next, the items were scored by popularity. Scoring by popularity simply means using the relative frequencies to score the SJQs. These were provided by the participants in the respective countries. For instance, consider the following responses to the SJQ item 29c ("I would encourage them to build something together"): 0.7% answered "I would definitely not do this", 9.0% replied I would probably not do this, 36.0% answered "I would probably this", and 54.2% responded "I would definitely do this". In this case, "I would definitely do this" is the modal response, and would get the highest score (3), while the three other responses would be ranked 0, 1, and 2, respectively. To this end, it makes it easier to compare alignment between for instance the ranking based on theory, and the ranking based on popularity.

Nominal scale

Item Response Theory (IRT) analysis models, or more specifically nominal response models, were estimated. These models were based on Bock's Nominal Response Model (1972[71]), which is an extension of the 2PL model Thissen, Cai and Bock (2010[72]). The nominal response model specified the probability that a respondent with a given value of the latent trait (e.g. behavioural management) selected a specific response option. The model specified the likelihood that a respondent of a given ability[1] (providing high process quality) will select a specific response option (e.g. *I would probably do this*) of a specific item (e.g. *I would speak firmly to child A while the other children are listening*). The thresholds provided by this model were used to score the items.

Expert

Experts scored the SJQs for both target populations. This scoring was compared with the ranking and the popularity score in order to search for any inconsistencies.

The different types of scoring were compared by means of reliability checks and by comparing country profiles. Three different reliability analyses were conducted. In the first analysis, the items of the SJQ were treated as a continuous scale tapping into the same underlying construct.

In the second approach, the items of the SJQ were treated as ordinal variables representing a ranking based on theory and analysed after recoding the items according to this ranking.

In the third approach, an IRT analysis was conducted for the Likert scale and the forced-choice option. This latter IRT approach used nominal response modelling (NRM) to estimate reliability. The IRT scores were chosen and used for further analysis.

The IRT analyses were conducted using the statistical software R, while SPSS was used for the other approaches.

In order to compare the different types of scoring, the reliability statistics for each SJQ was examined. The expert opinions were compared with the ranking based on theory. The modal responses were used to compare these two, as there were too few responses

[1] Although the word *ability* is commonly used in IRT analyses, SJQs in TALIS 3S are not a test of ability but rather a way to explore practices. Items are ranked from low to high process quality, and staff with high ability averages chose practices of high process quality (e.g. being able to manage the children during a conflict).

from the expert opinions to do a reliability analysis. The reliability was then compared across the three approaches: 1) the ordinal structure using the ranking of items based on theory; 2) the regular continuous scale; and 3) the NRM.

2.2.3. What is the reliability of the SJQs?

Using the NRM approach, the reliability of the SJQs was further examined in depth. This was done by estimating reliability (internal consistency), item difficulty and item discrimination parameters to evaluate the overall reliability. Factor loadings and explained variance were also calculated to help evaluate the overall reliability. The discrimination and item difficulty of the items were plotted as trace lines diagrams, which make how well the items function more visible.

The so-called "empirical reliability" was used to estimate reliabilities, which considered the observed variance of the expected a-posteriori (EAP) factor score estimates and the mean-square error variance of these estimates: $r_{xx'} = \frac{\mathrm{var}(EAP)}{\mathrm{var}(EAP)+\mathrm{mean}\left(\mathrm{var}\left(\mathrm{E}(EAP)\right)\right)}$. Empirical reliability was estimated using the empirical_rxx function in the mirt package in R.

2.2.4. To what degree are the SJQs comparable across the two Starting Strong 2018 target populations? To what degree are they comparable across countries?

Measurement invariance analysis was undertaken using NRM to check the comparability between the two target populations (i.e. staff working in centres for children under the age of three years and staff working in centres belonging to ISCED Level 02). It was not possible to do a measurement invariance analyses across countries due to too few respondents. The goodness of fit used to evaluate whether figural, metric or scalar invariance was met, followed those by Chen (2007[73]), originally used for ordinal factor analysis.

The IRT analyses were conducted using the statistical software R.

2.2.5. How valid are the SJQs in the assessment of ECEC process quality?

The questionnaire included several other items that were used to investigate the validity of the SJQs. To assess construct validity (i.e. "Are the SJQs measuring the constructs they are supposed to be measuring?") other items in the questionnaire that showed considerable conceptual overlap were identified (see Table 2.2).

Factor score estimates for each SJQ were extracted from the NRM analyses to examine associations with similar constructs measured in the questionnaire (construct validity) and associations with three staff reported measures pertaining to staff pre-service education content, self-efficacy, and perceived control over implementation of classroom practices (criterion validity).

Table 2.2. Overview of items included for construct validity

Situational Judgement Questionnaire (SJQ)	SJQ29 Encouraging pro-social behaviour	SJQ30 Behavioural management	SJQ31 Support for child-directed play
Questionnaire items	Encouraging sharing among children	I calm children who are upset	Not interfering when children play
	Encouraging children to help each other	I help children to follow the rules	If invited by the children, joining the children's play
	Encouraging children if they comfort each other	I help children understand the consequences if they do not follow the rules	Allowing children to take the lead when playing with children

For the purpose of convergent validity, the analysis focussed on two aspects of staff background that are considered to be important for their practices: specific training and self-efficacy. Several items were included to assess aspects that were part of staff pre-service education, some of which were deemed relevant topics related to the SJQs: child socio-emotional development, self-regulation, facilitating free play and playgroup management. In addition, staff reported on their self-efficacy on a number of aspects related to their work: adapt work to individual needs; help [children] to interact with each other; calm an upset child; help [children] develop self-confidence; and provide a feeling of security.

3. Results

3.1. How does process quality measured using Situational Judgement Questions (SJQs), vary within and across countries participating in TALIS Starting Strong Survey 2018?

The responses to some of the items in the SJQs showed great variation both within and across countries. For other items, the agreement between countries was high, with the exception of one or two countries with a different response pattern. The data for these countries would often act as outliers regardless of the type of items or SJQs, which could be interpreted as a sign of cultural bias. Table 3.1 shows an example of the frequencies and modal responses for the Likert option of SJQ item 29c (i.e. "*I would encourage them to build something together*") for staff working in centres belonging to ISCED Level 02 for the nine countries.

Table 3.1. Frequencies and modal responses for the Likert scale format of Situational Judgement Question item 29c "I would encourage them to build something together"

Based on data for staff working in centres belonging to ISCED Level 02

	I would definitely not do this	I would probably not do this	I would probably do this	I would definitely do this
Country 1	0.00	11.76	**47.06***	41.18
Country 2	1.20	1.20	24.70	**72.80***
Country 3	0.00	9.10	42.40	**48.50***
Country 4	0.00	3.00	33.30	**63.30***
Country 5	2.80	3.80	21.70	**71.70***
Country 6	1.40	40.60	**49.30***	8.70
Country 7	0.00	7.10	39.30	**53.60***
Country 8	0.00	4.70	45.30	**50.00***
Country 9	1.00	0.00	20.80	**78.20***
Experts		5.7	45.7	**48.60***

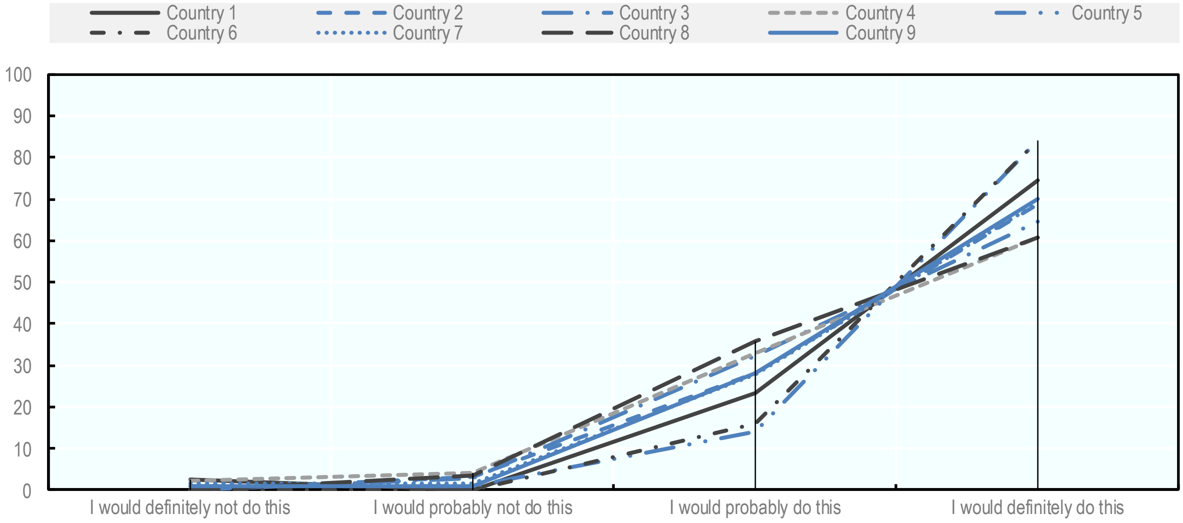
Note: Bold numbers with * denote modal response. The data are from the TALIS Starting Strong 2018 field trial, and hence the countries are anonymous.

Except for two countries, the modal response in the majority of countries was *I would definitely do this*. This is in line with the expert scoring, and it is in line with the ranking based on theory. Even though there is high agreement about this item, there is still some within- and cross-country variation, which is required for further analysis.

Likert scale format of SJQ item 30b (i.e. "*I would ask child A and B what happened*"), also demonstrated high agreement across countries, and was in line with the ranking based on theory and the expert scoring. One way to visualise the agreement is by making profiles for the items. Figure 3.1 illustrates the profile for Likert scale format of SJQ item 30b.

Figure 3.1. Profile of Likert scale format for Situational Judgement Question item 30b "I would ask child A and B what happened", by country

Percentage of staff giving the following answer, ISCED Level 2

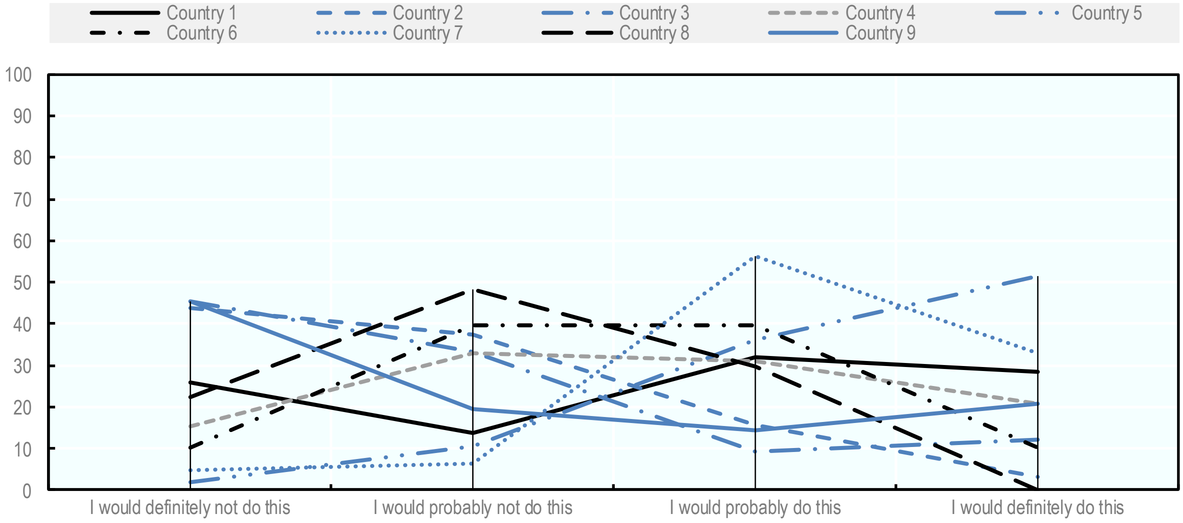


Note: The x-axis shows the response options, while the y-axis shows the relative frequency of staff responses (poularity).

In contrast, the profile of the Likert scale format of SJQ item 30g (i.e. "*I would include the other children in the discussion after the conflict"*) shows that there is high disagreement about this item (Figure 3.2).

Figure 3.2. Profile of Likert scale format for Situational Judgement Question item 30g "I would include the other children in the discussion after the conflict", by country

Percentage of staff giving the following answer, ISCED Level 02



Note: The x-axis shows the response options, while the y-axis shows the relative frequency of staff responses (popularity).

The items with the highest response frequencies for "I would definitely do this" and the highest agreement across countries (measured by modal responses) for the three Likert SJQs, were:

- SJQ29: I would encourage child A to share with child B.
- SJQ30: I would remind child A of our rules, that hitting is not allowed (see Table 3.1).
- SJQ31: I would play with the children by following their lead.

This is in line with the expert scoring, and it is in line with the ranking based on theory. The agreement across the two target populations was also high.

In general, the patterns were clearer for the Likert scale format than for the forced choice format and there was higher agreement within and across countries.

3.2. What type of scoring provides the highest reliability?

Table 3.2 provides an example of the different kind of scorings for the Likert scale format of SJQ item 29c (i.e. "*I would encourage them to build something together*") for ISCED Level 02. The last row of Table 3.2 (Nominal Response Modelling) provides the thresholds of going from "*I would definitely not do this*" to "*I would probably not do this*" (-3.83), and from "*I would probably not do this*" to *"I would probably do this",* etc.

Table 3.2. Results of the different types of scoring of Likert scale format for Situational Judgement Question item 29c "I would encourage them to build something together"

Based on data for staff working in centres belonging to ISCED Level 02

Option	I would definitely not do this	I would probably not do this	I would probably do this	I would definitely do this
Relative frequency (popularity)	0.7	9.0	36.0	54.2
Experts opinions	0	5.7	45.7	48.6
Rank scoring	0	1	2	3
Nominal Response Modelling		-3.83	-1.92	-0.09

The expert opinions were in line with the modal responses and the ranking for this and other SJQ items.

Results from the reliability analyses show that, compared to scoring based on a regular continuous scale, the ranking does not provide a consistent better estimate for the internal consistency of the SJQ, i.e. for question 29 it is higher, for question 30 it is the same, and for question 31 it is lower. However, the Nominal Response Modelling (NRM) provided higher reliability than the other types of scoring. These results are provided in Table 3.3 for the Likert scale format of the SJQs.

Table 3.3. Internal consistency of the Likert scale format of the three Situational Judgement Questions (SJQ) based on three different types of scoring

Based on data for both target populations (i.e. staff working in centres for children under the age of three years and staff working in centres belonging to ISCED Level 02

Type of Scoring	SJQ29	SJQ30	SJQ31
Regular	0.57	0.64	0.46
Ranking	0.61	0.64	0.40
Nominal Response Modelling	0.71	0.71	0.70

3.3. What is the reliability of the SJQs?

3.3.1. Internal consistency

Seeing how Nominal Response Modelling provided the best scoring approach, this approach was used to estimate internal consistency for the pooled model, and for each target population (see Table 3.4).

Table 3.4. Reliability for all Situational Judgement Questions (SJQs) based on the Nominal Response Modelling (NRM) scoring

Nominal scale

	SJQ29 Lik	SJQ29 FC	SJQ30 Lik	SJQ30 FC	SJQ31 Lik	SJQ31 FC
Pooled Model (both target populations)	0.71	0.41	0.71	0.37	0.70	0.47
Staff of children under 3 years old	0.71	0.36	0.66	0.40	0.74	0.59
Staff of children from 3-5 years-old	0.69	0.21	0.66	0.51	0.70	0.44

Note: “Lik” denotes Likert format, while “FC” denotes Forced choice.

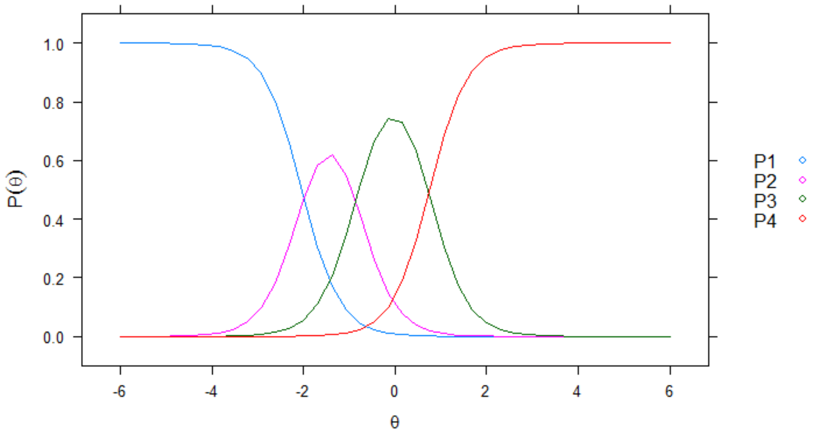
The results in Table 3.4 show that the Likert scale format of the SJQs have reliabilities around 0.7 which is fairly high, and also considerably and consistently higher than all forced choice formats.

3.3.2. Item difficulty and discrimination

Figure 3.3 provides the trace lines for the Likert scale format of SJQ item SQ31b (i.e. “*I would let children play by themselves and only intervene when they request it*“). The horizontal axis measures the respondents’ ability and the vertical axis measure the probability of choosing one of the response options (e.g. “*I would definitely do this*”). For this item, ability refers to the respondents’ ability to understand the importance of free play for children. The probability of choosing option 1 (i.e. “*I would definitely not do this”*) is illustrated by the blue line or P1, which is high for low ability scorers, and low for high ability scorers. In line with our hypothesis and theory, only staff with low abilities should choose this option, as this would reflect low process quality. Line P4 on the other hand, shows that scorers of high ability were likely to choose response option 4 (i.e. “*I would definitely do this”*). In other words, this item was able to discriminate between scorers of high and low ability. This is also evident when inspectingthe maxima of the four lines. The four peaks are separated, reflecting a well-functioning item.

Figure 3.3. Trace lines for Likert scale format of Situational Judgement Question item 31b "I would let children play by themselves and only intervene when they request it"

Both target populations

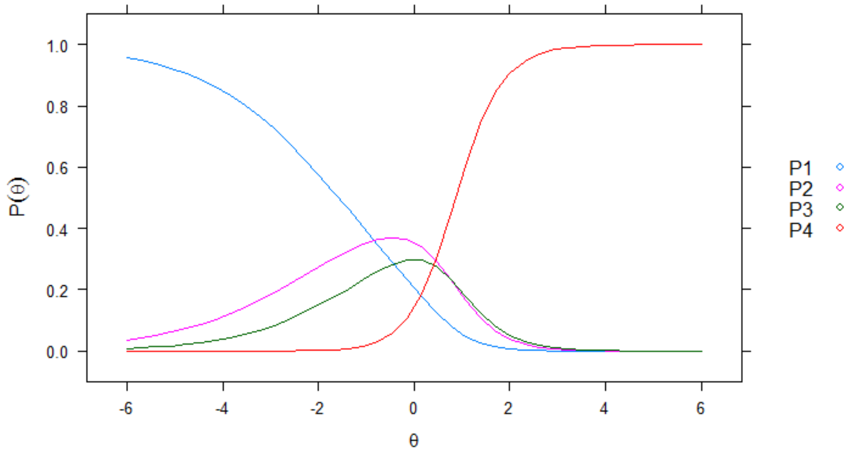


Note: The y-axis shows the probability of choosing one of the four options, while the x-axis reflects the ability of the scorers. P1 is the probability of choosing the first response option, P2 is the probability of choosing the second, etc.

In contrast, the trace line diagrams from the forced choice SJQ items show poor psychometric properties and low discrimination. For each forced choice SJQ item, there is one trace line diagram for the first choice (see Figure 3.4) and one for each of the remaining options left for the second choice (see Figure 3.5). If the respondent chooses a as first choice, then he/she is left with choice b, c, and d as second choice. There are thus four trace line diagrams for the second-choice option (depending on whether the respondent chose a, b, c, or d as first choice).

Figure 3.4. Trace lines for forced choice format of Situational Judgement Question 31, first choice

Both target populations.

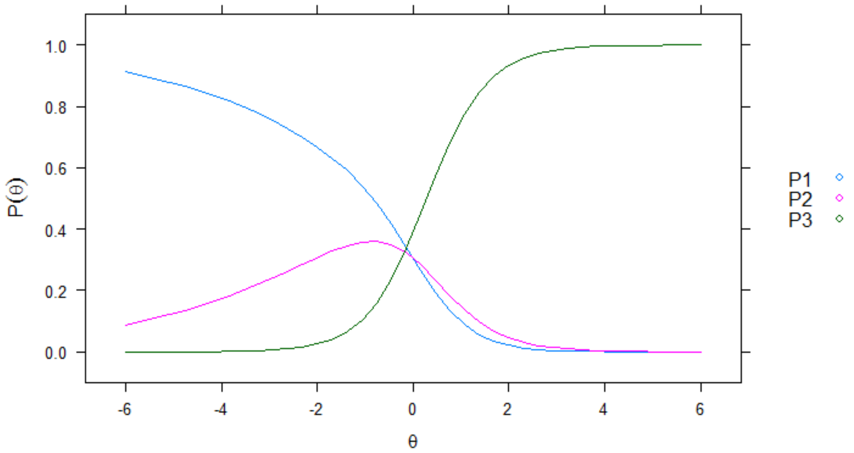


Note: The y-axis shows the probability of choosing one of the four options, while the x-axis reflects the ability of the scorers. P1 denotes the probability of choosing a as first choice, P2 is the probability of choosing the b, etc.

Figure 3.4 shows the trace lines for forced-choice option of SJQ31, first choice. The diagram shows poorer psychometric properties than that the Likert scale SJQ in Figure 3.3. Here, P1 is the probability of choosing a as first choice, P2 is the probability of choosing b as first choice, etc. Response option d (i.e. "*I would play along with the children and follow their lead*") is the one reflecting highest process quality (here supporting child-directed play). The item is not able to discriminate between respondents of low and high ability; the maxima are not separated and P2 and P3 have small maxima meaning that item b and c are bad at discriminating between high and low process quality.

Figure 3.5. Trace lines for forced choice option of Situational Judgement Question 31, second choice assuming 31a is the first choice

Both target populations



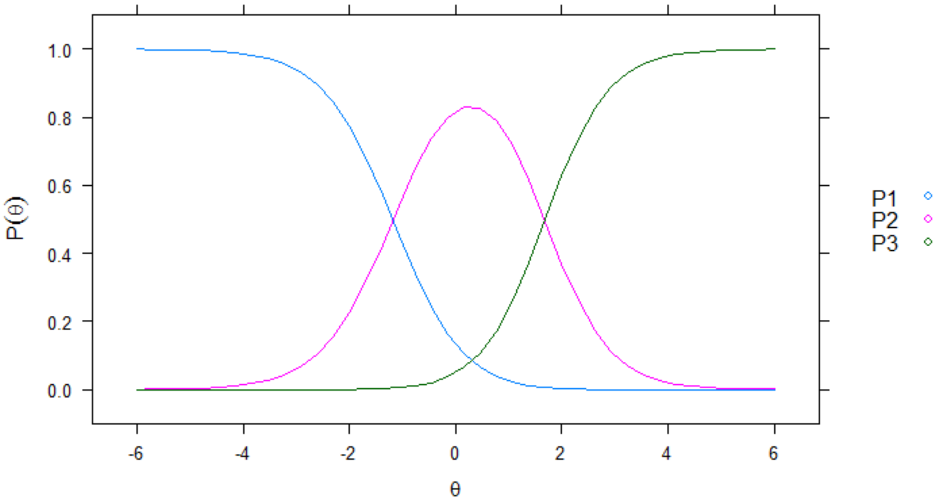
Note: The y-axis shows the probability of choosing one of the four options, while the x-axis reflects the ability of the scorers. P1 is the probability of choosing the b, when a is the first choice, P2 is the probability of choosing c, when a was the first choice, and P3 is the probability of choosing d when a was the first choice.

Figure 3.5 shows the trace lines for forced choice option of SJQ31, second choice. The trace line diagram shows the probabilities b (P1), c (P2) or d (P3) for the second choice when 31a (i.e. "*I would let children play by themselves without intervening*") was the first choice. Like most second-choice traces line diagrams, it exhibits very poor psychometric properties and low discrimination. However, the exceptions are the items within each second-choice SJQ that were ranked as the correct choice (according to theory, experts and popularity). These correct items were the forced choice formats (see Figure 3.5): SJQ item 29d (i.e. "*I would encourage them to build something together*"); SJQ item 30d (i.e. "*I would resolve the conflict together with child A and B*"); and SJQ item 31d (i.e. "*I would play along with the children and follow their lead*").

Figure 3.6 is an example of such a second-choice trace line diagram, and exhibits forced choice SJQ31 (Encouraging pro-social behaviour, sharing and collaboration among children) for the second choice, if the correct response 31d was chosen as first option. The diagram reflects excellent psychometric properties and good discrimination.

Figure 3.6. Trace lines for forced choice option of Situational Judgement TaskQuestion 31, second choice assuming 31d is the first choice

Both target populations



Note: The y-axis shows the probability of choosing one of the four options, while the x-axis reflects the ability of the scorers. P1 is the probability of choosing the b, when a is the first choice, P2 is the probability of choosing c, when a was the first choice, and P3 is the probability of choosing d when a was the first choice.

To summarise, the Likert-scale format of the SJQs showed overall better psychometric properties than the forced choice options of the SJQs. For the forced choice options of the SJQs, the first choice could not discriminate between low and high process quality and provided little information. The second-choice option had even poorer psychometric properties, unless the respondent had the correct first choice.

3.3.3. Factor loadings and explained variance

Table 3.5 shows the factor loadings (F1) of each item in Likert scale format of SJQ29 (i.e. Encouraging pro-social behaviour) as well as the communality (h2: explained variance in the variable by the common factors). The factor loadings range between 0.25 (which is low) and 0.94 (which is very high). Items a and b have much lower factor loadings than the rest, and these two items also reflect the lowest level of pro-social behaviour according to our ranking based on theory. These items also explain less of the variance of the underlying concept than all the other items.

Table 3.5. Factor loadings (F1) and explained variance (h2) of Situational Judgement Question 29, Likert scale format, for both populations

Suppose that you notice that two three-year old children are independently playing with building blocks. Child A has taken almost all the building blocks and is building things. Child B is shy, looks a bit sad and is struggling with his/her construction. What would you do?

	F1	h2
a) I would divide the building blocks in two equal piles, so that both children have an equal number of building blocks	0.25	0.06
b) I would help child B in building a construction	0.25	0.06
c) I would encourage them to build something together	0.67	0.45
d) I would talk to child A to try to make him/her aware of child B's feelings.	0.68	0.46
e) I would encourage child A to share with child B	0.94	0.89

Table 3.6 shows the factor loadings and explained variance of the same item but for the Forced choice scale format. The factor loadings and explained variances are much lower than for the Likert scale format. The first choice provides little information about the underlying trait, as do all the second choices, except for the second choice when d is chosen as first choice. This is the correct choice, and explains the higher factor loading and explained variance.

Table 3.6. Factor loadings (F1) and explained variance (h2) of item Situational Judgement Question 29, Forced choice scale format. Both populations

Suppose that you notice that two 3-year old children are independently playing with building blocks. Child A has taken almost all the building blocks and is building things. Child B is shy, looks a bit sad and is struggling with his/her construction. What would you do?

	F1	h2
First choice.	0.35	0.12
Second choice if a) was the first choice: I would divide the building blocks in two equal piles, so that both children have an equal number of building blocks.	0.38	0.15
Second choice if b) was the first choice: I would encourage child A to share with child B.	0.45	0.21
Second choice if c) was the first choice: I would help child B in building a construction.	0.32	0.10
Second choice if d) was the first choice: I would encourage them to build something together.	0.69	0.47

Table 3.7 shows the factor loadings (F1) of each item in Likert scale SJQ30 (two children quarrelling) as well as the communality (h2: explained variance in the variable by the common factors). The factor loadings range between 0.37 (which is fairly low) and 0.83 (which is high). Item a (i.e. "*I would speak firmly to child A while the other children are listening"*) has a lower factor loading than the others, and hence seems to measure something slightly different than the underlying concept of the scale. This item also explains less of the variance of the underlying concept than all the other items. The same may be said for item g. These two items are also the ones that reflect the worst behavioural management according to our ranking based on theory.

Table 3.7. Factor loadings (F1) and explained variance (h2) of item Situational Judgement Task 30, Likert scale format

Both populations

Suppose that you see two children of the same age (three years) and size quarrelling, and one child (child A) hits the other (child B). Child B is crying. Child A has previously regularly hit other children. Several other children are watching. What would you do?

	F1	h2
a) I would speak firmly to child A while the other children are listening	0.37	0.14
b) I would ask child A and B what happened	0.52	0.27
c) I would warn child A that if he/she hits again he/she will face negative consequences.	0.46	0.21
d) I would tell child A what he/she should have done differently in this situation	0.61	0.37
e) I would ask child A to apologise to child B.	0.75	0.56
f) I would remind child A of our rules, that hitting is not allowed.	0.83	0.68
g) I would include the other children in the discussion after the conflict.	0.38	0.15

The factor loadings of the same question, but with a forced choice format (Table 3.8), are in general lower than for the Likert scale format and explain less of the variance. This again provides evidence that the Likert scale format is the better format for these SJQs.

Table 3.8. Factor loadings (F1) and explained variance (h2) for Situational Judgement Task 30, Forced choice format

Both populations

Suppose that you see two children of the same age (three years) and size quarrelling, and one child (child A) hits the other (child B). Child B is crying. Child A has previously regularly hit other children. Several other children are watching. What would you do?

	F1	h2
First choice	0.31	0.09
Second choice if a) was the first choice: I would speak firmly to child A while the other children are listening.	0.65	0.42
Second choice if b) was the first choice: I would focus on child B and comfort him/her.	0.47	0.22
Second choice if c) was the first choice: I would tell the children who was wrong and who was right.	0.50	0.24
Second choice if d) was the first choice: would resolve the conflict together with child A and B.	0.41	0.16

The factor loadings for the last SJQ, SJQ31 are provided in Table 3.9 (Likert format) and Table 3.10 (Forced-choice). The item measures whether staff facilitate child-initiated play and learning by joining the children. For the Likert scale format, items b and d have lower factor loadings than the other items, and item b has a negative factor loading. These are also the only two items that do not reflect the ability to join child-initiated play. The item with highest factor loading and explained variance is item e. This item measures exactly what is aimed for, namely facilitating child-initiated play *and* learning by joining the children.

Table 3.9. Factor loadings (F1) and explained variance (h2) for Situation Judgement Task 31, Likert scale format

Both populations

Suppose that five 3-year old children are playing with different toys of their choosing. In an ideal situation where you could choose what to do during this time, what would you do?

	F1	h2
a) I would play with the children by following their lead.	0.54	0.29
b) I would let children play by themselves and only intervene when they request it.	-0.09	0.01
c) I would contribute to children’s play by asking questions or providing explanations.	0.82	0.67
d) I would encourage children to play together rather than joining in their play.	0.32	0.11
e) I would contribute to children’s play by providing new ideas or material.s	0.76	0.57

The Forced choice format of SJQ31 (Table 3.10) again shows that the factor loadings and explained variances are overall lower than for the Likert scale format, and also harder to interpret when comparing theory and empirical results.

Table 3.10. Factor loadings (F1) and explained variance (h2) for Situational Judgement Task 31, Forced choice format. Both populations

Suppose that five 3-year old children are playing with different toys of their choosing. In an ideal situation where you could choose what to do during this time, what would you do?

	F1	h2
First choice.	0.47	0.22
Second choice if a) was the first choice: I would let children play by themselves without intervening.	0.46	0.21
Second choice if b) was the first choice: I would let children play by themselves and only intervene when they request it.	0.44	0.19
Second choice if c) was the first choice: I would contribute to children’s play, for instance by asking questions or providing new ideas or materials.	0.40	0.16
Second choice if d) was the first choice: I would play along with the children and follow their lead.	0.41	0.17

3.4. To what degree are the Situational Judgement Questions (SJQs) comparable across the two Starting Strong 2018 target populations? To what degree are they comparable across countries? (RQ.4)

The results of the measurement invariance analyses showed that scalar invariance was met for the Likert format of the SJQs, except for SJQ29, which showed metric invariance. Hence, for SJQ29, only relations to other items or constructs may be compared across the age populations, while construct means may be compared across the populations for the other two SJQs. The results from the measurement invariance analysis for the forced choice SJQs were poorer, and neither metric nor scalar invariance was met. Hence, again the Likert scale SJQs seemed the more appropriate choice than the forced-choice format.

Table 3.11 shows the goodness of fit for Likert scale SJQ30 (see goodness of fit for the other SJQs in Annex B).

Table 3.11. Goodness of fit indices for measurement invariance across target populations for Likert scale format Situational Judgement Task SJQ30

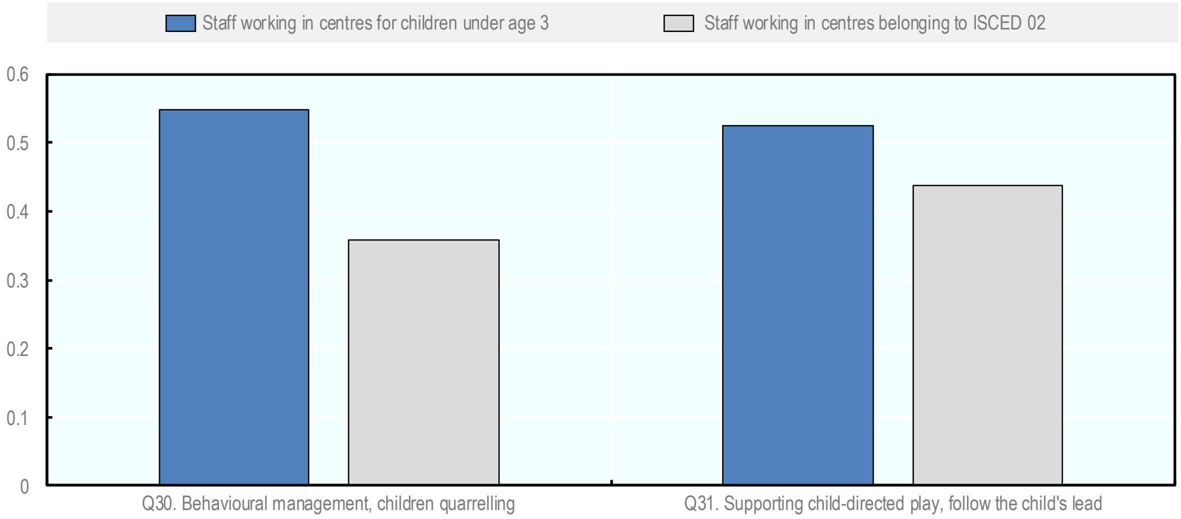
Level of invariance	AIC	AICc	SABIC	BIC	Loglikelihood	χ2	Df	P
Configural	47075.9	47077.42	47218.71	47374.4	-23489			
Metric	46997.57	46999.56	47160.77	47338.71	-23442.8	92.334	7	0
Metric	47318.68	47319.25	47406.11	47501.43	-23629.3			
Scalar	47075.9	47077.42	47218.71	47374.4	-23489	280.778	19	0
Configural	47318.68	47319.25	47406.11	47501.43	-23629.3			
Scalar	46997.57	46999.56	47160.77	47338.71	-23442.8	373.112	26	0

Notes: Akaike Information Criterion (AIC), Akaike Information Criterion small-sample equivalent (AICc), sample size adjusted Bayesian Information Criterion (SABIC), Bayesian Information Criterion (BIC), chi-square independence test (χ2), degrees of freedom (Df), p-value (P).

Seeing how scalar invariance was met across the two age groups for SJQ30 and 31, comparing means between the two groups is possible. The results exhibited in Figure 3.7 show the factor scores for each of the three Likert scale formats of the SJQs for the staff of the two different target populations. Staff working with the older children report higher behavioural management practices and support for child-directed play to a larger extent than the staff of the younger children. However, a t-test reveals that only the differences for behavioural management between the two target populations are statistically significant (p=.034).

Figure 3.7. Mean factor scores on SJQ30 and 31 for the two target populations

Mean factor scores



3.5. How valid are the Situational Judgement Questions (SJQs) in the assessment of early childhood education and care process quality?

3.5.1. Content validity

Table 3.12 shows the correlations of the SJQ items with each other and with other items in the TALIS Starting Strong Field Trial questionnaire that are measuring similar constructs. The correlations between the SJQs are moderate in size and all significant. The correlation coefficients ranged between 0.34 (between SJQ31 and SJQ29) and 0.40 (between SJQ30 and SJQ29). The correlations with the other constructs are small, but most are significant. Overall, the correlations appear the strongest for encouraging pro-social behaviour (i.e. "children sharing") with all other items. Moreover, the correlations are strongest for the items with the greatest conceptual overlap in the measured construct, supporting the discriminant validity. For the SJQ on behavioural management (i.e. "handling children quarrelling"), the correlation between this and encouraging pro-social behaviour (i.e. "children sharing") is smaller. This SJQ may appear to show less discriminatory power as there does not appear to be a stronger correlation with variables that are stronger aligned conceptually. The two variables most likely to correlate to the SJQ on behavioural management are "I help children to follow the rules" and "I help children understand the consequences if they do not follow the rules" (these two items are not part of the SJQs). While "I help children to follow the rules" has a weak, but significant correlation to the SJQ30 on behavioural management (.08), "I help children understand the consequences if they do not follow the rules" has a stronger and significant correlation to the SJQ30 (.23).

However, it must be noted that the staff items in general show less alignment with this SJQ in comparison to the SJQ on encouraging pro-social behaviour (i.e. "children sharing"). Moreover, it seems likely that items focusing on encouraging pro-social behaviour (i.e. "children sharing") and behaviour management (i.e. "handling children quarrelling") show more conceptual overlap between the two, than with the play items for instance, which is the case here. Therefore, the lower correlations of the behavioural management SJQ (i.e. "handling children quarrelling") with the play items are, in fact, supportive of discriminant validity. Lastly, for the SJQ on support for child-directed play (i.e. "following the child's lead"), the items with the strongest conceptual alignment again show the strongest correlations, except for one item. The non-interference of staff in children's play is unrelated to the SJQ on "Support for child-directed play", which makes sense when considering that this SJQ measures the active role of staff in play. Moreover, it is unrelated to the other SJQs as well, which might reflect that this is a poorly functioning item.

Table 3.12. Correlations among Situation Judgement Questions (SJQs) and between SJQs and questionnaire items measuring similar constructs

		SJQ30 Behavioural management (i.e. handling children quarrelling)	SJQ31 Support for child-directed play (i.e. following the child's lead)
SJQ29 Encouraging pro-social behaviour (i.e. "children sharing)		.40***	.34***
SJQ30 Behavioural management (i.e. handling children quarrelling)			.38***
Encouraging sharing among children	.24***	.20***	.13***
Encouraging children to help each other	.25***	.13**	.14***
Encouraging children if they comfort each other	.26***	.20***	.16***
I calm children who are upset	.14***	.20**	.10**
I help children to follow the rules	.20***	.08*	.02
I help children understand the consequences if they do not follow the rules	.09*	.23**	.08*
Not interfering when children play	.07	-.02	-.03
If invited by the children, joining the children's play	.16***	.11**	.23***
Allowing children to take the lead when playing with children	.17***	.09*	.17***

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

3.5.2. Criterion validity

Regarding the content of staff pre-service education it appeared that all aspects were most strongly related to the SJQ on encouraging pro-social behaviour (i.e. "children sharing") and to a lesser extent the SJQ on support for child-directed play (i.e. "following the child's lead"), and behavioural management (i.e. "handling children quarrelling") (see Table 3.13). Moreover, the strongest correlations were found for content that was more specific and/or more strongly aligned with the content of the SJQ. Pre-service education that covered different pedagogical approaches was unrelated to either one of the SJQs and a focus on children's special needs also revealed low and largely insignificant relations with the SJQs. These results provide support for the discriminatory validity of the SJQs.

Staff self-efficacy was positively related to all SJQs. This was especially the case for the SJQ on encouraging pro-social behaviour (i.e. "children sharing"), which shows the strongest alignment with the items on self-efficacy. The item "support special needs", which is an unrelated topic, showed the smallest association with the SJQs, thus supporting the discrimination of the SJQs.

Staff also reported on their perceived control over implementation of practices on a number of domains. The results show that having control over managing children's behavioural problems was only related to the SJQ on behavioural management (i.e. "handling children quarrelling"), whereas most other items were related to the SJQs on encouraging pro-social behaviour (i.e. "children sharing") and on support for child-directed play (i.e. "following the child's lead"), to at least some extent. The exception is having control over choosing play and learning materials, which appeared unrelated to all SJQs. This item covers more specifically the available materials and the way the environment is organised, and is thus not focusing on aspects of interactions as part of process quality.

Table 3.13. Correlations between the SJQs and staff formal education and self-efficacy

	SJQ29 Encouraging pro-social behaviour (i.e. “children sharing)	SJQ30 Behavioural management (i.e. handling children quarrelling)	SJQ31 Support for child-directed play (i.e. following the child’s lead)
Content of staff formal education			
Socio-emotional development	.28***	.11**	.13**
Self-regulation	.21***	.15**	.20**
Facilitating free play	.22***	.11**	.12**
Playgroup management	.24***	.19**	.21**
Children’s special needs	.12*	.04	.05
Different pedagogical approaches	.09	.02	-.02
Self-efficacy			
Adapt work to individual needs	.17***	.08*	.14*
Help to interact with each other	.27***	.07*	.09**
Calm an upset child	.24***	.08*	.12**
Help develop self-confidence	.23***	.11**	.16***
Provide feelings of security	.27***	.11**	.12**
Support special needs	.08*	-.02	.12**
Control over implementation			
Choosing play and learning materials	.06	-.02	.07
Choosing goals for development	.08*	.01	.10**
Monitoring children’s development and well-being	.12**	.06	.17**
Managing children’s behavioural problems	.09*	.11**	.14***
Adapting activities to children’s needs	.16***	.03	.13**

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

4. Discussion

4.1. Summary of the main findings

Overall, the findings from this study showed that the Situational Judgement Questions (SJQs) provided reliable and valid measures of process quality in TALIS Starting Strong 2018 Field Trial.

To summarise the findings from a conceptual perspective, there is high agreement between and within the participating countries as staff chose the following behaviour for each SJQ: supporting child-directed play by following children's lead (SJQ31); managing conflicts through behavioural management, by directing children's attention to the playroom or classroom rules (SJQ30); and supporting pro-social behaviour by encouraging sharing and collaboration among children (SJQ29). This agreement was also high across staff of the two target populations, and between experts, staff and theory.

Staff self-efficacy and formal education (as measured by non-SJQs) were found to be related to process quality (as measured by SJQs), especially to pro-social behaviour.

Regarding comparability and differences between staff of the two target populations, it was more common among staff of children between 3-5 years old to provide behavioural management during conflicts and support child-directed play than for staff of children between 0-2 years old. Comparisons with SJQ 29 (supporting pro-social behaviour) were not possible as the mean of this SJQ was found not to be comparable across the two target populations.

Given that process quality has never been measured in an International Large-scale Survey using SJQ formats, it is important to interpret the findings of this study in light of previous research.

Behavioural management was one of three key aspects of process quality measured by the SJQs. Behavioural management is considered to be an important aspect of staff process quality in early childhood education and care. This study shows high agreement within and between countries that staff would provide behavioural guidance during a conflict between children. More specifically, most participants across participating countries agreed that they would direct the children's attention to the classroom rules (i.e. the most appropriate practice given the context).Thus, there is strong evidence of validity for this SJQ.

It is especially relevant to provide behavioural guidance in the early years of development as children gradually learn to regulate their emotions and interact with peers (Bronson, 2000[20]; Kopp, 1982[22]). Yet, the findings of this study show that it was more common to provide behavioural management during conflicts among staff working in centres belonging to ISCED Level 02, than among staff working in centres for children under the age of three years. However, even though it may be important to teach children appropriate behaviour at an early age, staff may have higher expectations of older children's behaviour, and more actively practice behavioural management during conflicts between older children than younger children. It may also be that the SJQ describes a staff practice that is not suited to the behavioural management of the youngest of children in centres for children under the age of three years, i.e. it may not be age-appropriate to direct a one-year-old child's attention to the classroom rules. Therefore, it could be that, even though the context of the SJQ described a quarrel between two 3-year-old children, the SJQ may work better for staff of children between 3-5 years old.

The second key aspect of process quality measured by the SJQs was pro-social behaviour, more specifically staff support of pro-social behaviour by encouraging sharing and collaboration among children. This is an important aspect of children's social skills necessary for successful peer interactions (Howes, 2011[27]). Yet, there is little research on this child-centred perspective on facilitating peer interactions. The majority of the research focuses on adult-centred or one-to-one interactions between staff and children [e.g. Helmerhorst et al. (2014[31]), Williams et al. (2010[32])]. This study therefore contributes to the field; the related SJQ was the one showing the highest correlations with staff education and self-efficacy. Moreover, staff in most participating countries reported that they would support pro-social behaviour by encouraging sharing and collaboration among children.

The third key aspect of process quality measured by the SJQs was child-directed play. There was high agreement between and within the countries that staff would support child-directed play by following the children's lead. Previous research showed that staff use of child-centred strategies to facilitate peer interactions was positively related to children's peer sociability (Williams et al., 2010). The fact that staff in most countries chose the child-centred strategy ("I would play with the children by following their lead") provides evidence that such option maps on to higher process quality practices.

The findings also indicate a clear preference for the SJQs with response Likert-scale format, rather than the forced choice format. All reliability measures, including internal consistency, the degree to which the SJQs discriminate between staff providing high and low process quality, and factor loadings, showed that the Likert scale SJQs were more reliable than the forced choice format. Moreover, the Likert scale format was easier to interpret, as the second choice in the forced-choice format provided little additional information. With regards to the type of scoring, Nominal Response Modelling (NRM) scoring produced more reliable scores compared to ranking based on theory, popularity scores and expert scores.

Taken together, the evidence summarised in this paper indicates that the psychometric properties of the Likert items are overall better than the Forced choice items. This is in line with a more conceptual point of view: the Likert scale items make more sense as they are easier to interpret. Moreover, it is easier to align them with theory. One reason is that it is theoretically challenging to rank items in the forced choice format SJQs. For instance, the forced choice SJQ29 is meant to measure support of pro-social behaviour. The item "I would encourage them to build something together." is meant to reflect support of pro-social behaviour, while the other items in SJQ29 to a lesser degree reflects support of pro-social behaviour (e.g. "I would help child B in building a construction"). It is however, challenging to rank all the items that are meant to reflect a lower degree of support for pro-social behaviour. Moreover, more effort was made to create items that were not perceived as obviously "wrong choices", than to rank the items reflecting low levels of process quality. This was done to avoid social desirability.

In the forced choice SJQs, little information was provided by the second choice in the if the respondent's first choice was wrong. It seems challenging for the respondents to interpret which item reflects lower process quality than another. Although one item reflected higher process quality, the other choices were not ranked based on prior evidence. Another potential interpretation for why the second choice provided little information could be because the respondents' second choice was more or less arbitrary as they were not able to distinguish poor from good practices. Hence, the second choice has less meaning conceptually and psychometrically.

The examination of factor loadings showed that together, Likert scale items measure the same latent trait for a SJQ. In consequence, the mean of the construct reflects the level of a key aspect of process quality. The factor analysis done using Item Response Theory (IRT) showed that items with low factor loadings are the ones with the lowest level of process quality. The low factor loadings indicate that the items measure something slightly different conceptually than the others, i.e. these items do not tap into the underlying concept to the same degree as items with high factor loadings.

Diagrams exhibiting the results from both the descriptive analysis, as well as the Item Response Theory analyses, provided useful tools to compare results across the SJQs, across items within each SJQ, across formats and across methods of analysis. Generally, the different types of analyses produced the same results and emphasised the reliability and validity of the SJQs. For instance, an SJQ item that exhibited clear patterns and high agreement within and across countries in the profile diagrams based on descriptives, also exhibited desirable patterns in the trace line diagrams based on IRT (e.g. ability to discriminate between low and high abilities). The very same items also showed high factor loadings and explained variance.

Tests of measurement invariance for the two target populations (i.e. staff working in centres for children under the age of three years, and staff working in centres belonging to ISCED Level 02) demonstrated a high degree of comparability. The results show that scalar invariance is met for all Likert format SJQs but one (i.e. SJQ29 measuring support of pro-social behaviour), which showed metric invariance. In other words, for the two target populations it is possible to compare the means of all SJQs (but one), and for *all* SJQs it is possible to compare relationships to other constructs (e.g. staff education) across staff of the two target populations. Measurement invariance analysis of the Forced choice items produced poorer results and poorer fit than the Likert items.

Even though measurement invariance analysis across countries was not possible to test (i.e. because there were too few respondents in the Field Trial sample for some countries), results from the descriptive analysis show promising results, because the agreement across countries is high. The main study will include larger sample sizes, enabling investigations of measurement invariance across countries for the two target populations.

Upon comparing outcomes from measurement invariance analysis across the two target populations of the SJQs with all the 14 n on-SJQ constructs measuring process quality, only one of the other non-SJQ constructs was scalar invariant across the two target populations (i.e. Language stimulation and support for literacy learning), three were metric invariant (i.e. Staff support for literacy and numeracy) and some constructs did not reach invariance (e.g. Staff behavioural management, Facilitating play and child-initiated activities). This may indicate that the SJQs are more robust measures of process quality in terms of measurement invariance across target populations, and potentially less prone to cultural bias. Even so, it is advisable to cluster countries together according to culture, language, educational systems, etc, and perform further analysis in the main study within these groups of countries rather than across all countries. This goes for all types of scales, regardless of whether they are SJQs or not, as practices and cultural backgrounds may vary a lot across countries (Blömeke, Olsen and Suhl, 2016[74]).

Scoring based on Nominal Response Modelling (NRM) provided more reliable results. However, the reliability of the scoring based on ECEC experts could not be evaluated, as there were too few respondents, although descriptive statistics reflected a very high agreement between the experts and the ranking of the items based on theory. Another important result was the high agreement between scoring based on theory, scoring based on popularity, expert scoring and NRM. This result is promising with regards to cultural bias, as it may reflect agreement between theory and countries' own rankings of process quality. For these SJQs, it seems that countries agree what characterises high and low process quality. However, these results need to be validated by measurement invariance analyses of the data from the Main Survey.

Taken together, the results support validity of the SJQs. Overall, the strongest relations were found when the theoretical alignment between the measured concepts was higher, thus supporting the convergent validity of the SJQs. Other (unrelated) concepts showed no or only small correlations with the SJQs, underlining the discrimination validity of the SJQs.

The results also supported criterion validity, which is in line with a review of SJQs (Lievens, Peeters and Schollaert, 2008[9]).The SJQ on pro-social behaviour showed the strongest relations with other items of the questionnaire. One explanation is that this SJQ was most closely aligned in content to some of the items in the questionnaire. Another explanation might be that this SJQ covers a very basic aspect of process quality that is likely to be related to other aspects of quality as well. Altogether, the results show that the SJQs show sufficient construct and criterion validity, thus they can be useful in evaluating staff ECEC process quality, and more specifically the three key aspects of process quality; encouraging pro-social behaviour, behavioural management and support for child-directed play.

SJQs were also tried out in the TALIS study, but this was not successful. The reasons why the SJQs worked better in TALIS Starting Strong Survey 2018 than in TALIS are not entirely clear. However, throughout developing the items and implementing the methodology, there was a close follow-up and constant feedback and consultation from the TALIS Starting Strong Survey 2018 Technical Advisory Group, Questionnaire Expert Group, the National Project Managers, and experts in ECEC. This most likely contributed to enhancing the reliability and validity of the items.

4.2. Limitations

Despite its innovative nature, and promising results regarding the validity and reliability of the Situational Judgement Questions in the TALIS Starting Strong 2018 Field Trial, this study also presents some limitations.

First, the study presented a small sample size in the Field Trial, which meant that it could not examine measurement invariance across countries, and therefore there is limited evidence of reduced cultural bias.

Second, there is evidence to hypothesise that the forced choice format could have worked better psychometrically and been easier to interpret conceptually, had participants been asked to choose only one response, rather than having to rate the options by first and second choice. Moreover, it would have been easier to compare the Likert and the Forced choice formats, if they contained exactly the same items. This poor alignment between Likert and Forced choice format items limits the strength of the evidence of one format over the other.

Third, the paper shows no hard evidence of reduced social desirability bias because staff practices were not directly observed for these staff. SJQs are still self-reported behaviours and could be prone to potential social desirability.

Finally, and despite considerable collaborative efforts in the design of the initial six SJQs to include all the proposed subdomains of process quality, only three were included in the Field Trial. This means that as a measure of process quality, these three SJQs are limited and need to be used in collaboration with measures of other domains and subdomains of process quality.

Despite these limitations, the SJQs are still a promising, less time and resource consuming way to describe process quality while staying closer to what staff do than other self-report measures.

4.3. Key insights

There was high agreement between and within the countries on what the best practices are: to support child-directed play by following the children's lead; to manage conflicts through behavioural management by directing children's attention to the classroom rules; and to support pro-social behaviour by encouraging sharing and collaboration among children.

Staff self-efficacy and formal education were related to process quality, especially to pro-social behaviour.

Most items worked well, and the Likert format seemed to be the better choice over Forced choice format, just like the nominal response scoring showed better psychometric properties than rankings.

Overall the analyses provided thorough evidence for good reliability and validity of the SJQs, and hence a promising approach for the Main Survey data collection.

References

Ainsworth, M. et al. (1978), *Patterns of attachment: Assessed in the strange situation and at home*, Hillsdale, NJ: Erlbaum. [19]

Barros, S. et al. (2016), "Infant child care quality in Portugal: Associations with structural characteristics", *Early Childhood Research Quarterly*, Vol. 37, pp. 118-130, https://doi.org/10.1016/j.ecresq.2016.05.003. [11]

Bertram, T. and C. Pascal (2016), *Early childhood policies and systems in eight countries: Findings from IEA's early childhood education study*, IEA, Hamburg, https://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ECES-policies_and_systems-report.pdf. [4]

Bhavnagri, N. and R. Parke (1991), "Parents as direct facilitators of children's peer relationships: Effects of age of child and sex of parent", *Journal of Social and Personal Relationships*, Vol. 8/3, pp. 423-440, http://dx.doi.org/10.1177/0265407591083007. [29]

Blair, C. (2003), "Behavioral inhibition and behavioral activation in young children: Relations with self-regulation and adaptation to preschool in children attending Head Start", *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, Vol. 42/3, pp. 301-311, http://dx.doi.org/10.1002/dev.10103. [23]

Blömeke, S., R. Olsen and U. Suhl (2016), "Relation of Student Achievement to the Quality of Their Teachers and Instructional Quality", in Nilsen, T. and J. Gustafsson (eds.), *Teacher Quality, Instructional Quality and Student Outcomes*, Springer. [74]

Bock, R. (1972), "Estimating item parameters and latent ability when responses are scored in two or more nominal categories", *Psychometrika*, Vol. 37/1, pp. 29-51. [71]

Bolt, D., Y. Lu and J. Kim (2014), "Measurement and control of response styles using anchoring vignettes: A model-based approach", *Psychological Methods*, Vol. 19/4, p. 528, http://dx.doi.org/10.1037/met0000016. [57]

Broekhuizen, M. et al. (2015), *Stakeholders Study. Values, beliefs and concerns of parents, staff and policy representatives regarding ECEC services in nine European countries: First report on parents*, EU CARE project, http://ecec-care.org/fileadmin/careproject/Publications/reports/CARE_WP6_D6_2_European_ECEC__Stakeholder_study_FINAL.pdf. [26]

Bronfenbrenner, U. and P. Morris (2006), *Handbook of child psychology. Vol. 1: Models of human development*, Wiley Hoboken, NJ. [3]

Bronson, M. (2000), *Self-regulation in early childhood: Nature and nurture*, Guilford Press, New York, NY. [20]

Burchinal, M., K. Kainz and Y. Cai (2011), “How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings”, in Zaslow, M. et al. (eds.), *Quality Measurement in Early Childhood Settings*, aul H Brookes Publishing, Baltimore, MD. [34]

Burchinal, M. et al. (2010), “Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs”, *Early Childhood Research Quarterly*, Vol. 25/2, pp. 166-176, https://doi.org/10.1016/j.ecresq.2009.10.004. [43]

Burrus, J., B. Naemi and P. Kyllonen (2011), “Intentional and unintentional faking in education1”, in *New Perspectives on Faking in Personality Assessment*, Oxford University Press, New York, NY, http://dx.doi.org/10.1093/acprof:oso/9780195387476.003.0082. [70]

Cassidy, D. et al. (2005), “Revisiting the two faces of child care quality: Structure and process”, *Early Education and Development*, Vol. 16/4, pp. 505-520, https://doi.org/10.1207/s15566935eed1604_10. [35]

Catano, V., A. Brochu and C. Lamerson (2012), “Assessing the reliability of situational judgment tests used in high-stakes situations”, *International Journal of Selection and Assessment*, Vol. 20/3, pp. 333-346, https://doi.org/10.1111/j.1468-2389.2012.00604.x. [66]

Charlesworth, R. et al. (1993), “Measuring the developmental appropriateness of kindergarten teachers’ beliefs and practices”, *Early Childhood Research Quarterly*, Vol. 8/3, pp. 255-276, https://doi.org/10.1016/S0885-2006(05)80067-5. [48]

Chen, F. (2007), “Sensitivity of goodness of fit indexes to lack of measurement invariance”, *Structural Equation Modeling*, Vol. 14/3, pp. 464-504, https://doi.org/10.1080/10705510701301834. [73]

Colwell, N. et al. (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*, Vol. 28/2, pp. 218-233, http://dx.doi.org/10.1016/j.ecresq.2012.12.004. [36]

Creemers, B. and L. Kyriakides (2006), “Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model”, *School Effectiveness and School Improvement*, Vol. 17/3, pp. 347-366, http://dx.doi.org/10.1080/09243450600697242. [55]

Cryer, D. et al. (1999), “Predicting process quality from structural quality in preschool programs: A cross-country comparison”, *Early Childhood Research Quarterly*, Vol. 14/3, pp. 339-361, http://dx.doi.org/10.1016/S0885-2006(99)00017-4. [37]

de Kruif, R. et al. (2000), “Classification of teachers’ interaction behaviors in early childhood classrooms”, *Early Childhood Research Quarterly*, Vol. 15/2, pp. 247-268, http://dx.doi.org/10.1016/S0885-2006(00)00051-X. [38]

Emmer, E. and L. Stough (2001), “Classroom management: A critical part of educational psychology, with implications for teacher education”, *Educational psychologist*, Vol. 36/2, pp. 103-112, http://dx.doi.org/10.1207/S15326985EP3602_5. [24]

Ghazvini, A. and R. Mullis (2002), “Center-based care for young children: Examining predictors of quality”, *The Journal of Genetic Psychology*, Vol. 163/1, pp. 112-125, http://dx.doi.org/10.1080/00221320209597972. [12]

Gordon, R. et al. (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*, Vol. 49/1, p. 146, http://dx.doi.org/10.1037/a0027899. [39]

Guo, H., J. Zu and P. Kyllonen (2016), “Validation of the NRM-based scoring rules for a situational judgment test”, *Educational Testing Service*. [67]

Guo, H. et al. (2016), *Evaluation of Different Scoring Rules for a Noncognitive Test in Development*, http://dx.doi.org/10.1002/ets2.12089. [69]

Hamre, B. et al. (2014), *Classroom Assessment Scoring System (CLASS) Manual, Infant*, Paul H. Brookes Publishing, Baltimore, MD. [17]

Hamre, B. et al. (2013), “Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms”, *The Elementary School Journal*, Vol. 113/4, pp. 461-487, http://dx.doi.org/10.1086/669616. [18]

Hatfield, B. et al. (2016), “Thresholds in the association between quality of teacher–child interactions and preschool children’s school readiness skills”, *Early Childhood Research Quarterly*, Vol. 36, pp. 561-571, http://dx.doi.org/10.1016/j.ecresq.2015.09.005. [44]

He, J. and F. Van de Vijver (2016), “The motivation-achievement paradox in international educational achievement tests: Toward a better understanding”, in King, R. and A. Bernardo (eds.), *The Psychology of Asian Learners*, Springer, Singapore. [54]

He, J. and F. Van de Vijver (2013), “Methodological issues in cross-cultural studies in educational psychology”, in Liem, G. and A. Bernardo (eds.), *Advancing crosscultural perspectives on educational psychology: A festschrift for Dennis McInerney*, Information Age Publishing, Charlotte, NC. [53]

Helmerhorst, K. et al. (2014), “Measuring the interactive skills of caregivers in child care centers: Development and validation of the caregiver interaction profile scales”, *Early Education and Development*, Vol. 25/5, pp. 770-790, http://dx.doi.org/10.1080/10409289.2014.840482. [31]

Howes, C. (2011), “Social play of children with adults and peers”, in *The Oxford Handbook of the Development of Play*, http://dx.doi.org/10.1093/oxfordhb/9780195393002.013.0018. [27]

Howes, C. et al. (2008), “Ready to learn? Children’s pre-academic achievement in pre-kindergarten programs”, *Early Childhood Research Quarterly*, Vol. 23/1, pp. 27-50, https://doi.org/10.1016/j.ecresq.2007.05.002. [13]

Klieme, E., C. Pauli and K. Reusser (2009), “The pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms”, in Janik, T. (ed.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom*, Waxmann, Münster. [60]

Kopp, C. (1982), “Antecedents of self-regulation: A developmental perspective”, *Developmental Psychology*, Vol. 18/2, pp. 199-214, http://dx.doi.org/10.1037/0012-1649.18.2.199. [22]

Kuger, S. and K. Kluczniok (2008), “Process Quality in Kindergartens-Concepts, implementation and findings”, *Zeitschrift für Erziehungswissenschaft*, Vol. 10, pp. 159-178. [51]

Kyllonen, P. and J. Bertling (2013), “Innovative questionnaire assessment methods to increase cross-country comparability”, in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, Chapman and Hall/CRC, New York. [8]

Ladd, G. and C. Hart (1992), “Creating informal play opportunities: Are parents’ and preschoolers’ initiations related to children’s competence with peers?”, *Developmental Psychology*, Vol. 28/6, pp. 1179-1187, http://dx.doi.org/10.1037/0012-1649.28.6.1179. [30]

Layzer, J. and B. Goodson (2006), “The “quality” of early care and education settings: Definitional and measurement issues”, *Evaluation Review*, Vol. 30/5, pp. 556-576, http://dx.doi.org/10.1177/0193841X06291524. [40]

Lievens, F. (2017), “Construct-driven SJTs: Toward an agenda for future research”, *International Journal of Testing*, Vol. 17/3, pp. 269-276, http://dx.doi.org/10.1080/15305058.2017.1309857. [62]

Lievens, F., H. Peeters and E. Schollaert (2008), “Situational judgment tests: A review of recent research”, *Personnel Review*, Vol. 37/4, pp. 426-441, http://dx.doi.org/10.1108/00483480810877598. [9]

Melhuish, E. et al. (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).* [10]

Morgeson, F., M. Reider and M. Campion (2005), “Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge”, *Personnel Psychology*, Vol. 58, pp. 583 - 611, http://dx.doi.org/10.1111/j.1744-6570.2005.655.x. [63]

Motowidlo, S., M. Dunnette and G. Carter (1990), “An alternative selection procedure: The low-fidelity simulation”, *Journal of Applied Psychology*, Vol. 75/6, pp. 640-647, http://dx.doi.org/10.1037/0021-9010.75.6.640. [58]

Muijs, D. (2006), “Measuring teacher effectiveness: Some methodological reflections”, *Educational Research and Evaluation*, Vol. 12/1, pp. 53-74, http://dx.doi.org/10.1080/13803610500392236. [7]

Mussel, P., T. Gatzka and J. Hewig (2016), “Situational judgment tests as an alternative measure for personality”, *European Journal of Psychological Assessment*, Vol. 1/8, http://dx.doi.org/10.1027/1015-5759/a000346. [65]

Nilsen, T. and J. Gustafsson (eds.) (2016), *Teacher Quality, Instructional Quality and Student Outcome. Relationships Across Countries, Cohorts and Time*, Springer, Cham. [61]

OECD (2018), *Engaging Young Children: Lessons from Research about Quality in Early Childhood Education and Care*, Starting Strong, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264085145-en. [5]

OECD (2018), “Preliminary analysis of the structure of process quality in the TALIS Starting Strong Survey (meeting document)”*, Meeting Document EDU/EDPC/ECEC/RD(2018)3*, OECD, Paris. [15]

OECD (2014), *PISA 2012 Technical Report*, OECD, https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf. [59]

Oswald, F. et al. (2004), “Developing a biodata measure and situational judgment inventory as predictors of college student performance”, *Journal of Applied Psychology*, Vol. 89/2, pp. 187–207, https://doi.org/10.1037/0021-9010.89.2.187. [64]

Perlman, M., G. Zellman and V. Le (2004), “Examining the psychometric properties of the early childhood environment rating scale-revised (ECERS-R)”, *Early Childhood Research Quarterly*, Vol. 19/3, pp. 398-412, https://doi.org/10.1016/j.ecresq.2004.07.006. [41]

Pianta, R. and B. Hamre (2009), “Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity”, *Educational researcher*, Vol. 38/2, pp. 109-119, http://dx.doi.org/10.3102/0013189X09332374. [47]

Pianta, R. et al. (2005), “Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions?”, *Applied Developmental Science*, Vol. 9/3, pp. 144-159, http://dx.doi.org/10.1207/s1532480xads0903_2. [14]

Raver, C. (2004), “Placing emotional self-regulation in sociocultural and socioeconomic contexts”, *Child Development*, Vol. 75/2, pp. 346-353, http://dx.doi.org/10.1111/j.1467-8624.2004.00676.x. [25]

Ruzek, E. et al. (2014), “The quality of toddler child care and cognitive skills at 24 months: Propensity score analysis results from the ECLS-B”, *Early childhood research quarterly*, Vol. 29/1, pp. 12-21, http://dx.doi.org/10.1016/j.ecresq.2013.09.002. [45]

Sim, M. et al. (2019), “Starting Strong Teaching and Learning International Survey 2018 Conceptual Framework”*, OECD Education Working Papers*, No. 197, OECD Publishing, Paris, https://dx.doi.org/10.1787/106b1c42-en. [6]

Slot, P. (2018), “Structural characteristics and process quality in early childhood education and care: A literature review”*, OECD Education Working Papers*, No. 176, OECD Publishing, Paris, https://dx.doi.org/10.1787/edaf3793-en. [2]

Slot, P. et al. (2017), "Measurement properties of the CLASS Toddler in ECEC in The Netherlands", *Journal of Applied Developmental Psychology*, Vol. 48, pp. 79-91, https://doi.org/10.1016/j.appdev.2016.11.008. [16]

Slot, P. et al. (2015), "Associations between structural quality aspects and process quality in Dutch early childhood education and care settings", *Early Childhood Research Quarterly*, Vol. 33, pp. 64-76, http://dx.doi.org/10.1016/j.ecresq.2015.06.001. [33]

Sroufe, L. (2000), "Early relationships and the development of children", *Infant Mental Health Journal*, Vol. 21/1-2, pp. 67-74, http://dx.doi.org/10.1002/(SICI)1097-0355(200001/04)21:1/2<67::AID-IMHJ8>3.0.CO;2-2. [21]

Teig, N., R. Scherer and T. Nilsen (2018), "More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science", *Learning and Instruction*, Vol. 56, pp. 20-29, http://dx.doi.org/10.1016/j.learninstruc.2018.02.006. [56]

Thissen, D., L. Cai and R. Bock (2010), "The nominal categories item response model", in Nering, M. and R. Ostini (eds.), *Handbook of Polytomous Item Response Theory Models*, Routledge, Abingdon, http://dx.doi.org/10.4324/9780203861264.ch3. [72]

van de Vijver, F. and J. He (2016), "Bias assessment and prevention in noncognitive outcome measures in context assessments", in Kuger, S. et al. (eds.), *Assessing Contexts of Learning*, Springer, Cham. [52]

van Schaik, S., P. Leseman and S. Huijbregts (2014), "Cultural diversity in teachers' group-centered beliefs and practices in early childcare", *Early Childhood Research Quarterly*, Vol. 29/3, pp. 369-377, http://dx.doi.org/10.1016/j.ecresq.2014.04.007. [28]

Vermeer, H. et al. (2016), "Quality of child care using the environment rating scales: A meta-analysis of international studies", *International Journal of Early Childhood*, Vol. 48/1, pp. 33-60, http://dx.doi.org/10.1007/s13158-015-0154-9. [1]

Walston, J. and J. West (2004), *Full-day and half-day kindergarten in the United States: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99*, US Dept. of Education, Institute of Education Sciences, National Center for Education Statistics. [49]

Weiland, C. et al. (2013), "Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program", *Early Childhood Research Quarterly*, Vol. 28/2, pp. 199-209, http://dx.doi.org/10.1016/j.ecresq.2012.12.002. [46]

Whelpley, C. (2014), *How to Score Situational Judgment Tests: A Theoretical Approach and Empirical Test*, Virginia Commonwealth University, Richmond, VA. [68]

Williams, S., A. Mastergeorge and L. Ontai (2010), "Caregiver involvement in infant peer interactions: Scaffolding in a social context", *Early Childhood Research Quarterly*, Vol. 25/2, pp. 251-266, http://dx.doi.org/10.1016/j.ecresq.2009.11.004. [32]

Xue, Y. and S. Meisels (2004), “Early literacy instruction and learning in kindergarten: Evidence from the early childhood longitudinal study—kindergarten class of 1998–1999”, *American Educational Research Journal*, Vol. 41/1, pp. 191-229, http://dx.doi.org/10.3102/00028312041001191. [50]

Zaslow, M. et al. (2010), *Quality Dosage Thresholds and Features in Early Childhood Settings: A Review of the Literature*, Mathematica Policy Research, Princeton, NJ. [42]

Annex A. The Situational Judgement Questions and items in the TALIS Starting Strong 2018 Field Trial

This annex presents the SJQs that were included in the TALIS Starting 2018 Field Trial. Question numbers refer to the field trial and differ from those in the Main Survey. The Main Survey did not include SJQ30 and adopted the Likert scale format. Items for which the analysis has concluded that there is high agreement between and within the participating countries are indicated in bold.

SJQ29 Supporting pro-social behaviour, Likert scale format
Suppose that you notice that two three-year old children are independently playing with building blocks. Child A has taken almost all the building blocks and is building things. Child B is shy, looks a bit sad and is struggling with his/her construction.
What would you do?
For each suggestion, mark the option that best describes what you would do.

			I would definitely do this	I would probably do this	I would probably not do this	**I would definitely not do this**
	a)	I would divide the building blocks in two equal piles, so that both children have an equal number of building blocks.	□1	□2	□3	**□4**
	b)	I would help child B in building a construction.	□1	□2	□3	**□4**
	c)	**I would encourage them to build something together.**	□1	□2	□3	**□4**
	d)	I would talk to child A to try to make him/her aware of child B's feelings.	□1	□2	□3	**□4**
	e)	**I would encourage child A to share with child B.**	**□1**	**□2**	**□3**	**□4**

SJQ-A29 Supporting pro-social behaviour, forced choice format
Suppose that you notice that two three-year old children are independently playing with building blocks. Child A has taken almost all the building blocks and is building things. Child B is shy, looks a bit sad and is struggling with his/her construction.
What would you most likely do?
Please select your first and second choice.
Please mark two choices in total, one in each column.

		First choice	Second choice
a)	I would divide the building blocks in two equal piles, so that both children have an equal number of building blocks.	□1	**□1**
b)	I would encourage child A to share with child B.	□2	**□2**
c)	I would help child B in building a construction.	□3	**□3**
d)	**I would encourage them to build something together.**	**□4**	**□4**

SJQ30 **Behavioural management, Likert scale format**

Suppose that you see two children of the same age (three years) and size quarrelling, and one child (child A) hits the other (child B). Child B is crying. Child A has previously regularly hit other children. Several other children are watching.

What would you do?

For each suggestion, mark the option that best describes what you would do.

		I would definitely do this	I would probably do this	I would probably not do this	I would definitely not do this
a)	I would speak firmly to child A while the other children are listening.	□1	□2	□3	□4
b)	I would ask child A and B what happened.	□1	□2	□3	□4
c)	I would warn child A that if he/she hits again he/she will face negative consequences.	□1	□2	□3	□4
d)	I would tell child A what he/she should have done differently in this situation.	□1	□2	□3	□4
e)	I would ask child A to apologise to child B.	□1	□2	□3	□4
f)	**I would remind child A of our rules, that hitting is not allowed.**	□1	□2	□3	□4
g)	I would include the other children in the discussion after the conflict.	□1	□2	□3	□4

SJQ-A30 **Behavioural management, forced choice format**

Suppose that you see two children of the same age (three years) and size quarrelling, and one child (child A) hits the other (child B). Child B is crying. Child A has previously regularly hit other children. Several other children are watching.

What would you most likely do?

Please select your first and second choice.

Please mark two choices in total, one in each column.

		First choice	Second choice
a)	I would speak firmly to child A while the other children are listening.	□1	□1
b)	I would focus on child B and comfort him/her.	□2	□2
c)	I would tell the children who was wrong and who was right.	□3	□3
d)	**I would resolve the conflict together with child A and B.**	□4	□4

SJQ31 **Support for child-directed play, Likert scale format**

Suppose that five three-year old children are playing with different toys of their choosing.

In an ideal situation where you could choose what to do during this time, what would you do?

For each suggestion, mark the option that best describes what you would do.

		I would definitely do this	I would probably do this	I would probably not do this	I would definitely not do this
a)	**I would play with the children by following their lead.**	□1	□2	□3	□4
b)	I would let children play by themselves and only intervene when they request it.	□1	□2	□3	□4
c)	I would contribute to children's play by asking questions or providing explanations.	□1	□2	□3	□4
d)	I would encourage children to play together rather than joining in their play.	□1	□2	□3	□4
e)	I would contribute to children's play by providing new ideas or materials.	□1	□2	□3	□4

SJQ-A31 **Support for child-directed play, forced choice format**

Suppose that five three-year old children are playing with different toys of their choosing.

In an ideal situation where you could choose what to do during this time, what would you most likely do?

Please select your first and second choice.

Please mark two choices in total, one in each column.

		First choice	Second choice
a)	I would let children play by themselves without intervening.	□1	□1
b)	I would let children play by themselves and only intervene when they request it.	□2	□2
c)	I would contribute to children's play, for instance by asking questions or providing new ideas or materials.	□3	□3
d)	I would play along with the children and follow their lead.	□4	□4

Annex B. Goodness of fit indices for measurement invariance across the target populations

Table A B.1. Likert. SJQ 29

Column	AIC	AICc	SABIC	BIC	Loglikelihood	χ2	df	p
Configural	32241.63	32242.41	32343.93	32455.14	-16085.82	NaN	NaN	NaN
Metric	32250.88	32251.89	32367.79	32494.89	-16085.44	0.752	5	0.98
Metric	32374.48	32374.79	32438.78	32508.68	-16165.24	NaN	NaN	NaN
Scalar	32241.63	32242.41	32343.93	32455.14	-16085.82	158.843	13	0
Configural	32374.48	32374.79	32438.78	32508.68	-16165.24	NaN	NaN	NaN
Scalar	32250.88	32251.89	32367.79	32494.89	-16085.44	159.595	18	0

Note: Akaike Information Criterion (AIC), Akaike Information Criterion small-sample equivalent (AICc), sample size adjusted Bayesian Information Criterion (SABIC), Bayesian Information Criterion (BIC), chi-square independence test (χ2), degrees of freedom (Df), p-value (P).

Table A B.2. Forced Choice SJQ 29

Column	AIC	AICc	SABIC	BIC	Loglikelihood	χ2	df	p
Configural	46194.57	46194.87	46353.37	46477.3	-23058.28	NA	NA	NA
Metric	46204.57	46204.95	46383.72	46523.55	-23058.28	0.001	5	1
Metric	46170.01	46170.12	46267.73	46344	-23061	NA	NA	NA
Scalar	46194.57	46194.87	46353.37	46477.3	-23058.28	5.436	15	0.988
Configural	46170.01	46170.12	46267.73	46344	-23061	NA	NA	NA
Scalar	46204.57	46204.95	46383.72	46523.55	-23058.28	5.436	20	0.999

Note: Akaike Information Criterion (AIC), Akaike Information Criterion small-sample equivalent (AICc), sample size adjusted Bayesian Information Criterion (SABIC), Bayesian Information Criterion (BIC), chi-square independence test (χ2), degrees of freedom (Df), p-value (P).

Table A B.3. Forced choice SJQ 30

Column	AIC	AICc	SABIC	BIC	Loglikelihood	χ2	df	p
Configural	NA	NA	NA	NA	NA	NA	NA	NA
Metric	NA	NA	NA	NA	NA	NA	NA	NA
Metric	34357.67	34357.79	34455.39	34531.66	-17154.8	NaN	NaN	NaN
Scalar	34371.51	34371.81	34530.3	34654.24	-17146.8	16.164	15	0.371
Configural	NA	NA	NA	NA	NA	NA	NA	NA
Scalar	NA	NA	NA	NA	NA	NA	NA	NA

Note: Akaike Information Criterion (AIC), Akaike Information Criterion small-sample equivalent (AICc), sample size adjusted Bayesian Information Criterion (SABIC), Bayesian Information Criterion (BIC), chi-square independence test (χ2), degrees of freedom (Df), p-value (P).

Table A B.4. Likert SJQ 31

Column	AIC	AICc	SABIC	BIC	Loglikelihood	χ2	df	p
Configural	36389.6	36390.38	36491.68	36602.89	-18159.8	NaN	NaN	NaN
Metric	36361.93	36362.94	36478.59	36605.69	-18141	37.675	5	0
Metric	36474.03	36474.34	36538.19	36608.1	-18215	NaN	NaN	NaN
Scalar	36389.6	36390.38	36491.68	36602.89	-18159.8	110.426	13	0
Configural	36474.03	36474.34	36538.19	36608.1	-18215	NaN	NaN	NaN
Scalar	36361.93	36362.94	36478.59	36605.69	-18141	148.101	18	0

Note: Akaike Information Criterion (AIC), Akaike Information Criterion small-sample equivalent (AICc), sample size adjusted Bayesian Information Criterion (SABIC), Bayesian Information Criterion (BIC), chi-square independence test (χ2), degrees of freedom (Df), p-value (P).

Table A B.5. Forced choice SJQ 31

Column	AIC	AICc	SABIC	BIC	Loglikelihood	χ2	df	p
Configural	46194.57	46194.87	46353.37	46477.3	-23058.28	NA	NA	NA
Metric	46204.57	46204.95	46383.72	46523.55	-23058.28	0.001	5	1
Metric	46170.01	46170.12	46267.73	46344	-23061	NA	NA	NA
Scalar	46194.57	46194.87	46353.37	46477.3	-23058.28	5.436	15	0.988
Configural	46170.01	46170.12	46267.73	46344	-23061	NA	NA	NA
Scalar	46204.57	46204.95	46383.72	46523.55	-23058.28	5.436	20	0.999

Note: Akaike Information Criterion (AIC), Akaike Information Criterion small-sample equivalent (AICc), sample size adjusted Bayesian Information Criterion (SABIC), Bayesian Information Criterion (BIC), chi-square independence test (χ2), degrees of freedom (Df), p-value (P).