education sciences

MDPI

Article

Developing and Validating a Video-Based Measurement Instrument for Assessing Teachers' Professional Vision of Language-Stimulation Interactions in the ECE Classroom

Lien Dorme [1,*], Anne-Lotte Stevens [2], Wendelien Vantieghem [3], Kris Van den Branden [2] and Ruben Vanderlinde [1]

[1] Department of Educational Studies, Teacher Education & Professional Development, Ghent University, 9000 Ghent, Belgium; ruben.vanderlinde@ugent.be
[2] Centre for Language and Education, Faculty of Arts, KU Leuven, 3000 Leuven, Belgium; anne-lotte.stevens@kuleuven.be (A.-L.S.); kris.vandenbranden@kuleuven.be (K.V.d.B.)
[3] Department of Linguistics, Faculty of Arts & Philosophy, Ghent University, 9000 Ghent, Belgium; wendelien.vantieghem@ugent.be
* Correspondence: lien.dorme@ugent.be

Abstract: This study reports on the development and validation of a video-based instrument to assess early childhood education (ECE) teachers' professional vision (PV) of language-stimulation (LS) interactions. PV refers to noticing and reasoning about key classroom interactions, a skill that can be trained and distinguishes experts from novices. The instrument targets the PV of three language-stimulation (LS) strategies: language input (LI), opportunities for language production (OLP), and feedback (FB). The instrument measures noticing through comparative judgement (CJ) and reasoning through multiple-choice items. Construct validity was assessed using the AERA framework, using three samples: a sample of professionals (n = 22), a pre-service teachers' sample (n = 107), and a mixed sample with in- and pre-service teachers (n = 6). Reliability and validity were confirmed, with strong reliability scores for the CJ aggregated "master" rank orders (SRR: 0.827–0.866). Think-aloud procedures demonstrated that respondents' decisions during CJ were mainly based on LS-relevant video features. Decisions unrelated to LS require further study. Multiple-choice reasoning items were developed from professionals' open-ended feedback. Pre-service teacher reasoning scores showed no significant predictors. Using real classroom videos, this instrument provides an ecologically valid, scalable tool for assessing teachers' professional vision of LS interactions. This validated instrument offers a foundation for professional development programs aimed at addressing the theory–practice gap in early language education.

Keywords: early childhood education; teacher–child interaction; validation; comparative judgement; professional vision; noticing; reasoning; first language acquisition; videography; validity

check for updates

Academic Editor: Christopher Prilop

Received: 21 November 2024
Revised: 17 January 2025
Accepted: 21 January 2025
Published: 26 January 2025

Citation: Dorme, L., Stevens, A.-L., Vantieghem, W., Van den Branden, K., & Vanderlinde, R. (2025). Developing and Validating a Video-Based Measurement Instrument for Assessing Teachers' Professional Vision of Language-Stimulation Interactions in the ECE Classroom. *Education Sciences*, *15*(2), 155. https://doi.org/10.3390/educsci15020155

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the literature, the impact of early language development on children's later academic achievement have been well established (Hoff, 2013; Prevoo et al., 2016). Strong language skills are essential for nearly all aspects of learning, as they enable children to understand teachers' instructions, engage with peers, and express their thoughts clearly (Eadie, 2022). Moreover, the quality of teacher-child interactions in the classroom is crucial for supporting young children's language acquisition (European Commission, 2018; Pianta et al., 2014;

Sammons et al., 2015). For example, Mashburn et al. (2008) found that preschoolers who engage in high quality interactions with teachers showed greater gains in language and academic skills compared to peers who received less such interactions. Despite this understanding, classroom observations indicate that the quality and quantity of language interactions in most ECE classrooms is relatively low (Justice et al., 2008; Peleman et al., 2019; Zaslow et al., 2010). Teachers frequently rely on directive or managerial language rather than fostering rich, conversational exchanges with children. Moreover, teacher-child interactions are typically brief, often restricted to just one or two conversational turns, offering minimal opportunities for children to produce language. This theory-practice paradox underscores the need for targeted efforts to improve teachers' language-stimulation (LS) competencies, which are critical in the ECE context. To address these challenges, it is essential to equip teachers with the necessary tools to facilitate more effective LS interactions. Teacher education programs and professional development initiatives play a pivotal role in helping educators develop the skills to notice and interpret key aspects of teaching and learning that contribute to improved language stimulation (van Es & Sherin, 2002).

Central to these teaching skills is the concept of professional vision, first introduced in the context of teaching by Sherin (2001), which refers to the ability to observe and interpret significant features of classroom interactions. Professional vision is particularly important in the context of language-stimulation interactions, as teachers must be able to identify and capitalise on moments that foster language development in young children. Despite its significance, there are currently no reliable tools to assess how in-service or pre-service teachers perceive and interpret meaningful LS interactions in ECE classrooms. This study aims to bridge this gap by developing and validating a video-based instrument that empirically measures teachers' Professional Vision of Language-stimulation Interactions (ProLSIs). Given that both pre- and in-service teachers have much to gain when professionalising their teacher competencies, the instrument will be validated for use by both target groups.

2. Theoretical Framework

This theoretical framework explores the foundational concepts central to understanding how teachers can support language development in early childhood education (ECE). Specifically, it focuses on (1) the role of LS interactions in ECE, (2) the importance of professional vision in identifying and fostering such interactions, and (3) the methods that are currently used to measure these critical skills. These concepts provide the foundation for developing a video-based instrument for measuring teachers' Professional Vision of Language-stimulation Interactions (ProLSIs).

2.1. Language-Stimulation (LS) Interactions

High-quality language interactions are characterised by rich and responsive language exchanges between adults and children, a strong predictor of children's learning (Sabol et al., 2013). Researchers have identified three key language-stimulation strategies that help foster high-quality interactions: providing language input (LI), providing children with opportunities for language production (OLP), and providing language feedback (FB) (Peleman et al., 2020; Pianta et al., 2012; Verhallen & Walst, 2011). Each of these strategies contributes to creating high-quality adult–child interactions by ensuring that children are not only exposed to rich and meaningful language (LI) but also encouraged to actively participate in conversations (OLP) and receive constructive feedback that supports their language growth (FB).

The first strategy, providing language input, focuses on the quality and quantity of language exposure by providing "rich" language input to children. This "rich" language involves the use of diverse and stimulating vocabulary in full, meaningful sentences (Peleman

et al., 2020). Furthermore, teachers' language input should be understandable, for example, by making a connection between language that is spoken and visible objects or by repeating, paraphrasing, and using intonation (Algoet & Vandenbussche, 2024). This strategy also includes the selection of interesting and relevant subjects to discuss that are connected to the lifeworld and the capabilities of the child, while the provided language should be challenging by being beyond the learners' current abilities (Verhallen & Walst, 2011).

The second strategy is offering "opportunities for language production". In this strategy, the focus lies on actively engaging children in conversations with the teacher and facilitating discussions between children. It captures the extent to which teachers provide children with a language-stimulation environment, such as asking open-ended questions and encouraging children to verbalise their thoughts in multi-turn conversations (Algoet & Vandenbussche, 2024; Justice et al., 2008; Peleman et al., 2020; Pianta et al., 2014; Verhallen & Walst, 2011). Children should be given ample opportunities to speak, so that all children, including less proficient children, are given the chance to produce and experiment with language. It requires teachers to leave room for all children to speak, and to ask open-ended rather than closed questions.

In the third strategy, providing language-related "feedback", the key idea is that children receive feedback on the language they produce, to help them improve, expand, and refine their language abilities (Algoet & Vandenbussche, 2024; Peleman et al., 2020; Verhallen & Walst, 2011). In the broader research literature, this strategy is rather seen as the extent and quality of verbal evaluation provided by the teacher to children about their work, comments, and ideas (Justice et al., 2008), but in this study, the focus lies on language-related feedback. Feedback can take several forms, including the direct and indirect correction of children's errors, but also affirming correct language, helping children clarify what they mean or restating children's utterances in a more accurate or complex manner. This feedback exposes children to more advanced linguistic structures and guides them towards more precise and varied language use (Algoet & Vandenbussche, 2024; Peleman et al., 2020; Verhallen & Walst, 2011).

The three strategies are closely interconnected and support one another. For instance, effective feedback and open-ended questions from the teacher should ideally be "rich", adhering to the principles of high-quality language input (see earlier). Likewise, providing feedback can only be carried out when children are given opportunities for language production. If children are not encouraged to engage in conversation or share their thoughts, there is little opportunity for the teacher to offer feedback that enhances their learning.

2.2. Professional Vision of Language-Stimulation Interactions

The concept of professional vision was first introduced by Goodwin (1994) in the context of archaeologists and legal trials, and later applied by Sherin (2001) to teachers in the classroom context. The core idea is that teachers develop a professional vision of classroom events, enabling them to notice and interpret details that may be invisible to non-experts. As such, professional vision is recognised as a crucial component within the broader framework of teacher competencies (Vanderlinde et al., 2020). Authors van Es and Sherin (2010) define professional vision as teachers' ability to notice and reason about crucial aspects of classroom teaching. Noticing involves the teacher's skill to identify relevant and specific classroom events that are important for children's learning. Reasoning involves the teacher's skill to interpret and understand those situations in the light of their theoretical knowledge.

Pre-service teachers who have limited knowledge and experience are often not able to notice and interpret complex events in the classroom (Stürmer et al., 2013). Professional vision is therefore understood to be a distinctive feature of experts and can be learned and

developed during teacher education, throughout the professional career, or via specifically designed interventions (Gegenfurtner et al., 2020; Keppens, 2020; König et al., 2022). Training and supporting teachers to develop professional vision may have a positive impact on their language-stimulation competencies. Applied to LS competencies, professional vision involves recognising opportunities for meaningful language interactions (=noticing) and being able to interpret language interactions in the classroom, explain why they would be productive, and predict the consequences of these interactions (=reasoning). The ability to interpret and predict, however, requires not only observational skills but also a deep understanding of teaching and children's learning, i.e., teachers' theoretical knowledge (Seidel & Stürmer, 2014). This type of deep-level reasoning is a feature of expert teachers, who can integrate observations with their underlying knowledge structures (Gold & Holodynski, 2017; Seidel & Stürmer, 2014). Consequently, professional vision is hypothesised to develop over time, as experience and knowledge grow.

2.3. Measuring Professional Vision

While professional vision is challenging to measure, various instruments have been developed to assess how teachers perceive, interpret, and predict classroom situations, primarily in the context of classroom management and teaching effectiveness (Gold & Holodynski, 2017; Roose et al., 2019; Seidel & Stürmer, 2014). Traditionally, professional vision has been measured using classroom-based videos, which have the advantage of capturing the complexity and authenticity of real-life classroom situations (De Coninck et al., 2018; Gold & Holodynski, 2017; Roose et al., 2018; Seidel & Stürmer, 2014; Theelen et al., 2019; Weyers et al., 2023; Widjaja et al., 2018). Professional vision via videography can be paired with holistic or analytic measurement methods, which will be discussed in more detail below.

When professional vision is paired with analytic measurement methods, the construct of teachers' professional vision is divided into several measurable, predefined criteria. In summary, these methods break complex classroom interactions into discrete, measurable components, with each component containing multiple measurement items. Each item is rated separately and then summed to provide an overall score. Traditional tools, such as "The Observer" (Seidel & Stürmer, 2014), primarily rely on analytic assessment techniques using unique rating items per video to systematically assess teacher interactions. Other studies, including those by Gold and Holodynski (2017), Kersting (2008), and Meschede et al. (2017), similarly employ analytic methods to evaluate teachers' professional vision through video-based scoring systems. Note that in this system, noticing cannot be measured in an unbiased way because the predefined criteria guide respondents' attention to certain events in the video clip. To avoid this issue, open questions can be used to measure competencies (De Coninck et al., 2018; Gippert et al., 2022; Mathers, 2021). In this method, a coding framework is developed with predefined criteria to score respondents' answers.

Holistic scoring methods assess the overall quality of teachers' professional vision, rather than subdividing it into several components. A popular holistic approach used to measure teachers' professional vision is comparative judgement (CJ) (Keppens et al., 2019; Pollitt, 2012). CJ differs from analytical methods by evaluating objects through direct comparison rather than scoring them against a set of criteria. Comparative judgement distinguishes itself by evaluating the quality or value of videos through direct comparison with a video of another classroom interaction, rather than assigning scores to this video (Pollitt, 2012). Comparing is believed to be a more intuitive and reliable process requiring less cognitive effort than scoring against fixed criteria (Lesterhuis et al., 2015). CJ involves multiple assessors independently comparing pairs of classroom videos and deciding which one is better in the domain of interest (Pollitt, 2012). Through repeated comparisons, a

relative ranking of the objects is created based on collective judgements. This approach relies on the expertise of assessors to make nuanced decisions. When comparing a teacher's individual ranking with an aggregated expert ranking (a "master" ranking), a score for this teacher's professional vision can be obtained. This method has been applied to measure the noticing component of professional vision, for instance in the context of inclusive teaching or teacher–parent communication (Keppens et al., 2019; Roose et al., 2018; Vantieghem et al., 2020). In short, CJ provides a promising alternative to analytic scoring by using holistic comparisons to measure professional vision.

While analytic approaches are precise and reliable, they risk being reductionist because it is nearly impossible to formulate all relevant criteria in advance (Pollitt, 2012). In contrast, holistic approaches such as CJ would provide a more valid reflection of the subject as a whole (Barkaoui, 2007). In CJ, the "master" rank order is based on the decisions of several assessors and captures a shared understanding of quality (Pollitt, 2012). Furthermore, it is believed to provide greater consistency and reliability across raters due to the more intuitive nature of comparing versus scoring (Bramley, 2015; Keppens et al., 2019; Roose et al., 2018). It is therefore currently considered a promising method for measuring complex competencies (Lesterhuis et al., 2015). Despite these strengths, CJ is not without limitations. It lacks transparency in the way respondents score (e.g., do they base their judgement on criteria relevant to the construct that is being measured?) and may lead to inconsistency if raters' judgments vary widely.

3. Research Goal and Research Questions

This study aims to (1) develop a video-based instrument to measure (pre-service) teachers' ProLSIs—noticing and reasoning—and (2) investigate the validity and reliability of this instrument. The instrument combines holistic and analytical measurement methods using the method of comparative judgement for measuring teachers' noticing.

Ensuring the validity of an instrument is crucial to confirm that it measures what it intends to measure. Cook and Campbell (1979) refer to this as construct validity. Validation can be seen as a process of collecting arguments for and against the intended interpretation and use of a test score. To assess the construct validity of the scores resulting from the measurement instrument, the validity framework of the American Educational Research Association (2011) is used, which presents five types of evidence that can be collected to evaluate the intended test interpretation of the instrument. This framework was selected for its comprehensive approach to assessing validity, as it incorporates multiple sources of evidence to ensure a rigorous and holistic evaluation of the instrument. Its relevance is further demonstrated by prior applications in a related study that used the same framework, referred to as the Chan framework (Keppens et al., 2019). This study will investigate four sources of evidence to evaluate validity: (1) the content of the instrument, (2) response processes, (3) internal structure, and (4) relations to other variables. Table 1 summarises the decisions made within the AERA framework, providing an overview of the evidence sources and the measures evaluated for each type of validity. The rationale for the selected measures and methods is outlined below. To address these validation sources, evidence is collected through four research questions.

> RQ1: Evidence based on test content—Which videos and multiple-choice items are best suited to elicit respondents' noticing and reasoning of language-stimulation interactions (LSIs)?
>
> a. Which videos are suited to elicit teachers' noticing and reasoning regarding LI, OLP, and FB?
> b. Which video-based multiple-choice items accurately represent teachers' reasoning regarding LI, OLP, and FB?

Table 1. Overview of four sources of construct validity that were evaluated in this study and the measures collected to assess each source of evidence (American Educational Research Association, 2011). The fifth source of evidence (evidence based on consequences of testing) was not evaluated.

Source of Evidence Based On	Noticing	Reasoning
Test content (RQ1)	Sample 1 (professionals [1]) - Judge infit and representation infit - SSR calculation of aggregated "master" rank order	Sample 1 (professionals [1]) Open-ended feedback on reasoning aspects during the noticing comparison task.
Internal structure (RQ2)	Sample 1 (professionals [1]) - Correlations between LI, OLP, and FB	Not evaluated
Response processes (RQ3)	Sample 2 (mixed in- and pre-service teachers) - Think-aloud procedures (concurrent probing)	Sample 2 (mixed in- and pre-service teachers) - Perceived cognitive load
Relationships with other variables (RQ4)	Not evaluated [2]	Sample 3 (pre-service) Multiple regression analysis to determine the effect of background variables on pre-service teachers' reasoning scores

[1] Professionals were education professionals who have demonstrated experience in either language education or ECE or a combination of both. [2] The relationship between noticing skills and other variables was not evaluated because no background variables were collected with this measurement in samples 1 and 2.

In developing the video-based measurement instrument, classroom videos were carefully selected to represent clear, realistic, and recognisable examples of LI, OLP, and FB, so that videos can elicit teachers' noticing and so that multiple-choice reasoning items can be developed based on the unique elements of the classroom scenarios depicted in the videos. In this research question, (a) the content of the video-based instrument, i.e., the selected videos, is validated and (b) multiple-choice reasoning items are constructed, with care for content validity.

Professionals in the relevant field often play a key role in assessing the representativeness of test content (American Educational Research Association, 2011). In the method of comparative judgement, as mentioned earlier, we can obtain an individual score for teachers' noticing when an individual rank order is compared against a "master" rank order, which is an aggregated rank order based on professionals' evaluation. Professionals classify video clips according to relevant facets of LI, OLP, and FB, ensuring that each video clip is relevant to the construct to be measured, in this case, teachers' Professional Vision of Language-stimulation Interactions (ProLSIs). In this research question, we assess the alignment between the instrument's content and the construct it is supposed to measure, in this case (pre-service) teachers' ProLSIs, via the three language-stimulation strategies of LI, OLP, and FB.

> RQ2: Evidence based on internal structure—How do the three language-stimulation interaction strategies LI, OLP, and FB correlate with each other in the video instrument?

With this research question, we are interested in exploring how each of the LSI strategies, LI, OLP, and FB, are correlated with each other. Each strategy should ideally be internally consistent, yet distinct from the others to warrant separate assessments. From theory, we expect the correlations between the LSI strategies to be medium-sized, indicating that the strategies are related but still distinct enough from the others. Medium-size positive correlations between the strategies would indicate that the strategies are part of a common construct (LSIs) but that they are still three distinct dimensions of that construct. High correlations would indicate that the strategies are overlapping.

RQ3: Evidence based on response processes—How do respondents experience the use of the instrument in terms of interpretation and cognitive load?

a. What decision processes underlie the comparative judgement task to assess noticing?
b. How do respondents experience the use of the instrument in terms of cognitive demand?

The instrument aims to measure noticing and reasoning in the context of LSIs. It is thus crucial to determine whether respondents are genuinely engaging in reasoning. Specifically, we aim to ensure that respondents' judgements are grounded in one of the three targeted strategies of language-stimulation interactions: language input (LI), opportunities for language production (OLP), and feedback (FB). Respondents' judgements should not be influenced by unrelated factors, such as for example teacher style. Additionally, we will examine if there are differences in how respondents interpret the CJ task. We will also investigate the cognitive processes involved as respondents complete the instrument to evaluate cognitive load. Cognitive load refers to the mental demand placed on the cognitive system when answering questions (American Educational Research Association, 2011). Understanding this aspect is crucial, as excessive mental effort may result in decreased attention and lower accuracy in respondents' answers (Paas et al., 2003).

RQ4: Evidence based on relationship with other variables—How is pre-service teachers' reasoning related to respondents' background variables?

With this research question, we examine the relationship between respondents' reasoning skills and background variables such as, for example, age and educational background. Previous research has shown a positive relation between pre-service teachers' training, diploma, and their noticing ability (Keppens et al., 2019; Stürmer et al., 2013). By analysing the relationship between pre-service teachers' reasoning and relevant background variables, we could confirm whether test results align with our expectations and support the validity of the instrument by ensuring it behaves as predicted in relation to these factors (American Educational Research Association, 2011).

RQ5: Reliability—Can we establish a reliable "master" rank order of videos based on the judgements of a sample of professionals?

Besides assessing validity, this study will also examine the reliability of the video instrument. A reliable measurement of the respondents' noticing scores hinges on a reliable aggregated rank order since individual noticing scores are calculated based on the difference between their individual rank order with the master rank order. A reliable instrument ensures that it consistently produces stable and repeatable results, across time, across different conditions, and across different users. In analytic approaches, reliability is considered a measure of consistency over time and over similar samples (Cohen et al., 2009). In CJ, reliability refers to the accuracy with which scores for individual videos are calculated. This accuracy improves as more respondents provide judgements on the items being compared. So in CJ, an increase in reliability means that the scores of the individual videos can be calculated more accurately (Pollitt, 2012).

4. Methods

4.1. Instrument Description

The video-based instrument consists of two components: the noticing component and the reasoning component. Videography is central in both. In both components, participants have the option to rewatch the video clips as needed.

Noticing: In the noticing component, the respondents fill out a comparative judgement task. The respondents are shown a random pair of video clips and are asked to indicate in which video they think the ECE teacher best put a certain language interaction strategy to practice (LI, OLP, or FB). Brief definitions of each strategy are provided so respondents can understand the task at hand. The full CJ task comprises the comparison of ten video pairs. Appendix A provides a screenshot and translated transcript of a part of the CJ task.

Reasoning: Following the CJ task, the respondents complete a video-based exercise on three video clips from the CJ task, showcasing an authentic interaction between a teacher and child(ren) that clearly illustrates one of the three language interaction strategies, LI, OLP, or FB. Each video is preceded by a brief description of the class context depicted. Following the video, the respondents answer 15 multiple-choice items related to the specific language interaction strategy shown. The first 10 multiple-choice items present statements of advice that a teacher could use to improve their skills in the specified language interaction strategy (LI, OLP, or FB). The participants are asked to evaluate the advice based on the video they just watched, categorising it as (1) good advice, if the teacher could effectively use this advice to support the specified strategy; (2) redundant advice, if the teacher was already implementing it in the video clip; or (3) not good advice, if it would not contribute to the strategy for language-stimulation interaction in the given video. The second set of items included five dichotomous statements, to which respondents indicated true or false. These video-based items aimed to elicit explicit reasoning from respondents regarding the language stimulation observed in the video clip. Translated transcripts of the selected videos and their corresponding multiple-choice items can be found in Appendix B.

4.2. Instrument Development and Validation

Our research consisted of four phases, which are visualised in Figure 1. Phases 0–1 depict the development phase of the instrument, and phases 1–3 represent the validation of the instrument. In the figure, an overview can be found of the research questions per research phase (RQs) and the different samples that were used throughout the research. For an overview of the measures analysed per research question, we refer to Table 1. An overview of the samples used can be found in Table 2.

Table 2. A brief overview of the samples used.

	Sample 1	**Sample 2**	**Sample 3**
Type (sample size)	Professionals (n = 22)	Mixed in- and pre-service teachers (n = 6)	Pre-service teachers (n = 107)
Detailed sample description	Education experts with demonstrated experience in either language education or ECE or both (a diverse group of academics and practitioners)	ECE in-service and pre-service teachers	Pre-service teachers, i.e., students in ECE teacher education (students were from different years of teacher education).

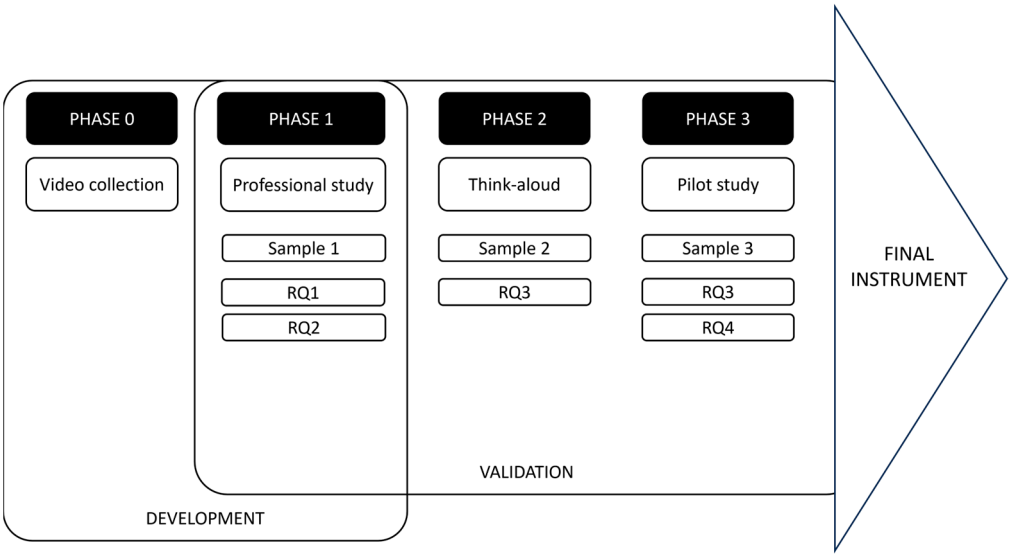


Figure 1. Overview of research phases, samples, and research questions (RQs).

4.2.1. Phase 0: Video Collection and Selection

In the preparatory phase, video material was collected from day-to-day activities in the classroom by 14 ECE teachers, focusing on the interactions of the teacher. Video materials were edited into short video clips, resulting in 134 video clips. Three independent coders coded these clips based on (a) the presence of each of the three language interaction strategies and (b) the extent to which the video was a good example of the interaction strategy. Three coding rounds were needed to obtain adequate inter-rater reliability (Kappa TF 0.638; Kappa TR 0.608; Kappa TA 0.608). During the coding process, 69 video clips were excluded. Fourty-six videos (67%) were excluded because one or more language interaction strategies were not clearly present, fourteen (20%) were excluded due to insufficient audio or video quality, five (7%) were excluded because the clip showed a reading activity in which all language input of the teacher originated from the book, two (3%) were excluded because more than one teacher was speaking, and finally two remaining clips were removed because the video clip had an unsuitable duration (shorter than 35 s or longer than 1 min 35 s). From the remaining 65 video clips, a final selection was made to ensure maximum diversity in terms of teacher characteristics, activity types (e.g., circle time, one-on-one interaction, free play, or small group activity), and the quality of the teacher's language interactions. Video clips that could distract from the language interaction strategy comparison—such as those involving sensitive topics like the death of a relative—were excluded. In the end, 21 video clips were chosen for inclusion in the video-based instrument.

4.2.2. Phase 1: Professionals Study

This study was conducted to address research questions RQ1 and RQ2, focusing on the validation of the test content and the validation of the internal structure of the noticing component of the video-based instrument. This study utilised the digital platform Comproved (*Comproved—Assess better. Learn more*, n.d.), a platform specialised in assessment via comparison judgement.

Participants: A convenience sample of 26 professionals were recruited, all with demonstrated expertise and experience in either education for preschoolers, language education, or both. The professionals' sample comprised a diverse group of academics and practitioners, including ECE teachers and pedagogical counsellors (n = 6), teacher educators (n = 6), trainers on language in education (n = 5), researchers with demonstrated experience in ECE, language education or language in ECE education (n = 32), and policy officers in language

education (n = 1). Invitations were sent via email, containing a personal online link to the Comproved platform. Participation was voluntary; a total of 23 professionals responded and participated in this study, 22 participants completed the full instrument.

Procedure: For validating the content of the video-based instrument (RQ1), two measures were evaluated: (a) the Scale Separation Reliability (SSR) for each language interaction strategy, which is a measure for comparative judgement, and (b) open answers on pros and cons of each video clip with regard to each language-stimulation strategy. For validating the internal structure of the professional vision model (RQ2), two additional measures were evaluated: (c) Infit statistics from the comparative judgement task for evaluating the consistency and reliability of the judges and video clips (infit statistics are explained in more detail below) and (d) correlations between the language interaction strategies noticing scores, LI, OLP, and FB. The first three measures (a–c) were obtained via the Comproved platform v3.0.1 and correlations were calculated using Microsoft Excel (Microsoft 365 MSO Version 2411 Build 16.0.18227.20082).

Measures for comparative judgement: Professionals were invited to complete a comparison task on the Comproved platform. Professionals were shown a random pair of video clips and asked to indicate in which video they think the ECE teacher best put a certain language interaction strategy (LI, OLP, or FB) to practice. Each professional compared fourteen pairs of videos for each of the three language interaction strategies.

Open answers: After each comparison, professionals were invited to document the specific elements that influenced their choice of one video over the other regarding the language interaction strategy in question. This was prompted immediately after each decision to capture detailed feedback. For participants, providing these written comments was optional to avoid overburdening participants and to minimise cognitive load. We received written feedback on at least one of the strategies on 57% of the videos.

Data analysis: Comparative judgement (CJ) is a method for assessing the quality of objects by comparing them in pairs. Each video clip has an underlying quality, and the method predicts the likelihood that one video clip will "win" or "lose" over another based on this perceived quality. CJ data are modelled according to the Bradley–Terry–Luce (BTL) model, which is comparable to the Rasch model (Bradley & Terry, 1952). Through an iterative process, the quality estimates are refined and updated until they stabilise, ultimately producing a reliable ranking of the video clips. In simple terms, video clips are ranked from "worst" to "best" based on how frequently they are judged as superior in pairwise comparisons. This ranking is then represented on a logit scale, quantifying the relative differences in quality between the video clips, with higher scores indicating better-perceived quality (Pollitt, 2012). A rank order was produced for each of the strategies, LI, OLP, and FB.

To evaluate the research questions on content validity (RQ1) and reliability (RQ5), a quality measure was calculated to assess the quality of the video clips (Lesterhuis et al., 2015; Pollitt, 2012). This quality measure is the Rasch separation reliability, which is expressed as the Scale Separation Reliability (SSR). This index can best be compared with Cronbach's alpha in classical test theory and provides information on the extent to which the spread in scores is due to actual differences between items rather than measurement error (Bramley, 2015). A high SSR thus indicates a stable and reliable ranking of the video clips and implies that the relative position of the video clips on the ranking is fixed and that the video clips can be effectively differentiated in terms of quality. Furthermore, the consistency of the judges and the video clips can be evaluated by calculating infit statistics. Judge infit was measured to determine the consistency of each judge's decisions. If a judge's infit statistic deviated by more than two standard deviations from the mean—either above or below—it indicated significant inconsistency or judge misfit. Such judges were removed

from the data set, as reliable assessments depend on judges maintaining consistency with one another (Pollitt, 2012). Judges with an infit value exceeding this criterion indicate that they interpreted the video clips differently from the others. Next, a similar consistency measure can be calculated for the video clips under analysis, the representation infit, which evaluates how uniformly judges rate the clips. For the video clips, a representation misfit means that judges find it difficult to consistently judge the video: This might be due to a distracting feature in the video, which means that the video can be interpreted in different ways concerning LI, OLP, or FB. Misfitting clips were excluded, ensuring that the final analysis was based on reliably judged items.

To address the second research question (RQ2) concerning the internal structure of the data, correlations were calculated between the three strategies—the language interaction strategies LI, OLP, and FB—to assess whether these were sufficiently distinct from one another to warrant separate assessments.

Finally, evidence for test content (RQ1) for the reasoning component of the instrument was collected by extracting open answers from the professionals on the platform. These open answers were grouped per language interaction strategy and used to develop the video-based analytical items for the reasoning component of the instrument.

4.2.3. Phase 2: Think-Aloud Study

This study addresses research question RQ3 and focuses on respondents' response processes during the completion of the instrument.

Participants: A mixed convenience sample of 6 participants was recruited to perform think-aloud procedures. The sample consisted of pre-service teachers (n = 4) and in-service teachers (n = 2). Three pre-service teachers were in their last bachelor year of teacher education and one pre-service teacher was in their second bachelor year. Both in-service teachers had over 25 years of experience: one in a rural area and the other in a big city in Flanders.

Procedure: Think-aloud interviews were carried out to investigate the response processes during the noticing task. All think-aloud interviews followed the same procedure. The respondents were asked to fill out the instrument and encouraged to make their thought processes explicit while completing the instrument, according to the method of concurrent probing (Young, 2005). To facilitate this process, stimuli such as "What are you thinking about as you make this decision?" and "Can you explain why you chose this option?" were posed during the task to elicit more detailed reflections on their decisions. Think-aloud interviews were carried out until data saturation was reached, ensuring that no new insights were emerging. The think-aloud interviews were audio-recorded. The average duration time to complete the instrument was measured as an estimate of the cognitive load. This is based on the premise that time on task reflects the cognitive capacity required to process and complete the task (Paas et al., 2003).

Data analysis: Think-aloud sessions were recorded, transcribed, and coded in Microsoft Word (Microsoft 365 MSO Version 2411 Build 16.0.18227.20082) to identify key themes. The analysis focused on whether strategies of LI, OLP, and FB played a role in their decisions, or if other factors influenced their choices.

4.2.4. Phase 3: Pilot Study (Of the Reasoning Component of the Instrument)

A pilot study was conducted to answer research questions RQ3, and RQ4. This study focuses on validating the reasoning component of the instrument, aiming to explore its response processes (RQ3) and relationships with other variables (RQ4).

Participants: Second- and third-year bachelor students from two Flemish university colleges of teacher education were invited to participate in the pilot study via an invitation

link hosting a background survey and the reasoning component of the instrument. The participants were informed about this study via their teachers. In total, 238 students were invited to participate, with participation being voluntary. Ultimately, 107 pre-service teachers completed the instrument—denoting a response rate of 45%—of whom 105 were female and two were male. The respondents filled out the instrument independently and remotely.

The respondents' ages ranged from 19 to 51, with an average age of 23 years. Most respondents (n = 71) were in their second year of teacher education, while the remaining 36 were in their third and final year. A total of 29 respondents had gained teaching experience through short internships (less than two weeks), whereas 78 had experience through a longer internship (more than two weeks of active teaching experience in ECE settings).

Procedure: Pre-service teachers were invited to complete a video-based exercise on the digital Qualtrics platform (Qualtrics software, Version 04/2024 of Qualtrics). The participants viewed one of the video clips from research phase 0, showcasing an authentic interaction between a teacher and child(ren) that clearly illustrated one of three language interaction strategies, LI, OLP or FB. Following the video, the respondents answered 15 multiple-choice items related to the specific language interaction strategy shown. Using a criterion-referenced approach, correct answers for all video-based items were determined by a panel of professionals (n = 7). After independently completing the reasoning task, the most selected response was marked as correct. Items with less than 80% agreement were reviewed, and if no consensus was reached, the item was excluded. The reasoning score for each participant is calculated as a sum score of their correct answers. For validating the reasoning component of the instrument, two measures were evaluated: (a) a self-perceived cognitive load estimate ranging from 1 (=not difficult at all) to 9 (=very difficult) and (b) relationships between participants' background variables (like age and educational background) and their reasoning scores, which were assessed using regression coefficients.

Data analysis: All analyses in this research phase were performed on the pilot study cross-sectional data. To answer RQ3 on respondents' response processes, descriptive analyses were performed on the cognitive load measures to capture the mental effort of respondents while filling out the instrument. To answer research question RQ4 on the relationship between respondents' ability to reason and other variables, multiple linear regression analyses were carried out on the impact of respondents' background variables. All data analyses were carried out in IBM SPSS Statistics (Version 29.0.2.0).

5. Results

5.1. RQ1: Evidence Based on Test Content

Noticing: To answer the question of which videos are best suited to elicit noticing of LI, OLP, and FB, the judge infit and representation infit were calculated after the pairwise comparisons in phase 1 of the research study. With the infit measure, we can assess whether any professionals (judge infit) or video clips (representation infit) deviated from the consensus. Infit scores of the judges are presented in Table 3 and representation infit scores can be found in Table 4. A stepwise procedure was applied to remove misfitting judges and video clips with infit values exceeding the criterion of 2. As a result, two misfitting video clips and four misfitting judges were excluded. Importantly, judges were removed only from the specific strategy where misfit was observed. For example, if a judge exhibited inconsistent ratings for videos in the LI strategy but showed no misfit in OLP or FB, only the LI-related comparisons for that judge were removed (e.g., judge 13). If a video misfitted for one of the strategies, the misfitting video clip was removed from the entire instrument, in order to maintain the possibility to judge all videos for all three strategies.

The stepwise procedure resulted in three rank orders, consisting of 19 video's each, based on the comparisons of 20 judges for LI and 21 judges for OLP and FB.

Table 3. Judge infit measures for LI, OLP, and FB after the stepwise deletion process was applied to the misfitting judges and video clips.

Judge	LI	OLP	FB
1	−0.99	−1.47	−0.24
2	−0.89	−0.65	−0.60
3	0.10	0.98	1.21
4	−0.68	−1.22	0.94
5	1.21	−0.38	0.07
6	−0.21	0.89	−0.22
7	0.62	1.20	1.05
8	0.96	0.27	0.80
t9	0.42	0.80	0.28
10	0.47	1.07	−0.12
11	−1.07	1.20	0.76
12	−0.72	0.35	−1.46
13	-	0.91	−1.09
14	−0.84	0.55	−1.82
15	0.50	0.36	0.44
16	0.45	0.30	1.01
17	−0.16	0.90	0.90
18	0.05	−0.90	0.54
19	1.06	1.09	−1.93
20	0.76	−0.12	0.70

Table 4. Representation infit measures for LI, OLP, and FB after the stepwise deletion process was applied to the misfitting judges and video clips.

Video Clip	LI	OLP	FB
1	−0.11	−0.46	−0.32
2	0.04	0.11	0.62
3	−1.28	1.27	0.24
4	0.00	0.50	1.06
5	0.79	0.65	0.69
6	1.89	−0.73	−0.54
7	1.07	0.45	−0.02
8	−0.10	1.75	0.66
9	−0.49	0.38	−0.77
10	0.27	0.24	−0.66
11	0.84	0.32	0.61
12	1.89	−0.02	−0.67
13	0.01	0.78	−0.55
14	−0.46	0.33	0.20
15	0.24	−0.41	−0.18
16	−1.04	1.71	1.18
17	0.41	0.19	−1.43
18	0.06	0.33	0.93
19	−1.65	1.11	0.94

In Table 5, the rank orders of the video clips can be found as a result of the pairwise comparisons in phase 1 of this study. The ability scores in the Bradley–Terry–Luce model are expressed in logit scores and express the relative strength or ability of the video clips, where higher scores indicate a higher likelihood of success in the pairwise comparisons. Negative scores indicate a lower success rate in the pairwise comparison. For example, video 11 is judged among the professionals as the worst video for LI. For OLP, it scores third last, and for FB it is the penultimate video in the ranking.

Table 5. Rank order, true score, and standard error of the video clips for LI, OLP, and FB.

Ranking LI			Ranking OLP			Ranking FB		
Repres.	**True Score**	**SE**	**Repres.**	**True Score**	**SE**	**Repres.**	**True Score**	**SE**
Video 2	1.87	0.59	Video 9	3.64	0.75	Video 4	3.05	0.80
Video 1	1.76	0.60	Video 13	3.59	1.08	Video 12	3.01	0.81
Video 10	1.62	0.51	Video 12	3.00	0.73	Video 18	1.44	0.56
Video 19	1.08	0.51	Video 19	2.43	0.67	Video 13	1.02	0.54
Video 5	0.78	0.56	Video 4	1.26	0.57	Video 1	0.76	0.52
Video 18	0.75	0.51	Video 18	0.91	0.62	Video 8	0.74	0.51
Video 16	0.72	0.54	Video 8	0.59	0.55	Video 17	0.64	0.53
Video 6	0.54	0.49	Video 17	0.40	0.54	Video 2	0.34	0.47
Video 12	0.32	0.50	Video 6	0.16	0.47	Video 14	0.11	0.46
Video 14	0.24	0.47	Video 14	−0.39	0.55	Video 5	−0.04	0.52
Video 8	0.01	0.50	Video 10	−0.45	0.47	Video 19	−0.05	0.46
Video 7	−0.07	0.46	Video 1	−0.57	0.55	Video 6	−0.15	0.45
Video 13	−0.09	0.52	Video 2	−0.79	0.52	Video 16	−0.55	0.49
Video 17	−0.35	0.49	Video 16	−1.00	0.52	Video 9	−0.58	0.53
Video 3	−0.88	0.52	Video 3	−1.00	0.55	Video 10	−1.05	0.49
Video 4	−1.06	0.52	Video 5	−1.77	0.70	Video 15	−1.41	0.50
Video 9	−2.18	0.82	Video 11	−2.21	0.59	Video 3	−1.66	0.56
Video 15	−2.76	0.69	Video 15	−2.30	0.63	Video 11	−2.22	0.64
Video 11	−2.94	0.82	Video 7	−5.40	2.21	Video 7	−2.47	0.76

Next, the Scale Separation Reliability (SSR) was calculated to assess the reliability of the professionals' judgements of the videos concerning LI, OLP, and FB. The SRR is comparable to Cronbach's alpha and was very good for TA with 0.827. For TR, the reliability was even higher (SSR = 0.866). For FB, the SSR was 0.849. From these results, we can conclude that the professionals' reliability of judgements of LI, OLP, and FB was very good, indicating that the professionals judged the videos consistently for all three strategies.

Reasoning: From each rank order, one video was chosen for the reasoning task. The chosen videos were clear examples of the concerned language interaction strategy and were chosen through internal consensus discussions among two of the coders from phase 0 and a third, independent professional. For LI, video 16 was selected. For OLP and FB, respectively, video 13 and video 12 were selected. Video 16 shows a small group activity led by the teacher, where the teacher is hollowing out a pumpkin while the children sort its contents into the correct columns on a worksheet. Video 13 features a teacher explaining the next activity on emotions during circle time while interacting with a few children in the group. Video 12 shows a teacher engaging with two children during a creative small-group activity: one child uses an incorrect word to refer to scissors, and another child approaches the teacher to present and explain their artwork in an unclear way. Video-specific items were developed based on the written comments from professionals that documented the specific elements that influenced their choice of one video over the other in the comparison task. This resulted in 15 items per strategy. Translated transcripts of the selected videos and their corresponding multiple-choice items can be found in Appendix B.

5.2. RQ2: Evidence Based on Internal Structure

A Pearson correlation analysis was conducted to examine the relationships among the strategies LI, OLP, and FB in the 19 videos in the instrument based on the judgements of professionals in research phase 1. The correlations between LI and OLP (r = 0.02) and between LI and FB (r = 0.30) were not significant. A strong, positive, and statistically significant correlation was found between OLP and FB (r = 0.69; p = 0.001), indicating a substantial relationship between these two strategies. These correlations were calculated

based on final rank order true scores from a comparative judgement task completed by professionals on each of the three strategies (LI, OLP, and FB).

5.3. RQ3: Evidence Based on Response Processes

To investigate the decision-making processes involved in the comparative judgement task and open the underlying "black box", four think-aloud interviews were carried out with both in-service and pre-service teachers (sample 2 in Table 2).

In the task, the respondents chose the best video clip from two randomly presented video clips, in terms of LI, OLP, and FB. To do so, all respondents used the provided descriptions of each strategy and occasionally referred back to them during the task. For LI, the respondents focused on the materials teachers used to introduce new vocabulary (e.g., a real pumpkin) and on the complexity of vocabulary used by the teacher. The respondents tended to favour videos with complex vocabulary (e.g., using "pulp" instead of simpler terms), rather than those with lengthy but simpler phrasing. When judging the OLP strategy, teachers asking broad, open questions were rewarded in the comparative judgement task. The respondents also took the perceived language skills of the children into account when making their decisions. FB was the most challenging strategy for respondents to identify, as was mentioned explicitly by all of the respondents. One respondent mentioned that she experienced the strategies LI, OLP, and FB as very similar, with there often being a "thin line" between what falls under one or the other strategy.

Challenges in comparing videos were noted, particularly when the type of activity, children's age, or language abilities varied between clips. For instance, one respondent consistently chose well-prepared activities over spontaneous ones and language-oriented activities (e.g., "classical show-and-tell") over others (e.g., weighing objects), regardless of the quality of language interaction in these activities. When comparing videos with similar LSI quality or two clips from the same teacher, the respondents reported difficulty in distinguishing subtle language-stimulation nuances. In such cases, the respondents sometimes relied on non-LSI-related criteria (e.g., child engagement, teaching style) or, in rare cases, made random choices and justified them as such.

To determine how respondents experience the use of the instrument in terms of cognitive demand, the cognitive load was assessed. For the noticing component, this was done via the think-aloud procedures. The respondents reported that while the task was lengthy, it was manageable, and they appreciated video subtitles, which reduced the effort needed to understand spoken content. The average duration time of the respondents in the think-aloud procedures was 58 min. One respondent highlighted the advantage of comparative judgement, noting it was easier to assess LSI quality by comparing clips rather than evaluating them in isolation. The respondents perceived the videos as authentic. For the reasoning component of the instrument, among the pre-service sample (i.e., sample 3; n = 122), perceived cognitive load scores were medium-high, with an average and median score of 5 out of 9 (SD = 1.87). All scores on the spectrum were found, from 1 (min) to 9 (max). The median duration to fill out the reasoning component was 17 min.

5.4. RQ4: Evidence Based on Relationship with Other Variables

To assess the relationship between pre-service teachers' reasoning and their demographic variables, a multivariate regression analysis was conducted with reasoning scores as the dependent variables (reasoning scores of LI, OLP, and FB) and background variables as independent variables (type of secondary school diploma, year of bachelor, ECE teaching experience, university college affiliation, and age). Due to an insufficient sample size of male participants (n = 2), gender was not included as a variable.

The multivariate analysis revealed no significant effect of the background variables on reasoning scores across all iterations of the model. Step-wise exclusion of non-significant factors, including ECE teaching experience, year of bachelor's degree, and university college affiliation, also showed no significant effects (Wilks' Lambda values ranged from 0.831 to 0.986, all $p > 0.05$). In the final reduced model, which included only the type of secondary school diploma, there was still no significant relationship between diploma type and reasoning scores (Wilks' Lambda = 0.856, $p = 0.197$). Similarly, a regression analysis using the total reasoning score (sum of LI, OLP, and FB scores) as the dependent variable showed no significant predictive value for any background factor, with all predictors yielding $p > 0.05$. Collinearity diagnostics confirmed that multicollinearity was not a concern (Tolerance > 0.8 for all variables).

Additional independent samples t-tests for cognitive load and reasoning scores also indicated no significant differences related to demographic variables. These findings suggest that the demographic characteristics we included (age, educational level, type of diploma, teaching experience, and university affiliation) are not significantly related to pre-service teachers' reasoning within the context of this instrument.

6. Discussion

6.1. General Discussion

Research in early childhood education (ECE) is crucial yet underrepresented due to the unique challenges of effectively measuring outcomes in young children (Phillips et al., 2017). Despite these challenges, early language development is a promising area of focus, as enhancing language interactions at this stage can yield substantial long-term benefits (Shonkoff & Phillips, 2000; World Health Organization, 2023). This study addresses the theory–practice gap in ECE (Peleman et al., 2019, 2020) by developing a video-based measurement instrument aimed at assessing pre-service and in-service teachers' Professional Vision of Language-stimulation Interactions (ProLSIs). This is a competency critical for supporting children's language development growth. This instrument represents an important step toward strengthening teachers' professional vision, defined as the ability to notice and reason about significant aspects of classroom interactions (Sherin & van Es, 2009).

This study sought to validate this measurement instrument across four sources of validity evidence as outlined by the American Educational Research Association (2011), with a focus on three language-stimulation strategies: language input (LI), opportunities for language production (OLP), and feedback (FB). Following insights from research on professional vision, we developed separate measures for teachers' noticing and reasoning abilities by employing a hybrid approach that combines holistic and analytic methods. Video-based CJ was used to assess the teacher's noticing, while reasoning was assessed using an analytic method with multiple-choice items. Although CJ here involves three subdimensions—LI, OLP, and FB—no predefined criterion lists were provided for each strategy. Instead, assessors interpreted each strategy as a whole, preserving the intrinsic holistic nature of CJ while offering some structure to guide judgements. Reasoning items were tied to specific videos to reflect the inherently context-dependent nature of interpreting classroom scenarios (van Es & Sherin, 2010). This approach leveraged the strengths of a holistic approach for assessing noticing with an analytic approach for assessing reasoning, making use of videography for both components of the measurement instrument. The authenticity of the video clips used in this study, garnered from video documentation of real-life and non-scripted interactions in the classroom, was confirmed by respondents and adds to the ecological validity of the instrument. However, we keep in mind that videos will always remain an interpretative lens rather than a complete classroom representation

since they inherently direct attention to certain aspects over others (Sherin, 2001). This instrument was validated in a mixed sample with in-service and pre-service teachers.

Overall, our results confirmed the reliability and validity of the video-based instrument. Content validity was ensured by collecting and analysing professionals' opinions in the CJ task. This resulted in a reliable "master" rank order for each language interaction strategy (LI, OLP, and FB) with high Scale Separation Reliability, safeguarding the reliability and content validity of the noticing component of each language strategy within the instrument. Individual respondents' scores can be compared against this rank order to calculate their infit with this rank order. Content validity for the reasoning component was ensured via open-ended feedback from professionals.

Think-aloud procedures in a mixed sample further strengthened the validity of response processes for the noticing component by "opening the black box" of CJ. Prior research (Keppens et al., 2019; Young, 2005) has questioned whether the CJ method truly reflects teachers' professional vision. Through think-aloud interviews, we confirmed that the respondents primarily based their judgments on the intended LSI characteristics, supporting the instrument's validity in capturing the constructs it aims to measure. However, overall, the respondents sometimes relied on non-LSI-related criteria (e.g., child engagement, teaching style), particularly when comparing videos with similar LSI quality. Additionally, we observed judge-specific foci unrelated to ProLSIs, which may contribute to judge misfit. This behaviour could also indicate poor noticing abilities of the language-stimulation strategies, rather than instrument limitations. These findings highlight the need for further research to determine whether judge misfit reflects inherent instrument challenges—such as the situations in the videos being overly dissimilar to compare or insufficient differentiation in LSI quality of the teachers in the videos—or whether it reveals respondents' limited noticing skills. Pairing think-aloud protocols with respondents' noticing infit scores on a larger scale could provide more clarity on this issue. Although the current think-aloud procedures were limited by a small sample size and the recording context, they offer valuable qualitative evidence for the validity of the instrument's response processes.

In validating the internal structure of the ProLSI instrument, our findings indicate a strong positive relationship between the OLP and FB strategies in the video clips. Video clips that "win" for the strategy OLP, are more likely to "win" for the strategy FB. This aligns with theoretical expectations, as opportunities for language production naturally create more occasions for feedback. However, no significant correlations were found between the other LSI strategies. From our theoretical framework, we anticipated moderate correlations among all strategies, reflecting their interdependence in effective language-stimulation interactions. It is important to emphasise, however, that these results do not reflect actual classroom interactions, as the analyses were limited to how the subdimensions OLP, FB, and LI are related in the video clips in the instrument, as found in the master rank order. These clips were deliberately selected to highlight all three strategies, to optimally measure ProLSI. This selection process likely influenced the observed relationships, making them distinct from how these three strategies would interact in more spontaneous classroom situations.

Interestingly, contrary to our initial hypothesis, no significant predictors were identified for pre-service teachers' reasoning scores. This finding contrasts with studies in related fields, such as Keppens et al.'s (2019) work on inclusive teaching, where factors like teacher training and diploma type were predictive of reasoning abilities in a pre-service sample. The absence of such relationships in our study may indicate that reasoning in the context of language-stimulation interactions (LSIs) in early childhood education (ECE) operates differently, with demographic factors playing a lesser role. Alternatively, this lack of significant predictors could suggest that the instrument does not fully capture aspects

of reasoning that are sensitive to demographic influences. Another possibility is that the relatively homogeneous nature of our pre-service sample limits the variability needed to detect such relationships. Future research with larger and more diverse samples could help clarify whether demographic variables genuinely hold limited predictive value for LSI reasoning or whether this finding is specific to our study design and sample.

6.2. Limitations and Suggestions for Further Research

This study provides significant insights into the development and validation of the ProLSI instrument. However, several limitations should be noted and future research opportunities should be explored to strengthen its utility and generalisability.

Limitations specific to the noticing component: The medium-high perceived cognitive load and duration of the instrument might indicate a potential risk of excessive mental effort, with the risk of decreased attention and lower accuracy in respondents' answers (Paas et al., 2003). This challenge arises primarily due to the repetitive nature of tasks for distinguishing between the three strategies, thereby increasing cognitive load. Future research could address this issue through adaptive comparative judgement (CJ) methods (Pollitt, 2012), which minimise the cognitive load by reducing the number of necessary comparisons. For instance, informative comparisons (e.g., between videos that have won the same number of times) could replace random pairing, or respondents could review the same videos multiple times to expedite the task. Investigating the feasibility and effectiveness of such adaptive methods would enhance the instrument's practicality (Bramley, 2015; Pollitt, 2012). Additionally, while this research provides important initial insights into decision-making processes in CJ tasks, further exploration could yield greater clarity on the mechanisms behind CJ. Eye-tracking studies, for example, could compare attention patterns of expert and novice respondents, offering empirical support for differences in professional vision.

Limitations specific to the reasoning component: The internal structure of the reasoning component was not evaluated in this study. Applying item response theory (IRT) could provide more insights into the item functioning of each of the multiple-choice reasoning items. Moreover, future research with larger sample sizes could explore differential test functioning (DTF) to detect whether respondents from varying backgrounds interpret and respond to the video clips consistently. Such analyses could reveal potential measurement biases and enhance the reasoning items' quality by evaluating their difficulty, discriminative value, and reliability.

General limitations: The ProLSI instrument was developed specifically for use in Flemish ECE settings, which may not generalise to other contexts. However, as Blömeke's competence continuum (Blömeke et al., 2015) emphasises, measuring situation-specific skills inherently requires contextualised tools to capture the unique demands of specific environments and situations. Furthermore, as Starcke and colleagues (Starcke et al., 2024) highlight, working with children requires more stringent ethical considerations. As videos were collected with young children, all necessary ethical clearances were obtained to ensure compliance with these standards. Another limitation pertains to implementation challenges that may arise. For optimal results, the instrument should be completed on a computer with a stable internet connection in a quiet, stimulus-free environment, preferably with headphones. These conditions may present logistical challenges in some settings. Additionally, the estimated total time to complete the instrument—approximately 75 minutes—demands sustained concentration, which could be taxing for respondents. Finally, while this study thoroughly assessed the instrument's validity, it did not evaluate the potential educational impact of the instrument on respondents' learning, referred to as the consequences of testing in the validity framework (American Educational Research Association, 2011).

Future research should explore how using this instrument affects teachers' learning and instructional practices.

6.3. Significance of the Research

The video-based measurement instrument presented in this study was developed in the context of an inter-university research and development project, which aims to enhance understanding of language-stimulation practices in early childhood education and identify effective professional development approaches. In this study, a quantitative and holistic measurement instrument was developed to empirically assess teachers' ProLSI. The instrument enables the measurement of ProLSIs in large samples simultaneously, distinguishing it from qualitative designs more suited to smaller sample sizes. This efficient and validated tool facilitates large-scale assessments, which will expand empirical knowledge of language-stimulation interactions in the classroom and shed light on two challenging, under-researched areas: language interactions and early childhood education (ECE).

While this instrument is rooted in a fundamental research project and is primarily designed for use in a specific Flemish ECE context, its development process still offers valuable insights for researchers interested in adapting similar tools to other settings. By detailing the decisions and procedures undertaken in the development process, this publication provides a framework that could inspire the creation of comparable instruments tailored to other countries' ECE contexts. Moreover, the instrument is intended for use by researchers, ensuring rigorous application in empirical studies of early language-stimulation practices.

In the research field, the need has been expressed for targeted (video-based) professional development initiatives (Santagata, 2024). The current study helps address this need by serving as a foundation for targeted professional development initiatives. These programs can, for example, focus on teachers' noticing ability for a specific language interaction strategy, on teachers' reasoning ability in general, or even on practising reflective practices, depending on the specific challenges identified by the assessment. It also offers higher education institutions valuable insights into the starting point of their students. By tracking these insights across different stages of teacher education, institutions can identify critical areas for improvement and tailor their teacher training curricula accordingly. By promoting reflective practices and providing insights into teachers' professional vision in authentic classroom settings, the instrument has the potential to help bridge the theory–practice gap of LSIs in ECE.

Subsequent research can build on this instrument to explore how the different components of LSI competencies, such as dispositions and actual classroom behaviour, are linked to each other.

Author Contributions: Conceptualization, L.D. and R.V.; methodology, L.D., A.-L.S. and R.V.; software, L.D.; validation, L.D.; formal analysis, L.D.; investigation, L.D. and A.-L.S.; resources, L.D. and A.-L.S.; data curation, L.D.; writing—original draft preparation, L.D.; writing—review and editing, L.D., R.V., A.-L.S., W.V. and K.V.d.B.; visualization, L.D. and A.-L.S.; supervision, R.V., K.V.d.B. and W.V.; project administration, L.D.; funding acquisition, R.V., W.V. and K.V.d.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research Foundation Flanders (FWO) under grant number S007223N.

Institutional Review Board Statement: Ethical clearance was sorted and obtained from the Ethics Committee of the Faculty of Arts and Philosophy at Ghent University on 26 May 2023 under approval number 2023-22.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw, anonymized quantitative data supporting the conclusions of this article (Samples 1 and 3) are publicly available at https://doi.org/10.5281/zenodo.14675612. Data from Sample 2 cannot be shared publicly, as participant consent for the disclosure of qualitative data was not obtained.

Acknowledgments: We thank Victoria Van Oss for her role as the third independent coder, as well as for coordinating the project and facilitating communication with the teachers to ensure their involvement. Maarten Goossens and Michiel Devrindt are acknowledged for their technical assistance with the Comproved platform, and Dina R. for conducting the think-aloud sessions. Finally, we thank the teachers who contributed to the video recordings and all the respondents who participated in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Noticing Component

In the noticing component of the video-based instrument, the respondents complete a comparative judgement task. The respondents are shown a random pair of video clips and are asked to indicate in which video they think the teacher best put each of the language interaction strategies to practice: language input (LI), opportunities for language production (OLP), and feedback (FB). The task involves analysing ten video pairs, with each pair being compared across the three different language interaction strategies. Figure A1 provides a screenshot of one comparison pair, with key screen elements explained in text boxes. Note that while the screenshot shows blurred faces for data protection, this blurring is absent in the actual instrument.

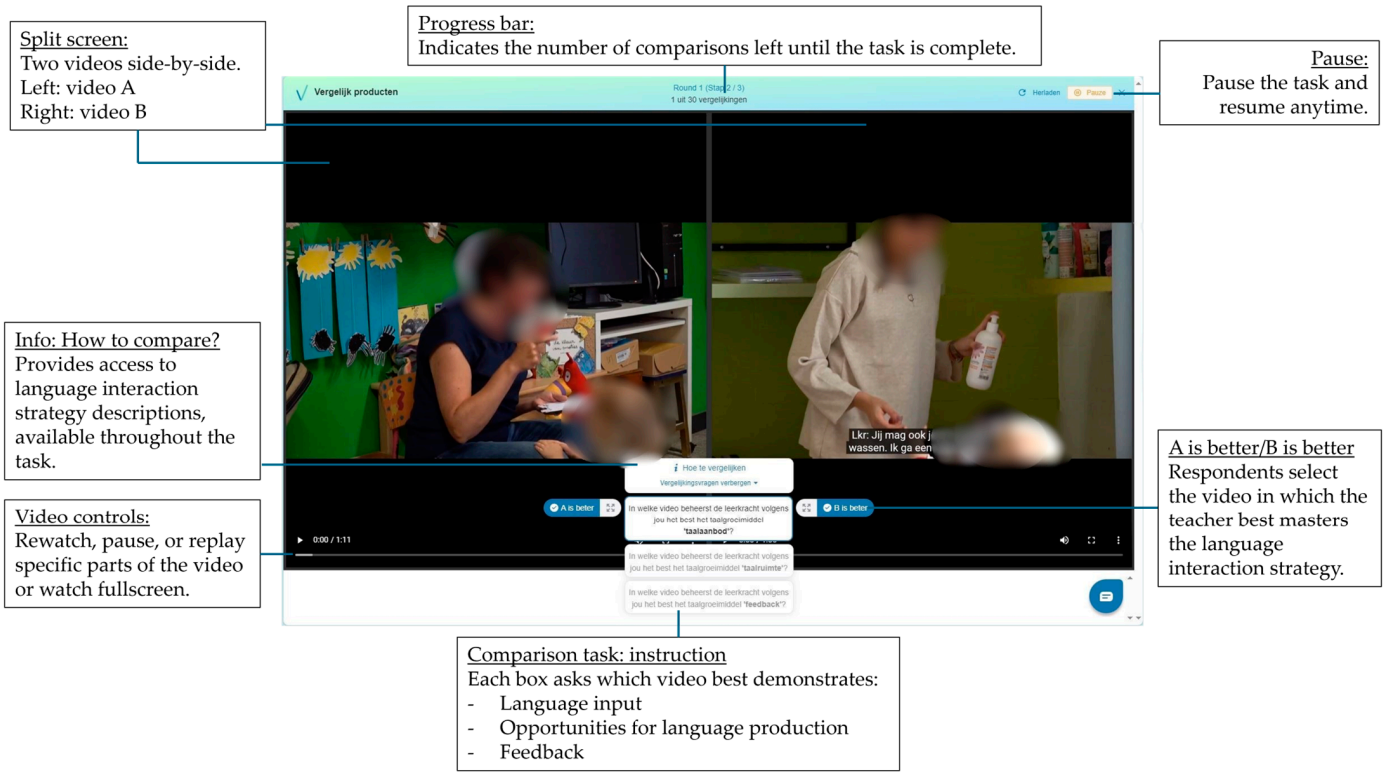


Figure A1. Screenshot of one pair of videos in the noticing comparison task.

Video clip A (translated transcript)	**Video clip B (translated transcript)**
The teacher holds a blank colouring page of a monster in front of the full class. She is pointing at it while she talks.	*In the video, several children are shown lining up to wash their hands as the teacher holds a can of soap.*
Teacher: You can feel calm; that is green. I can add a bit of yellow to it (*the drawing*) because today I feel happy and calm. *(Puts the drawing down)* **Teacher**: Good, you may do that (*making the drawing*) later. **Ralph**: I'm blue. **Teacher**: Ah, blue? Are you sad, Ralph? **Ralph**: No. **Teacher**: Ah, not you, but Monster was sad? *Ralph says something that cannot be heard clearly.* **Teacher to class**: Yes, but you can be sad, right? **Anna:** I am sad! **Teacher:** Oh, you are sad? And can you tell me why you are sad? **Other child**: *(Interrupting)* I want . . . **Teacher to other child**: Wait, wait, wait. First, we'll listen to Anna. **Teacher to Anna:** Anna, why are you sad? *(Marcus tries to interrupt.)* **Teacher**: Marcus, wait. **Teacher to Anna**: Tell me. **Anna**: Sad and angry. Sad and angry. **Teacher:** You are sad and angry? And can you tell me why you're sad? Did you lose something, like Mona? **Anna:** Yes. **Teacher:** Have you lost something? What did you lose? **Anna:** A cup. **Teacher:** Also a cup? Oh no. Ah, and you haven't found it yet? *(Another child tries to interrupt. The teacher continues with Anna.)* **Teacher:** And why are you angry? *(Anna mumbles something that cannot be understood.)* **Teacher to Anna**: Are you angry about something that happened? **Teacher to the class:** You know what? You can tell me later at the table.	**Teacher to child**: You can wash your hands too. I will help. Your hands are still wet. I'm going to dry them on the towel. Yes. **Teacher to other child**: A bit of soap. I will roll op your sleeves, so that it is not too wet. There you go. *(Teacher helps child with rolling her sleeves)* **Livia**: Me help too **Teacher to Livia**: Help? *(Teacher helps Livia with rolling her sleeves)* **Teacher to Michael:** Calm, Michael. If you are ready, you may put your box in your satchel. **Teacher to other child**: Oops, your crown is falling off. *(The teacher puts the child's crown back on her head)*. There you go. **Teacher to Todd**: Are you ready Todd? **Todd:** *(nods)* **Teacher**: We will wash your hands. **Teacher to Livia**: yes, Livia, you may wash your hands. A bit of soap *(adds soap in Livia's hands).* There you go. **Teacher to Todd**: Here, your box can go in the satchel and I will then wipe your mouth one more time, your mouth is dirty. **Teacher to Livia:** Ready? Yes. You may dry your little hands on the towel, Livia. **Todd**: This satchel. **Teacher**: Yes, just put it in the satchel.

Transcripts for both videos in the screenshot are provided on the next page, translated from Dutch to English with names altered for data protection. Each transcript is preceded by a brief description of the scene depicted in the video. It is important to note that in the comparative judgement task, no additional context is provided within the videos. However, the videos in the instrument include subtitles.

Appendix B. The Reasoning Component

Following the comparative judgement task, the respondents complete a video-based exercise on three video clips that are familiar to the respondents from the previous task. The videos selected for this task showcase an authentic interaction between a teacher and child(ren) that clearly illustrates one of three language interaction strategies: language input (LI), opportunities for language production (OLP), or feedback (FB). Each video is preceded by a brief description of the class context depicted. After the respondents have seen the video, the respondents answer two sets of multiple-choice items related to the specific language interaction strategy shown.

The first set of multiple-choice items presents statements of advice that a teacher could use to improve their skills in the specified language interaction strategy. The participants are asked to evaluate the advice based on the video they just watched, categorising it as (1) good advice, if the teacher could effectively use this advice to support the specified strategy; (2) redundant advice, if the teacher was already implementing it in the video clip; or (3) not good advice, if it would not contribute to the strategy for language-stimulation interaction in the given video.

In the second set of multiple-choice items, the respondents are shown several dichotomous statements about the video they have seen, to which respondents either agree or disagree.

Translated transcripts of each of these videos are provided below, followed by the two sets of multiple-choice items per video. Only items where the professionals' panel reached a consensus on the correct answer are included in Appendix A.

Appendix B.1. Video 1: Language Input (LI)

Brief description of the class context: In teacher Sandra's class, the central theme of the week is pumpkins. Some children hollow out a pumpkin with the teacher so they can see what it looks like on the inside. In this clip, you will see how teacher Sandra introduces this activity.

Instruction: Watch the clip of teacher Sandra and pay close attention to her language input. Later, you will provide tips to the teacher on how to improve her language input.

LI description: Language input means challenging all children with rich language by using full sentences and incorporating an extensive vocabulary, both from everyday language and abstract school language.

Four children sit around a table with the teacher. There is a pumpkin and a worksheet on the table.

Teacher: I brought these little spoons. *(Teacher shows the spoon).* What are they normally for? What do we do with these?
Child: For ice cream.
Teacher: For ice cream, right. But now we're not going to use them for ice cream. Now we're going to use them for the pumpkin.
Look, like this. *Teacher scoops inside the pumpkin*. With the scoop inside. Look.
Child: Can we do that too?
Teacher: And then we can get it out . . . You're going to be able to do that too. And I brought a sheet with me. *Teacher shows a sheet to the children with three columns on it: a column with a picture of pumpkin'flesh',a picture of 'pulp', and a picture of 'seeds'.*
And we have . . . Oh, wait a second. We have pulp, seeds, and the flesh of the pumpkin. (*Teacher points with her fingers to the different categories as she says them*). And look, I already have one here. *(The teacher holds up a pumpkin seed in front of the children).* What's that?
Child: A seed.
Teacher: A seed. Where does it go (*on the sheet*)?
(Child shows the placement on the sheet.)
Teacher: Yes, there, like this.
Child: You're not supposed to eat that, right?
Teacher: And this? What was this here? (*The teacher holds up a piece of pulp in front of the children).* The . . .
Child: That's also a seed, look.
Teacher: It's not the flesh, but. . . The pulp.
Child: Look!
Teacher: And we have to put the pulp here. (*The teacher shows placement on the sheet).* And the seeds. . .
Child: Look, seeds.
Teacher: Another seed? The seed goes here. (*The teacher shows placement on the sheet).*
Teacher: Our hands are going to get a little dirty, but that's okay.
Child: And then we'll wash our hands.

Instruction: Imagine you are an expert observing teacher Sandra's class and you see this short clip. What advice would you give the teacher to enhance her language input?

	The teacher is already doing this	This is good advice for the teacher that enhances language input	This is not good advice because it does not enhance language input
Verbally express in full sentences what the children are doing (=parallel talk).			
Explain out loud why you are doing what you are doing (=explain your reasoning).			
Avoid using gestures or visual aids as they may distract the children's attention from your language.			
Explain the difference between pulp, seed, and flesh using words, so children can hear the differences in language, not just see them.			
Simplify your language (words and sentence structure) so that all children, including those with weaker language skills, can follow.			
Use diminutives and child-friendly language to keep the children's attention.			

Instruction: Answer the following statements about teacher Sandra's approach in this clip with "agree" or "disagree".

	Agree	Disagree
The teacher's approach ensures that the children will not forget the new words.		
The teacher's approach ensures that the children now know and can use these new words.		
The teacher works with the words introduced by the children.		
The teacher's approach is mainly focused on helping the children understand what 'flesh', 'seed', and 'pulp' are, rather than teaching the words themselves.		
Because the teacher links the new words 'seed', 'flesh', and 'pulp' to a familiar word (the pumpkin), the children learn these words more easily.		

Appendix B.2. Video 2: Opportunities for Language Production (OLP)

Brief description of the class context: Teacher Selene has just read a book to the entire class about a monster's emotions. In this story, each feeling has a colour. Soon, the children will colour a paper monster in the colour that matches how they feel. The teacher explains what the children will be doing.

Instruction: Watch the clip of teacher Selene. Pay close attention to how she stimulates interactions in the classroom by creating opportunities for the children to produce language. Later, you will provide tips for the teacher on how to expand her opportunities for language production in the classroom.

OLP description: Giving opportunities for language production means encouraging all children to engage in thinking and talking with you and with each other by asking good questions and giving them enough time to think and speak.

The teacher holds a blank colouring page of a monster in front of the full class. She is pointing at it while she talks.

Teacher: You can feel calm; that is green. I can add a bit of yellow to it (*the drawing*) because today I feel happy and calm. (*The teacher puts the drawing down).*
Teacher: Good, you may do that (*making the drawing*) later.
Ralph: I'm blue.
Teacher: Ah, blue? Are you sad, Ralph?
Ralph: No.
Teacher: Ah, not you, but Monster was sad?
(Ralph says something that cannot be heard clearly).
Teacher to class: Yes, but you can be sad, right?
Anna: I am sad!
Teacher: Oh, you are sad? And can you tell me why you are sad?
Other child: I want . . .
Teacher to other child: Wait, wait, wait. First, we'll listen to Anna.
Teacher to Anna: Anna, why are you sad?
(Marcus interrupts).
Teacher: Marcus, wait.
Teacher to Anna: Tell me.
Anna: Sad and angry. Sad and angry.
Teacher: You are sad and angry? And can you tell me why you're sad? Did you lose something, like Mona?
Anna: Yes.
Teacher: Have you lost something? What did you lose?
Anna: A cup.
Teacher: Also a cup? Oh no. Ah, and you haven't found it yet?
(Another child tries to interrupt).
Teacher to Anna: And why are you angry?
(Anna mumbles something that cannot be heard).
Teacher to Anna: Are you angry about something that happened?
(Anna remains silent).
Teacher to the class: You know what? You can tell me later at the table.

Instruction: Imagine you are an expert observing teacher Selene's class and you see this short clip. What advice would you give the teacher to increase opportunities for language production in the classroom?

	The teacher is already doing this	This is good advice for the teacher that helps create opportunities for language production	This is not good advice because it does not help create opportunities for language production
Make sure the speaking turns of children who are speaking are not interrupted by other children.			
Ask another child to help Anna express her answer.			
Make eye contact with the children you are speaking with.			
Look at the children expectantly.			
Ask open-ended questions to Anna.			
Wait long enough between the questions "Why are you sad?" and "Did you lose something like Mona?" so the child has a chance to answer.			
Involve a child, who doesn't spontaneously engage in the conversation, into the conversation with Anna.			

Limit the time for spontaneous conversations to leave more room for the planned activity, which is drawing.			
Ask fewer questions to children with weaker language skills so they do not feel uncomfortable if they cannot answer.			

Instruction: Answer the following statements about teacher Selene's approach in this clip with "agree" or "disagree".

	Agree	Disagree
Both Ralph and Anna get a challenging speaking opportunity in this clip.		
Teacher Selene's approach stimulates the children to think and talk about their feelings.		
By limiting her own speaking time, teacher Selene creates an inviting atmosphere for the children to speak.		
Teacher Selene's approach ensures that quiet children also get a chance to speak.		

Appendix B.3. Video 3: Feedback

Brief description of the class context: In this clip, the children in teacher Jake's class are doing different activities across the classroom. At teacher Jake's table, they are crafting. It is autumn, and they are allowed to decorate a branch with glitter and feathers.

Instruction: Watch the clip of teacher Jake closely, paying particular attention to how he stimulates language interaction in the classroom by providing feedback on what the children say. Later, you will provide tips to teacher Jake on how he can improve his feedback.

FB description: Feedback involves responding carefully to the children's expressions, enabling them to enhance, expand, and refine their language and understanding of the world.

A few children and the teacher are sitting around the table as they are crafting.

Minnie: I need "cut".
Teacher: Is that a "cut"? Those are . . .
Minnie: Scissors.
Teacher: Scissors. But with scissors you can . . . cut. Take these scissors, Minnie. Minnie, take these scissors.
(Ryan comes to stand next to the teacher and shows a drawing)
Ryan to teacher: Look!
Teacher to Ryan: What did you make, my friend?
Ryan: This is a playground . . . If you can play outside, also, uh . . . play with it, but actually not with a little friend because the dog was lost.
Teacher: Oh no, where was the dog lost? What happened to the dog?
Ryan: The dog was . . . lost . . . by that man.
Teacher: Aaah, so the man lost his dog. That's unfortunate, isn't it? That's not very nice, is it? How do you think the man felt?
Ryan: *(Ryan mumbles something that cannot be understood).*
Teacher: Not good. And when you lose someone or something, then you feel a bit . . .
Ryan: Uhmm . . .
Teacher: How do you feel then?
Ryan: Sad.
Teacher: Sad, yes. I would also feel very sad if my dog ran away.

Instruction: Imagine you are an expert observing teacher Jake's class and you see this short clip. What advice would you give the teacher to improve his feedback on the children's expressions?

	The teacher is already doing this	This is good advice for the teacher that is helpful for providing feedback	This is not good advice because it is not helpful for providing feedback
Expand Minnie's incorrect sentence by adding new language or a follow-up question (=expanding).			
Leave Minnie's incorrect statement as it is to maintain her confidence.			
Clarify the difference between "cutting" and "scissors" for Minnie by explaining the link between the two words.			
Check if you have understood the message from Ryan (the child with the dog) correctly.			
Describe what you see in the picture before Ryan tells his story, so he receives immediate feedback.			
Summarize in your own words what you think Ryan means (=paraphrasing).			
Encourage Ryan to rephrase what he's saying to clarify his meaning, for example by asking: "Can you explain in other words?"			

Instruction: Answer the following statements about teacher Jake's approach in this clip with "agree" or "disagree."

	Agree	Disagree
The way the teacher responds to what the children say helps them refine and improve their language.		
The interaction with the teacher enhances Ryan's sense of language competence.		
In this clip, the teacher corrects the children both directly (explicitly) and indirectly (implicitly).		

References

Algoet, M., & Vandenbussche, L. (2024). Interactievaardigheden van de kleuterleraar: De motor die de zeven principes doet draaien. In *Boordevol taal. Didactiek nederlands voor het kleuteronderwijs* (pp. 67–92). Owl Press. ISBN 978-94-64788-85-3.

American Educational Research Association (Ed.). (2011). *Report and recommendations for the reauthorization of the institute of education sciences*. American Educational Research Association. ISBN 978-0-935302-35-6.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. [CrossRef]

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13. [CrossRef]

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324. [CrossRef]

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment*. Cambridge Assessment.

Cohen, L., Manion, L., & Morrison, K. (2009). *Research methods in education* (6th ed., reprint). Routledge. ISBN 978-0-203-02905-3.

Comproved—Assess better. Learn more. (n.d.). Available online: https://comproved.com/en/ (accessed on 20 November 2024).

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin. ISBN 0-395-30790-2.

De Coninck, K., Valcke, M., & Vanderlinde, R. (2018). A measurement of student teachers' parent–teacher communication competences: The design of a video-based instrument. *Journal of Education for Teaching*, *44*(3), 333–352. [CrossRef]

Eadie, P. (2022). Oral language skills as a foundation for learning to learn. In J. Law, S. Reilly, & C. McKean (Eds.), *Language development: Individual differences in a social context* (pp. 397–419). Cambridge University Press. ISBN 978-1-108-64371-9. [CrossRef]

European Commission. (2018). *Proposal for a council recommendation on high quality early childhood education and care systems*. European Commission.

Gegenfurtner, A., Lewalter, D., Lehtinen, E., Schmidt, M., & Gruber, H. (2020). Teacher expertise and professional vision: Examining knowledge-based reasoning of pre-service teachers, in-service teachers, and school principals. *Frontiers in Education*, *5*, 59. [CrossRef]

Gippert, C., Hörter, P., Junker, R., & Holodynski, M. (2022). Professional vision of teaching as a focus-specific or focus-integrated skill—Conceptual considerations and video-based assessment. *Teaching and Teacher Education*, *117*, 103797. [CrossRef]

Gold, B., & Holodynski, M. (2017). Using digital video to measure the professional vision of elementary classroom management: Test validation and methodological challenges. *Computers & Education*, *107*, 13–30. [CrossRef]

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633. [CrossRef]

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*(1), 4–14. [CrossRef]

Justice, L. M., Mashburn, A. J., Hamre, B. K., & Pianta, R. C. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly*, *23*(1), 51–68. [CrossRef] [PubMed]

Keppens, K. (2020). *Videography in teacher education: A study on inclusive teaching competences* [Ph.D. thesis, Ghent University].

Keppens, K., Consuegra, E., Goossens, M., De Maeyer, S., & Vanderlinde, R. (2019). Measuring pre-service teachers' professional vision of inclusive classrooms: A video-based comparative judgement instrument. *Teaching and Teacher Education*, *78*, 1–14. [CrossRef]

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, *68*(5), 845–861. [CrossRef]

König, J., Santagata, R., Scheiner, T., Adleff, A.-K., Yang, X., & Kaiser, G. (2022). Teacher noticing: A systematic literature review of conceptualizations, research designs, and findings on learning to notice. *Educational Research Review*, *36*, 100453. [CrossRef]

Lesterhuis, M., De Maeyer, S., Van Daal, T., Van Gasse, R., Coertjens, L., Verhavert, S., Mortier, A., Coenen, T., Vlerick, P., Vanhoof, J., & Van Petegem, P. (2015). Competenties kwaliteitsvol beoordelen: Brengt een comparatieve aanpak soelaas? *Tijdschrift voor Hoger Onderwijs 33*, 55–67.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749. [CrossRef] [PubMed]

Mathers, S. J. (2021). Using video to assess preschool teachers' pedagogical knowledge: Explicit and higher-order knowledge predicts quality. *Early Childhood Research Quarterly*, *55*, 64–78. [CrossRef]

Meschede, N., Fiebranz, A., Möller, K., & Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: On its relation and differences between pre-service and in-service teachers. *Teaching and Teacher Education*, *66*, 158–170. [CrossRef]

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. [CrossRef]

Peleman, B., Vandenbroeck, M., & Van Avermaet, P. (2019). *De overgang naar de kleuterschool voor kinderen uit gezinnen in armoede*. Universiteit Gent.

Peleman, B., Vandenbroeck, M., & Van Avermaet, P. (2020). Early learning opportunities for children at risk of social exclusion. Opening the black box of preschool practice. *European Early Childhood Education Research Journal*, *28*(1), 21–42. [CrossRef]

Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., Duncan, G., Dynarski, M., Magnuson, K., & Weiland, C. (2017). *The current state of scientific knowledge on Pre-Kindergarten effects* (p. 104). Brookings Institution.

Pianta, R. C., Burchinal, M., Jamil, F. M., Sabol, T., Grimm, K., Hamre, B. K., Downer, J., LoCasale-Crouch, J., & Howes, C. (2014). A cross-lag analysis of longitudinal associations between preschool teachers' instructional support identification skills and observed behavior. *Early Childhood Research Quarterly*, *29*(2), 144–154. [CrossRef]

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2012). *Classroom Assessment Scoring System (CLASS) manual PRE-K* (8th ed., print). Paul H. Brookes Publishing Co. ISBN 978-1-55766-941-4.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. [CrossRef]

Prevoo, M. J. L., Malda, M., Mesman, J., & Van IJzendoorn, M. H. (2016). Within-and cross-language relations between oral language proficiency and school outcomes in bilingual children with an immigrant background: A meta-analytical study. *Review of Educational Research*, *86*(1), 237–276. [CrossRef]

Roose, I., Goossens, M., Vanderlinde, R., Vantieghem, W., & Van Avermaet, P. (2018). Measuring professional vision of inclusive classrooms in secondary education through video-based comparative judgement: An expert study. *Studies in Educational Evaluation*, *56*, 71–84. [CrossRef]

Roose, I., Vantieghem, W., Van Damme, K., Lambert, P., Vanderlinde, R., & Van Avermaet, P. (2019). Measuring teachers' professional vision of inclusive classrooms through video-based comparative judgement. What does it mean to misfit? *International Journal of Educational Research*, *98*, 257–271. [CrossRef]

Sabol, T. J., Soliday Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-k programs predict children's learning? *Science*, *341*(6148), 845–846. [CrossRef] [PubMed]

Sammons, P., Toth, K., Sylva, K., Melhuish, E., Siraj, I., & Taggart, B. (2015). The long-term role of the home learning environment in shaping students' academic attainment in secondary school. *Journal of Children's Services*, *10*(3), 189–201. [CrossRef]

Santagata, R. (2024). Avenues for future research on teacher noticing and professional vision. In *Teacher professional vision: Empirical perspectives* (pp. 287–301). Routledge. ISBN 978-1-003-37060-4.

Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal*, *51*(4), 739–771. [CrossRef]

Sherin, M., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20–37. [CrossRef]

Sherin, M. G. (2001). Developing a professional vision of classroom events. In T. L. Wood, B. S. Nelson, & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 75–93). Erlbaum. ISBN 978-0-8058-3570-0.

Shonkoff, J. P., & Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early child development*. National Academy Press. ISBN 978-0-309-50488-1.

Starcke, L., Barenthien, J., Heinze, A., Steffensky, M., & Dunekacke, S. (2024). Illustrations as item stimuli to assess early childhood teachers' situation-specific skills—A promising approach? In *Teacher professional vision: Empirical perspectives* (pp. 251–268). Routledge. ISBN 978-1-003-37060-4.

Stürmer, K., Könings, K. D., & Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology*, *83*(3), 467–483. [CrossRef]

Theelen, H., Beemt, A., & Brok, P. (2019). Using 360-degree videos in teacher education to improve preservice teachers' professional interpersonal vision. *Journal of Computer Assisted Learning*, *35*(5), 582–594. [CrossRef]

van Es, E., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Information Techology for Teacher Education*, *10*, 571–596.

van Es, E., & Sherin, M. G. (2010). The influence of video clubs on teachers' thinking and practice. *Journal of Mathematics Teacher Education*, *13*(2), 155–176. [CrossRef]

Vanderlinde, R., Godaert, E., & Keppens, K. (2020). Competentiegericht opleiden en beoordelen in de lerarenopleiding: Naar een nieuw conceptueel model. *Tijdschrift Voor Lerarenopleiders*, *41*, 238–252.

Vantieghem, W., Roose, I., Gheyssens, E., Griful-Freixenet, J., Keppens, K., Vanderlinde, R., Struyven, K., & Van Avermaet, P. (2020). Professional vision of inclusive classrooms: A validation of teachers' reasoning on differentiated instruction and teacher-student interactions. *Studies in Educational Evaluation*, *67*, 100912. [CrossRef]

Verhallen, M., & Walst, R. (2011). *Taalontwikkeling op school: Handboek voor interactief taalonderwijs* (2nd. ed., print). Uitgeverij Coutinho. ISBN 978-90-469-0254-7.

Weyers, J., König, J., Santagata, R., Scheiner, T., & Kaiser, G. (2023). Measuring teacher noticing: A scoping review of standardized instruments. *Teaching and Teacher Education*, *122*, 103970. [CrossRef]

Widjaja, W., Xu, L., & Jobling, W. (2018). Examining primary school teachers' professional noticing through a video-based methodology. In L. Xu, G. Aranda, W. Widjaja, & D. Clarke (Eds.), *Video-based research in education* (pp. 66–82). Routledge. ISBN 978-1-315-10921-3. [CrossRef]

World Health Organization. (2023). *Improving early child hood development: WHO Guideline* (pp. 1–6). World Health Organization.

Young, K. A. (2005). Direct from the source: The value of "think-aloud" data in understanding learning. *Journal of Educational Enquiry*, *6*(1), 19–33.

Zaslow, A., Anderson, R., Redd, Z., Wessel, J., Tarullo, L., & Burchinal, M. (2010). *Quality dosage, thresholds, and features in early childhood settings: A review of the literature*. U.S. Department of Health and Human Services.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.