HOW WOULD YOU DO IT?

ASSESSING PROBLEM-SOLVING COMPETENCIES IN THE PRESCHOOL CLASSROOM

JORIS VAN ELSEN

HOW WOULD YOU DO IT?

Assessing problem-solving competencies in the preschool classroom

JORIS VAN ELSEN

Nederlandstalige titel

Hoe zou jij dat doen? Probleemoplossend vermogen evalueren in de kleuterklas

Proefschrift voorgelegd tot het behalen van de gezamenlijke graad van doctor in de Onderwijswetenschappen aan de Universiteit Antwerpen en doctor in de Pedagogische Wetenschappen aan de KU Leuven te verdedigen door Joris Van Elsen

Declaration of integrity

The PhD researcher and supervisors declare that the PhD research was conducted according to the principles of scientific integrity, as mentioned in the general PhD regulations and charter for PhD researchers of University of Antwerp and the integrity charter for PhD researchers and supervisors affiliated with the University of Antwerp.

Cover illustration

Frederike (6 years old)

Printing

Ridderprint | www.ridderprint.nl

University of Antwerp
Faculty of Social Sciences
Training and Education Sciences

KU LEUVEN
Faculty of Psychology and Educational Sciences
Centre for Instructional Psychology and Technology

HOW WOULD YOU DO IT?

Assessing problem-solving competencies in the preschool classroom

To be defended by

JORIS VAN ELSEN

Dissertation submitted to obtain the joint degree of
Doctor of Education Sciences — University of Antwerp
and Doctor of Educational Sciences — KU Leuven

Supervisors:
Prof. dr. Sven De Maeyer
Prof. dr. Joke Torbeyns

Antwerp, 26 August 2026

fwo

Supported by FWO
[Grant Number 1S03525N]

Composition of the doctoral jury

Supervisors

Prof. dr. Sven De Maeyer	University of Antwerp	
Prof. dr. Joke Torbeyns	KU Leuven	

Doctoral jury

Prof. dr. David Gijbels	University of Antwerp	*Chair doctoral committee*
Dr. Goedele Vandommele	KU Leuven	*Member doctoral committee*
Prof. dr. Vincent Donche	University of Antwerp	
Dr. Terhi Vessonen	Southern Methodist University	
Prof. dr. Ilonca Hardy	Goethe University Frankfurt	

SUMMARY

This dissertation focuses on the development of a valid, reliable, feasible, and usable instrument for the formative assessment of preschoolers' problem-solving competencies. Solving problems requires the coordination of a wide range of cognitive, metacognitive, motivational, social, emotional, and motor processes. This makes it particularly challenging to assess and support the development of problem-solving competencies in young children. Formative assessment enables teachers to identify preschoolers' strengths and instructional needs, allowing them to tailor learning activities accordingly.

Problem solving concerns the transformation of a given problem situation into a desired goal state when the solution path is not straightforward. Problem solving is studied across research fields and with various types of problems, target groups, and aims. The first challenge addressed in this dissertation was to map the diversity of research on problem solving in young children (aged 3–8 years). This was approached through a bibliometric review. The results demonstrate that problem solving is a multidimensional construct that is studied from multiple perspectives. The first study provides a framework to situate research on early problem solving within the broader scientific literature.

The next challenge was to determine an appropriate method for assessing the problem-solving competencies of preschool children (aged 3–6 years). Drawing on a scoping review, the second study offers an overview of the characteristics, scope, and psychometric properties of such methods used in scientific research. The findings highlight the potential of structured observation of young children as they solve engineering design problems to assess their problem-solving competencies.

The third challenge addressed in this dissertation concerned making preschoolers' problem solving visible by translating it into observable behaviours. To this end, preschoolers (aged 5–6 years) were recorded as they solved three engineering design problems. Subsequently, preschool teachers and fellow researchers were invited to identify relevant problem-solving behaviours and classify them according to the steps of the PISA problem-solving framework. The outcome of this study is an observation scheme that can be used to document young children's problem-solving behaviours.

Finally, we developed a method for the formative assessment of preschoolers' problem-solving competencies and asked teachers to apply it in their classrooms. Based on the collected observation data of 53 children (aged 5–6 years) and teachers' feedback, we constructed an argumentation that identifies the most critical aspects related to the validity of the interpretation and use of the assessment results. The findings of this study point to the potential of formative assessment for identifying preschoolers' strengths and instructional needs in relation to the development of their problem-solving competencies.

In addition, the study highlights the importance of context as it influences both problem-solving behaviours and its evaluation.

This dissertation brings together existing scientific evidence on assessing the problem-solving competencies of young children and bridges the gap to classroom practice. In summary, the four studies show that problem solving is a multifaceted construct; that young children display a wide repertoire of problem-solving behaviours; that these behaviours can be validly assessed using a scientifically grounded observation instrument; and that formative assessment provides teachers with a comprehensive picture of young children's problem-solving competencies, enabling them to tailor their teaching to the children's needs.

SAMENVATTING

Dit proefschrift richt zich op de ontwikkeling van een valide, betrouwbaar, haalbaar en bruikbaar instrument voor de formatieve evaluatie van het probleemoplossend vermogen van kleuters. Problemen oplossen vereist coördinatie van een breed scala aan cognitieve, metacognitieve, motivationele, sociale, emotionele en motorische processen. Dit maakt het bijzonder uitdagend om de ontwikkeling van het probleemoplossend vermogen bij kleuters te evalueren en te ondersteunen. Formatieve evaluatie kan leerkrachten helpen om de sterktes en ondersteuningsnoden van kleuters in kaart te brengen en hun leeractiviteiten daarop af te stemmen.

Probleemoplossing draait om het omvormen van een bestaande probleemsituatie naar een gewenste doelsituatie waarbij de weg ernaartoe niet eenduidig is. Probleemoplossing wordt bestudeerd in tal van onderzoeksgebieden en met verschillende soorten problemen, doelgroepen en doelstellingen. De eerste uitdaging die we met dit proefschrift aangingen, was het in kaart brengen van de diversiteit aan onderzoek naar probleemoplossing bij jonge kinderen (3–8 jaar). Dit deden we door middel van een bibliometrische review. De resultaten tonen aan dat probleemoplossing een multidimensionaal construct is dat vanuit verschillende perspectieven wordt benaderd. De eerste studie biedt een kader om onderzoek naar probleemoplossing bij jonge kinderen te situeren binnen de bredere onderzoeksliteratuur.

Een volgende uitdaging was het identificeren van een geschikte methode om het probleemoplossend vermogen van kleuters te evalueren. Dit deden we met een scoping review waarmee we de kenmerken, focus en psychometrische eigenschappen in kaart brachten van evaluatiemethoden die in wetenschappelijk onderzoek worden gebruikt voor het bestuderen van het probleemoplossend vermogen bij kleuters (3–6 jaar). De resultaten wijzen op het potentieel van gestructureerde observatie van kleuters tijdens het oplossen van engineering design problemen voor de evaluatie van het probleemoplossend vermogen van kleuters.

De derde uitdaging van dit proefschrift bestond erin probleemoplossing van kleuters zichtbaar te maken door het te vertalen naar observeerbaar gedrag. Hiervoor filmden we kleuters (5–6 jaar) tijdens het oplossen van drie engineering design problemen. Vervolgens vroegen we kleuterleerkrachten en collega-onderzoekers om relevant probleemoplossingsgedrag te identificeren en te plaatsen binnen de stappen van het PISA–kader voor probleemoplossing. Het resultaat van deze studie is een observatieschema dat kan worden gebruikt om het probleemoplossingsgedrag van kleuters te documenteren.

Tot slot ontwikkelden we een methode voor formatieve evaluatie van het probleemoplossend vermogen van kleuters en vroegen we leerkrachten om die toe te passen in hun klas. Op basis van de verzamelde observatiedata van 53 kleuters (5–6 jaar)

en de feedback van de leerkrachten bouwden we een argumentatie op waarmee we de meest kritische aspecten met betrekking tot de validiteit van de interpretatie en het gebruik van de resultaten blootlegden. De resultaten van deze studie tonen het potentieel van formatieve evaluatie voor het identificeren van sterktes en ondersteuningsnoden van kleuters met betrekking tot de ontwikkeling van hun probleemoplossend vermogen. Daarnaast benadrukt de studie het belang van context, aangezien de context een invloed heeft op zowel het probleemoplossingsgedrag als op de evaluatie ervan.

Dit proefschrift brengt de bestaande wetenschappelijke evidentie rond het evalueren van het probleemoplossend vermogen van jonge kinderen samen en maakt de brug naar de klaspraktijk. Kort samengevat tonen de vier studies aan dat probleemoplossing een veelzijdig construct is, dat kleuters een breed repertorium aan probleemoplossingsgedrag vertonen, dat dit gedrag valide kan worden geëvalueerd met behulp van een wetenschappelijk onderbouwd observatie-instrument en dat formatieve evaluatie leerkrachten een uitgebreid beeld geeft van het probleemoplossend vermogen van kleuters dat hen in staat stelt hun onderwijsaanbod af te stemmen op de noden van de kinderen.

DANKWOORD | ACKNOWLEDGMENTS

Dit proefschrift zit boordevol vol problemen: kleine en grote problemen, problemen voor kleine en grote mensen, gestructureerde en ongestructureerde problemen, engineering en designproblemen, conceptuele problemen en meetproblemen, publicatieproblemen en eRRoRs. Maar problemen oplossen doe je niet alleen. Tijdens het hele traject kon ik rekenen op mensen die me bijstonden wanneer ik handen of ogen te kort kwam, wanneer ik kinderen en leerkrachten nodig had om alles uit te testen, wanneer ik nood had aan inspiratie of een sparringpartner, wanneer ik mijn weg moest zoeken in de academische wereld of wanneer problemen even onoplosbaar leken. Enkele van die mensen wil ik hier graag bedanken.

Ik begin met mijn promotoren, Sven en Joke. Jullie geloofden in mij en mijn project vanaf het eerste uur en gaven me de kans te doctoreren. Dankzij jullie kon ik me op mijn onderzoek storten en de academische wereld leren kennen. Jullie leerden me snoeien om te groeien en te bloeien. Jullie stonden altijd achter, voor, of naast me en hielpen me grenzen te verleggen tot ver voorbij wat ik ooit haalbaar achtte. Elke doctoraatsstudent verdient een promotorenteam als jullie.

Pirjo, mijn EARLI-mentor. Voor ik naar de biënnale EARLI-conferentie in Thessaloniki vertrok, had ik mijn promotoren beloofd dat ik een onderzoeksverblijf in het hoge noorden zou ‘fiksen’. Jouw keynote in de expo hal bood me die kans. Met de steun van de EARLI mentoring grant reisde ik naar Helsinki waar mijn onderzoek een internationale dimensie kreeg. Je stimuleerde me te doen wat nodig was, zonder omwegen, zonder compromissen, gewoon doen.

Veel dank ook aan mijn Finse ‘babysitters’ Natalia, Anne, Anssi en aan de professoren Kati, Heidi en Anu om me wegwijs te maken in de Finse gebruiken, me uit te nodigen in Vaasa en me te doen beloven om het probleemoplossingsproces nooit meer op een lineaire manier voor te stellen.

Pirjo, my EARLI mentor. Before travelling to the EARLI Biennial Conference in Thessaloniki, I had promised my supervisors that I would arrange a research stay in the North. Your keynote in the Expo Hall created that opportunity. Supported by the EARLI mentoring grant, I travelled to Helsinki, where my research gained an international dimension. You encouraged me to do what was necessary, directly and without compromise, just do it.

I am also grateful to my Finnish “babysitters”, Natalia, Anne, and Anssi, and to the professors Kati, Heidi and Anu, for introducing me to Finnish academic practices, inviting me to Vaasa, and making me promise never again to represent the problem-solving process as linear.

Lieven, mijn “promotor tot de tweede macht”. Eerlijk is eerlijk, het idee om jou te betrekken bij een conceptuele studie over probleemoplossing bij jonge kinderen kwam van Pirjo. Dat je op emeritaat bent, was voor haar geen valabel argument. En jij zei ja. Een conceptuele studie bleek even moeilijk af te slaan als een kopje koffie, dus combineerden we beide in Leuven, Gent of Graz. Je gaf mijn voorliefde voor concepten een nieuwe dimensie.

David. Als voorzitter van mijn doctoraatscommissie hielp je mee de weg uit te stippelen van mijn onderzoek en onderzoeksloopbaan. Je leerde mij en mijn collega's navigeren in de academische wereld en gunde me een blik achter schermen van het publicatieproces ("editors zijn ook mensen"). Eén advies zal ik altijd meedragen: dat strategische keuzes nooit alleen strategisch mogen zijn, maar altijd in lijn moeten liggen met wat je zelf wil.

Goedele. Als lid van mijn doctoraatscommissie ben je blijven hameren op het belang van validiteit en het integreren van het leerkrachtenperspectief. De beide thema's zijn hierdoor uitgegroeid tot leidmotief van dit proefschrift.

Leen. Als promotor van mijn masterproef heb je alles mee in gang gezet. Je moedigde me aan de stap te zetten en gewoon dat mailtje te sturen naar Sven om hem te vragen of hij mijn promotor wilde zijn. Het resultaat heb je nu in handen.

Bram en Fien. De PhD Masterclass heeft me meer kansen geboden dan jullie mogelijk vermoeden. Zowat alles wat ik tijdens mijn doctoraat heb gedaan, van schrijven en spreken in het Engels tot netwerken, lag aanvankelijk ver buiten mijn comfortzone. De PhD Masterclass creëerde een veilige omgeving waar ik als startende doctoraatsstudent zelfvertrouwen kon tanken en lotgenoten uit Gent, Leuven en daarbuiten kon ontmoeten. De PhD Masterclass heeft heel wat muren gesloopt en deuren geopend.

Lisanne. Je deed zoveel meer dan het werk van een stagiaire. Je nam het project mee in handen. Toen ik je op dag één van je stage vroeg of we kleuters individueel of per twee zouden observeren, koos je voor het laatste. Een goede keuze, zo bleek, die het onderzoek weliswaar niet eenvoudiger, maar wel authentieker maakte. Je hielp niet alleen bij het verzamelen van data, maar ontpopte je tot allround onderzoeker die mee organiseerde, observeerde, codeerde en analyseerde.

Collega's van Edubron en CIP&T. Jullie deelden ervaringen, projectvoorstellen, posters en proefschriften, organiseerden barbecues en CIP&T trips, schreven kaartjes en vroegen hoe het ermee ging, ook toen ik me enkele maanden opsloot in "mijn grot" om dit proefschrift af te werken. Speciale dank aan Charlotte, Bea en Katrijn om jullie kinderen en living tijdelijk af te staan aan de wetenschap.

Anja. Ik wens elke onderwijsonderzoeker iemand toe zoals jij. Vanaf het begin van mijn doctoraat was ik altijd welkom in jouw klas om dingen uit te proberen, om data te verzamelen of om kleuters te vragen een cover te tekenen voor dit proefschrift. Ik heb ongelooflijk veel van jou geleerd.

Ann (x3), Magali, Evy, Lieve. Door opnames te bekijken en het probleemoplossingsgedrag van kleuters te beschrijven, brachten jullie het leerkrachtenperspectief mee in mijn onderzoek. Het observatieschema en evaluatie-instrument zou er zonder jullie inzet en inzicht niet hetzelfde hebben uitgezien.

Lotte, Mieke, Christel en Anja (opnieuw). Jullie gingen de uitdaging aan om het instrument dat we ontwikkelden als eersten uit te testen in de klas. Jullie inspanningen leverden rijke data op die een gedetailleerde analyse van de mogelijkheden en hindernissen van formatieve evaluatie in de kleuterklas mogelijk maakte.

Konijn, Vliegtuig, Beker, Bloem, Vlinder, Druif, Druif, Helikopter, Taart, Robot (de pseudoniemen zijn authentiek) en alle andere kleuters die deelnamen aan het onderzoek. Jullie hebben mij verrast, ontroerd, doen lachen en vooral doen inzien dat jullie creatieve probleemoplossers zijn.

Eddy, loodgieter en tegelijk één van de beste leermeesters die ik ooit heb gehad. Mijn fascinatie voor problemen oplossen en mijn waardering voor loodgieters als ultieme probleemoplossers is wellicht ontstaan toen we samen in de kruipkelder verwarmingsbuizen aan het leggen waren. Je leerde me dat voor elk probleem een oplossing bestaat. De vraag is alleen hoeveel kap- en breekwerk eraan te pas zal komen.

Dries, Marie en Fie. Misschien zijn toch niet alle problemen op te lossen met kap- en breekwerk. In de loop van mijn onderzoek werd jullie pad gekruist door problemen waar jullie, elk op jullie eigen manier, mee trachtten om te gaan. Dat kan al eens botsen, en dat is misschien zo slecht nog niet. Breek je wereld open en zoek je weg. Alles komt meestal goed. En anders maken we er samen het beste van.

Daan en Lien. Jullie waren de kroongetuigen van mijn onderzoek. Als kleuter gaven jullie me de inspiratie voor dit doctoraat. Nu staan jullie klaar om naar het middelbaar te gaan. Als ‘kleinste steekproef’ losten jullie al mijn problemen op. Hopelijk blijven jullie dat doen zodat ik een zorgeloze toekomst tegemoet kan gaan.

Astrid. Dank je wel om te zijn wie je bent en er altijd voor me te zijn.

Tenslotte wil ik mijn ouders bedanken die me alle kansen gaven om mijn eigen weg te zoeken en te gaan. Want kinderen kansen geven, daar draait dit proefschrift om.

Vijf jaar onderzoek leerde me dat problemen kansen bieden: kansen om bij te leren, kansen om te tonen wat je kan, kansen om nieuwe dingen te ontdekken, kansen om elkaar beter te leren kennen. Ik leerde ook dat problemen vooral kansen bieden als anderen je aanmoedigen, je ondersteunen en je vertrouwen geven. Tijdens mijn doctoraatsonderzoek was ik omringd door mensen die me inspireerden, die in me geloofden, die me hielpen bergen werk te verzetten, die kansen creëerden en me hielpen ze te grijpen. Ik wens iedereen hetzelfde toe.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GENERAL INTRODUCTION

What's the problem?

"A problem arises when a living organism has a goal, but does not know how this goal is to be reached" (Duncker, 1945, p. 2). The world is filled with problems, ranging from small, everyday challenges, such as preparing a cheesecake for guests with gluten and lactose intolerances, to major global issues, including the impacts of climate change and artificial intelligence.

Problem solving is widely recognised as a fundamental 21st-century competence across everyday, educational, and professional contexts (Csapó & Funke, 2017b; Jonassen, 2000; Scherer & Beckmann, 2014; Suto, 2013). Problem solving is essential for navigating today's complex and rapidly changing societies (OECD, 2013), it is a prerequisite for lifelong learning (European Commission, 2019), and is highly valued in STEM (science, technology, engineering, and mathematics) careers (Jonassen et al., 2006; Tytler, 2020).

Learning to solve problems is an important goal of education (Jonassen, 2010b; Schoenfeld, 2016; van Merriënboer, 2013). Problem-solving competencies take years, even decades, to develop (Csapó & Funke, 2017a; Molnár et al., 2013) and the preschool years constitute a critical period for laying their foundations (Burns et al., 2025; Demir & Öztürk, 2024; Malcok & Ceylan, 2022; Rogers, 2004; Wheatley & Wheatley, 1984). Preschool teachers can support the development of problem-solving competencies through targeted pedagogical activities (Diamond, 2018; Erol, 2024; O'Reilly et al., 2022).

Young children are confronted with numerous problems every day. Due to their limited knowledge base and life experience, they cannot yet rely on readily available schemas or strategies to solve them. Young children compensate for their limited knowledge and lack of experience through rapidly, albeit non-linear and irregular (Mascolo & Fischer, 2015), developing abilities such as curiosity (Engel, 2011; Jirout et al., 2024), flexibility (Bobrowicz & Thibaut, 2023), creativity (Evans, Schlesinger, et al., 2021), and a strong drive for exploration (Evans, Todaro, et al., 2021) and experimentation (Jirout & Zimmerman, 2015).

Problem-solving competencies do not develop by learning about it, but through engaging in problem-solving activities (Jonassen, 2010a). By solving problems, young children learn about phenomena, apply knowledge and skills in an integrated way, explore features and functions of materials and tools, and develop general abilities such as collaboration, persistence, and planning. However, children do not automatically become better problem solvers simply by working on problems independently. They require appropriate adult support (Vygotsky, 2017). For this support to be effective, adults, and more specifically teachers, need to understand how to support the development of problem-solving competencies in young children and what support is needed for every child (Diamond, 2018). Teachers therefore need ways to make problem solving visible in the preschool classroom and to identify children's strengths and instructional needs to tailor their instruction.

What's the goal?

With this dissertation, we aimed to investigate how preschool children's problem-solving competencies can be assessed. More specifically, we sought to examine the validity, reliability, feasibility, and usability of formative problem-solving assessment in the preschool classroom. We start from the premise that scientifically substantiated assessments can support teachers in gaining insight into the problem-solving competencies of children, identifying children's strengths and instructional needs, adjusting instructional practice accordingly, and evaluating the effects of instructional interventions.

To realise this aim, this dissertation addresses the following four key questions:

1. What is early problem solving?
2. How can preschoolers' problem-solving competencies be assessed?
3. What observable behaviours do preschoolers demonstrate when solving problems?
4. What are the critical factors that enable or hinder valid, reliable, feasible, and usable formative assessment of problem-solving competencies in preschoolers?

What do we know?

This research project is built upon three cornerstones: problem solving, STEM, and formative assessment.

Problem solving

The first cornerstone is problem solving. Problem solving is often defined as "cognitive processing directed at transforming a given situation into a goal situation when no obvious method of solution is available to the problem solver" (Mayer, 1994, p. 284). Problem solving should not be regarded as a single process but rather as "the amalgam of many different cognitive processes that are orchestrated to achieve a certain goal that could not be reached, at least obviously, by simply applying a well-known procedure, process, routine or algorithm from a single subject area" (OECD, 2004, p. 159). Both definitions place strong emphasis on cognitive processes. However, problem solving also draws on metacognitive, affective, conative (Jonassen, 2000), social (Vygotsky, 2017), and motor (Keen, 2011) skills and dispositions.

Problem solving is not a unitary construct. Without aiming to be exhaustive, the scientific literature distinguishes between simple and complex problem solving (Funke & Frensch, 2007), well-structured and ill-structured problem solving (Jonassen, 1997; Simon, 1978), and domain-specific and domain-general problem solving (Greiff et al., 2014; Molnár et al., 2013). In addition, problem solving is conceptualised and studied from various perspectives, such as mathematical thinking (Ginsburg, 1983), computational thinking

(Wing, 2006), engineering thinking (Lippard et al., 2019), design thinking (Yalçın & Erden, 2021), creative thinking (Guilford, 1967; OECD, 2019), and critical thinking (O'Reilly et al., 2022).

Numerous problem-solving models can be found in scientific literature (e.g., Bransford & Stein, 1993; Dewey, 1910; Jonassen, 1997; Schoenfeld, 1985). These models are commonly phase models that describe or prescribe the different stages of the problem-solving process (Rott et al., 2021; Salminen-Saari et al., 2021). This dissertation draws on the PISA problem-solving framework. Although the PISA framework was originally developed to assess the problem-solving competencies of 15-year-olds (OECD, 2004, 2013), it is highly suitable as a framework for the formative assessment of these competencies in preschoolers. The PISA model offers several advantages over other widely used models, such as those proposed by Polya (1945), Newell and Simon (1972), Schoenfeld (1985), and Jonassen (1997). First, unlike existing normative problem-solving models that prescribe idealised problem-solving processes or descriptive models that represent actual problem-solving processes (Rott et al., 2021), the PISA model was specifically developed as an assessment framework to support the measurement of problem-solving competencies (OECD, 2013). Second, the PISA framework divides the problem-solving process into eight steps: exploring, understanding, representing, formulating, planning, executing, monitoring, and reflecting. This makes it more fine-grained than most other models (e.g., Polya, 1945; Schoenfeld, 1985) and, therefore, more informative for teachers seeking to identify preschoolers' strengths and support needs. Third, the PISA framework focuses on cross-disciplinary, real-life problems that do not require extensive domain-specific knowledge or skills (OECD, 2013), which young children may not yet have fully developed (Schäfer et al., 2024). Fourth, the PISA framework emphasises exploratory and creative processes that are central to young children's successful problem solving (Evans, Todaro, et al., 2021), but are often not explicitly addressed in models developed for older students or adults (e.g., Mayer, 1994; Newell & Simon, 1972; Polya, 1945). Finally, in contrast to models developed with a focus on well-structured problems (e.g., Polya, 1945; Schoenfeld, 1985), the PISA framework can be used for authentic, ill-structured problems, thereby enabling the observation of problem-solving behaviours in play-based settings, aligning with preschoolers' everyday experiences (Jonassen, 2010a; Moore et al., 2020).

STEM

Problems are encountered in nearly all aspects of human life. In this dissertation, the focus is on problem solving within STEM education, the second cornerstone of this research project. STEM is an acronym for science, technology, engineering, and mathematics, and is used to refer to educational activities within one discipline or across all four disciplines (English, 2016). Problem solving is indispensable for all four STEM disciplines (Dubosarsky et al., 2018; Priemer et al., 2020). Moreover, integrated STEM problems are meaningful, authentic, hands-on problems that provide children with opportunities to learn how to deal with problems and how to solve them (Aldemir & Kermani, 2017; MacDonald et al., 2022).

More specifically, this dissertation places its emphasis on engineering, the E in STEM. Engineering or engineering design is the 'interdisciplinary glue' between the STEM disciplines as it draws on the application of knowledge and skills from science, technology, engineering, and mathematics (Moore et al., 2018). Engineering design is a problem solving process directed at "making 'things' that work and making 'things' work better" (Lucas et al., 2014, p. 2) under certain constraints (Grubbs & Strimel, 2015; Lippard et al., 2019). Engineering design problems are authentic and open-ended (Purzer & Douglas, 2018) and closely resemble children's everyday building, construction, and crafts activities (Gold et al., 2021; Heroman, 2017; Shechter et al., 2021). This makes engineering design problems easy to implement in the daily practice of preschool classrooms and highly suitable for investigating preschoolers' problem-solving competences in a natural context.

Formative assessment

Formative assessment or assessment for learning is the third cornerstone of this dissertation. Formative assessment can be defined as any attempt to collect, interpret and use evidence on children's competencies in a certain domain to improve learning and instruction (Black & Wiliam, 2009; Hopster-den Otter et al., 2019). In that sense, formative assessment focuses on determining where learners are in their learning, where they are heading, and what actions are needed to support their progress (Black & Wiliam, 2009). Unlike summative assessment or assessment of learning, formative assessment is not intended to take high-stakes decisions. Formative assessment is better understood as part of the learning process itself rather than an evaluation of its outcomes. It is important to note that formative assessment is not primarily about collecting data on student performance; rather, it concerns how teachers use these data to inform their instructional decisions (Schildkamp et al., 2020).

Formative assessment in preschool can take many forms. It can range from unplanned and informal to planned and curriculum embedded (Hopster-den Otter et al., 2019). Formative assessment may be conducted individually or in groups, within the regular classroom or in another setting, and either by the classroom teacher or by a support teacher. It can be based on drawings, children's work samples, or systematic observations. Despite this diversity, there is broad consensus on several core principles. High-quality formative assessment in early childhood education should be child-centred, developmentally appropriate, embedded in everyday practice, dynamic, ongoing, and multifaceted (Brassard & Boehm, 2007; Deluca et al., 2019).

Central to all forms of assessment are the concepts of reliability and validity. Reliability is generally conceptualised as the extent to which test results are consistent across measurement occasions (AERA et al., 2014; Lee & Harris, 2025; Popham, 2017) or as the extent to which a measurement is free from random measurement error (AERA et al., 2014;

Mokkink et al., 2016). However, this conceptualisation is developed with a focus on standardised tests and is not well suited to formative assessment (Dochy, 2009; Parkes, 2013). Formative assessment does not aim to generalise, but rather to provide teachers with timely information that they can use to adapt their instruction (Wiliam, 2011). Consequently, reliability in terms of consistency is a less primary concern in formative assessment than in high stakes summative assessment (Shepard, 2001; Smith, 2003). Therefore, reliability in formative assessment is sometimes viewed as minimising errors during the assessment process (Parkes, 2013) or as sufficiency of information, indicating that the teacher has sufficient information to make a reasonable instructional decision (Smith, 2003).

Validity is defined as the extent to which the interpretations and uses of test scores are supported by evidence (AERA et al., 2014; Kane & Wools, 2019). In the context of formative assessment, validity can be understood as the extent to which assessment outcomes positively support children's learning (Stobart, 2012). The argument-based approach to validation provides a unified framework that integrates various types of validity and evidence and enables the systematic evaluation of claims based on assessment outcomes (Kane, 2006; Messick, 1989). As such, it provides an appropriate framework for examining the inferences and assumptions teachers make when interpreting and using assessment results to inform instructional decisions (Kane, 2013a).

How to solve it?

Four studies

To examine and determine the validity, reliability, feasibility, and usability of formative problem-solving assessment in the preschool classroom, this dissertation reports on four studies. The four studies form a coherent whole and build upon one another, progressively narrowing the focus along three dimensions. A first dimension concerns the target population. The educational level and age group narrows from early childhood (3–8 years; Study 1) to preschool children (3–6 years; Study 2), to the oldest preschoolers preparing for the transition to primary school (5–6 years; Study 3 and Study 4). A second dimension concerns the focus of problem solving. Here, the focus shifts from early problem-solving in general (Study 1), to ill-structured problems in STEM (Study 2), to engineering design problems (Study 3 and Study 4). A third dimension concerns the type of the assessment. Study 1 does not exclusively focus on assessment, Study 2 includes different types of assessments, Study 3 lays the foundations for observation-based assessment, and Study 4 addresses observation-based formative assessment.

Study 1 – Mapping the field of early problem-solving research

Study 1 aims to provide conceptual clarity regarding the construct of problem solving in early childhood education (aged 3–8 years). Problem solving is a broad construct used across multiple research domains. Therefore, we mapped the different research fields in which early problem solving is investigated through a bibliometric review. Across these research fields, we identified distinct types and dimensions of early problem solving. The study resulted in a framework that can be used by researchers to delineate the construct of early problem solving for their specific research context and to position their works within the broader scholarly literature.

Study 2 – Measuring problem solving competencies in preschoolers

The second research question addressed in this dissertation concerns how problem solving can be assessed in preschool (aged 3–6 years). Through a scoping review, we identified methods used in the scientific literature to assess problem solving competencies in the context of ill-structured STEM problems. Analyses of the publications and instruments provided insights into the characteristics, scope, and psychometric properties of these methods. The findings of this study guided the development of the problems and protocols for the subsequent studies.

Study 3 – Developing an observation scheme

The third study aimed to identify observable behaviours associated with the different steps of the problem-solving process. To this end, 12 preschool children (aged 5–6 years) were recorded as they individually solved three engineering design problems. Preschool teachers and fellow researchers were invited to review the videos and identify and describe all behaviours they associated with problem solving. Coding these behaviours resulted in a structured observation scheme that enables documentation, examination, and evaluation of preschoolers' problem-solving behaviours.

Study 4 – Developing a validity argument for formative preschool assessment of problem solving

Drawing on the observation scheme resulting from Study 3, we designed an instrument for formative assessment of preschoolers' problem-solving competencies. Teachers were invited to implement the instrument in their classrooms. Based on the quantitative and qualitative observation data of 53 children (aged 5–6 years) and teachers' feedback, a validity argument was constructed to identify the critical aspects concerning the validity, reliability, feasibility, and usability of the assessment and the interpretations and uses of the results.

Structure of the dissertation

The four studies outlined above form the core of this dissertation. The first four chapters each report on one of these studies. As these chapters have been written as scientific articles, they can be read as self-contained works, which makes some overlap inevitable. The final chapter summarises the main findings and discusses their implications for research and practice.

STUDY 1 | MAPPING THE LANDSCAPE OF RESEARCH ON EARLY PROBLEM SOLVING: A BIBLIOMETRIC REVIEW

This chapter is currently under review as:

Van Elsen, J., Verschaffel, L., De Maeyer, M., & Torbeyns, J. (2026). Mapping the landscape of research on early problem solving: A bibliometric review [Manuscript submitted for publication].

Abstract

Background: Problem solving is an important, transversal 21st century skill with early childhood providing a critical window to support the development of problem-solving competencies. However, most theoretical and empirical studies on problem solving focus on older students.

Aims: With this bibliometric review, we aimed to provide a comprehensive overview of research on early problem solving across disciplines by mapping the research fields in which it is studied, examining sample diversity, and identifying the types and dimensions of early problem solving.

Methods: A systematic search of Scopus, Web of Science, ERIC, PubMed, and Sociological Abstracts resulted in 1480 publications on problem solving in children aged 3–8 years, enrolled in pre-primary education programmes through Grade 3, published between 1928 and 2025. Included publications were classified using the OpenAlex classification system.

Results: Early problem solving is examined across multiple research fields, with over half of the publications concentrated in Psychology and Mathematics. Research attention systematically increases with age. Approximately one-fifth of the publications focus on identifying or supporting early problem solving in diverse populations. We identified 168 early problem-solving types that could be categorised into four clusters: mathematical, social, creative, and collaborative problem solving, and into five dimensions: problem type, problem topic, problem-solving skills, problem-solving processes, and problem-solving context.

Conclusions: This bibliometric review highlights the diversity of research on early problem solving and provides a comprehensive map that enables researchers to identify and position this work within the broader scholarly literature.

Introduction

Way back in the 1940s, Karl Duncker (1945) stated that “a problem arises when a living creature has a goal but does not know how this goal is to be reached” (p. 1) with problem solving being the realisation of that goal. Around the same time, William Brownell (1942) noted that problem solving can be so broadly defined that it is essentially synonymous with learning. These broad interpretations of problem solving as any goal directed activity or learning align with Karl Popper that “all life is problem solving” (Popper, 1999, p. 99). Consequently, it is not surprising that problem solving is examined across research fields resulting in a rich yet fragmented research landscape. This fragmentation makes it challenging for researchers to navigate the scientific literature on problem solving, position their own work within the scientific literature, and communicate their findings to a broad and diverse audience. It also limits cross-fertilisation within and across research fields and hampers the accumulation of insights into the nature, characteristics, and development of problem solving.

The substantial body of theoretical and empirical research reflects the recognised importance of developing problem-solving competencies. Numerous highly influential publications like those from Polya (1945), Newell and Simon (1972), Schoenfeld (1985), and Jonassen (2010a) are written with high school students, college students, or adults in mind. For early childhood education, the literature on problem solving is considerably more limited. However, early childhood represents a critical window for developing problem-solving competencies (Diamond, 2018; Rogers, 2004). A growing body of research demonstrates that young children are capable problem solvers and that early childhood represents a critical period for cultivating problem-solving competencies (Bobrowicz & Thibaut, 2023; Diamond, 2018; Keen, 2011; O'Reilly et al., 2022; Rogers, 2004).

Evidence from studies involving older students, as well as theories and models developed for them, cannot be readily generalised to young children. Compared to older learners, young children’s limited knowledge base constrains their reliance on domain-specific knowledge, skills, or strategies. In addition, young children need their hands to think (Ata-Aktürk, 2023) and their developing executive functions and self-regulation hinder their capacity to engage in planning before acting (Otero & Barker, 2014). The problems young children encounter in play and daily life also differ significantly from the well-structured textbook problems that dominate classes in high school and college (Jonassen, 2010a). Finally, the broader physical and social context in which young children solve problems plays a significant role (Wood et al., 1976). Generalising theoretical constructs and empirical findings about older learners to young children risks promoting a deficit perspective that emphasises what they cannot yet do, while neglecting their emerging strengths. For example, young children might compensate their lack of domain-specific knowledge and skills through general problem-solving strategies (English, 1992; Schäfer et al., 2024). Rather than relying on established routines, procedures, or algorithms, they might adopt a more exploratory and creative approach (Evans, Todaro, et al., 2021). Instead of planning in advance, they might monitor the effects of their spontaneous actions (Study 3; English, 1992). Finally, children might be more responsive to stimuli from their environment to adjust their problem-solving process (Study 3).

Given that problem solving manifests differently in young children, a comprehensive overview of the existing scientific literature on problem solving in early childhood is warranted. The current study aims to map the research landscape of problem solving in early childhood education. In this context, the term *early problem solving* is used to refer to problem solving in children aged 3 to 8 years enrolled in pre-primary education programmes as classified under ISCED 02 (UNESCO Institute for Statistics, 2012) through Grade 3.

Mapping the field of research on early problem solving

To provide an overview of research on early problem solving, we employ science mapping through bibliometric analysis. Science mapping refers to the classification and visualisation of publications on a particular topic, research field, or domain (Boyack & Klavans, 2014). Science mapping can be achieved through bibliometric network analysis, which maps relationships between publications based on bibliographic coupling, co-citation, co-authorship, and keyword co-occurrence (Donthu et al., 2021; van Raan, 2019). The purpose of science mapping and bibliometric analyses is not to synthesise empirical evidence, but rather to identify the structure, features, and trends in scientific output related to the phenomenon under investigation (Hallinger, 2023; Jing et al., 2024).

Bibliometric science mapping can be approached in various ways. First, a distinction can be made between journal-level and document-level classifications (Shu et al., 2019). At the journal level, metadata from journals are used, assuming that all publications within a journal belong to the same discipline. At the document level, metadata from individual papers are analysed for classification. Web of Science and Scopus classify publications at the journal level. However, document-level classification is more accurate than journal-level classification (van Raan, 2019). Shu et al. (2019) estimated that 51.1% of papers were published in journals outside their primary discipline, leading to substantial misclassification when research is mapped using journal-level systems.

Second, bibliometric science mapping can be conducted through local mapping and global mapping (Klavans & Boyack, 2011). In local mapping, a selected set of relevant publications is organised into clusters based on their relationships. In global mapping, the global structure of scientific output is mapped first, and the target literature is subsequently embedded within this structure. According to Klavans and Boyack (2011), global mapping ensures higher coherence among documents within clusters compared to local mapping. It also makes smaller clusters or isolated publications within a selected subsample of publications more visible. However, global mapping requires a map of the global structure of science. Such a global map was created by Van Eck and Waltman (2024) based on a snapshot of OpenAlex (Priem et al., 2022), an openly accessible index of scholarly works. The classification system, adapted by OpenAlex, is based on the Scopus All Science Journal Classification (ASJC) system and consists of 4516 topics, 252 subfields, 26 fields, and 4 domains (OpenAlex, 2024b). Large language models were used to create labels, write a summary, and assign ten keywords to every topic, based on the 250 most cited publications in the cluster of publications (Van Eck & Waltman, 2024). OpenAlex assigns topics and corresponding subfields, fields, and domains to new records using a comprehensive deep learning model that calculates a probability score for each of the 4516 topics based on the title, abstract, journal name, and links with 'gold citations'; the highest cited works for each topic (OpenAlex, 2024b).

The following example illustrates the OpenAlex classification system. The article "Interpersonal problem solving as a mediator of behavioral adjustment in preschool and kindergarten children" by Shure and Spivack (1980), published in the *Journal of Applied Developmental Psychology*, was classified under the topic *Child and Adolescent Psychosocial and Emotional Development* within the subfield *Clinical Psychology*, in the field *Psychology*, and in the domain *Social Sciences*. In the OpenAlex classification system, the named topic is summarised as follows:

> This cluster of papers explores the intricate relationship between emotion regulation, psychopathology, and developmental trajectories in children and adolescents. It delves into the impact of various factors such as parenting, anxiety disorders, depression, and behavioral problems on the mental health and well-being of young individuals. (OpenAlex, 2024a)

The ten keywords assigned to this topic are: Emotion Regulation; Psychopathology; Children; Adolescents; Developmental Trajectories; Mental Health; Parenting; Anxiety Disorders; Depression; Behavioral Problems (OpenAlex, 2024a).

Recent bibliometric reviews have attempted to map specific parts of the scientific corpus on problem solving, covering mathematical problem solving (Hendriani et al., 2025; Suseelan et al., 2022), creative problem solving (Sucilestari et al., 2025), collaborative problem solving (Cheruiyot & Molnár, 2026), complex problem solving in engineering (Azizan et al., 2025), and problem-based learning (Yuhana & Fajari, 2025). Typically, these reviews focus on one specific type of problem solving, extract records from one single database (Scopus or Web of Science), and analyse document-level networks of keywords, authors, and citations with VOSviewer (local mapping). None of these studies addresses problem solving in early childhood education.

The present study

The present bibliometric review aims to systematically map the scientific publications on early problem solving across disciplines. Three research questions are addressed in this study:

RQ 1: What are the research fields and topics in which early problem solving is studied?

RQ 2: What are the target populations of the publications on early problem solving within research fields?

RQ 3: What types and dimensions of early problem solving can be identified across research fields?

By addressing the first research question, we aim to establish the current state of research on early problem solving and identify distinct lines of inquiry across research fields. The resulting overview will help researchers to navigate the literature on early problem solving within and beyond their own disciplines.

The second research question addresses the demographic focus of early problem-solving research. Mapping target populations across research fields clarifies how research attention is distributed between preschool and primary school children and across typically developing, neurodiverse, linguistically diverse, and socially diverse groups. This provides insight into the educational, developmental, and societal concerns that drive research efforts on early problem solving in different fields, facilitating the identification of populations that receive relatively more or less attention and highlighting gaps to inform directions for future research.

The third research question intends to deepen our understanding of the diversity of perspectives on early problem solving by identifying different types of problem solving and the dimensions along which they vary. The resulting framework can support researchers in clearly delineating their work and positioning it more precisely within the broader research literature.

Method

Design

To map the research on early problem solving, document-level global mapping based on the OpenAlex classification system was employed. The pre-registered protocol, data, and scripts are provided through Open Science Framework [https://osf.io/3pqde/overview?view_only=c605b50baf6547a4a36984ebf455da25]. Text mining, quantitative analyses, and visualisations were conducted in R (Version 4.4.2, R Core Team, 2023) and VOSviewer (Version 1.6.20, van Eck & Waltman, 2010).

Search strategy

To ensure broad coverage across research fields, records were retrieved from five databases: Scopus, Web of Science, ERIC, PubMed, and Sociological Abstracts. The main search was conducted on June 25, 2025. We specifically searched for publications in which the authors explicitly identified problem solving as a focus by mentioning the term “problem solving” in the title or keywords. The search was restricted to publications that referenced children aged 3 to 8 years enrolled in pre-primary education programmes through Grade 3 as the target population in the title, keywords, or abstract, and to publications written in English. The following search string was applied (Scopus syntax):

> [Scopus] (TITLE ("Problem-solving" OR "problem solving") OR AUTHKEY ("Problem-solving" OR "problem solving")) AND (TITLE-ABS-KEY (child* OR boy OR boys OR girl* OR young* OR preteen* OR preschool* OR pre-school* OR kindergarten* OR prekindergarten OR pre-kindergarten OR k-12* OR k12 OR "Day care" OR daycare OR nursery OR "Primary education" OR "Elementary education" OR "Elementary school" OR "Grade 1" OR "1st grade*" OR "First grade*" OR "Grade 2" OR "2nd grade*" OR "Second grade*" OR "Grade 3" OR "3rd grade*" OR "Third grade*")) AND (LIMIT-TO (LANGUAGE , "English"))

Screening and selection

Eligibility criteria

All publications including or addressing problem solving in children aged 3 to 8 years or enrolled in pre-primary education programmes through Grade 3 were considered. No restrictions were imposed on the publication type, study design, or publication date. publications were only excluded if they were not written in English or did not explicitly address the learning, instruction, teaching, development, measurement, or assessment of problem solving in early childhood.

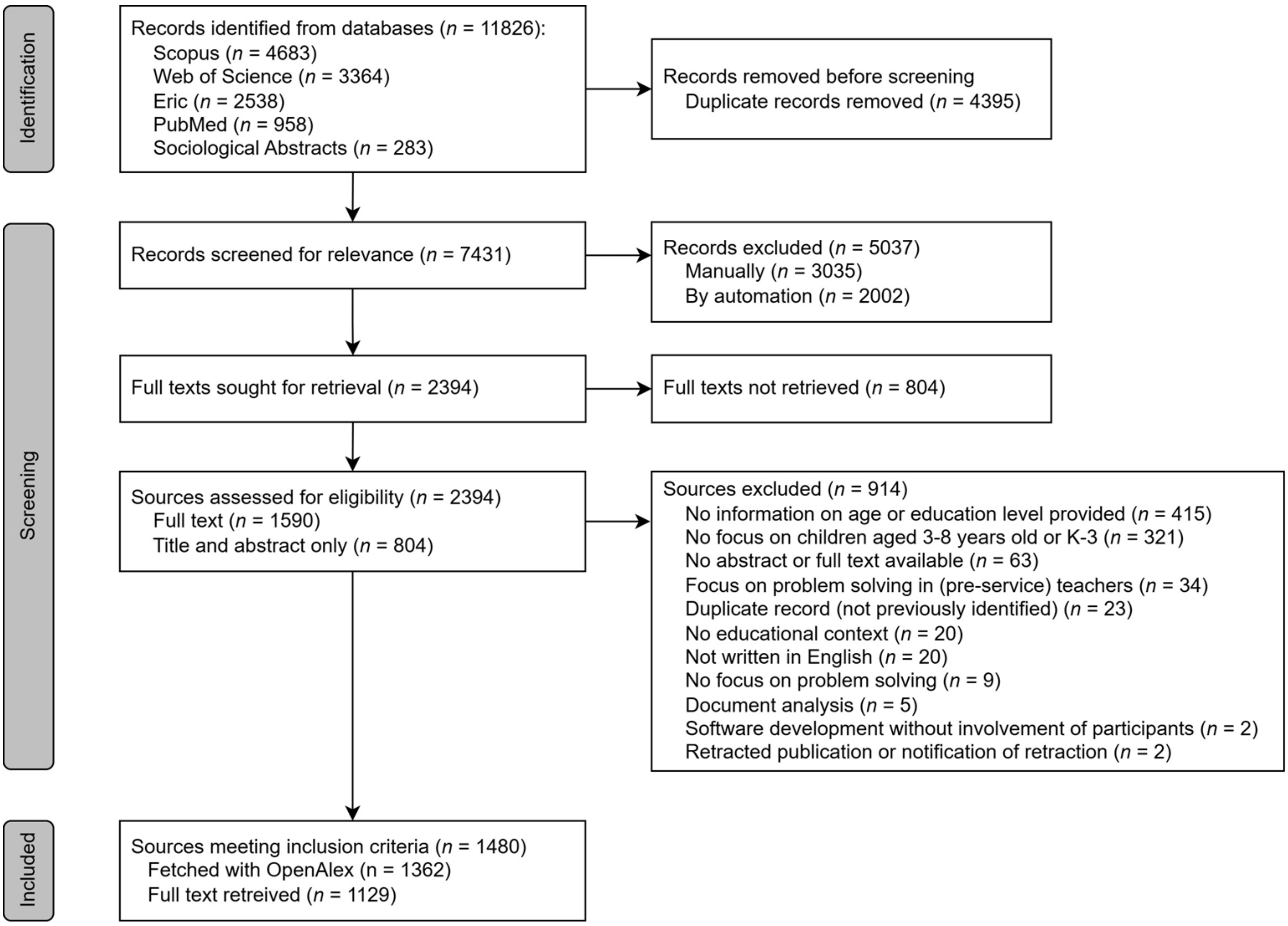
Screening title and abstract

Titles and abstracts were screened for eligibility using ASReview (Version 2.1), a researcher-in-the-loop active-learning application that helps prioritising relevant publications (van de Schoot et al., 2021). The SAFE procedure proposed by Boetje and van de Schoot (2024) was applied to minimise the risk of missing relevant papers. This procedure provides guidelines for active-learning assisted systematic reviews, aligns with the PRISMA statement (Page et al., 2021), and complies to the Open Science principles (Lombaers et al., 2024). The ELAS U4 and ELAS h3 models were used for screening (for more information on the models, see https://asreview.readthedocs.io/en/stable/lab/models.html).

Screening full text

Full texts of eligible records were retrieved through EndNote, doi-link, database-link, ResearchGate, and web search. All accessible publications were screened for eligibility by the first author. When full texts could not be retrieved, publications were still included if their title, abstract, or keywords contained sufficient information to assess their eligibility concerning the focus on problem solving as a concept of interest and a target population including early childhood. Excluded publications were coded for the reason of exclusion. In total, 1480 records meeting eligibility criteria were identified of which 1362 could be matched in the OpenAlex database for classification (Figure 1).

Figure 1. PRISMA flow diagram of the search and selection process (Page et al., 2021)



Data charting and data items

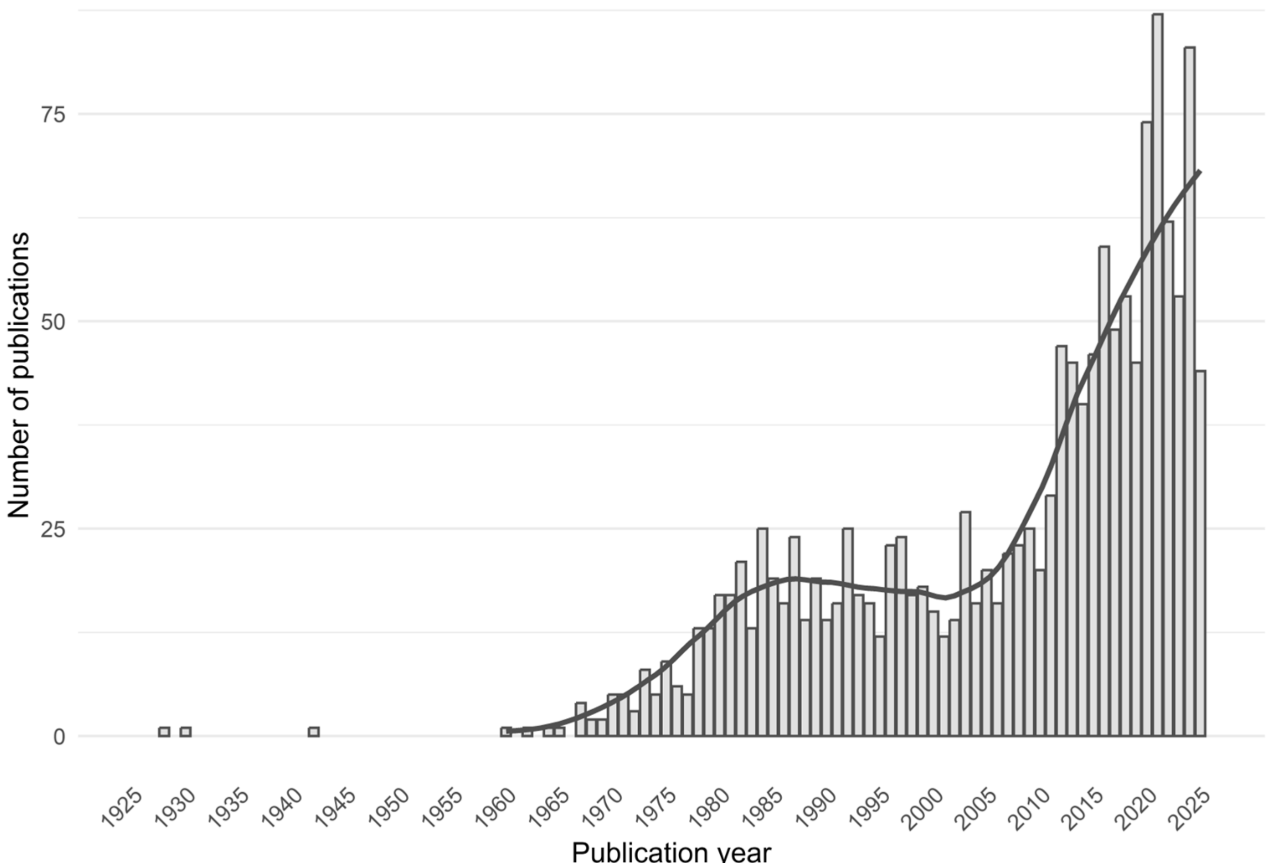
A spreadsheet was created to collate the metadata of the records provided by the databases. The R package openalexR (Version 2.0.2, Aria et al., 2024) was used to automatically add OpenAlex metadata and topic classifications to the records of the included publications, including OpenAlex topics, subfields, fields, and domains (Priem et al., 2022). All records were manually checked and mismatches were corrected. Data on the specific age, education level, and characteristics of the target population as well as the country of the first author's institution were charted manually from the title, abstracts, keywords, and full texts, and added to the spreadsheet.

Results

Characteristics of the publications

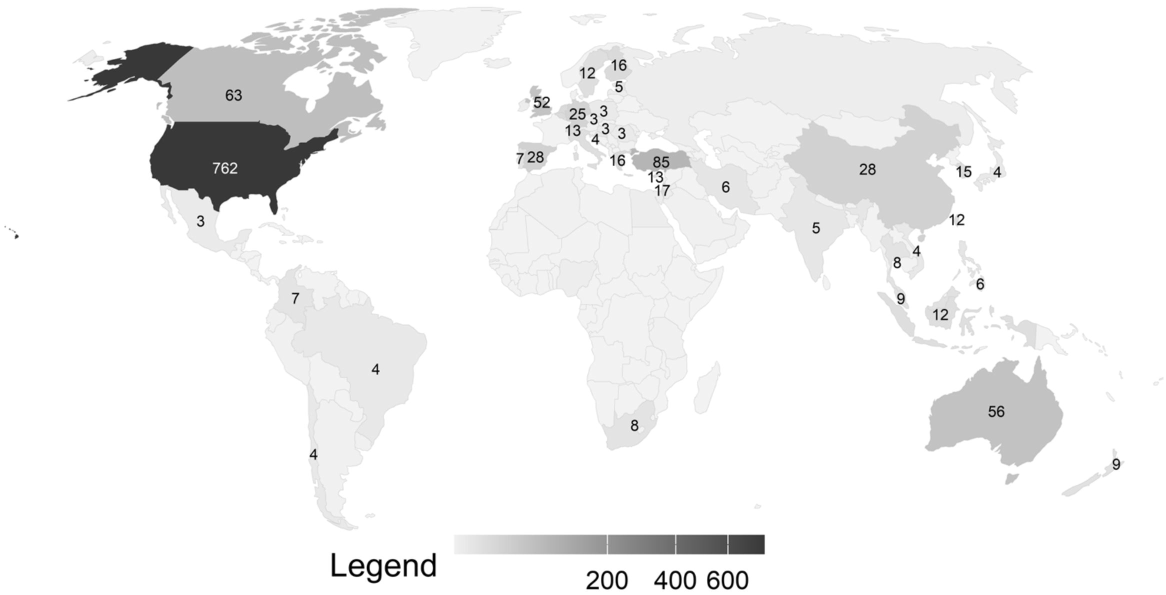
The corpus of publications on early problem solving comprises scholarly journal articles (n = 1098), conference papers (n = 126), research reports (n = 111), theses and dissertations (n = 45), general documents (not specified; n = 45), book chapters (n = 32), and books (n = 21), and magazine articles for a broader audience (n = 2). The included publications date back to 1928 (see Figure 2). However, research on early problem solving only began to develop in the late 1960s and 1970s. A significant surge occurred around the turn of the century, with an average annual growth rate of 7% since 2000.

Figure 2. Number of publications on early problem solving by year from 1928 to 2025 (N = 1480)



Geographically, research on early problem solving is largely concentrated in the Anglosphere (USA, Canada, Australia, UK), with over half of the publications (n = 762) originating from the United States (see Figure 3). Turkey is the only non-English-speaking country among the top 5 (n = 85). Africa and South America are only scarcely represented in the corpus.

Figure 3. Geographical distribution of publications on early problem solving (N = 1453)



Note. Numbers are based on the affiliation of the first author.

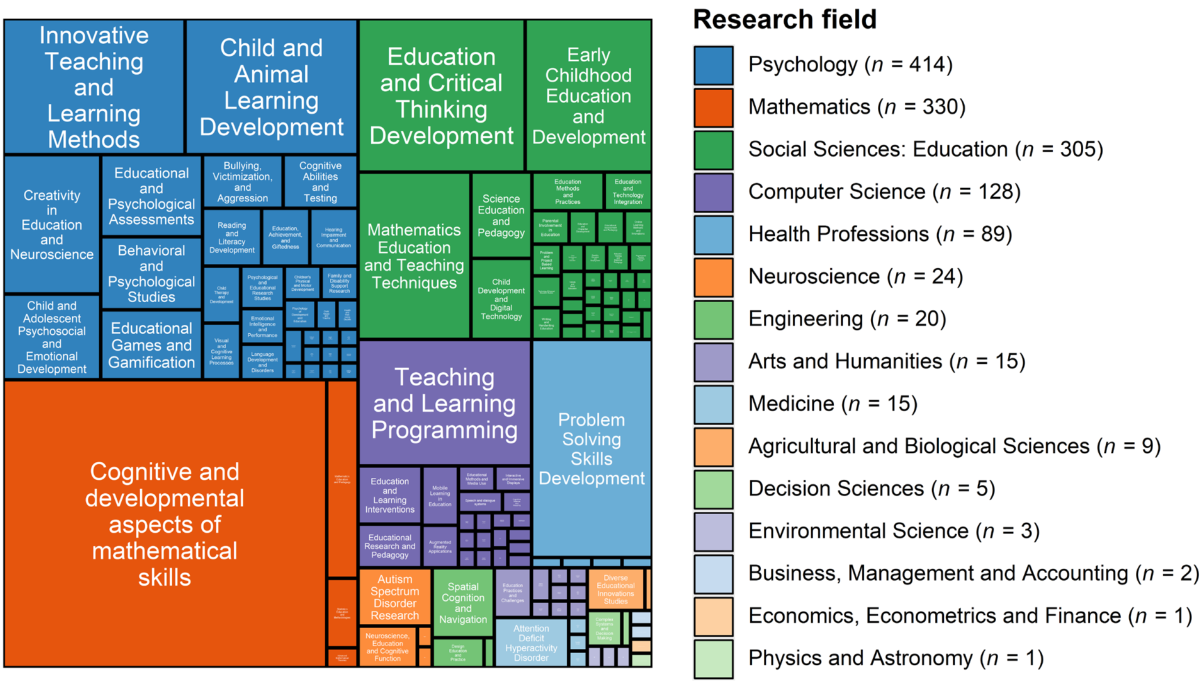
Research fields and topics in which early problem solving is studied

To identify the research fields in which early problem solving is studied, the records from the database searches were fetched with the OpenAlex database. The highest-ranked topic and its corresponding subfield, field, and domain were used to locate each record within the hierarchical OpenAlex classification system. Assigning topics to the records was possible for 1361 of the 1480 selected records (92.0%). One publication was not assigned to a topic in the OpenAlex database. The remaining 118 records, primarily comprising grey literature such as conference papers, reports, theses, and older works published before 2000, were not included in the OpenAlex database and therefore could not be classified.

The publications in the corpus were classified in 4 domains, 15 fields, 51 subfields, and 132 topics. Figure 4 presents a hierarchically structured treemap showing all topics within their respective research fields. For clarity, the subfield and domain levels were omitted. Rectangle size represents the proportion of publications assigned to each topic or research field relative to the entire corpus. A complete overview of topics and research fields with frequencies is available on Open Science Framework [https://osf.io/3pqde/overview?view_only=c605b50baf6547a4a36984ebf455da25].

The research field with the highest number of publications on early problem solving is Psychology (n = 414), followed by Mathematics (n = 330), Social Sciences: Education (n = 305), Computer Science (n = 128), and Health Professions (n = 89). Psychology and Social Sciences: Education fall under the domain of Social Sciences. Mathematics and Computer Science fall under the domain of Physical Sciences, and Health Professions falls under the domain of Health Sciences. In the following sections, we focus on the five main research fields, accounting for 93.0% of the publications that were classified (85.5% of the total corpus). The smaller fields and publications that could not be classified are addressed together.

Figure 4. Hierarchical treemap illustrating the distribution of OpenAlex topics in publications on early problem-solving within research fields ($N = 1361$)



Note. Rectangle size reflects the proportion of publications assigned to the topic and field relative to the total number of publications in the corpus.

To provide a clear understanding of research on early problem solving within the research fields, we describe the main topics under which the included publications are classified. As noted earlier, OpenAlex provides a summary and ten keywords for each topic (Van Eck & Waltman, 2024). In this section, we present a concise overview of the primary topics addressed in the publications within each research field. For clarity, we indicate the subfields in which these topics are situated for fields with a large number of topics (not included in Figure 4). We also present five representative topic-level keywords selected from the main topics in each field.

Psychology

Psychology constitutes a broad and multifaceted research field comprising 33 topics, which can be further organised into six subfields: *Developmental and Educational Psychology* ($n = 263$), *Experimental and Cognitive Psychology* ($n = 80$), *Clinical Psychology* ($n = 44$), *Social Psychology* ($n = 23$), *Applied Psychology* ($n = 3$), and *Neuropsychology and Physiological Psychology* ($n = 1$). The main topics addressed within Psychology are

Innovative Teaching and Learning Methods (*n* = 80), *Child and Animal Learning Development* (*n* = 76), and *Creativity in Education and Neuroscience* (*n* = 44). Five representative keywords for research on early problem solving within these topics are: Metacognition, Motivation, Executive Functions, Creativity, and Collaborative Learning.

Mathematics

Mathematics includes publications classified in four topics with the topic *Cognitive and developmental aspects of mathematical skills* accounting for 91% (*n* = 302) of all publications in the field. The other topics are *Mathematics Education and Pedagogy* (*n* = 20), *Statistics Education and Methodologies* (*n* = 7), and *Advanced Mathematical Theories* (*n* = 2). Representative topic-level keywords for Mathematics are: Early Numeracy, Executive Functioning, Developmental Dyscalculia, Spatial Representation, Cognitive Development,

Social Sciences: Education

In the OpenAlex classification, the same label Social Sciences is used for the field and for the domain. Given that 96% of the publications assigned to the field of Social Sciences are classified under the subfield Education and to distinguish the field from the domain, we refer to the research field as Social Sciences: Education. Publications classified under Social Sciences: Education primarily address topics such as *Education and Critical Thinking Development* (*n* = 83), *Early Childhood Education and Development* (*n* = 63), *Mathematics Education and Teaching Techniques* (*n* = 61), *Science Education* (*n* = 17), and *Child Development and Digital Technology* (*n* = 16). Representative keywords for the field of Social Sciences: Education are: 21st Century Skills, Cognitive Psychology, Social and Emotional Learning, Mathematics Instruction, and STEM Education.

Computer Science

Computer Science includes publications classified under 23 topics. The majority (55%) of publications in this field are classified under the topic *Teaching and Learning Programming* (*n* = 71). *Education and Learning Interventions* (*n* = 12) is the second largest topic. All other topics contain fewer than 10 publications. Representative keywords for the topics in the field of Computer Science are: Computational Thinking, Programming Education, Educational Technology, Maker Movement, STEAM Education.

Health Professions

Health Professions consists almost exclusively (95.5%) of publications related to the topic *Problem Solving Skills Development* (*n* = 85). The topic summary provided by OpenAlex indicates that publications classified under this topic primarily concern solving and managing problems in social contexts. This is also evident from the keywords assigned to the topic, including: Social Skills, Stress Management, Coping Strategies, Mental Health, and Communication Skills. The four remaining topics each contain only one publication from the corpus.

Other research fields

Beyond these five main research fields, publications on early problem solving were also classified into several other fields, each represented by a smaller number of publications in the corpus. Neuroscience (n = 24) primarily focuses on cognitive neuroscience, including *Autism Spectrum Disorder Research* (n = 14) and Neuroscience, Education and Cognitive Function n = 8). Engineering (n = 20) emphasises Spatial Cognition and Navigation (n = 14). Arts and Humanities (n = 15) primarily address *Education Practices and Challenges* (n = 6) and Creative Drama in Education (n = 2). Medicine (n = 15) concentrates on *Attention Deficit Hyperactivity Disorder* (n = 12). Six additional fields were identified, each represented by fewer than 10 records (see Figure 4).

Target populations of research on early problem solving

In addition to the metadata provided by the databases and OpenAlex, we manually charted data on the target population's age and educational level from the title, abstract, keywords, or full text. Based on this information, publications were categorised as pre-primary (education levels categorised under ISCED 02 or ages 3–5), primary education (Grade 1–3, or ages 6–8), or both. If the education level or grade of the target population was specified in the title, abstract, keywords, or methodology section, this information was used. Otherwise, the publication was classified based on age. For 53 publications (3.6%), it was not possible to identify the specific target audience, apart from a general focus on early childhood. Table 1 summarises the number and percentage of publications for each group and research field.

About half of the publications in the corpus focus exclusively on children in primary education. In contrast, about one quarter of the publications address pre-primary education exclusively. Another 22.0% address both levels. The overall predominance of primary education is most pronounced in Mathematics (71.8%), with Grade 3 accounting for 46.7% of all mathematics-related publications. Health Professions is the only field where the focus is primarily on pre-primary education.

Table 1. Publications by education level and research field

	Psychology	Mathematics	Social Sciences: Education	Computer Science	Health Professions	Other / NA	Total
Pre-primary	125 (30.2%)	38 (11.5%)	104 (34.1%)	23 (18.0%)	39 (43.8%)	55 (25.7%)	384 (25.9%)
Primary	166 (40.1%)	237 (71.8%)	130 (42.6%)	62 (48.4%)	30 (33.7%)	93 (43.5%)	718 (48.5%)
Both	113 (27.3%)	50 (15.2%)	59 (19.3%)	34 (26.6%)	19 (21.3%)	50 (23.4%)	325 (22.0%)
Not specified	10 (2.4%)	5 (1.5%)	12 (3.9%)	9 (7.0%)	1 (1.1%)	16 (7.5%)	53 (3.6%)
Total	414	330	305	128	89	96	1480

Note. Percentages are relative to the total number of included publications per research field. Other / NA: Other field or not assigned to any field.

In addition to the educational level, we documented characteristics related to neurodiversity, linguistic diversity, and socioeconomic diversity of the target populations. Publications concerning children with disabilities, disorders, or learning difficulties were categorised as *Neurodiverse learners*. Publications concerning children's linguistic backgrounds, such as second-language learners or bilingual children, were categorised as *Language learners*. Publications concerning gifted, talented, or high-ability children were categorised as *Gifted learners*. Publications concerning children who are socially, economically, or culturally disadvantaged were categorised as *Disadvantaged learners*.

Approximately one in five publications (n = 274; 18.5%) address neurodiverse, linguistically diverse, or socioeconomically diverse populations. Two fifth of these publications is assigned to the field of Mathematics. Table 2 presents an overview of the number and percentages of publications addressing problem solving in neurodiverse, language, gifted, and disadvantaged learners. Of these groups, neurodiverse learners are most frequently studied, especially in the field of Mathematics. Neurodiverse learners are also represented in the fields of Psychology (14.7%) and Health Professions (10.1%). Language learners, gifted learners, and disadvantaged learners are studied in less than 3% of the publications.

Table 2. Publications by diversity in target population and research field

	Psychology	Mathematics	Social Sciences: Education	Computer Science	Health Professions	Other / NA	Total
Neurodiverse learners	61 (14.7%)	85 (25.8%)	9 (2.9%)	4 (3.1%)	9 (10.1%)	41 (19.2%)	208 (14.1%)
Language learners	5 (1.2%)	24 (7.3%)	2 (0.7%)	1 (0.8%)	0 (0%)	5 (2.3%)	37 (2.5%)
Gifted learners	11 (2.7%)	5 (1.5%)	4 (1.3%)	3 (2.3%)	2 (2.2%)	2 (0.9%)	27 (1.8%)
Disadvantaged learners	8 (1.9%)	1 (0.3%)	5 (1.6%)	1 (0.8%)	2 (2.2%)	0 (0%)	17 (1.1%)

Note. Percentages are relative to the total number of included publications per research field. Other / NA: Other field or not assigned to any field.

Types and dimensions of early problem solving

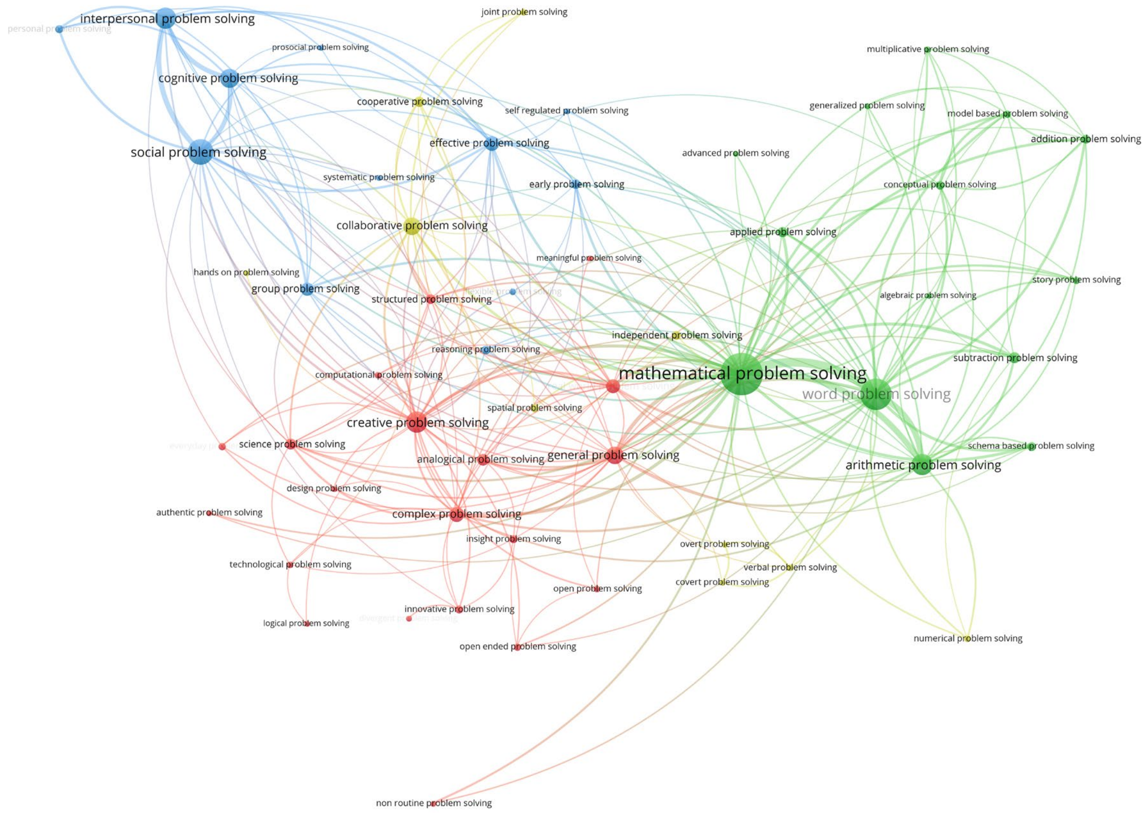
Types of early problem solving

Early problem solving is a multidimensional construct. By adding a modifier, typically an adjective preceding the term problem solving, authors delineate the type of problem solving under investigation. The types of early problem solving were analysed in several steps. During the screening of titles, abstracts, and full texts, all encountered modifiers were listed in a spreadsheet. Based on this overview, a thesaurus for equivalent terms was created. Subsequently, the modifiers were automatically extracted from the title, abstract, keywords, and full text using an R script and added to the metadata of the records. Frequency tables of the early problem-solving types were generated and co-occurrence networks were visualised with VOSviewer (van Eck & Waltman, 2010).

We identified 168 different early problem-solving types. Mathematical problem solving emerged as the predominant type (479 publications), followed by word problem solving (n = 265). These two types often co-occur or are combined under the label mathematical word problem solving (n = 228). Social (n = 180), interpersonal (n = 123), arithmetic (n = 127), creative (n = 123), cognitive (n = 100), collaborative (n = 87), general (n = 87), and complex problem solving (n = 62) complete the top 10.

Figure 5 depicts the network of co-occurrence of early problem-solving types. Four clusters can be identified in the visualisation, each named after its most prevalent type: mathematical problem solving, social problem solving, creative problem solving, and collaborative problem solving. Although social problem solving and collaborative problem solving may appear closely connected at first glance, they represent two distinct clusters. Social problem solving addresses the resolution of social issues (e.g., conflicts between children), whereas collaborative problem solving concerns forms of cooperation during the resolution of problems of any kind.

Figure 5. Co-occurrence network of types of early problem solving (VOSviewer)



Note. Only types that occur in more than 10 records are included. Font size and circle size reflect the number of publications in which the type occurs. Line thickness reflects the number of publications in which both terms occur. Colours indicate the cluster in which the type is classified. Green: mathematical problem solving; Blue: social problem solving; Red: creative problem solving; Yellow: collaborative problem solving.

Next, we examined which early problem-solving types were most frequently referenced in publications classified in the different research fields (Table 3). Mathematical problem solving is the most frequently referenced type of problem solving in all research fields, except in Health Sciences, where interpersonal problem solving and social problem solving dominate. Social problem solving holds the second position in Psychology, Social Sciences: Education, and Other Fields. Furthermore, creative problem solving and collaborative problem solving both feature among the top 5 in Psychology, Social Sciences: Education, and Computer Science.

Table 3 reveals the substantial overlap of types of early problem solving across the different research fields. The strong presence of mathematical problem solving in the field of Social Sciences: Education can be explained by the inclusion of the topic *Mathematics Education and Teaching Techniques* ($n = 61$). Psychology includes the topics *Educational and Psychological Assessments and Cognitive Abilities and Testing*. As a result, publications on children with (risk of) math disabilities might be classified under Mathematics (about developmental aspects of mathematical skills), Social Sciences: Education (about education), or Psychology (about testing).

The two research fields most clearly distinguishable based on the types of early problem solving are Mathematics and Health Professions with, respectively, a focus on mathematically oriented forms of problem solving (e.g., word problem solving, addition problem solving) and socially oriented forms (e.g., interpersonal problem solving). Computer Science distinguishes itself from the other fields through its more domain-specific focus on computational problem solving.

Table 3. Five most frequent early problem-solving types per research field

Psychology	Mathematics	Social Sciences: Education	Computer Science	Health Professions
Mathematical problem solving (91)	Mathematical problem solving (234)	Mathematical problem solving (74)	Mathematical problem solving (27)	Interpersonal problem solving (48)
Social problem solving (58)	Word problem solving (180)	Social problem solving (43)	Creative problem solving (22)	Social problem solving (46)
Creative problem solving (46)	Arithmetic problem solving (87)	Creative problem solving (32)	Computational problem solving (14)	Cognitive problem solving (38)
Interpersonal problem solving (36)	Subtraction problem solving (25)	Collaborative problem solving (23)	Collaborative problem solving (13)	Personal problem solving (15)
Collaborative problem solving (34)	General problem solving (23)	Cognitive problem solving (20)	General problem solving (13)	Effective problem solving (8)

Note. Numbers between brackets indicate the number of publications referencing the problem-solving type in the title, abstract, keywords, or full text.

Dimensions of early problem solving

Each problem-solving type refers to a specific aspect of early problem solving, thereby reflecting different dimensions. To map these dimensions, the early problem-solving types were coded inductively and iteratively. This process led to the identification of five dimensions within which the early problem-solving types can be classified: (a) problem type, (b) problem topic, (c) problem-solving skills, (d) problem-solving process, and (e) problem-solving context. Table 4 presents an overview of the five dimensions, a description, and examples of early problem-solving types.

It should be noted that these dimensions are not mutually exclusive. Early problem-solving types may refer to multiple dimensions. For example, spatial problem solving can be viewed as related to the topic (solving spatial problems) or as a skill (applying spatial reasoning when solving a problem).

Table 4. Dimensions of early problem solving

Dimension	Description	Examples of early problem-solving types
Problem type	Describes the main characteristics and presentation format of the problems of interest	Word problem solving, Open-ended problem solving, Hands-on problem solving
Problem topic	Describes the subject or domain of the problems of interest	Arithmetic problem solving, Spatial problem solving, Design problem solving
Problem-solving skills	Describes the general skills that underlie problem solving	Creative problem solving, Self-regulated problem solving Computational thinking problem solving
Problem-solving process	Describes the structure of the process(es) required to solve the problem	Systematic problem solving, Flexible problem solving Step-by-step problem solving
Problem-solving context	Describes the broader setting in which the problems are solved, related to the physical environment, access to support, and modes of collaboration	Real-world problem solving, Collaborative problem solving, Independent problem solving

Discussion

Through bibliometric analysis, we aimed to map the research landscape on early problem-solving across disciplines. Drawing on a comprehensive search across five databases, we identified and analysed 1480 publications on early problem solving. We found that early problem solving is most often studied within the fields of Psychology, Mathematics, Social Sciences: Education, Computer Science, and Health Professions. Across these fields, early problem solving is investigated within a broad range of topics. We also found that early problem solving is studied more frequently in the first years of primary education than in pre-primary education, and that nearly one in five publications specifically targets diverse learners. In total, 168 types of early problem solving were identified, which were subsequently grouped into four clusters: mathematical, social, creative, and collaborative problem solving. Finally, we identified five dimensions of problem solving within these early problem-solving types: problem type, problem topic, problem-solving skills, problem-solving process, and problem-solving context.

Early problem solving is studied across different research fields and topics

For this study, we compiled an extensive corpus of publications on early problem solving and categorised them by research field. We found that research on early problem solving has been conducted across a wide range of research fields. The majority of publications were classified in five fields: Psychology, Mathematics, Social Sciences: education, Computer Science, and Health Professions. Beyond these five major research fields, the landscape is highly fragmented, with studies on early problem solving in fields such as neuroscience, engineering, and arts and humanities receiving only sporadic attention.

The fragmentation of the research literature becomes even more apparent at the topic level with publications on early problem solving classified under 132 different topics. The topic *Cognitive and Developmental Aspects of Mathematical Skills* clearly stands out, with by far the largest number of publications. Our mapping of the research literature also reveals topics within which only a single study addresses early problem solving, such as *Digital Storytelling and Education* (Mohamed & Maker, 2011), *Virtual Reality Applications and Impacts* (Araiza-Alba et al., 2021), and *Urban Green Space and Health* (Kim & Koo, 2019).

The presented treemap (Figure 4) provides researchers with a framework to position their work within the broader landscape of research on early problem solving. It also offers them a tool to identify research beyond their own field. By having mapped the full body of literature, we aim to make the research domain more accessible and foster cross-fertilisation.

Target populations

The results also reveal that research attention to problem solving increases with educational level. This pattern is particularly pronounced in mathematics research, which is strongly oriented toward primary education with emphasis on Grade 3. By contrast, children in pre-primary education receive far less attention, except for studies on social problem solving. Given young children's idiosyncratic developmental trajectories (Reilly et al., 2022) and the importance of understanding how problem-solving skills develop and can be supported in early childhood (Keen, 2011), the relatively limited number of publications on problem solving in pre-primary education is somewhat problematic. Therefore, there is a clear need for more research on the youngest target group.

Approximately one in five studies addresses learners with diverse characteristics, with neurodiverse learners forming the most frequently studied group of diverse learners, particularly within the domain of mathematics. This includes research on the early detection of and support for children with mathematics disabilities, dyscalculia, dyslexia, another home language, and other learning difficulties or disabilities. The concentration of studies on problem solving in diverse learners within mathematics can be explained by the observation that solving mathematical word problems, found in curricula worldwide at all educational levels, is highly complex, requiring knowledge, skills, and executive functions beyond mathematics (Geary, 2004; Verschaffel et al., 2020). Therefore, the early identification of children at risk for mathematics difficulties and the support of the development of their mathematical problem-solving skills have emerged as important themes in research. Moreover, leading scholars such as H. Lee Swanson and Lynn and Douglas Fuchs have made substantial contributions to research on the relationship between mathematical problems and learning disabilities.

Problem solving is examined not only in the context of children experiencing difficulties but also among gifted learners. Although the proportion of publications focusing on this target group is relatively small, their presence underscores the potential of problem solving in identifying and supporting not only struggling learners but also gifted once (e.g., Kuo et al., 2010). When thoughtfully designed, problem-solving activities embody the principle of "low floor, high ceiling" (English, 2017; Resnick & Silverman, 2005) meaning that all children can engage and everyone is challenged.

Types and dimensions

The diversity of the research landscape on early problem solving is further illustrated by the 168 distinct types of early problem solving identified in the publications. Using co-occurrence network analyses, these types were categorised into four clusters: mathematical, social, creative, and collaborative problem solving. These four clusters can be considered as an overlay map, an additional layer on top of the OpenAlex classification, specifically focused on early problem-solving types. When this overlay is applied to the research fields, mathematical problem solving is found to be the most dominant type across the main research fields. The only exception is Health Professions, where mathematical problem solving does not appear among the top five; instead, interpersonal problem solving is predominant. Moreover, social, creative, and collaborative problem solving also appear frequently across multiple research fields. Overall, these findings suggest that mathematical, social, creative, and collaborative problem solving constitute a common core of research on early problem solving across research fields. This highlights the need for researchers to look beyond the boundaries of their own field, as relevant work may be found elsewhere.

Out of the 168 early problem-solving types, we distilled five dimensions: problem type, problem topic, problem-solving skills, problem-solving process, and context. These dimensions can be combined in various ways. Examples include the omnipresent mathematical word problem solving (topic and type), collective creative problem solving (context and skill, Spoon et al., 2021), real-life mathematical problem solving (context and topic, Fuchs et al., 2006), and systematic social problem solving (process and topic, Vlachou & Stavroussi, 2016). Awareness of these dimensions can help researchers delineate the scope of their studies and position them more accurately within the broader research literature. It can also guide authors in selecting appropriate keywords to enhance discoverability of their work.

Limitations

Despite its benefits, science mapping with OpenAlex has notable limitations. The OpenAlex search engine does not support wildcards (see https://docs.openalex.org/), and issues with author disambiguation persist (Culbert et al., 2025). Therefore, we relied on established commercial databases for the main search of publications, which we subsequently fetched with the OpenAlex database to enable source classification. During this process, we observed that not all publications could be automatically linked to the OpenAlex database and that some records were incorrectly matched. Consequently, all links had to be manually verified. During this quality check, we linked 56 records and corrected 10 misclassified records.

For the creation of the overview map and subsequent analysis by research field, we relied on the *primary topic* provided by OpenAlex. In OpenAlex, a probability score is calculated for all publications across all topics, and the database records the three highest-scoring topics together with their associated subfields, fields, and domains. In some cases, the probability score of the primary topic is very close to that of other topics. Moreover, automated systems always carry a risk of misclassification, although we are not aware of any studies that have systematically documented misclassifications in OpenAlex. Therefore, the classification is not always conclusive at the level of the individual study.

Pathways for further research

The present study lays the foundation for further research on early problem solving. First, our findings highlight the need for conceptual work examining commonalities and differences in how early problem solving is theorised, conceptualised, and operationalised across research fields. Such analyses could clarify whether early problem solving constitutes a unified construct across research fields or should be understood as different, domain-specific constructs that vary fundamentally by field or even topic or type. Ultimately, the results of such research could provide the research and educational community with a more nuanced understanding of early problem solving's complex, multifaceted nature. Second, our bibliometric review helps researchers to identify gaps and directions for further investigation. This includes systematic reviews and meta-analyses examining the effects of early problem-solving interventions, reviews on developmental aspects of early problem solving, and reviews investigating what exactly is measured when authors claim to measure problem solving. Our openly available database can serve as a foundation for such reviews, saving researchers weeks or even months of screening work.

To facilitate navigating early problem-solving literature within the OpenAlex classification, we developed the *Early Problem-Solving Compass*, an openly accessible R Shiny app. This tool integrates the OpenAlex topic classifications with the metadata of records included in this study, enabling researchers to identify relevant publications across disciplines based on author, title, keywords, or topic. Researchers can use the compass when exploring available literature, identifying the state-of-the-art in a certain field, or designing empirical studies or systematic reviews. Editors can use the compass to identify suitable reviewers, and early childhood educators can use it to trace the available evidence on a specific aspect of problem solving. The *Early Problem-Solving Compass* is freely available at https://earlyproblemsolvingcompass.shinyapps.io/EarlyProblemSolvingCompass/.

STUDY 2 | METHODS FOR ASSESSING PROBLEM-SOLVING COMPETENCIES IN PRESCHOOLERS: A SCOPING REVIEW

This chapter has been published as:

Van Elsen, J., De Maeyer, S., & Torbeyns, J. (2026). Methods for assessing problem-solving competencies in preschoolers: A scoping review. *Review of Education*, *14*, Article e370181. https://doi.org/10.1002/rev3.70181

Abstract

Background: Problem solving is an important 21st century skill and an indispensable part of early STEM (science, technology, engineering, and mathematics) education. To effectively study the development of children's problem-solving competencies and to evaluate the effectiveness of classroom interventions and programmes, researchers and educators need reliable and valid assessment instruments.

Aims: This scoping review aims to identify and discuss methods that are developed, studied, or used in scientific research to assess preschoolers' competencies in solving ill-structured problems in and across the four STEM disciplines.

Methods: A systematic search of Web of Science, Scopus, and ERIC identified 24 studies published between 1974 and 2024. These studies reported 18 unique methods for assessing problem-solving competencies in preschool children aged 3 to 6 years.

Results: Results indicate that most assessment methods rely on observation of hands-on tasks or oral examination, typically focusing on a single step of the problem-solving process within one STEM discipline. Only one instrument was found for which the psychometric properties were adequately examined.

Conclusions: Observing children solve engineering design problems and engaging in dialogue with them appears appropriate for integrated problem-solving assessment. Further research is needed to develop reliable and valid assessment methods to provide researchers and educators with age-appropriate methods to investigate and monitor preschoolers' problem-solving competencies.

Introduction

Problem solving is a key 21st century competency that is required for lifelong learning and empowers individuals to navigate our increasingly complex society (European Commission, 2019; Suto, 2013). Early childhood offers a critical window for developing problem-solving competencies through pedagogical activities (Burns et al., 2025; Diamond, 2018; Malcok & Ceylan, 2022; O'Reilly et al., 2022; Rogers, 2004). Developing problem-solving competencies takes years, even decades (Molnár et al., 2013). Therefore, stimulating these competencies from preschool (3–6 years old) onwards is recommended (Fusaro & Smith, 2018; Lopes et al., 2017). Problem solving is key within science, technology, engineering, and mathematics (STEM, Priemer et al., 2020). Integrating STEM-related problem solving in early childhood education provides ample opportunities to develop problem-solving skills (MacDonald et al., 2022).

Researchers and educators need reliable and valid assessment methods to investigate, monitor, and support the development of children’s problem-solving competencies and to evaluate the effectiveness of classroom interventions and programmes (Snow & Van Hemel, 2008). Assessment refers to the process of collecting, interpreting, and using evidence about student learning (Harlen, 2007; McMillan, 2013) and includes different kinds of measurements and testing (AERA et al., 2014; Popham, 2017). For researchers and educators, it is important to understand how problem-solving competencies are being assessed in preschoolers, including the methods used, their characteristics, their scope, the procedures, and their psychometric properties.

Researchers and educators can only rely on assessment outcomes if the methods used are thoughtfully developed, considering crucial elements related to reliability and validity. In their review on reporting practices related to reliability and validity of early childhood assessment, Barghaus et al. (2022) found that only a small fraction of assessment developers report on all required issues concerning reliability and validity as proposed in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). In their systematic review and meta-analysis on approaches to teaching problem-solving skills in early childhood education and care settings, Burns et al. (2025) emphasise the importance of comprehensive, reliable, and valid instruments to assess young children’s problem-solving skills. However, it remains unclear to what extent reliable and valid instruments were used in the included studies.

The present scoping review aims to provide an overview of methods developed, studied, or used in early childhood education research to assess problem-solving competencies in preschool children (aged 3–6 years). More specifically, the study aims to identify and analyse the characteristics, scope, procedures, and psychometric properties of these methods.

Before describing the methodology and results, we elaborate on the conceptual definition of problem solving in the context of STEM and the steps of the problem-solving process. We conclude with discussing the implications of our findings for research and practice.

Problem solving in STEM

Problem solving is the process of dealing with a problem where no obvious solution method is available (OECD, 2004). It encompasses a wide range of cognitive and non-cognitive competencies, such as information processing, representation, reasoning, self-regulation, planning, and decision making (Greiff et al., 2014). Given the diversity of potential problems and the observation that a routine job for one person may be an impossible problem for another person (Bransford & Stein, 1993), a situation or assignment is considered a problem when the subject is able to understand it, but currently lacks a procedure to solve it (Brownell, 1942).

Problem solving is an indispensable part of STEM education (Priemer et al., 2020). STEM refers to educational activities within science, technology, engineering, and mathematics (Bryan & Guzey, 2020). According to the International Technology and Engineering Educators Association (ITEEA, 2020), science can be understood as “Investigating and understanding the phenomena in the natural world” (p. 159). Technology involves altering the environment through the creation of tools, systems, or processes designed to fulfil human needs and desires. Engineering refers to the application of scientific knowledge and mathematical reasoning to design and refine technological solutions that meet specific criteria under constraints. Mathematics is thinking and reasoning about patterns, order, measurement, properties, quantities, numbers, operations, and geometry (see also Milburn et al., 2019). STEM education can be approached in various ways, ranging from teaching discipline-specific knowledge and skills in isolation to integrating and applying them across multiple disciplines (English, 2016).

Authentic STEM-problems tend to be ill-structured (Nadelson & Seifert, 2017). Ill-structured problems are “fuzzy” problems that occur in everyday life (Simon, 1978). They are open-ended and allow for multiple solution paths, as there are no clearly defined rules or preferred algorithms to solve them (Frensch & Funke, 1995). Well-structured problems, on the other hand, typically have one or a limited number of correct solutions and can be solved by applying certain procedures or algorithms (Jonassen, 1997; Simon, 1973). Solving ill-structured problems involves different cognitive processes than solving well-structured ones and requires distinct learning environments (Jonassen, 2010a). Therefore, findings derived from solving well-structured problems such as Tower of Hanoi-style puzzles or routine mathematical (word) problems cannot simply be generalised to solving ill-structured problems (Funke & Frensch, 2007).

Steps of the problem-solving process

Different problems require different solution methods and strategies. Although most researchers emphasise that problem solving is not necessary a linear process, many scholars (e.g., Dewey, 1910; Polya, 1945; Schoenfeld, 1985) have decomposed the process to operationalise problem solving in measurable and teachable phases or steps (Rott et al., 2021). The PISA Problem-Solving Framework builds on Polya's (1945) four phase model and is developed to measure the problem-solving competencies of 15-year-olds across the globe. The framework has been used in the PISA 2003 cycle (OECD, 2005), the PISA 2012 cycle (OECD, 2013, 2014) and the PISA 2015 cycle on collaborative problem solving (OECD, 2017a, 2017b). The PISA framework divides the problem-solving process into eight steps: exploring, understanding, representing, formulating, planning, executing, monitoring and reflecting (OECD, 2013).

Exploring is conceptualised as observing and interacting with the problem, searching for information, and finding limitation and obstacles. *Understanding* involves understanding given and obtained information and demonstrating understanding of relevant concepts. *Representing* refers to building a mental representation and constructing tabular, graphical, symbolic or verbal representations. *Formulating* is defined as generating hypotheses by identifying relevant elements of the problem and their interrelationships and critically evaluating information. *Planning* relates to setting and clarifying (sub-)goals and devising a plan or strategy to reach that goal. *Executing* refers to carrying out the plan. *Monitoring* refers to monitoring progress towards the goal, checking (intermediate) results, identifying unexpected events, and adjusting as required. Finally, *Reflecting* involves critically evaluating solutions from different perspectives, identifying additional information needs, and communicating progress.

The present study

To monitor the development of children's problem-solving competencies and to evaluate the effectiveness of classroom interventions and programmes, researchers and educators require reliable and valid assessment instruments (Burns et al., 2025). A comprehensive overview of available instruments, detailing their key characteristics, scope, and psychometric properties, would provide essential guidance for researchers and educators in selecting, developing, or implementing appropriate assessment methods.

Two recent review studies aimed to identify assessment methods to enhance students' problem-solving skills, but did not address preschool education (Ukobizaba et al., 2021; Wicaksono & Korom, 2022). The same applies for recent review studies on assessment of related topics such as computational thinking (Cutumisu et al., 2019; Lu et al., 2022; Tang et al., 2020; Varghese & Renumol, 2021) and creative thinking (Bolden et al., 2020; Said-Metwaly et al., 2017; Suherman & Vidákovich, 2022). Systematic reviews examining assessment methods for preschool children address related concepts such as executive functions (Mehsen et al., 2022; Silva et al., 2022) or self-regulation (Philpott-Robinson et al., 2023), but not problem solving. The recent systematic review and meta-analysis of Burns et al. (2025) addresses problem solving in early childhood education (including preschool), but only pays limited attention to assessment. To the best of our knowledge, no comprehensive overview of methods developed, studied, or used in scientific research to assess problem-solving competencies in preschoolers currently exists.

The present scoping review aims to fill this gap in scientific research by providing an overview of the methods developed, studied, or used to assess the problem-solving competencies of preschool children (aged 3–6 years), and to map the characteristics, scope, procedures, and psychometric properties of these methods. A scoping review is an appropriate approach to map how a phenomenon is studied across research with varying methodological designs (Aromataris & Munn, 2020; Munn et al., 2018). Unlike systematic reviews, scoping reviews do not aim to answer a single, narrowly defined question. Instead, they aim to identify the characteristics of a particular phenomenon or to provide an overview of the existing evidence (Munn et al., 2018).

Synthesising the characteristics of existing assessment methods offers valuable insights into how problem-solving competencies are currently assessed in research and can support researchers and educators in selecting, developing, or implementing appropriate assessment methods. Mapping the scope of these methods helps identify which aspects of the problem-solving process and STEM disciplines are being addressed, while also revealing gaps that warrant further investigation. Additionally, examining the procedures and psychometric properties of the methods provides important information about their reliability and validity. The following research questions guide this study:

RQ 1: What are the characteristics of the methods developed, studied, or used in scientific literature to assess problem-solving competencies in preschoolers?

RQ 2: Which steps of the problem-solving process and which STEM-disciplines are being assessed with these methods?

RQ 3: What are the procedures and psychometric properties of these methods?

Method

Design

The study was conducted according to the scoping review guidelines framework of the Joanna Briggs Institute (JBI, Aromataris & Munn, 2020). The scoping review methodology was chosen for a number of reasons. First, the present study aims to explore the field of research concerning methods for assessing problem-solving competencies in preschoolers, which is in line with the exploratory nature of scoping reviews (Christou et al., 2025; Munn et al., 2018; Tricco et al., 2018). Second, the scoping review approach provides an inclusive framework for bringing together diverse literature, including quantitative, qualitative, and mixed-method studies (Aromataris & Munn, 2020). This approach enables inclusion of development studies, validation studies, and intervention studies. Third, scoping reviews are particularly well-suited to identify characteristics of a phenomenon and how it is being investigated (Munn et al., 2018). This aligns with our aim to map and analyse the characteristics and psychometric properties of the methods that have been developed, studied, or used in scientific research to assess problem-solving competencies in preschoolers. Finally, scoping reviews allow to identify gaps in scientific knowledge and can serve as a foundation for future systematic reviews (Munn et al., 2018). By mapping assessment methods developed, studied, or used in scientific research, this study provides an overview of the current state of research on assessing problem-solving competencies in preschoolers, and provides direction for future research into reliable and valid assessment methods.

Reporting is based on the Preferred Reporting Items for Systematic reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) statement (Tricco et al., 2018). A review protocol was pre-registered on Open Science Framework (OSF, https://doi.org/10.17605/OSF.IO/QPJAT). The protocol outlines the objectives, search and selection procedures, data extraction, and analysis. Any deviations from the protocol are addressed in the Discussion section.

Search strategy

A three-stage search strategy was applied to identify relevant sources (Aromataris & Munn, 2020). First, a search string was derived from the research questions. A preliminary search on Web of Science was conducted and keywords of relevant studies were screened to identify additional search terms. In consultation with a university librarian, the search string was refined, and three databases were selected: Web of Science, Scopus, and Eric. Second, a main search on title, abstract, and keywords was conducted across all databases, using the following search string:

> (TITLE-ABS-KEY (problem* W/5 solv*) AND TITLE-ABS-KEY (preschool* OR kindergarten* OR "pre-k" OR "early educat*" OR "early school*" OR ecec OR "early childhood educat*" OR "nursery school") AND TITLE-ABS-KEY(assess* OR monitor* OR evaluat* OR testing* OR measur*)) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar")).[1]

[1] This search string conforms to the syntax of Scopus. The W/5 in the first argument is a proximity operator to allow a distance of five words between both terms.

Finally, reference lists of selected reports were screened for additional sources (backward search) and sources that refer to the identified reports were traced through cited reference searching in Web of Science and Scopus (forward search).

Eligibility criteria

Only peer-reviewed journal articles presenting empirical quantitative, qualitative, and mixed method studies were considered. Restricting the sources to peer-reviewed articles was intended to ensure that the methods are grounded in scientific evidence. Only sources written in English were considered. Inclusion criteria were conceived according to the PCC framework (Participants/Population, Concept, and Context) (Aromataris & Munn, 2020). We included studies that focus on children enrolled in preschool programmes with an intentional education component classified under ISCED 02: Early childhood education from age 3 to the start of primary education (UNESCO Institute for Statistics, 2012). Specifically, studies were retained if the data were collected in an institution described as 'preschool' or 'kindergarten', or if the data were collected in a school or classroom and at least some of the participants were in the 3–6 age group. To be eligible, the studies must include problem-solving assessment that focus on ill-structured STEM problems. Because structuredness of a problem is a multidimensional continuum (Jonassen, 2010a), studies were retained if at least one problem presented to the participants was non-routine and provided opportunities for multiple solutions or multiple solution paths. Studies that focus only on solving well-structured problems or problems with only one correct solution such as simple tasks (e.g., conservation tasks, balance beam tasks), prescribed operations (e.g., mathematical operations, coding problems), or puzzles (e.g., Tower of Hanoi, labyrinths) were excluded. Since this review focuses on STEM problems, only studies in which the problem-solving assessment relates to at least one of the STEM disciplines were considered eligible, thereby excluding social problems (e.g., interpersonal problems, conflict handling) and reasoning problems (e.g., problems related to causality, analogy, syllogisms). Finally, only studies that include educational assessment administered in preschool were eligible. Clinical studies, labo-experiments conducted outside the school, and screenings for disabilities were excluded. Consequently, we did not include studies that make use of tests such as the Raven's Colored Progressive Matrices (RCPM), Ages & Stages Questionnaires® (ASQ) or the Dimensional Change Card Sort (DCCS) for problem-solving assessment since they only pay limited attention to problem solving and don't focus on STEM.

Study selection

All identified sources were collated in EndNote (version 20.1) and duplicates were removed. Titles and abstracts were screened for eligibility by the first author using ASReview (version 1.0), a researcher-in-the-loop active-learning application that helps prioritising relevant studies (van de Schoot et al., 2021). The default Naïve Bayes model (TF-IDF, Naive Bayes, Maximum, Dynamic resampling) was used for screening. The model was trained on two previously identified relevant studies meeting the inclusion criteria and ten randomly selected studies of which one was marked relevant and nine were rejected. During the screening process, all records labelled irrelevant were tagged with a reason why they were excluded. Records not presented by the algorithm were considered not eligible.

Two mutual independent stopping criteria were applied: at least 33% of the total dataset has been screened (van de Schoot et al., 2021) and 100 consecutive records were labelled irrelevant (Wagner et al., 2022). The second criterion is more stringent than the recommendation of Ros et al. (2017) to stop after 50 consecutive rejections.

Risk of bias through automation was assessed by having the second and third authors independently screen a random sample of 10% of the records (*n* = 250). The results of the screening process by the first author were compared with those of the second and third authors. Percentage of agreement between raters was 93%. This aligns with the expected error rates for human reviewers as reported by Wang et al. (2020). All discrepancies were analysed and discussed within the research team.

During the first round of screening, we noticed that the concept problem solving is used as an umbrella term to refer to resolving a wide range of tasks. We therefore decided to further specify the inclusion criteria and align them more closely with the conceptualisation of problem solving described in the theoretical framework of this study. Based on a list of all encountered problem types that was composed by the first author during screening, the inclusion criteria were further specified. Studies focussing on well-structured problems, problems with only one correct answer, routine tasks, operations, or screenings were excluded. For convenience, the eligibility criteria presented in the previous section are the final criteria.

The adjustments to the inclusion criteria required us to redo the screening process. In scoping reviews, an iterative team approach is advised (Levac et al., 2010). ASReview was used to screen titles and abstracts. Six relevant studies from the previous round and five randomly selected studies marked irrelevant were provided as prior knowledge. Once again, 10% randomly selected records were screened manually by the second and third author. Percentage of agreement between raters was 91%, which is again consistent with expected error rates in manual screening (Wang et al., 2020). Discrepancies were analysed and discussed. The differences were due to unclear or missing information in the abstract and all differently rated sources were rejected after discussion within the research team or full text screening.

Full texts of eligible sources were retrieved through EndNote, databases, journal webpages, and ResearchGate. If a full text was not readily accessible, authors were contacted through ResearchGate, email, or LinkedIn. During full text screening, assessment methods were scrutinised. If the study failed to provide sufficient information about the assessment method, additional information was looked up through the provided references and authors of the sources and instruments were contacted. All excluded studies were coded by the main reason of exclusion and doubtful cases were discussed with the research team.

Data extraction

A digital data extraction form was drafted in Qualtrics (https://www.qualtrics.com), pilot-tested, and refined after discussion with the research team. To avoid errors, charting was performed twice. During the second round, relevant data were charted in a pre-structured spreadsheet. Differences between the two rounds were double checked.

From the selected sources, information was extracted about the study characteristics (year of publication, educational system, study design, sample size, participants' age, research questions, theoretical framework), characteristics of the assessment method (description of the assessment, assessment type, assessment materials, objectives of the assessment, data types, and scoring), scope of the problem-solving assessment (addressed problem-solving steps and STEM disciplines), details about the assessment procedures (administrator, location, duration, and whether the session was done individual, in pairs, or with the entire classroom), and psychometric properties of the included assessment instruments (types of reliability and validity, metrics). Only the psychometric properties that were explicitly articulated by the authors were included (for an overview of types of reliability and validity, see Cohen et al., 2018; Mokkink et al., 2016). If the scope of the assessment method was unclear, authors were contacted to obtain the relevant items or scale. All data on the sources and assessment methods are presented in Tables 5–9.

Synthesis

For each included assessment method, a short description was written by the first author in a standardised format. Frequency tables were generated to provide a summary of assessment types, materials used, data collection methods, and outcomes, highlighting prevalence of the assessment characteristics. Typical and atypical methods were described qualitatively and reported in more depth. All included assessment methods were coded regarding to the problem-solving steps being assessed and the STEM-disciplines being addressed. Where authors indicated that the method used addressed integrated STEM, all four STEM disciplines were checked. Frequency tables were created and prevalence of the steps and disciplines was analysed. The reported statistics on the psychometric properties of the assessment methods were collated in a spreadsheet and categorised by type of evidence.

Results

Literature search

The main database search was conducted on 21 September 2022 and resulted in 2979 records. Figure 6 provides an overview of the selection process according to the PRISMA 2020 statement (Page et al., 2021). After deduplication (n = 521), 2458 records were retained for title and abstract screening. The stopping rules were met after screening 1389 records of which 213 were marked relevant. The 1069 records that were not presented by the algorithm were labelled irrelevant.

Fourteen, mainly older articles ($n_{< 2000}$ = 12) could not be retrieved. After full text screening, 17 reports were included. Four additional sources were included through backward search. Forward search yielded no new records. On 2 October 2024, an additional search was conducted to retrieve sources that were published during the course of the study. This updated literature search resulted in 349 new reports of which three met the inclusion criteria. The 24 reports included in the final selection contain information about 18 different assessment methods.

Study characteristics

Study characteristics of the 24 studies included in this review are presented in Table 5. The studies were published between 1978 and 2024 with more than half of the studies published in the last ten years (*n* = 16). Studies were conducted in nine different countries. The Middle East is best represented in the selected studies (*n* = 10), followed by North America (*n* = 7), and Europe (*n* = 5). Three studies were conducted in East Asia. There were no studies from Central and South America, South and Southeast Asia, Africa, or Australia. Different designs were used in the studies, with most studies being (quasi-)experimental (*n* = 9), correlational (*n* = 4), or descriptive (*n* = 5). The age of the children that participated in the studies ranged from 3 to 7 years. The study of German (2000) involved children aged 5, 6, and 7 of whom only the 5- and 6-year-olds are considered preschoolers. A total of 2219 preschoolers participated in the studies with an average of 92 (*min = 5, max. =* 328) preschoolers per study.

The included studies report on 18 different methods to assess problem-solving competencies in preschool. Four methods were used in two or three studies. In one study, two methods were used to assess problem-solving competencies (Leikin & Tovli, 2014). Most studies include an assessment method purposefully designed to meet the needs of the study (*n* = 14). Five of these methods were adaptations of, or based on, existing methods. Ten studies made use of an existing method and two studies aimed at developing a new assessment method or instrument.

Figure 6. PRISMA flow diagram of the search and selection process (Page et al., 2021)

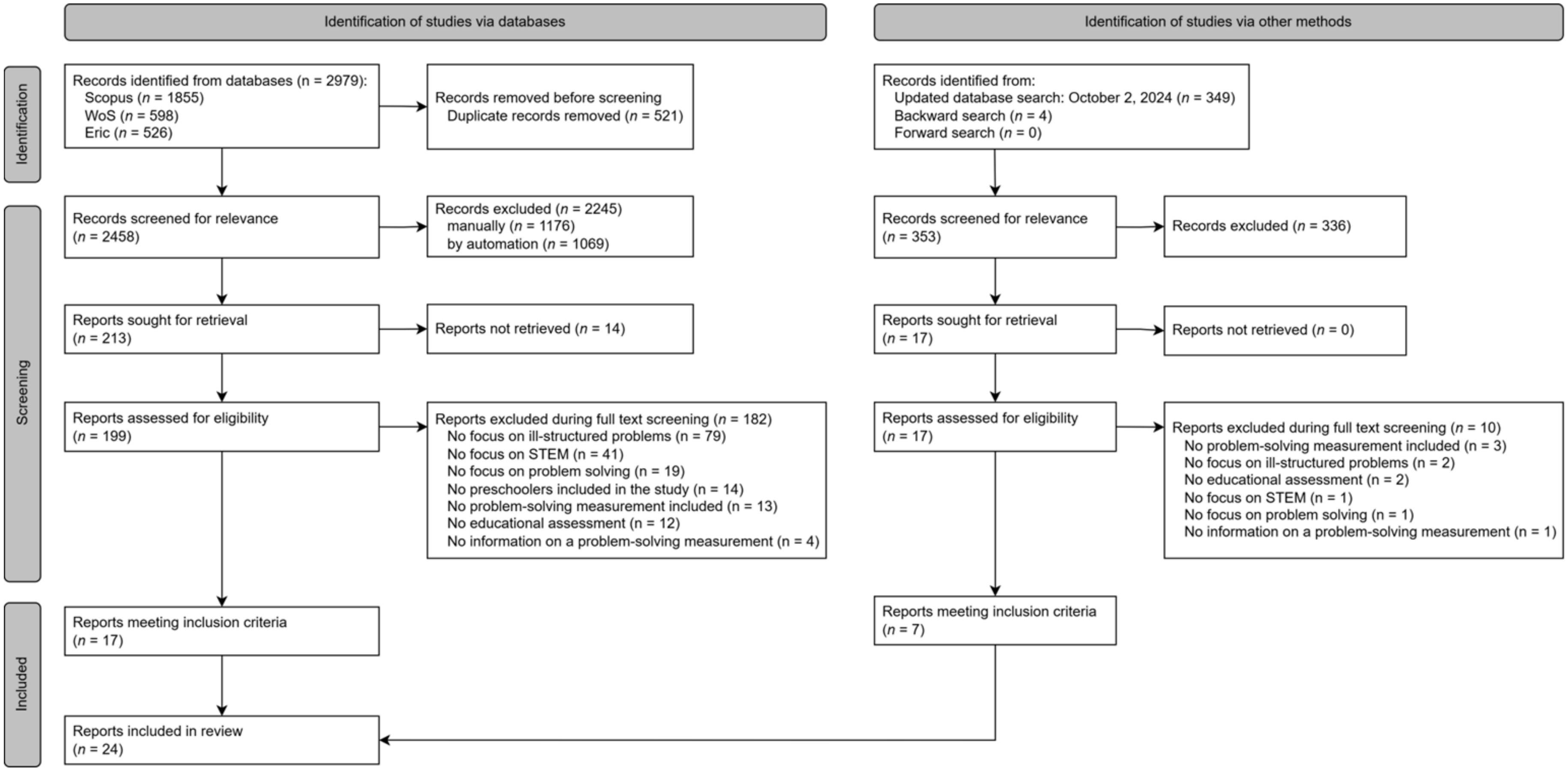


Table 5. Characteristics of included studies

Study	Country	Design	Age range	*N* preschoolers	Origins of the assessment method
Alexander et al. (1994)	USA	Descriptive	*NA*	61	Ad hoc, based on Torrance (1974)
Anggoro et al. (2021)	USA	Test development	3–5	238	Ad hoc
Annevirta and Vauras (2006)	Finland	Longitudinal	6	43	Ad hoc
Bartholomew et al. (2019)	USA	Descriptive	5–6	55	Ad hoc
Cakir et al. (2021)	Türkiye	Experimental	5	40	Oğuz and Köksal- Akyol (2015)
Demirel and Deretarla Gul (2021)	Türkiye	Quasi-experimental	4–5	32	Oğuz and Köksal- Akyol (2015)
Fusaro and Smith (2018)	USA	Correlation	4–5	24	Ad hoc
Gardiner et al. (2012)	USA	Experimental	2–3	46	Ad hoc
German and Defeyter (2000)	UK	Experimental	5–7	60	Ad hoc, based on Duncker (1945)
Guevara et al. (2020)	Netherlands	Repeated measures	4–6	8	Ad hoc
Hong and Milgram (1991)	USA/Israel	Descriptive	4–6	48	Ad hoc and Ray (1955)
Leikin (2013)	Israel	Descriptive	3–5	37	Ad hoc, based on Leikin (2009) and Torrance (1974); Tsamir et al. (2010)
Leikin and Tovli (2014)	Israel	Correlation	5–6	31	Leikin (2013), based on Torrance (1974); Tsamir et al. (2010)
Lin et al. (2021)	China	Quasi-experimental	4–5	122	Fusaro and Smith (2018)
Lin et al. (2023)	China	Correlation	5	226	Fusaro and Smith (2018)
Nam et al. (2019)	South Korea	Quasi-experimental	5–6	53	Adapted from Ward (1993)
Parsonson and Baer (1978)	USA	Multiple baseline	3–6	5	Ad hoc
Shechter et al. (2021)	Israel	Correlation	4–6	228	Ad hoc
Tsamir et al. (2010)	Israel	Descriptive	5–6	163	Ad hoc
Unal and Aral (2014a)	Türkiye	Test development	5	174	Ad hoc
Unal and Aral (2014b)	Türkiye	Quasi-experimental	5	42	Unal and Aral (2014a)
Unal and Saglam (2018)	Türkiye	Quasi-experimental	5	25	Unal and Aral (2014a)
Van Elsen et al. (2024)	Belgium	Quasi-experimental	5–6	130	Ad hoc, based on Leikin (2013)
Vessonen et al. (2024)	Finland	Descriptive	3–5	328	Ad hoc

Characteristics of the identified assessment methods

To answer the first research question, we shift the unit of analysis from the individual study to the assessment methods that are used in these studies. Table 6 provides a short description of the assessment methods and their characteristics. Nearly all assessment methods make use of hands-on assignments (*n* = 10) or are based on oral examination (*n* = 7). The Design Portfolio's (Bartholomew et al., 2019) apply portfolio assessment.

Hands-on assignments

Generally, hands-on assignments make use of commonly available crafts materials (e.g., cardboard, tape, glue, bottle caps, pipe cleaners) (*n* = 5). Other materials used are building bricks (e.g., Lego), toys, tools, and consumer goods. Two methods make use of a specially designed apparatus. Hands-on assignments are most often accompanied by observation for data collection. These observations were conducted directly (*n* = 5), via video recordings (*n* = 4), or combined both approaches (*n* = 1). The Stringent Solution-Standard Measures (Hong & Milgram, 1991) makes use of oral examination to collect data during the hands-on assignment. Observations of hands-on assignments typically result in binary outcomes (*n* = 5) or count scores (*n* = 6). Binary outcomes indicated whether participants successfully solved the problem, utilised a specific tool, performed a particular action, or whether the solution met predefined criteria. Count scores relate to the number of attempts to solve a problem or the number of different solutions. Other quantitative outcomes are time on task (*n* = 3) and scale scores (*n* = 4).

An illustrative example of a hands-on problem-solving assignment is the Open-Ended Problem-Solving Construction Task (Shechter et al., 2021). To assess preschoolers' problem-solving competencies, a map of a river was placed on a table and the participants were challenged to build a bridge over the river with standard Lego (Duplo) building bricks, meeting four requirements: (1) the bridge must be structurally stable, (2) the bridge must connect both riverbanks, (3) the bridge must not have any pillar in the water, and (4) the bridge must be high enough to allow a paper boat to pass underneath.

The Tool Improvisation Tasks (Parsonson & Baer, 1978) is another example of a hands-on problem-solving assessment method. Participants were challenged to hammer wooden pegs in a peg board. Instead of providing participants with a hammer, the children could choose between seven functional (e.g., a brick) and ten non-functional items (e.g., a Styrofoam hammer) to pound down the pegs. Some of the items could be combined to effective tools (e.g., by putting a rod in a drilled block). A similar challenge required participants to find effective containers to carry 80 marbles. Again, participants could choose between functional (e.g., a glove) and non-functional items (e.g., a torn bag, recipients that were too small) or could combine items to create useful containers (e.g., by blocking the hole of a bottomless Styrofoam cup).

Table 6. Short description and characteristics of problem-solving assessment methods

Assessment method	Studies	Short description	Type	Materials	Data collection	Outcomes / scores
Air Pressure Tasks	Guevara et al. (2020)*	Children are asked to solve a problem with an air pressure apparatus, using provided functional and non-functional objects.	Hands-on assignment	Apparatus	Video observation	Binary outcome
Box Tasks	German and Defeyter (2000)*	Children are asked to come up with a solution to help a toy bear reach an object, using provided functional and non-functional objects.	Hands-on assignment	Craft materials (general), toys	Direct observation	Binary outcome + time
Creating Equal Number Task (CEN)	Leikin (2013); Leikin and Tovli (2014); Tsamir et al. (2010)*	Children are asked to make an equal number of bottle caps (N=10) on each side of a table in multiple ways.	Hands-on assignment	Craft materials (bottle caps)	Direct observation	Count score
Design Portfolio	Bartholomew et al. (2019)*	Children are presented open-ended design challenges, based on popular children's rhymes.	Portfolio assessment	Craft materials (general), photo camera, portfolio	Portfolio	Qualitative coding
Engineering Preschool Children Observation Tool (EPCOT)	Anggoro et al. (2021)*	Children are asked to create a solution to a character's problem, based on the concepts of light and shadow.	Hands-on assignment	NA	Video observation	Qualitative coding
Mathematical Problem-Solving Measure	Vessonen et al. (2024)*	Children are asked to solve 21 mathematical problems presented in pictures or materials.	Oral examination	Pictures, crafts materials, toys	Individual interview	Binary outcome + count score
Metacognitive Skill Tasks	Annevirta and Vauras (2006)*	Children are asked to build constructions with crafts materials and Lego bricks, meeting given criteria.	Hands-on assignment	Craft materials (general), building bricks (Lego)	Video observation	Qualitative coding + count score

Assessment method	Studies	Short description	Type	Materials	Data collection	Outcomes / scores
Open-Ended Problem-Solving Construction Task	Shechter et al. (2021)*	Children are asked to build a bridge with Lego bricks over a river (pictorial map), meeting certain conditions.	Hands-on assignment	Building bricks (Lego), paper	Direct observation + video observation	Qualitative coding + count score + scale score + binary outcome
Open-Ended Story Problems	Alexander et al. (1994)*	Children are asked to finish a story about rescuing a character that has fallen in a deep hole or is locked up in a high tower.	Oral examination	Stories	Individual interview	Scale score
Pictorial Multiple Solutions Tasks (PMST 1)	Leikin (2013)*; Leikin and Tovli (2014)	Children are asked to provide as many solutions as possible to a visually presented problem.	Oral examination	Picture	Individual interview	Count score
Pictorial Multiple Solutions Tasks (PMST 2)	Van Elsen et al. (2024)*	Children are asked to provide as many solutions as possible to a visually presented problem.	Oral examination	Pictures	Individual interview	Count score
Picture Problem Solving Task	Fusaro and Smith (2018)*, Lin et al. (2021), Lin et al. (2023)	Children are asked to provide solutions to a visually presented problem and to elaborate on their solution.	Oral examination	Pictures	Individual interview + video observation	Qualitative coding
Problem-Solving Performance Instrument	Nam et al. (2019)*	Children are asked to solve a series of mathematical problems, using provided functional and non-functional objects.	Hands-on assignment	Craft materials (general)	Direct observation	Binary outcome + Count score
Problem-Solving Scale in Science Education (PSSSE)	Unal and Aral (2014a)*; Unal and Aral (2014b); Unal and Saglam (2018)	Children are asked to formulate possible solutions to 16 drawings of science-related problem situation.	Oral examination	Pictures	Individual interview	Scale score

Assessment method	Studies	Short description	Type	Materials	Data collection	Outcomes / scores
Problem-Solving Skill Scale (PSSS) (Oğuz & Köksal- Akyol, 2015)*	Cakir et al. (2021) ; Demirel and Deretarla Gul (2021)	Children are asked to formulate possible solutions to 18 drawings of problem situation.	Oral examination	Pictures	Individual interview	Scale score
Stringent Solution-Standard Tasks	Hong and Milgram (1991)*	Children are asked to formulate as many solutions as possible to four problems in real-life set-up.	Hands-on assignment	Consumer goods (general)	Individual interview	Count score
Tool Improvisation Tasks	Parsonson and Baer (1978)*	Children are asked to solve three problems, using provided functional and non-functional objects.	Hands-on assignment	Craft materials (general), tools	Direct observation	Count score
Toy-Retrieval Tasks	Gardiner et al. (2012)*	Children are asked to get a toy out of a tube, using provided functional and non-functional objects.	Hands-on assignment	Apparatus	Direct observation	Binary outcome

Note. *Original study in which the assessment method was developed or first described.

Oral examinations

Most methods that are based on oral examinations make use of pictures to illustrate problem situations (n = 6). In the Open-Ended Story Problems (Alexander et al., 1994), the problems were presented as a story. In most cases, the children were asked to come up with multiple solutions to the presented problems. The Mathematical Problem-Solving Measure (Vessonen et al., 2024) required children to solve mathematical problems presented in pictures and materials. Oral examinations typically result in count scores (n = 3) or scale scores (n = 2). Some items in the Mathematical Problem-Solving Measure were scored binary (whether or not the children solved the problem) and the Picture Problem Solving Task (Fusaro & Smith, 2018) was processed by qualitative coding.

A typical example of an oral examination is Leikin's Pictorial Multiple Solution Task (PMST; 2013). A black and white picture of an anthropomorphic kitten who can't reach its cap on a high shelf was presented to the children. The picture contains some objects that can be of help, such as a chair, a table, and a stick. Children were asked to come up with as many possible ways to get the cap from the shelf. The answers were scored for fluency (the number of different solutions), flexibility (the number of different strategies employed in the formulated solutions), and originality (a score for unusual or original solutions). This way of scoring solutions is based on the Torrance Test of Creative Thinking (Torrance, 1974) and was also applied in the Pictorial Multiple Solution Task developed by Van Elsen et al. (2024) and the Open-Ended Story Problems (Alexander et al., 1994).

Problem-Solving Skills Scale (PSSS, Cakir et al., 2021; Demirel & Deretarla Gul, 2021) and Problem Solving Scale in Science Education (PSSSE, Unal & Aral, 2014a) follow a similar approach as the PMST's, but result in scale scores. In the PSSS, Participants were asked to formulate multiple solutions for 18 depicted problems. A five-point scale was used to indicate the number of formulated solutions, ranging from 0 (no solution) to 4 (more than three solutions). In the PSSSE, children were asked to formulate a solution to 16 depicted problems related to life sciences, health sciences, and physical sciences. The answers were scored on a four-point scale indicating whether (1) the child didn't provide a solution or a wrong solution; (2) the answer had a false focus; (3) the answer didn't include a next step; or (4) the answer was right and complete.

Design Portfolio

Finally, Design Portfolio (Bartholomew et al., 2019) makes use of a portfolio to assess preschoolers problem-solving competencies. Children were asked to design solutions to problems that appear in popular children's rhymes (Itsy Bitsy Spider, Baa Baa Black Sheep and Little Boy Blue). A worksheet was developed for each problem on which the children could draw their ideas at the different phases of the problem-solving process. In addition, the children were encouraged to create a prototype of their solution using crafts materials. At the end, portfolios and prototypes were photographed and all worksheets and prototypes were coded holistically.

Scope of the assessment methods

To address the second research question, we first analyse the problem-solving steps included in the identified methods. Next, we examine which STEM disciplines are emphasised within these methods. Table 7 provides an overview of the problem-solving steps covered by the assessment methods. Table 8 shows the STEM disciplines addressed.

Steps of the problem-solving process

Nine identified methods focus on one step of the problem-solving process. Seven of these methods solely address formulating possible solutions. This is the case for all oral examinations, except for the Problem-Solving Performance Instrument (Nam et al., 2019) which assesses executing. Three methods capture nearly the entire problem-solving process: Design Portfolio (Bartholomew et al., 2019), Mathematical Problem-Solving Measure (Vessonen et al., 2024), and Open-Ended Problem-Solving Construction Task (Shechter et al., 2021). These three methods all contain hands-on problems and make use of tangible objects. Overall, formulating ($n = 12$) and executing ($n = 11$) are the most assessed steps of the problem-solving process in the methods included in this review. Exploring ($n = 2$), understanding ($n = 2$) and representing ($n = 1$) are assessed least.

STEM disciplines

Regarding the STEM disciplines, most methods focus exclusively on one STEM discipline ($n = 13$). Five studies are concerned with more than one STEM discipline of which two address all STEM disciplines: Design Portfolio's (Bartholomew et al., 2019) and Picture Problem Solving Task (Fusaro & Smith, 2018).

Technology is the most addressed discipline ($n = 11$). Assessment methods related to technology typically require children to retrieve hard-to-reach objects, as in the Box task (German & Defeyter, 2000), Stringent Solution-Standard Tasks (Hong & Milgram, 1991), Toy Retrieval Task (Gardiner et al., 2012). Other methods require children to generate tool-based solutions for object retrieval, such as the Open-Ended Story Problems (Alexander et al., 1994) and PMST's (Leikin, 2013; Van Elsen et al., 2024). The Picture Problem Solving Task (Fusaro & Smith, 2018) requires children to come up with ideas to solve a problem such as getting a strawberry out of an ice cube or finding a way to discriminate between broccoli seeds and carrot seeds that are dropped on the ground and got mixed up. In the Tool Improvisation Tasks (Parsonson & Baer, 1978), children are challenged to use unconventional materials to replace conventional tools.

Table 7. Steps of the problem-solving process addressed in problem-solving assessment methods

Assessment method	Studies	Exploring	Under-standing	Repre-senting	Formu-lating	Planning	Execu-ting	Moni-toring	Reflec-ting	Total (N)
Air Pressure Tasks	Guevara et al. (2020)*						x			1
Box Tasks	German and Defeyter (2000)*						x			1
Creating Equal Number Task (CEN)	Leikin (2013); Leikin and Tovli (2014); Tsamir et al. (2010)*				x		x			2
Design Portfolio	Bartholomew et al. (2019)*			x	x	x	x	x	x	6
Engineering Preschool Children Observation Tool (EPCOT)	Anggoro et al. (2021)*	x	x		x	x	x	x	x	7
Mathematical Problem-Solving Measure	Vessonen et al. (2024)*						x			1
Metacognitive Skill Tasks	Annevirta and Vauras (2006)*					x	x	x	x	4
Open-Ended Problem-Solving Construction Task	Shechter et al. (2021)*	x	x		x	x	x	x	x	7
Open-Ended Story Problems	Alexander et al. (1994)*				x					1
Pictorial Multiple Solutions Tasks (PMST 1)	Leikin (2013)*; Leikin and Tovli (2014)				x					1
Pictorial Multiple Solutions Tasks (PMST 2)	Van Elsen et al. (2024)*				x					1

Assessment method	Studies	Exploring	Under-standing	Repre-senting	Formu-lating	Planning	Execu-ting	Moni-toring	Reflec-ting	Total (N)
Picture Problem Solving Task	Fusaro and Smith (2018)*, Lin et al. (2021), Lin et al. (2023)				x					1
Problem-Solving Performance Instrument	Nam et al. (2019)*						x			1
Problem-Solving Scale in Science Education (PSSSE)	Unal and Aral (2014a)*; Unal and Aral (2014b); Unal and Saglam (2018)				x					1
Problem-Solving Skill Scale (PSSS) (Oğuz & Köksal- Akyol, 2015)*	Cakir et al. (2021) ; Demirel and Deretarla Gul (2021)				x					1
Stringent Solution-Standard Tasks	Hong and Milgram (1991)*				x					1
Tool Improvisation Tasks	Parsonson and Baer (1978)*				x		x			2
Toy-Retrieval Tasks	Gardiner et al. (2012)*					x	x			2
	Total (N)	2	2	1	12	5	11	4	4	

Note. *Original study in which the assessment method was developed or first described.

Table 8. STEM disciplines addressed in problem-solving assessment methods

Assessment method	Studies	Science	Technology	Engineering	Mathematics	Total (N)
Air Pressure Tasks	Guevara et al. (2020)*	x	x			2
Box Tasks	German and Defeyter (2000)*		x			1
Creating Equal Number Task (CEN)	Leikin (2013); Leikin and Tovli (2014); Tsamir et al. (2010)*				x	1
Design Portfolio	Bartholomew et al. (2019)*	x	x	x	x	4
Engineering Preschool Children Observation Tool (EPCOT)	Anggoro et al. (2021)*	x		x		2
Mathematical Problem-Solving Measure	Vessonen et al. (2024)*				x	1
Metacognitive Skill Tasks	Annevirta and Vauras (2006)*			x		1
Open-Ended Problem-Solving Construction Task	Shechter et al. (2021)*			x		1
Open-Ended Story Problems	Alexander et al. (1994)*		x			1
Pictorial Multiple Solutions Tasks (PMST 1)	Leikin (2013)*; Leikin and Tovli (2014)		x			1
Pictorial Multiple Solutions Tasks (PMST 2)	Van Elsen et al. (2024)*		x			1
Picture Problem Solving Task	Fusaro and Smith (2018)*, Lin et al.	x	x	x	x	4

Assessment method	Studies	Science	Technology	Engineering	Mathematics	Total (N)
	(2021), Lin et al. (2023)					
Problem-Solving Performance Instrument	Nam et al. (2019)*				x	1
Problem-Solving Scale in Science Education (PSSSE)	Unal and Aral (2014a)*; Unal and Aral (2014b); Unal and Saglam (2018)	x				1
Problem-Solving Skill Scale (PSSS) (Oğuz & Köksal- Akyol, 2015)*	Cakir et al. (2021) ; Demirel and Deretarla Gul (2021)	x	x			2
Stringent Solution-Standard Tasks	Hong and Milgram (1991)*		x			1
Tool Improvisation Tasks	Parsonson and Baer (1978)*		x			1
Toy-Retrieval Tasks	Gardiner et al. (2012)*		x			1
	Total (N)	6	11	5	5	

Note. *Original study in which the assessment method was developed or first described.

Three out of four methods that include engineering are hands-on and require children to build or construct solutions to a problem. For the Engineering Preschool Children Observation Tool (EPCOT, Anggoro et al., 2021), children were asked to create a solution for Panda who wants to play outside when it is too hot and bright. In the Metacognitive Skill Task (Annevirta & Vauras, 2006), children are asked to build a fence to protect inhabitants of an island against a dangerous wolf and, in another task, to separate the island in three parts using building bricks (Lego) and crafts materials. As discussed earlier, the Open-Ended Problem-Solving Construction Task (Shechter et al., 2021) required children to build a bridge over a river using standard Lego (Duplo) building bricks. The Picture Problem Solving Task (Fusaro & Smith, 2018) is the only method including engineering in which data is collected by oral examination.

Science-oriented problem solving is most often combined with technology or engineering. In the Air Pressure Task (Guevara et al., 2020) children had to build an air pressure mechanism with syringes and tubes of different sizes and shapes to reach a predefined goal. Solving the problem requires applying scientific and technological knowledge and skills. The Air Pressure Task is the only included science related method which is hands-on. The other three science-related methods are based on oral examination.

Among the large number of well-structured mathematical problems (e.g., mathematical operations, mathematical word problems, seriation problems, unambiguous classification problems) that came up during screening, we found three methods that made use of ill-structured mathematical problems or problems with more than one correct answer: Creating Equal Number Task (CEN, Tsamir et al., 2010), Mathematical Problem-Solving Measure (Vessonen et al., 2024), and Problem-Solving Performance Instrument (Nam et al., 2019). CEN (Tsamir et al., 2010) required children to distribute a number of bottle caps so that there were an equal number of bottle caps on each side of a table. After the first rearrangement was done, the children were asked if they could think of another way to rearrange the bottle caps. Not only the solutions the children came up with were assessed, but also the strategies they applied to rearrange the bottle caps. In addition to well-structured mathematical problems related to categorising, comparing, patterning, and numbering, the Problem-Solving Performance Instrument (Nam et al., 2019) requires children to compare the length of lines using crafts materials (e.g., pieces of paper, paper clips, strings, pipe cleaners). After finding a solution, children are asked to think of multiple ways to find out which line is longer. Likewise, the Mathematical Problem-Solving Measure (Vessonen et al., 2024) includes both open-ended problems in addition to closed ones. For the open-ended problems, children were asked to sort different objects into groups in different ways and to explain the criteria they used for sorting. In another item they were asked to organise different coloured ice cream balls in a paper cone in multiple ways or to match shoes with monsters in multiple combinations.

The Design Portfolio's (Bartholomew et al., 2019) explicitly address problem solving in integrated STEM. As mentioned above, children are asked to design a solution to problem that emerges in popular children's rhymes. This requires them to move beyond individual STEM disciplines and navigate a largely unconstrained problem space.

Assessment procedures and psychometric properties of the assessment methods

The third research question relates to the procedures and psychometric properties of the test. This section presents the administration procedures (including format, administrator, location, and duration) and reported types and metrics related to reliability and validity. Table 9 provides an overview of key aspects concerning procedures and psychometric properties of the assessment methods as reported in the studies.

Assessment procedures

In most studies, the assessment was administered by an external assessor, not affiliated with the school of the participants, including researchers, (*n* = 13), graduate students (*n* = 2), and research assistants (*n* = 1). In two studies, the preschoolers' own teacher conducted the assessment (Nam et al., 2019; Van Elsen et al., 2024). Six studies did not specify who administered the assessment.

In seven of the ten studies that reported the location, the assessment took place in a quiet, separate classroom. The three other assessments were conducted in the participant's own classroom. More than half of the studies (58%), did not provide any information on the assessment location.

The reported duration of assessment administration ranged from 3 to 180 minutes. Approximately half of the studies that provided information on duration reported a time frame between 10 and 30 minutes. Three studies indicated a duration of more than 30 minutes. In two studies, participants were not subject to any time restrictions.

In all but two studies, the assessment was administered individually. Guevara et al. (2020) assessed the children in pairs and in the study of Anggoro et al. (2021) the assessments were carried out classically.

Table 9. Assessment procedures and types of reliability and validity reported in the included studies

Study	Assessment method	Adminis-trator	Location	Duration (min.)	Format	IRR	TRR	Internal Consist.	Face Val.	Construct Val.	External Val.	Ecol. Val.
Alexander et al. (1994)	Open-Ended Story Problems	Researcher	*NA*	*NA*	Individual	x						
Anggoro et al. (2021)	Engineering Preschool Children Observation Tool (EPCOT)	Researcher	*NA*	*NA*	Classical	x						
Annevirta and Vauras (2006)	Metacognitive Skill Tasks	Researcher	*NA*	35–45	Individual	x		x				
Bartholomew et al. (2019)	Design Portfolio	Researcher	*NA*	Ca. 60	Individual	x						
Cakir et al. (2021)	Problem-Solving Skill Scale (PSSS)	Researcher	*NA*	*NA*	NA		x	x	x	x		
Demirel and Deretarla Gul (2021)	Problem-Solving Skill Scale (PSSS)	Researcher	*NA*	*NA*	NA		x	x	x	x		
Fusaro and Smith (2018)	Picture Problem Solving Task	Researcher	*NA*	*NA*	Individual	x					x	
Gardiner et al. (2012)	Toy-Retrieval Tasks	Researcher	Quiet room	Unlimited	Individual							
German and Defeyter (2000)	Box Tasks	Researcher	Classroom	Max. 180	Individual	x					x	

Study	Assessment method	Adminis-trator	Location	Duration (min.)	Format	IRR	TRR	Internal Consist.	Face Val.	Construct Val.	External Val.	Ecol. Val.
Guevara et al. (2020)	Air Pressure Tasks	Researcher	Quiet room	20–25	In pairs							
Hong and Milgram (1991)	Stringent Solution-Standard Tasks	*NA*	*NA*	Unlimited	Individual							
Leikin (2013)	Creating Equal Number Task (CEN); Pictorial Multiple Solutions Tasks (PMST 1)	*NA*	*NA*	*NA*	*NA*							
Leikin and Tovli (2014)	Creating Equal Number Task (CEN); Pictorial Multiple Solutions Tasks (PMST 1)	*NA*	*NA*	*NA*	*NA*							
Lin et al. (2021)	Picture Problem Solving Task	Graduate students	*NA*	*NA*	Individual	x						
Lin et al. (2023)	Picture Problem Solving Task	Graduate students	Quiet Room	10–15	Individual	x						
Nam et al. (2019)	Problem-Solving Performance Instrument	Teacher	*NA*	Ca. 25	*NA*	x			x			

Study	Assessment method	Adminis-trator	Location	Duration (min.)	Format	IRR	TRR	Internal Consist.	Face Val.	Construct Val.	External Val.	Ecol. Val.
Parsonson and Baer (1978)	Tool Improvisation Tasks	Researcher	Quiet room	Ca. 15	Individual	x						
Shechter et al. (2021)	Open-Ended Problem-Solving Construction Task	Researcher	Quiet room	Ca. 20	Individual	x			x			
Tsamir et al. (2010)	Creating Equal Number Task (CEN)	Researcher	Classroom	NA	Individual							
Unal and Aral (2014a)	Problem-Solving Scale in Science Education (PSSSE)	*NA*	Quiet room	10–15	Individual	x	x	x	x	x		
Unal and Aral (2014b)	Problem-Solving Scale in Science Education (PSSSE)	*NA*	*NA*	*NA*	*NA*		x	x		x		
Unal and Saglam (2018)	Problem-Solving Scale in Science Education (PSSSE)	*NA*	*NA*	*NA*	*NA*		x	x				

Study	Assessment method	Adminis-trator	Location	Duration (min.)	Format	IRR	TRR	Internal Consist.	Face Val.	Construct Val.	External Val.	Ecol. Val.
Van Elsen et al. (2024)	Pictorial Multiple Solutions Tasks (PMST 2)	Teacher	Classroom	Ca. 3	Individual	x						x
Vessonen et al. (2024)	Mathematical Problem-Solving Measure	Researcher	Quiet room	Ca. 30	Individual			x				
Information provided	*N* (%)	18 (75)	10 (42)	13 (54)	17 (71)	13 (54)	5 (21)	7 (29)	5 (21)	4 (17)	2 (8)	1 (4)
NOT provided	*N* (%)	6 (25)	14 (58)	11 (46)	7 (29)	11 (46)	19 (79)	17 (71)	19 (79)	20 (83)	22 (92)	23 (96)

Note. This table only presents psychometric properties that were provided by at least one of the included studies. IRR: Interrater reliability; TRR: Test-retest reliability; Val: Validity; Consist: Consistency; Ecol: Ecological; *NA*: Not assigned.

Psychometric properties of the assessment methods

As shown in Table 9, interrater reliability was assessed and reported in more than half of the studies (54%). In these studies, interrater reliability was analysed by comparing scores or codes provided by multiple raters and calculating Cohen's kappa (n = 4), percentage of agreement (n = 3), or intraclass correlation (ICC, n = 2). Unal and Aral (2014a) reported Wilcoxon Signed Rank Test Scores as a measure for interrater reliability and German and Defeyter (2000) stated that there were no disagreements between two raters. The reported kappa's range from .60 (Bartholomew et al., 2019) to .81 (Lin et al., 2021). The percentage of agreement ranged from 67 to 100% (Parsonson & Baer, 1978) and the ICC ranged from 0.69 to 1 (Van Elsen et al., 2024). Van Elsen et al. (2024) observed better interrater agreement between four teachers (ICC = 1) than between four researchers (ICC = .69-.89). Test-retest reliability was discussed in five studies. The reported correlations between tests ranged from .60 (Cakir et al., 2021) to .96 (Unal & Aral, 2014a).

Internal consistency was reported in seven studies. Cronbach's alpha ranged from .70 to .90 (Shechter et al., 2021). Vessonen et al. (2024) reported that internal consistency was acceptable for children aged 4 and 5 but fell below acceptable levels for 3-year-olds.

Content validity was established by involving two (Nam et al., 2019), six (Shechter et al., 2021), or ten (Unal & Aral, 2014a) experts in preschool education and research to evaluate the instrument (face validity). Two studies calculated the Content Validity Index which ranged from .96 to .99 (Cakir et al., 2021; Demirel & Deretarla Gul, 2021). The remaining studies (n = 19) did not report on content validity. Construct validity was assessed with exploratory factor analysis (EFA) in four studies. External validity was briefly addressed by German and Defeyter (2000) and Fusaro and Smith (2018). Only Van Elsen et al. (2024) discussed issues related to ecological validity. Six included studies did not report any issues or statistics concerning validity and reliability.

Discussion

Problem solving is a key competence for lifelong learning (European Commission, 2019). Learning to solve problems takes time (Molnár et al., 2013). Therefore, teaching children how to solve problems best starts from preschool onwards (Burns et al., 2025; Diamond, 2018; Rogers, 2004). Problem solving is a fundamental component of STEM education and STEM activities offer numerous opportunities for young children to engage with ill-structured problems and develop essential problem-solving competencies. (Priemer et al., 2020). To monitor preschoolers' development of problem-solving competencies and evaluate the effectiveness of classroom interventions, researchers and educators need reliable and valid assessment methods. Up to now, an overview of methods that are developed, studied, or used in scientific research to assess problem-solving competencies within the context of early STEM education research was lacking. With this scoping review, we aimed to provide such an overview and analysed the characteristics (RQ 1), scope

(RQ 2), procedures, and psychometric properties (RQ 3) of these methods. The results can be used by researchers and educators to select, implement, or develop reliable and valid methods for assessing STEM-based problem solving in preschool.

A systematic search in Web of Science, Scopus and Eric resulted in 24 eligible studies in which 18 different methods related to problem solving were identified. We found that the typical method is based on direct observation of hands-on problem assignments or is based on oral examination, resulting in binary outcomes or count scores and focuses on a single step of the problem-solving process in a single STEM discipline. In addition, we found that the assessments are generally conducted individually, in a separate classroom, by an external researcher, and last about 10 to 30 minutes. Finally, it became evident that most assessment instruments had been developed on an ad hoc basis, with limited consideration given to reliability and validity. In the following sections, we elaborate on the main findings and discuss some limitations of our study. We end with implications and recommendations for research and practice.

Characteristics of assessment methods

Our scoping review of the scientific literature revealed that researchers most commonly use either observations of hands-on activities or oral examinations to assess the problem-solving competencies of preschoolers. Both methods have their merits. Observation allows teachers and researchers to record what preschoolers effectively do in a real-world context. It does not depend heavily on a certain level of verbal skills, it provides opportunities to record both verbal and non-verbal behaviour, and it offers opportunities to map social interactions during performance (Azevedo, 2009). Brassard and Boehm (2007) argue that observation of meaningful, unstructured tasks is an essential part of any assessment of young children.

Oral examination, on the other hand, can be used to assess (meta-)cognitive processes and competencies that are difficult to observe, such as coming up with multiple solutions or reflecting. Engaging in dialog with children can elicit reflective thinking (Stoll et al., 2012) and allows researchers and educators to verify whether preschoolers have correctly understood the problem and supports the interpretation of their observed behaviour (Snow & Van Hemel, 2008).

Hands-on assignments often make use of easily available and familiar crafts materials (e.g., cardboard, tape, glue, bottle caps, pipe cleaners) or popular toys (e.g., Lego building bricks). This allows children to build on their previous experience with the material's properties and possible uses. When unfamiliar materials are being used, it is suggested to allow them more or even unlimited time to explore the materials and their features (Gardiner et al., 2012). Another benefit of using ubiquitous materials for the assessment is that this makes the assessment method applicable across preschool classrooms, which facilitates the use of the methods in multiple studies and daily practice.

The use of visual representations, such as pictures depicting problem situations, emerged as a common practice in problem-solving assessments. Given their developing language skills, pictures can help preschoolers correctly interpret the problem. As stated by Polya (1945), a problem that is not well understood, is impossible to solve. Whereas older students can be encouraged to create their own representation of the problem (OECD, 2013), preschoolers may benefit from visual support.

Most assessment methods, whether observation-based or based on oral examination, result in quantitative outcomes (binary score, count score, or scale score). Although five studies applied qualitative coding, the resulting categories were in most cases used for quantitative analysis. Among the reviewed studies, only the Design Portfolios (Bartholomew et al., 2019) included qualitative outcomes such as children's drawings and photographs. In addition to other methods, portfolio assessment can be a powerful tool for researchers and educators who want to monitor the development and learning of young children (Seitz & Bartholomew, 2008).

Scope of the assessment methods

Altogether, the assessment methods included in this review covered all the steps of the PISA problem-solving framework (OECD, 2013). However, our review reveals that the individual methods generally focus on one or at most two steps of the problem-solving process. We only found three methods that include roughly the entire problem-solving process. However, problem solving is more than the sum of isolated knowledge and skills (Jonassen, 1997; Maker, 2001; Martín-Páez et al., 2019). It requires metacognitive and practical skills as well as certain attitudes and the willingness to engage with the problem that cannot be assessed with simple tasks (Mayer, 1998; OECD, 2013). To become proficient problem-solvers, children must learn to integrate the different steps of the problem-solving process (Polya, 1945). This highlights the need for a more integrated approach to assessing problem-solving competencies in preschoolers.

The same applies for the assessed STEM disciplines. Although STEM is often used as an umbrella term for educational activities within the four STEM disciplines (Martín-Páez et al., 2019), in the context of solving ill-structured STEM problems, STEM is best conceptualised as an integrative approach to assessing and teaching science, technology, engineering, and mathematics (Wan et al., 2021). Ill-structured problems are embedded in their context and often require knowledge and processes from more than one STEM discipline. However, most assessment methods included in this review focus on one discipline. We found five methods that address more than one STEM discipline, but only the Design Portfolio (Bartholomew et al., 2019) is explicitly embedded in an integrated STEM approach. These results are in line with the concern of Douglas et al. (2020) that educational assessment in STEM is often directed at domain-specific knowledge and skills. In formal education, segregated STEM problems (i.e., problems involving only knowledge and skills from one STEM discipline) are typically well-structured, resulting in one best and known solution, and are associated with direct instruction and a focus on content (Nadelson & Seifert, 2017). According to Martín-Páez et al. (2019), an isolated approach to the STEM disciplines is far removed from authentic, ill-structured STEM problems. Therefore, there remains a need for assessment methods that incorporate ill-structured problems and integrate the STEM disciplines.

Although initially underrepresented (Kelley & Knowles, 2016), technology and engineering have gained attention in recent years and are now seen as an important component of science and mathematics (Bryan & Guzey, 2020). Technology is by far the most addressed STEM discipline in the assessment methods included in this review. Of all STEM disciplines, technology is perhaps the most visible aspect of real-life problems (Aydin Gunbatar et al., 2022). Technology is also the most 'tangible' STEM discipline, making it more accessible to preschoolers. Technology is instrumental for assessing and teaching other STEM disciplines.

Engineering, on the other hand, is somewhat less represented in this review. According to Martín-Páez et al. (2019), engineering can bring all STEM-disciplines together. Engineering is strongly related to design. Two studies included in this review explicitly relate problem solving to the engineering design process (Anggoro et al., 2021) or design thinking (Bartholomew et al., 2019). Both methods cover nearly the entire problem-solving process. Although design is not part of the STEM-acronym, engineering design is “the interdisciplinary glue” between all STEM disciplines (Tank et al., 2018, p. 183) and has the potential to integrate all steps of the problem-solving process (English, 2023).

Assessment procedures and psychometric properties of assessment methods

Most assessment methods described in the included studies were conducted individually in a quiet, controlled environment under the supervision of a researcher or assistant. This aligns with the observation of Rhodes et al. (2024) that problem-solving competencies in preschoolers are typically assessed in controlled settings, removed from the authentic classroom context. However, Brassard and Boehm (2007) and DeLuca et al. (2020) emphasise the importance of embedding preschool assessment in authentic learning environments.

Consistent with the findings of Cutumisu et al. (2019), most studies implemented a new assessment method tailored to their specific needs, while giving limited consideration to key issues related to the method’s reliability or validity. In line with Barghaus et al. (2022) and Burns et al. (2025), our review reveals that limited attention is given to the psychometric properties of instruments that are used in research to assess problem-solving competencies in preschoolers. A quarter of the reviewed studies lack any reporting on the reliability and validity of the assessment instruments used. Although half of the studies report statistics on interrater reliability, other aspects related to reliability and validity are often overlooked.

Consistent with Burns et al. (2025), we found that reliability and validity were only systematically examined for the Problem Solving Scale in Science Education (PSSSE) developed by Unal and Aral (2014a). Here we add that Cakir et al. (2021) and Demirel and Deretarla Gul (2021) also address multiple aspects concerning reliability and validity in their respective studies, albeit in a more limited and less systematic manner. These findings underscore the need for reliable and valid problem-solving assessment methods in early childhood, as recently emphasised by Burns et al. (2025).

Limitations

For this review, we followed the rigorous JBI and PRISMA standards for systematic reviews. Nevertheless, this scoping review is not without limitations. We only included peer reviewed journal articles. Although scoping reviews allow for the inclusion of grey literature (e.g., Pollock et al., 2021), it was not possible to provide a systematic and exhaustive overview of problem-solving assessments based on grey literature due to limited accessibility and variety of this type of evidence. In addition, grey literature does not always meet the required quality criteria concerning reliability and validity. As a result, this review may be subject to publication bias since we may have missed interesting assessment methods that were not published in scientific journals or indexed by Scopus, Web of Science, and Eric.

There is also a potential risk of bias due to the study selection. For this review, we only included studies that used assessment methods involving ill-structured or open-ended problems. However, the classification of problems in ill-structured and well-structured problems is not unambiguous. As Jonassen (2010a) points out, all problems can be situated on a continuum of structuredness. Consequently, we had to make informed judgments about the inclusion of studies that focused on problems situated midway along the continuum. To ensure consistency, we compiled a list of types of problems encountered during the initial screening round. This list was then discussed with the research team, and decisions about inclusion and exclusion were made by consensus. The resulting list was subsequently used during the second round of screening.

Finally, it is worth noting that we deviated from our initial protocol in two ways. In scoping reviews, study selection and data extracting is often an iterative process (Aromataris & Munn, 2020; Levac et al., 2010; Pollock et al., 2021). Given the diversity of problems used to assess problem-solving competencies, we slightly adjusted the eligibility criteria by specifying eligible problem types and conducted a second round of screening. After the first round of screening, we also decided to adjust the stopping criteria during title and abstract screening. Rather than stopping after 50 consecutive rejections as initially planned and recommended by Ros et al. (2017), we chose to continue until 100 consecutive rejections to minimise the risk of overlooking relevant studies.

Implications for practice and research

The current review illustrates the tendency to approach problem-solving assessment in a segregated STEM structure. In most K-12 education systems, STEM is discipline-based, while STEM problems are not (Nadelson & Seifert, 2017). Rethinking the historically rooted division of STEM disciplines can be conducive to assessing and teaching competencies that are needed for solving authentic, ill-structured STEM-problems. According to Nadelson and Seifert (2017), researchers and educators also need the knowledge and willingness to introduce ill-structured or open-ended problems in the classroom and feel comfortable with the associated uncertainties. Allowing children to engage in ill-structured problems means giving up discipline-boundaries and control of the problem space.

We also note that there is still a need for assessment methods that can be administered multiple times and can be used for monitoring the development of preschoolers' problem-solving competencies in longitudinal studies or studies with a pretest, post-test, retention test design. Most methods included in this review can only be administered ones. Problems previously encountered and solved, are no longer real problems because preschoolers can build on prior experience and are expected to already know a solution or a solution strategy.

The study revealed a notable lack of assessment instruments for which reliability and validity have been thoroughly examined. This impedes research on the development of children's problem-solving competencies and the effectiveness of problem-solving interventions. To move the field forward, there is a clear need for developmentally appropriate, reliable, and valid assessment methods that reflect children's real-world problem-solving competencies. We also found that researchers often develop ad hoc assessment tools tailored to the specific needs of their studies. Reusing existing instruments can facilitate the validation of these instruments. Alternatively, new instruments can be developed in accordance with the guidelines outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Among the 18 methods we found, Bartholomew's' (2019) Design Portfolio stands out in several ways. First, it captures nearly the entire problem-solving process by asking children to draw their ideas (representing and formulating), plan their actions (planning), effectively build the design (executing), and evaluate if the design worked (monitoring and reflecting). Although exploring and understanding are not included in the assessment, they can easily be added since the assignments already elicits exploratory behaviour by asking children explicitly to explore the materials. Preschoolers' understanding can be derived from their drawings and follow-up questions. Second, the Design Portfolio explicitly addresses STEM in an integrated way. Third, it makes use of tangible materials and follow-up questions to get a better understanding of what has been observed. Although the Design Portfolio as such may not be directly applicable as an assessment method in preschool practice due to the necessity of qualitative coding, it offers a promising mix of elements that can be further refined and developed into effective assessment tools.

Observing children solve engineering design problems and engaging in dialogue with them seems another promising way to assess their problem-solving competencies. Engineering design problems are motivating, open-ended, hands-on problems that require children to construct or craft solutions by going through all steps of the problem-solving process (Cunningham et al., 2018; Dubosarsky et al., 2018; Isabelle et al., 2021), incorporating all STEM disciplines (Tank et al., 2018). These types of problems closely align with preschoolers' play and daily activities in preschool classrooms, which makes them suitable for authentic preschool assessment.

Conclusion

This scoping review provides an overview of the characteristics, scope, procedures, and psychometric properties of assessment methods that are developed, studied, or used in scientific research to assess preschoolers' problem-solving competencies. The findings support researchers and educators in selecting, implementing, or developing reliable and valid instruments, ultimately contributing to a more comprehensive understanding of these competencies. Our review of the scientific literature revealed that there is still a need for scientifically substantiated, reliable, and valid methods to assess STEM-based problem solving in an integrated way. The review also highlights the need for increased attention to and transparency concerning the procedures and psychometric properties of assessment methods. Based on our findings, we propose combining observation of hands-on problem assignments with oral examination as an age-appropriate method for assessing preschoolers' problem-solving competencies.

STUDY 3 |
PROSPER: A COMPREHENSIVE, VALID, AND RELIABLE INSTRUMENT TO OBSERVE PROBLEM-SOLVING BEHAVIOURS IN PRESCHOOLERS

This chapter has been published as:

Van Elsen, J., Buckers, L., Van Tricht, C., Torbeyns, J., & De Maeyer, S. (2025). PROSPER: A comprehensive, valid, and reliable instrument to observe problem-solving behaviours in preschoolers. *Thinking Skills and Creativity, 45*, Article 101940. https://doi.org/10.1016/j.tsc.2025.101940

Abstract

Background: Problem solving is an important yet challenging competence that is required in STEM education and is highly valued in STEM careers. Learning to solve problems should start at an early age. However, to optimally support young children in developing their problem-solving competencies, there is a need for instruments to measure these competencies in an authentic context.

Aims: The present study aims to develop the Problem-solving Observation Scheme for Playful Engineering and the Real-world (PROSPER), a comprehensive, valid, and reliable observation scheme to make problem solving visible in preschool children.

Methods: Twelve preschool children aged 4–6-years individually solved three engineering design problems. All sessions were videotaped, resulting in 17.33 hours of recordings. Five researchers and five preschool teachers identified 1150 problem-solving behaviours through video-based observations. Identified behaviours were coded iteratively, based on the PISA problem-solving framework. Psychometric properties of the observation scheme were analysed.

Results: PROSPER consists of 33 behaviours, covering the whole problem-solving process. The PISA framework was extended with four generic, overarching behaviours: commitment, perseverance, autonomy, and responsiveness. Strong intercoder reliability was established. Across task consistency was high with small differences between the three problems.

Conclusions: PROSPER proved to be a comprehensive, valid, and reliable instrument to capture preschoolers' problem-solving behaviours in an authentic setting. Future studies can rely on PROSPER to investigate children's development of problem-solving competencies and the effectiveness of classroom interventions.

Introduction

Problem solving is an important and challenging set of competencies that is required for lifelong learning (European Commission, 2019) and highly valued STEM (Science, Technology, Engineering, and Mathematics) careers (Tytler, 2020). Developing problem-solving competencies takes years. Therefore, fostering problem-solving competencies best starts from early on (e.g., Malcok & Ceylan, 2022; Rogers, 2004). Young children can learn how to solve problems with tailored adult support (Diamond, 2018; Vygotsky, 1978). However, much is still unclear about how problem-solving competencies develop during the preschool years and what interventions are effective in supporting this development, especially when it concerns authentic, real-world problems.

To study the development of problem-solving competencies and the effectiveness of educational interventions, researchers need comprehensive, valid and reliable instruments to assess these competencies. Likewise, educators can benefit from such instruments to systematically monitor the development of preschool children's problem-solving competencies and to tailor their classroom activities to the children's needs.

Since Herbert Simon's (1973) article "The structure of ill-structured problems", there is an ongoing debate about the transferability of findings derived from solving well-structured problems (e.g., Tower of Hanoi) to ill-structured problems (e.g., authentic STEM-problems). Findings resulting from research based on well-structured problems may be foundational for studying ill-structured problems (Reed, 2016), but both types of problems require different cognitive processes (Jonassen, 2010a). Nowadays, researchers recognise the importance of integrating the context in which the problem is embedded (Csapó & Funke, 2017b). Nevertheless, problem-solving competencies are typically assessed using well-structured problems in controlled environments, often removed from the authentic settings of classrooms or homes (Rhodes et al., 2024). There remains a need for instruments that enable researchers and educators to assess real-world problem-solving competencies in context.

Systematic observation of hands-on problem-solving assignments enables the documentation of children's behaviours in naturalistic contexts, capturing both verbal and non-verbal actions without relying too much on children's language skills (Azevedo, 2009). Researchers emphasise the importance of incorporating purposeful observation of children within familiar environments as an integral component of any preschool assessment (Brassard & Boehm, 2007; Monteira & Jiménez-Aleixandre, 2016).

Engineering design problems are particularly well-suited for observing children's problem-solving behaviours. Engineering design problems are authentic, real-world problems that require the application of STEM ideas to be solved (Arık & Topçu, 2022; Moore et al., 2014). They bear strong similarities to the everyday activities of children as they try to build structures with building blocks and outdoor materials (Geist, 2025; Shechter et al., 2021; Weiser, 2022) or craft things (Kangas et al., 2022).

To the best of our knowledge, there are currently no validated instruments that enable the systematic collection of evidence on preschoolers' problem-solving behaviours, that encompass the full problem-solving process, that can be applied across different problems, and that can be used repeatedly to monitor the development of preschoolers' problem-solving competencies. Therefore, the present study aims to make problem solving visible by developing a comprehensive, valid, and reliable observation instrument that researchers and educators can use to collect evidence on preschoolers' problem-solving competencies in the classroom.

Problem solving

Problem solving is a complex process that requires a variety of (meta-)cognitive, affective, and conative processes (Jonassen, 2000). It is often defined as "cognitive processing directed at transforming a given situation into a goal situation when no obvious method of solution is available to the problem solver" (Mayer, 1994, p. 248). According to Mayer, problem solving takes place in the solver's mind and can only be inferred from the solver's behaviour. Other authors argue that the solver's behaviour is not only a proxy of their internal problem-solving processes, but the very essence of problem solving itself (Skinner, 1953). Wheatley and Wheatley (1984) conceptualise problem solving as "What (…) you do when you don't know what to do" (p. 22). In many cases, both internal processes and physical actions (non-verbal and verbal) are needed to solve real-world problems. From the perspectives of embedded and embodied cognition, cognitive, physical, and environmental resources should be in concert with each other to efficiently solve problems (Pouw et al., 2014). In addition, the willingness to engage in problem solving and to persist when faced with difficulties are critical components of the problem-solving process (Jonassen, 2010a). In this study, we therefore conceptualise problem solving as the dynamic interplay between cognitive, metacognitive, physical, volitional, and social processes involved in resolving an unfamiliar task or situation.

Following Dewey (1910) and Polya (1945), many scholars have decomposed the problem-solving process into separate steps or phases. Most problem-solving phase models address mathematical problems (for an overview see: Rott et al., 2021; Salminen-Saari et al., 2021). In contrast to these domain-specific models, the PISA model is designed to measure cross-disciplinary problem-solving competencies (OECD, 2004, 2013). The PISA model addresses real-world problems that do not rely on domain-specific knowledge and skills. The PISA model is a more fine-grained elaboration of Polya's four phase model. It is particularly suitable for assessing problem solving in young children because it emphasises exploratory and creative processes that are essential for young children's successful problem solving (Evans, Todaro, et al., 2021). Moreover, the PISA framework's focus on authentic, cross-disciplinary problems aligns well with the goals of early STEM education (Moore et al., 2020). Table 10 presents the eight steps of the PISA problem-solving framework (OECD, 2013)

Table 10. Eight steps of the PISA problem-solving framework (adapted from OECD, 2013)

Step	Description
Exploring	Observing and interacting with the problem, searching information, and finding limitation and obstacles
Understanding	Understanding given and obtained information and demonstrating understanding of relevant concepts
Representing	Building a mental representation and constructing tabular, graphical, symbolic or verbal representations
Formulating	Generating hypothesis by identifying relevant elements of the problem and their interrelationships and critically evaluating information
Planning	Setting and clarifying (sub-)goals and devising a plan or strategy to reach that goal
Executing	Carrying out a plan
Monitoring	Monitoring progress towards the goal, checking (intermediate) results, identifying unexpected events, and making adjustments as required
Reflecting	Critically evaluating solutions from different perspectives, identifying additional information needs, and communicating progress

Observing problem-solving behaviours in preschoolers

Problem-solving behaviours can be defined as any observable, physical, verbal, or non-verbal activity that is triggered by the solver’s intention to solve a problem (Gunawan et al., 2019; Uher, 2016). Systematically observing preschoolers’ behaviours provides a window through which internal cognitive processes can be inferred (Mayer, 1994). It also provides educators with opportunities to assess preschoolers’ problem-solving competencies to improve and tailor their learning activities to fit the children’s needs (Brassard & Boehm, 2007). Ideally, assessment of preschoolers' competencies is embedded in authentic, play-based activities (DeLuca et al., 2020; Geist, 2025).

Playful STEM activities are particularly suitable to observe and develop problem-solving competencies in preschoolers (Geist, 2025; MacDonald et al., 2022). STEM is the interdisciplinary approach to learning and teaching science, technology, engineering, and mathematics (Johnson et al., 2020). Problem solving is inherent to all four STEM disciplines (Priemer et al., 2020) and is considered “the heart of STEM education” (Dubosarsky et al., 2018, p. 259). Engineering design is “the interdisciplinary glue” in STEM (Tank et al., 2018, p. 183). It is a creative process in which the engineer tries to craft or construct a solution that requires the application of mathematical and scientific principles and the use of technical devices. Engineering design problems are typically authentic, hands-on, ill-structured problems with multiple solutions, including specific constraints and success criteria (Arık & Topçu, 2022; Cunningham et al., 2018; Purzer & Douglas, 2018). Engineering design problems tap into children's natural curiosity and urge to explore the world around them (Isabelle et al., 2021; Moore et al., 2018). These problems also share many similarities with real-life situations during children’s construction play (Shechter et al., 2021). Therefore, Gold et al. (2021) argue that observing children’s engineering play is an integrative way to assess their cognitive abilities. Despite its potential, engineering has long been underrepresented in early STEM education (Bustamante et al., 2018; English, 2018).

The engineering design process parallels the problem-solving process in several ways, as it starts with establishing a problem or a need and aims to find a solution that meets certain criteria under certain constraints (Cunningham et al., 2018; Dubosarsky et al., 2018; Gold et al., 2021). According to Isabelle et al. (2021), the engineering design process provides a framework to help children develop their problem-solving skills. Recent studies revealed that preschoolers spontaneously engage in engineering-based problem solving during playful activities (e.g., Lippard et al., 2019; Yuan et al., 2024). Shechter et al. (2021) observed how 4- to 6-year-olds brainstormed, planned, tested, and improved their solution to a bridge building task.

In previous studies, various observation instruments and coding schemes have been developed to map aspects of preschool-aged children’s problem-solving processes in STEM. Escolano-Pérez et al. (2019) created a coding scheme to capture 5-year-olds’ planning, monitoring, and evaluating skills during solving a tangram puzzle. Bryce and Whitebread (2012) asked 5–7-year-olds to build a train track in a specified shape and observed planning, monitoring, and evaluating behaviours. Planning, monitoring, and evaluating behaviours are also at the heart of the MetaSCoPE coding scheme used by Marulis and Nelson (2021) to observe 3–5-year-olds building complex structures from multidimensional Wedgits blocks according to a depicted model. These three coding schemes make use of well-structured puzzle-problems and are embedded in research on metacognition, rather than problem solving, which explains the focus on monitoring and control.

Other coding schemes focus on problem solving within an engineering context. The Early EHoM Coding Scheme created by Shechter et al. (2021) addresses engineering habits of mind of 5–6-year-olds during a well-structured, open-ended bridge-building task, including elements related to understanding, formulating, planning, executing, and monitoring. Although Engineering Habits of Mind (EHoM) shares many similarities with problem solving, the EHoM Coding Scheme is specifically designed for this bridge-building task, making it difficult to apply the coding scheme to other tasks. The Engineering Preschool Children Observation Tool (EPCOT), developed by Anggoro et al. (2021), captures the problem-solving behaviours of 3–5-year-olds during STEM-related classroom activities including behaviours related to exploring, understanding, formulating, planning, executing, and monitoring. As one of the aims of EPCOT was to evaluate the use of vocabulary related to the engineering design process, it largely focuses on oral expression (e.g., repeat, describe, explain, articulate, propose, predict, discuss, state). In addition, the coding scheme largely focuses on knowledge about scientific phenomena.

None of the above-mentioned studies aimed to develop an observation scheme that systematically covers all the activities involved in the problem-solving process. For all mentioned instruments, intercoder reliability was investigated. Some studies also investigated correlations between codes (Shechter et al., 2021) or with other measures (Bryce & Whitebread, 2012; Marulis & Nelson, 2021). But none of the studies provide further elaboration on the psychometric properties of the observation scheme. This same lack of psychometric soundness has been identified by Lippard et al. (2017) as a major obstacle in early engineering research. Scrutinising critical aspects of the validity and reliability of an observational instrument is crucial to its use in research and practice. Therefore, there remains a need for a comprehensive observation instrument that researchers can use to investigate problem-solving behaviours in preschoolers across various, authentic contexts, with clearly established psychometric properties for key aspects of validity and reliability.

The present study

This study is part of a larger research project that investigates how formative assessment can support the development of problem-solving competencies in preschoolers (ISCED 02; UNESCO Institute for Statistics, 2012). To effectively assess and monitor this development, there is a clear need for comprehensive, psychometrically sound assessment methods that can be used repeatedly in classroom settings.

The present study aims to develop and establish the psychometric properties of the Problem-solving Observation Scheme for Playful Engineering and the Real-world (PROSPER), a comprehensive observation instrument for systematically observing preschoolers' problem-solving behaviours. PROSPER operationalises the PISA problem-solving framework for preschool-aged children by mapping and describing observable behaviours across the different steps of the problem-solving process. It offers researchers and educators valuable insights into how preschool children approach and solve ill-structured, real-world problems. This study elaborates on the development of the instrument and the establishment of structural validity (construct validity), content validity, ecological validity, intercoder reliability, and across-task consistency (generalisability).

Method

Design

PROSPER was developed based on videotaped observations of preschoolers individually solving three different engineering design problems. Five researchers and five teachers identified problem-solving behaviours in the recordings. The identified problem-solving behaviours were then coded and categorised according to the steps of the PISA framework. Finally, psychometric properties of the PROSPER were assessed.

Participants

Participants were 12 children (5 girls and 7 boys) between 4 and 6 years old (mean age = 66 months, *SD* = 5.45) enrolled in a Flemish (Belgium) preschool. Only children whose parents had given written consent were included. Parents were also asked to provide information on the children's gender, age, and linguistic background. Alongside the language spoken at school, four children spoke another language at home (2 English, 1 Arabic, 1 Polish).

The five authors identified problem-solving behaviours from the recordings. Additionally, five preschool teachers (all female) with 6 to 41 years of experience participated in identifying problem-solving behaviours from the recordings.

Data collection

Engineering design problems

Three engineering design problems were developed for this study. The Volcano Bridge Problem required participants to build a bridge over a papier-mâché volcano with only beer mats and toilet rolls (inspired by Shechter et al, 2021). In the Hook-a-Duck Problem, participants were asked to create a tool to get a ringed rubber duckling out of a transparent tube with functional and non-functional craft materials, cutlery, and tools (inspired by Evans, Todaro, et al., 2021). The Storm Problem required participants to build a stable for two sheep that can withstand the wind of a hairdryer using only natural materials (adapted from Thunder et al., 2022). Examples of solutions to the problems can be found in Appendix 1.

The three problems were purposefully selected and developed to be solved by building or crafting with readily available materials, to include three constraints or success criteria, and to allow for multiple solutions and solution paths (Cunningham et al., 2018; Dubosarsky et al., 2018). We assured that the three problems were diverse, enjoyable, required no prior domain-specific knowledge, and presented an appropriate challenge for preschoolers, aligned with their "zone of proximal development" (Vygotsky, 1978). The problems were designed based on the design parameters provided by Cunningham et al. (2018), discussed with preschool teachers, pilot tested, and fine-tuned with five children aged 4 to 6 to ensure that all children can get started and get stuck (low floor, high ceiling, see English, 2017; Resnick & Silverman, 2005). To ensure ecological validity, the problems closely resembled typical preschool classroom activities. All details about the problems are available on the Open Science Framework (OSF; https://osf.io/k6x5r/?view_only=da1682d0dbd94153b70c10ef188a7f3b).

Procedure

All participants solved the problems individually in a spare classroom at the participants' school during school hours. Participants were presented with the three problems at one-week intervals. No time limit was set to solve the problem. All data were collected by the first author.

After demonstrating the technical equipment, the participants were invited to explore the available materials and were asked to name all the objects. Next, the problem, the goal state, and success criteria were explained and visually demonstrated with assistance of Mr Rabbit, a soft bunny hand puppet. The hand puppet was used to put the children at ease, to motivate them, and to elicit talk (Dorie et al., 2013). Mr Rabbit acted as a naïve 'alien' puppet who didn't know how to solve the problem and asked for the child's help.

During the actual problem-solving process, interventions were kept to a minimum. The researcher only intervened when the participants lost engagement with the assignment, did not act in accordance with the assignment (e.g., participants violated constraints multiple times), asked for help, got really stuck (e.g., participants repeated the same action several times without success, or indicated that they really did not know what to do), or showed clear signs of upcoming frustration (e.g., participants became angry or sad). In these cases, the researcher followed a predefined protocol, responding with prompts that became progressively more concrete and directive. Prompts were given as questions (Wood et al., 1976). Only when this was not sufficient to overcome the impasse, more direct support was offered by demonstrating part of a potential solution.

Once a solution was found, the participants were asked to recall the problem and reflect on the outcome and the process (Anggoro et al., 2021; Epstein, 2003; Stoll et al., 2012), guided by Mr Rabbit. As in the Dorie et al. (2013) study, the puppet was asleep during the process and didn't know what had happened. This encouraged children to explain what they did and what happened to 'someone' who was not present. The reflective questions gradually increased in difficulty from simple closed questions (e.g., "Was it difficult?") to more complex open questions (e.g., "What was difficult?", "Why didn't it work?", "What else did you try?"). From the second problem onwards, the children were also asked to recall the previous problem(s) and solution(s) before the new problem was explained. The observations lasted on average 29 minutes (*range* = 12–47 min, *SD* = 9 min) resulting in 17.33 hours of footage. The full protocol can be found on OSF.

Recording materials

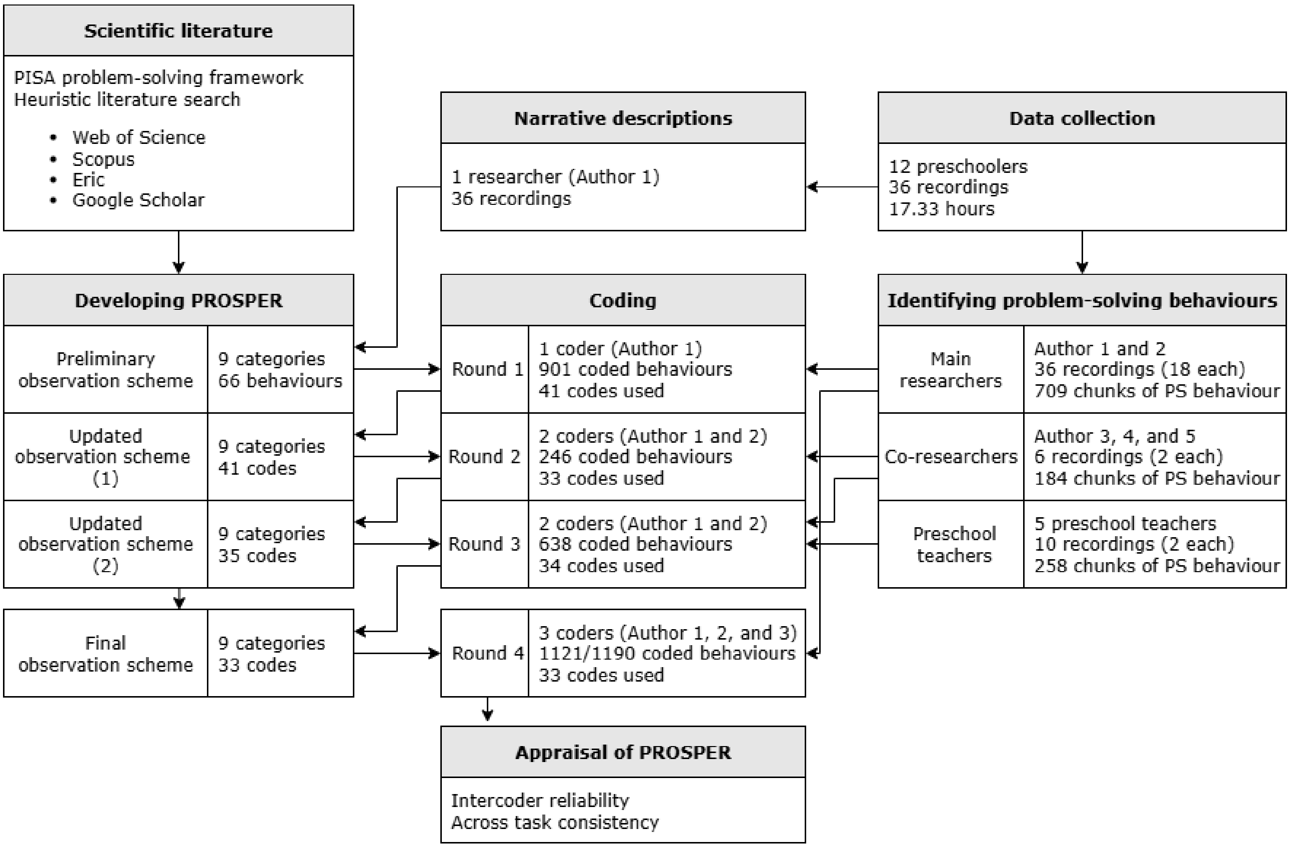
All sessions were recorded with two cameras (Logitech C922 Pro Stream Webcam) which captured the working space from two angles. The inbuild omnidirectional microphone of one of the camera's was used to register all speech and sounds. Footage from both cameras was processed by the open-source video recording application OBS Studio (version 30.2.2, www.obsproject.com).

Developing PROSPER

Developing PROSPER involved several stages. First, a preliminary observation scheme was drafted based on the existing scientific literature (e.g., Anggoro et al., 2021; Bryce & Whitebread, 2012; Marulis & Nelson, 2021; Shechter et al., 2021) and open narrative descriptions made during the initial screening of recordings. Second, problem-solving behaviours were identified from the recordings and briefly described. Then, these chunks of identified problem-solving behaviours were coded in several rounds (Saldaña, 2013). Figure 7 illustrates the different stages of the developmental process of the PROSPER.

As problem solving is a process, the eight steps of the PISA problem-solving framework where conceived as the main categories and the specific behaviours as the codes (Saldaña, 2013). As the codes refer to observable actions, gerunds (the "-ing" form of verbs) were used (Thornberg & Charmaz, 2014). The observation scheme was developed iteratively. Over the first three rounds of coding, codes were created, removed, moved, merged, and split (Thornberg & Charmaz, 2014). The final round of coding was used to establish intercoder reliability. This section outlines the development of PROSPER, focusing on the steps taken to ensure the instrument's comprehensiveness, validity, and reliability.

Figure 7. Developing PROSPER: flowchart



Creating a preliminary observation scheme

The eight steps of the PISA problem-solving framework were used as the main categories in the observation scheme (OECD, 2013). For every category, related problem-solving behaviours were looked up through heuristic literature search in the databases of Web of Science, Scopus, Eric, and Google Scholar. The list of problem-solving behaviours was further completed based on the field notes taken during data collection and the first author's open, narrative descriptions of all recordings (Saldaña, 2013). All behaviours were categorised across the eight PISA steps. A residual category was created for generic, overarching problem-solving behaviours that could not be assigned to any of the PISA steps.

Identifying problem-solving behaviours

A protocol and a structured observation form for identifying problem-solving behaviours were developed and piloted. The first two authors each watched half of the recordings. Two recordings were processed by both researchers. Doubtful cases were discussed and resolved in consensus to ensure uniform interpretation of the protocol. Problem-solving behaviours were described and categorised according to the PISA steps.

To ensure content validity and to avoid blind spots, the last three authors, and five preschool teachers also identified problem-solving behaviours in the recordings. All observers were trained and were provided with a protocol and a digital and paper version of the observation form. They each processed two different recordings independently. Recordings were purposefully chosen to maximise variation in problems, gender and home language of the preschoolers, and duration of the session. Divergent perspectives were analysed by comparing behaviours identified by the first two authors and the other observers.

Coding problem-solving behaviours

In the first round of coding, the preliminary observation scheme was applied by the first author to code all chunks of problem-solving behaviours identified by the first two authors. The aim of this first round of coding was to identify missing behaviours and to update the preliminary observation scheme. After coding and in consultation with the last two authors, the observation scheme was updated by adding, splitting, merging, moving, and deleting codes.

In Round 2, the updated observation scheme was used to code the behaviours identified by the three last authors. In this round, the first two authors independently coded all chunks of identified problem-solving behaviours. If the chunks described more than one action or behaviour, multiple codes were assigned, resulting in as many coded behaviours as actions described in the chunks of identified behaviours.

After coding, intercoder reliability was assessed by calculating percentage of agreement, Cohen's kappa based on pairwise deletion, and Cohen's kappa with missing values as a regular category (De Raadt et al., 2019). When one of the coders assigned more codes to a chunk, missing values were produced for the other. Cohen's kappa with missing values as a regular category is considered the lower bound as it is heavily biased downward (De Raadt et al., 2019). We aimed to achieve a percentage of agreement > 80% and a strong intercoder agreement with $\kappa >. 80$ as proposed by McHugh (2012) and a percentage of agreement > 80%. Intercoder reliability was assessed on the level of the categories as well as on the level of the codes. Intercoder reliability statistics are presented in Table 11. Differences between the coders were discussed between the two coders. Again, the observation scheme was adjusted by adding, splitting, merging, moving, or deleting codes.

In Round 3, the updated observation scheme was used by the first two authors to code all chunks of identified behaviours by the three last authors and the preschool teachers. Once again, intercoder reliability was assessed, discrepancies were discussed between the first two authors, and the observation scheme was adjusted. This resulted in the final observation scheme (PROSPER).

Psychometric properties of PROSPER

The psychometric properties being addressed for PROSPER include structural validity (construct validity), content validity, ecological validity, intercoder reliability, and across task consistency (generalisability). Structural validity, as part of construct validity, refers to the degree to which the observation scheme reflects the theoretical structure of the construct under investigation (Mokkink et al., 2010). Structural validity was established by grounding the observation scheme in the PISA problem-solving framework (OECD, 2013). As many other problem-solving models found in the scientific literature, the PISA model builds on the well-established phase model of Polya (1945).

Content validity refers to the degree to which the content of an instrument reflects the construct (Terwee et al., 2018). Content validity was ensured by aligning PROSPER with the eight steps of the PISA problem-solving framework (OECD, 2013) and by having five experienced preschool teachers observe two video recordings each and identify problem-solving behaviours. The involvement of practice experts ensured that the behaviours included in the observation scheme were relevant, representative, and meaningful within the preschool context (Boateng et al., 2018).

Ecological validity is part of external validity and refers to the degree to which results can be generalised to real-world contexts. Ecological validity was ensured by creating problems closely related to preschoolers' play activities, by using materials familiar to the participants, by mimicking authentic teacher-child interactions, by conducting the sessions in a familiar classroom environment at the participants' own school, and by involving teachers in the identification of problem-solving behaviours (Rhodes et al., 2024).

Intercoder reliability refers to the consistency between independent coders. This was established by having the first two authors independently code the problem-solving behaviours they had previously identified using the final observation scheme (Round 4). As the two coders had developed the observation scheme and were familiar with the data, a third coder (Author 3), who was not involved in developing the observation scheme and was unfamiliar with the data, coded the same dataset. Intercoder reliability was assessed by calculating percentage of agreement and Cohen's kappa for coding conducted between Author 1 and Author 2, and between Author 1 and Author 3. As before, we aimed to achieve a strong intercoder agreement with $\kappa > .80$ (McHugh, 2012) and a percentage of agreement > 80%. Intercoder reliability statistics of the final observation scheme are presented in Table 11.

Finally, across-task consistency or task dependency was assessed to determine whether the three problems elicited comparable levels of problem-solving behaviours. This was achieved by comparing the frequencies of coded problem-solving behaviours and their corresponding categories assigned by the first author in Round 4.

Quantitative analyses were conducted in R (Version 4.3.2, R Core Team, 2023). The package *irr* (Version 0.84.1, Gamer et al., 2019) was used to calculate intercoder reliability measures. All protocols are available on Open Science Framework (https://osf.io/k6x5r/?view_only=da1682d0dbd94153b70c10ef188a7f3b).

Results

A preliminary observation scheme

Based on the scientific literature and the narrative descriptions of the recordings, a list of 66 problem-solving behaviours was composed by the first author. The list consisted of 46 behaviours categorised in the eight PISA steps. Twenty behaviours were initially not categorised in one of the eight PISA categories and related to identifying problems, goal setting, decision making, engagement, persistence, autonomy, self-efficacy, motor skills.

Identifying problem-solving behaviours

Together, the first two authors identified and described 709 chunks of problem-solving behaviours in the 36 recordings (see Table 11). On average, 19.69 behaviours were identified per recording with a range of 15 to 27 ($SD = 3.28$). The last three authors identified 184 problem-solving behaviours in 6 recordings with an average of 30.67 described behaviours ($SD = 15$, $Min = 18$, $Max = 60$). The preschool teachers identified 258 behaviours in 10 recordings with an average of 25.8 described behaviours per recording ($SD = 7.8$, $Min = 14$, $Max = 41$). Sometimes all behaviours were neatly recorded separately as intended, sometimes detailed descriptions of sequences of actions combining several behaviours were provided. Two examples illustrate the differences in the registration of identified problem-solving behaviours between observers. The numbers between square brackets indicate the time stamp of the registered behaviour. Between brackets we indicate whether the behaviours were identified by a researcher or by a teacher, the observation number, and the problem.

> [33:25] Tests kitten on bridge.
> [34:00] Puts card on toilet roll (Trial and error).
> [34:34] Tests kitten on bridge.
> [35:55] Another toilet roll.
> [36:00] Tests kitten on bridge.
> [36:10] Places card on bridge.
> [36:55] Tests kitten on bridge.
> [37:27] Evaluates that this is a good solution. (Researcher, Observation 16, Volcano Bridge Problem)

> [7:30] Starts for the third time. Makes another attempt. Rebuilds the same tower again. Now places a beer coaster on top so the cat can sit on it. It falls over... Persists but does not come to new experiments or plans. Keeps making the same mistakes. (Teacher, Observation 18, Volcano Bridge Problem)

Table 11. Summary of the four rounds of coding with intercoder reliability statistics on the level of the codes and on the level of the categories (between brackets)

	Round 1	Round 2	Round 3	Round 4a	Round 4b
Dataset (Behaviours identified by)	Author 1-2	Author 3-5	Author 3-5 + teachers	Author 1-2	Author 1-2
Coders	Author 1	Author 1-2	Author 1-2	Author 1-2	Author 1/3
Codes used	41	33	34	33	33
Coded chunks	709	184	441	709	709
Coded behaviours	901	246	638	1121	1190
Missing values	25	88	145	245	290
Percentage of agreement	*NR*	86.71% (92.05%)	84.18% (88.57%)	89.84% (92.12%)	82.00% (86.50%)
κ pairwise deletion	*NR*	.856 (.893)	.827 (.858)	.893 (.907)	.810 (.841)
κ missing values as regular category	*NR*	.553 (.714)	.664 (.723)	.711 (.706)	.624 (.632)

Note. *NR*: Not Relevant; κ: Cohen's kappa.

A rich variety of problem-solving behaviours was observed, covering seven of the eight PISA steps. Exploring is the first step in the PISA framework and refers to deliberately gathering information about the problem, materials, and possibilities by observation and interaction with the problem and environment. Participants explored by observing, manipulating, describing, and comparing materials and associating them with other materials, potential application, or past experiences. We also observed children rummaging through provided materials to see what they can use and asking questions for more information. However, some children needed encouragement to explore.

> Names objects; describes if she does not know the name (bark); Associates (straw with cowboy); manipulates only some objects. (Researcher, Observation 31, Storm)

> Stares at the box of materials without touching or saying anything. Needs encouragement to explore. (Researcher, Observation 3, Volcano Bridge)

Demonstrating understanding of the problem, its key elements, constraints and success criteria is the second step. Although understanding is a mental process, we identified behaviours from which understanding can be inferred. We found that some participants listened and watched attentively during the problem description while others were already up and running before the problem was even properly explained. Children also confirmed understanding by nodding or verbal expression and by repeating or demonstrating the problem and criteria. Finally, understanding could be derived from the extent to which the children acted in accordance with the criteria.

> Listens and watches Mr Rabbit and Kitten/Vulcano; Gives verbal affirmation or nods to questions; Touches the volcano and pretends it burns him (demonstrates criteria); Can repeat and demonstrate criteria. (Researcher, Observation 17, Volcano Bridge)

> Indicates he is ready and demonstrates how the kitten is shot over the volcano using his canon. This is, of course, not in accordance with the assignment. (Researcher, Observation 34, Volcano Bridge)

Representing, creating a mental, verbal, visual, or physical representation of the problem or goal, was difficult to observe. The few behaviours that could possibly be attributed to representing, were more associated with demonstrating understanding, as in the excerpt below.

> Demonstrates what the kitten wants to do by going over the volcano with the kitten in the air. (Researcher, Observation 15, Volcano Bridge)

Formulating refers to coming up with possible ways to solve the problem and evaluating alternatives. Generating new ideas can be observed when preschoolers verbalise their ideas or when preschoolers try out a new solution. While children frequently generated new ideas throughout the entire problem-solving process, considering was only observed to a limited extent.

> Suggests what he can try (sticking tape to the knife, piercing [a skewer] at the bottom of the duckling, using tape to attach a pair of scissors 'to that', tying pipe cleaner around the stick) and demonstrates this with the materials or gestures. (Researcher, Observation 7, Hook-a-Duck)

> Says ‘I have an idea’ (sucking up the kitten), evaluates it and concludes it will not work. (Researcher, Observation 15, Volcano Bridge)

Planning behaviours involve preparing for and anticipating actions. As with representing, planning is often a mental process. However, we observed some behaviours that reflect planning such as collecting materials, clearing the workspace, and demonstrating or verbalising next steps.

> Prepares, lays out the collected materials and sets his goal ‘Attaching to each other’. (Teacher, Observation 28, Hook-a-Duck)

> ‘First, I have to make a wall ... And then with a stick ....’ (Researcher, Observation 12, Storm)

Executing refers to working towards a solution or trying to do so. Executing is about doing things and doing things well. Besides building and crafting solutions, we also saw preschoolers refining their solutions. Sometimes preschoolers lacked a hand or motor skills to perform a desired action and asked for help.

> Tries; trial & error with different materials. (Researcher, Observation 27, Hook-a-Duck)

> Tinkers. Tries to make various things, but with no real clear purpose. (Researcher, Observation 6, Hook-a-Duck)

> Tries first by herself and then asks the teacher to help her tear the tape; tries again a 2nd time first by herself; 3rd time she immediately asks for help; 4th time she succeeds by herself! (Researcher, Observation 23, Hook-a-Duck)

Monitoring means keeping track of progress, checking in between to see if the criteria are being met, and changing plans if necessary. We observed children testing, checking, measuring, and counting. Children established additional problems, adjusted elements of the solution, and indicated that a solution had been reached.

> Spontaneously tests with the hairdryer and says, ‘Here comes the storm again.’ One stone falls over. He says, ‘Not such a sturdy little house.’ But then adds, ‘These ones work well,’ referring to the two stones that stayed up. ‘But this stone doesn´t work,’ and puts the stone aside. (Researcher, Observation 32, Storm)

> Shouts ‘Finished!’, but she has NOT tested the criteria: her house has only two walls, the sheep are elsewhere, and she hasn’t tested with the hairdryer. (Researcher, Observation 10, Storm)

Reflecting refers to evaluating the solution and the process. To elicit reflective thoughts, the reflection phase was guided be the researcher by means of Mr Rabbit. Children were asked to recall elements of the process and evaluate elements of the process and the solution. Some children were able to recall and evaluate several attempts, others struggled with it or said things that did not reflect the reality. When asked if they enjoyed doing it, almost all the children answered yes, even if this does not always seem consistent with their expressions during the process. When the researcher inquired further, some children were able to identify situations they found unpleasant.

> Tells Mr Rabbit it's a good house because it doesn't blow away. He checks with the hairdryer. He also indicates that it has a roof, three walls and the sheep can get out easily. (Researcher, Observation 14, Storm)

> Recalls several failed trials and explains why it failed (e.g. the pencil was too short). (Researcher, Observation 4, Hook-a-Duck)

> Says it was not difficult, but he needed help with everything. (Researcher, Observation 12, Storm)

> Indicates that he liked it, but on the video, it seems to be different. (Researcher, Observation 11, Storm)

> Says she didn't like the bridge collapsing that much. (Researcher, Observation 2, Volcano Bridge)

The observers also identified some generic behaviours that span the whole problem-solving process. Although participants often were focused on solving the problem, some children quickly became tired of it, started playing, or exhibited disruptive behaviour (*commitment*). Similarly, some children persisted after several failed attempts, while others gave up quickly or got frustrated (*perseverance*). The observers noted that some participants took ownership of the problem and worked autonomously without needing or asking for support, while others relied heavily on the researcher for support (*autonomy*). Finally, the observers noted how the children responded to the researchers' prompts, which could either be to follow the prompt, ignore the prompt, or respond to the prompt but do something else than what was suggested (*responsiveness*).

> Commitment: Starts playing/throwing the kitten. Loses engagement. (Researcher, Observation 36, Volcano Bridge)

> Perseverance: Gets visibly frustrated (hits around, turns away) when rolls keep falling: ‘Aaarg! I can't do this.’ (Researcher, Observation 36, Volcano Bridge)

> Autonomy: Repeats that he does not know how to do it and keeps repeating this when investigator invites (several times) to think of other things; does nothing, waits. (Researcher, Observation 8, Storm)

> Responsiveness: Ignores the researcher’s suggestion to make the house smaller. (Teacher, Observation 30, Storm)

Coding problem-solving behaviours

After identifying problem-solving behaviours in the recordings, the chunks of identified behaviours (as the excerpts above) were coded. After the first round of coding, in which the first author coded the behaviours identified by the first two authors, the preliminary observation scheme was discussed within the research team and adapted, resulting in 41 codes. The observation scheme was adjusted by rewording code names (*n* = 11), merging codes (*n* = 30), relocating codes to other categories (*n* = 8), or removing codes (*n* = 14). Codes were reworded to active verbs (e.g., “Associates materials with other things or situations” became “Associating”). Codes were merged when they were difficult to distinguish (e.g., “Touches materials” and “Makes movements with the materials” were merged into “Manipulating”) or when they related to similar processes (e.g., “Thinking ahead” and “Sequencing”). Codes were relocated to another category when they turned out to be more related to another category. Codes were removed when they turned out to be difficult to observe or interpret (e.g., “Thinks”).

For the second round of coding, the identified behaviours by the last three authors were used. In this round, 184 chunks of identified behaviours were coded, resulting in 246 coded behaviours. All metrics are presented in Table 11. Strong intercoder agreement was established with 86.71% agreement between both coders and Cohen’s kappa with pairwise deletion .856 ($p < .001$). Cohen’s kappa with missing values as a regular category was below our threshold ($\kappa = .553$, $p < .001$). In many chunks, multiple behaviours were described, resulting in multiple codes per chunk. When both coders assigned a different number of codes to a chunk, missing values were produced for the other coder. Therefore, we clarified the guidelines to deal with multiple behaviours in one chunk. In addition, the observation scheme was discussed within the research team and adapted. Three codes were adjusted, six were merged, one was split, and three were removed. We also added a code for responsiveness to the category ‘Generic’.

For the third round of coding the first two authors coded 441 chunks of behaviour identified by the last three authors and the teachers, resulting in 638 coded behaviours. Agreement between both raters was comparable with Round 2 (percentage of agreement = 84.18%; $\kappa_{\text{pairwise deletion}} = .827$, $p < .001$). Cohen’s kappa with missing values as a regular category improved up to .664 ($p < .001$). Once again, the observation scheme was discussed within the research team and adapted. We adjusted one code, merged seven codes, and added three new codes: complying, asking ideas (which was a spin-off from the more general asking help), and declaring that a solution has been found.

Psychometric properties of PROSPER

The final observation scheme (PROSPER), including definitions and examples, is presented in Appendix 2. PROSPER consists of nine categories representing the eight steps of the PISA framework and a generic category for overarching behaviours. Thirty-three behaviours were assigned to seven of the eight PISA categories: exploring (8 behaviours), understanding (4 behaviours), formulating (3 behaviours), planning (2 behaviours), executing (3 behaviours), monitoring (6 behaviours), and reflecting (3 behaviours) and four overarching behaviours were assigned to the generic category: commitment, perseverance, autonomy, and responsiveness. No behaviours were assigned to the step of representing because we were not able to identify representing behaviours that could not equally well be attributed to one of the other categories.

Intercoder reliability

The results of the intercoder reliability analysis of the final observation scheme are presented in Table 11. Percentage of agreement (89.84%) and intercoder reliability ($\kappa_{pairwise\ deletion}$ = .893, p < .001) between the first two authors who developed the observation scheme was strong. Cohen's kappa with missing values as regular category was still good ($\kappa_{missing\ values\ as\ regular\ category}$ = .711, p < .001). Here we recall that missing values are produced when two coders assign a different number of codes to a chunk of identified behaviours. Sometimes it was difficult to assess whether one or more behaviours were described, especially when generic behaviours were involved. For example, "Tests by taking the sheep out of the house once (at teacher's suggestion)." (Researcher, Observation 8, Storm) was coded 'testing' by Author 1 and 'testing' and 'responsiveness' by Author 2.

Although slightly less, the percentage of agreement (82.00%) and intercoder reliability ($\kappa_{pairwise\ deletion}$ = .810, p < .001) between Author 1 and Author 3, who was not involved in developing the observation scheme, was still strong. This indicates that other coders can reliably use the observation scheme to code problem-solving behaviours.

Most discrepancies between coders relate to the code doing ($N_{Discrepancies}$ = 34). Doing is defined as working towards a solution through physical actions. Therefore, it is sometimes difficult to differentiate doing from testing ($N_{Discrepancies}$ = 9) or modifying ($N_{Discrepancies}$ = 7). For example, "Tries to fish out the duckling with a long dessert spoon." (Researcher, Observation 4, Hook-a-Duck) was coded 'doing' by Author 1 and 'testing' by Author 2.

Across-task consistency

Table 12 shows the number of coded behaviours for each problem and category based on the codes provided by the first author in coding Round 4. Although the numbers in the table should be interpreted with some caution, we note that an equivalent number of behaviours was identified for all three problems. However, we observed some differences between the three problems. The Volcano Bridge Problem elicited least exploring behaviours and most formulating behaviours. To solve the Volcano Bridge Problem, the participants could only use uniform toilet rolls and beer mats, leaving little to explore. On the other hand, participants tried different constructions, which revealed a lot of formulating behaviours. During the Hook-a-Duck problem, preschoolers revealed most exploring behaviours and least planning behaviours. Here, participants were challenged to explore materials by providing a small toolbox filled with many materials in different sizes and shapes. Planning was difficult to observe since most children tried to solve the problem by trial and error. In the Storm Problem, preschoolers demonstrated most monitoring behaviours and least formulating behaviours. We observed participants counting (e.g., the number of walls), measuring (e.g., whether the walls were sufficiently high), checking stability (e.g., blowing with the hairdryer), and adjusting (e.g., increasing the height of the walls). As most participants continued to build on their initial structure, there was less incentive to come up with or consider new ideas.

Table 12. Number (and column percentages) of coded behaviours per problem and per category

Category	Volcano Bridge	Hook-a-Duck	Storm	Total
Exploring	46 (13.98%)	89 (26.25%)	53 (16.51%)	188 (19.01%)
Understanding	54 (16.41%)	36 (10.62%)	40 (12.46%)	130 (13.14%)
Representing	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Formulating	37 (11.25%)	26 (7.67%)	20 (6.23%)	83 (8.39%)
Planning	12 (3.65%)	3 (0.88%)	15 (4.67%)	30 (3.03%)
Executing	39 (11.85%)	45 (13.27%)	35 (10.90%)	119 (12.03%)
Monitoring	58 (17.63%)	72 (21.24%)	86 (26.79%)	216 (21.84%)
Reflecting	46 (13.98%)	52 (15.34%)	39 (12.15%)	137 (13.85%)
Generic	37 (11.25%)	16 (4.72%)	33 (10.28%)	86 (8.70%)
Total	329	339	321	989

Note. The total number of assigned codes differs from the total number of coded behaviours in Table 11 because we only used the codes from author 1 and we excluded missing values to make the percentages easier to compare.

Discussion

Studying the development of young children's problem-solving competencies requires comprehensive, valid, and reliable instruments to observe problem solving in an authentic context. To date, such an instrument is lacking. The aim of the present study was to address this gap by developing a comprehensive, valid, and reliable instrument to make preschoolers' problem solving visible in an authentic context. The Problem-solving Observation Scheme for Playful Engineering and the Real-world (PROSPER) is grounded in the PISA framework and was developed through an in-depth analysis of 36 recordings of 4- to 6-year-olds involving five researchers and five preschool teachers. PROSPER is the first validated observation instrument that covers the whole problem-solving process. It consists of 33 behaviours, categorised across seven of the eight PISA steps. In addition, four generic, overarching problem-solving behaviours were identified. Construct and content validity were established by drawing on a longstanding research tradition on problem-solving phase models and involving preschool teachers in the analysis of recordings. Ecological validity was supported by designing problems closely related to children's play and creating a setting that mimicked a classroom corner and authentic teacher-child interactions. Intercoder reliability was strong and across task consistency was good.

In this section, we first discuss the observed behaviours per step of the problem-solving process in relation to existing scientific literature. Then, we delve into the psychometric properties of the instrument, highlighting the strengths and limitations of the study. Finally, we discuss the implications of our findings for research and education.

Preschoolers' problem-solving behaviours

We observed a rich variety of problem-solving behaviours involving seven of the eight PISA steps. Exploring is the first step in any problem-solving process and is a crucial element of successful problem solving for preschool children (Evans, Todaro, et al., 2021; Fusaro & Smith, 2018). Preschoolers observed, manipulated, described, and compared materials and associated them with other materials or past experiences. This finding is consistent with the observation of Cunningham et al. (2018) that preschool children aged 5 to 6 engage in and benefit from observing, manipulating, and discussing properties of materials.

The second step is understanding the problem. A problem not well understood is impossible to solve (Polya, 1945). However, it can be difficult to determine whether a preschooler understands the problem, criteria, constraints, and prompts. In developing PROSPER, we have worked around this problem by including behaviours that indicate attempts to understand the problem and the instructions. These behaviours include focusing on instructions (listening and confirming), repeating them and acting on them (complying). Although PROSPER can be helpful to identify relevant behaviours related to (attempts to) understanding, making understanding visible remains challenging. Demonstrating understanding requires strong receptive and expressive language skills (Vessonen et al., 2024). Many children experienced difficulties with repeating the criteria and constraints immediately after the problem was stated by the researcher. This raises the question of whether they did not understand the criteria and constraints or just had trouble articulating them. Further research is needed to find ways to make understanding visible. The use of visual aids can be helpful, such as showing children pictures of solutions that either meet or violate the constraints and criteria and asking them to identify the acceptable and unacceptable solutions.

Step three is representing. In our study, we were unable to make representing visible through observation. First, representing is typically an internal mental process (Mayer, 1994), impeding its operationalisation. In older students, mental representations can be externalised by tabular, graphical, symbolic or verbal representations (OECD, 2013). In preschool, children may be asked to externalise their representation of the problem or solution through drawing and verbal explanation (Bartholomew et al., 2019; Cartwright, 2024). Although 4–6-year-olds can be taught to graphically express their representations (Shiakalli & Zacharos, 2012), both drawing and interpreting their drawings remains challenging for them (Bat Or et al., 2022). Second, distinguishing between behaviours related to representing and understanding proved difficult. Representing and understanding are conceptually distinct yet closely related aspects of the problem-solving process. This is illustrated by Hayes (1981) who defines representing as “understanding the nature of the gap to be crossed.” (p.11). Pretz et al. (2003) define representing as “the manner in which the information known about a problem is mentally organized.” (p. 6). Whereas understanding is purely about grasping the problem situation, goal situation, criteria, and constraints, representing refers to connecting all the elements into a coherent mental model or problem space (Jonassen, 2010a). For ill-structured or real-world problems, integrating all pieces into a coherent representation proves particularly challenging (Pretz et al., 2003). Articulating the relationships between all elements of the problem and problem space is even more demanding. Given preschoolers' difficulties in externalising their problem representations, the strong conceptual relationship between representing and understanding, and the inherent complexity of representing ill-structured, real-world problems, the question arises whether and how representing can and should be distinguished from understanding in preschool-aged children.

Formulating possible solutions is the fourth step of the problem-solving process. Formulating consists of divergent thinking (thinking up) and convergent thinking (considering). Although strongly related, these are two distinct cognitive processes (Guilford, 1967). Generating new ideas can be observed when preschoolers verbalise their ideas or inferred when preschoolers try out a new solution. While children frequently generated new ideas throughout the entire problem-solving process, considering was only observed to a limited extent. Follow-up research can make preschoolers work together to solve to create a context in which preschoolers are encouraged to spontaneously evaluate and weigh up solutions to make decisions together.

Planning is the fifth step of the problem-solving process. We did not observe many planning behaviours. We observed planning behaviours when participants prepared their work (e.g., by collecting materials) or expressed their thoughts by explained next steps (thinking ahead). Since planning is mainly a mental process, we cannot rule out the possibility that preschoolers made plans internally. However, planning takes time and often participants started doing something immediately, even before the problem was fully explained to them. Therefore, we question whether preschoolers do plan in advance. Although previous research points out that children can plan their actions at the age of five (e.g., Warneken et al., 2014), planning behaviours do not always occur spontaneously. Possible explanations can be a lack of prior knowledge, failure to appreciate the importance of planning, or difficulties with remembering and implementing plans (e.g., Zelazo et al., 1997). Hudson et al. (1997) advocate that external support such as reminders can be beneficial to elicit preschoolers’ planning activities.

Executing is the sixth step and refers to acting towards a solution. The PISA framework defines executing as “carrying out a plan” (OECD, 2013, p. 126). According to this interpretation, one needs to have a plan prior to executing. Since preschoolers often start without a clear plan, as discussed above, and only start planning on-the-go (English, 1992), less systematic behaviours with the intention of solving the problem were also categorised as executing (e.g., trial and error and tinkering). Executing includes all physical actions to transform the problem state into the goal state. To distinguish between 'just doing' and 'trying to do it well', we added a code for refining. This covers all behaviours that are intended to refine the solution beyond the formulated criteria (e.g., make the solution more attractive).

Step seven of the PISA framework is monitoring. We identified a variety of monitoring behaviours, including testing, measuring, determining whether something works or not, modifying, and declaring that a solution has been reached. This observation is in line with the study of Bryce and Whitebread (2012) who observed a dozen different monitoring behaviours during a train track building task. English (1992) argues that children's monitoring behaviour is a compensation for their lack of domain-specific knowledge and initially non-planned, trial-and-error behaviour. As such, monitoring appears to be the central hub through which the problem-solving process unfolds. This contrasts with problem solving in older learners, where planning appears to play a primary role in guiding the process (e.g., Polya, 1945).

The last step is reflecting. PROSPER distinguishes between three reflective actions: recalling (Epstein, 2003), evaluating the result (e.g., Bonner et al., 2021; Stoll et al., 2012; Zimmerman & Moylan, 2009), and evaluating the process (e.g., Anggoro et al., 2021; Epstein, 2003). We found that preschoolers' reflections were not always consistent with what we observed. At the same time, some preschoolers could recall failed attempts and even elaborated on possible causes of them. Further research can provide deeper insights into preschoolers' reflective abilities by comparing their reflections with observed behaviours.

Our observations also revealed four generic, overarching behaviours that cut across the whole process and cannot be attributed to any of the PISA steps: commitment, perseverance, autonomy, and responsiveness. Commitment is the willingness to solve the problem and can be inferred from the eagerness to start and focussed on-task behaviour as opposite to swirling around and engaging in off-task behaviour (e.g., Liljedahl, 2016). Without commitment, children demonstrate little problem-solving behaviours. Therefore, commitment is a prerequisite for reliable observation of problem-solving competencies.

In contrast to Liljedahl (2016), we differentiate between commitment (engagement) and perseverance (persistence). While commitment is about the dedication to solve a problem, perseverance is about the ongoing efforts and resilience to overcome obstacles. Perseverance can be derived from children's reactions to unsuccessful attempts which can be positive (e.g., rebuilding or reconstructing pillars of the bridge after collapse) or negative (e.g., giving up after the first collapse).

Autonomy is the extent to which a child takes ownership on solving the problem. According to Deci and Ryan (1987), autonomy is about choice and feeling responsible (Geist, 2025). It can be derived from the degree to which a child relies on external support. It is important to emphasise that it is not because a child asks for or needs help that they cannot be autonomous. For instance, asking for help in tearing the tape illustrates agency. The same applies to children who need help to overcome an impasse after several failed attempts to solve the problem.

Finally, responsiveness refers to the degree to which a child responds appropriately or in accordance with the support provided (see also McLeod et al., 2024). Responsiveness can be inferred from a change in children's actions after a prompt is given, such as changing the structure of a building or, conversely, ignoring the prompt. Although there is a paucity of research on children's responsiveness to teachers' prompts, it has emerged as an important theme during identification of problem-solving behaviours by both researchers and preschool teachers. We observed some children immediately following up on the suggestions, while others simply ignored the suggestions. Further research is needed to investigate preschoolers' responsivity to teachers' prompts in relation to successful problem solving.

Commonly used problem-solving phase models such as those of Polya (1945) and Schoenfeld (1985) don't cover overarching behaviours that span the entire process. Based on our observations and the identified problem-solving behaviours by the teachers, we argue that generic behaviours such as commitment, perseverance, autonomy, and responsiveness should be part of any instrument that maps preschoolers' problem-solving behaviours.

In PROSPER, special attention is given to asking questions which is often associated with curiosity (e.g., Fusaro & Smith, 2018; Luce & Hsi, 2015). Young children tend to ask a lot of questions (e.g., Jirout & Klahr, 2020; Wong et al., 2024). For them it is a way to explore, solve problems, and learn. Initially, asking questions was conceived as a separate, generic behaviour. Through iterative coding and discussion of discrepancies within the research team, it emerged that questions had different aims and that they were related to different steps in the problem-solving process. Therefore, asking questions was broken down in asking information (exploring), asking ideas (formulating), asking motor help (executing), and asking confirmation (monitoring). In addition, we observed children who kept on asking for help in only general terms. This indicated a lack of taking responsibility and was integrated in the generic code autonomy.

Psychometric quality of PROSPER

In developing PROSPER, we ensured strong structural validity, content validity, and ecological validity and assessed intercoder reliability and across task consistency. Construct validity was ensured by firmly rooting the instrument in the PISA problem-solving framework which is in turn embedded in a rich research tradition on problem-solving phase models (OECD, 2013). Content validity was established by involving five researchers and five preschool teachers in identifying problem-solving behaviours. To strengthen ecological validity, three engineering design problems were developed that are closely related to preschoolers' construction play and crafting activities (e.g., Shechter et al., 2021). Additionally, data were collected in a familiar classroom environment following a protocol that simulated authentic teacher-child interactions. It would be interesting to further validate the instrument with a larger sample, in other education systems and settings (cross-cultural validity, ecological validity), with other STEM-problems (external validity), against other problem-solving measures (concurrent validity), and to examine how well the instrument predicts later learning outcomes (predictive validity) (see also AERA et al., 2014).

As stated above, commitment is a prerequisite for reliable observation of problem-solving skills We observed that most participants were highly motivated to take part in the research. Between sessions, the children often asked when they could participate again and showed clear excitement when it was their turn. While this enthusiasm may have been partly driven by the individual attention they received, we asked them after each session whether they had enjoyed the activity and what they liked or disliked about it. Almost without exception, they reported positive experiences, although this did not always align with our observations. Achieving a solution may have mitigated some frustration experienced during the problem-solving process.

Strong intercoder reliability and across task consistency were established. This indicates that PROSPER can reliably be used to observe problem-solving behaviours across different tasks. For this study, we only focused on intercoder reliability. In future research, it would be interesting to apply the observation scheme repeatedly to assess test-retest reliability.

Limitations and future directions

Despite the strengths outlined above, this study is not without its limitations. First, the study is based on the analysis of 36 recordings in a sample of 12 preschoolers including boys and girls and children with diverse linguistic background who were offered three different engineering problems. Although we ensured the representativeness of the sample and aimed for maximum heterogeneity between participants, it would be beneficial to include more participants to further investigate the generalisability of the observation scheme. In addition, applying the observation scheme with other STEM-problems can provide more evidence on the external validity of the instrument.

Second, data collection was conducted by a researcher. Although the researcher was involved in classroom activities to familiarise with the children in the days before and during data collection, and the interactions closely mimicked authentic teacher-child dialogue, there remains a gap compared to a real classroom situation. Having data collected by the teachers in the children’s own classroom would be even more authentic and could potentially elicit different behaviours. It is therefore desirable to involve teachers in data collection in follow-up studies.

Third, we decided to minimise guidance throughout the process to observe what preschoolers did on their own, while separating exploring and reflecting from the rest of the process. Closer involvement may be beneficial to better observe representing and planning behaviours.

Finally, the current observation scheme does not provide a systematic approach to translate observed behaviours into assessments of preschoolers' problem-solving competencies. Follow-up research could explore how the observation scheme can be used for assessment purposes.

Conclusion

This study contributes to the field of early childhood research and education by operationalising the PISA framework for use in preschool by assigning observable problem-solving behaviours to the different steps of the problem-solving process. It extends the PISA framework with four generic, overarching behaviours: commitment, perseverance, autonomy, and responsiveness. Although the present study focuses on preschool children aged 4–6 years solving engineering design problems, the observation instrument can be applied to other types of STEM problems and throughout early childhood (3–8 years).

This study is foundational for further research on problem solving in early childhood education by developing PROSPER, a comprehensive, valid, and reliable observation instrument that makes preschoolers' problem-solving behaviours visible in authentic contexts. Researchers can use the instrument to further investigate the development of preschoolers' problem-solving competencies, to investigate the effectiveness of problem-solving interventions, and to investigate the relationship between problem solving and other constructs. However, given the complex nature of the problem-solving process and the comprehensiveness of the observation scheme, researchers should be aware that implementing this observation scheme is both challenging and time-consuming, necessitating training and substantial practice.

For educational practitioners, PROSPER provides a useful lens to observe problem solving as it occurs in the classroom. Focused observation can help preschool teachers monitor the development of preschoolers' problem-solving competencies, identify areas that require attention, and guide them in selecting or developing appropriate learning activities. Presenting children with challenging problems, observing their problem-solving behaviours, and engaging with the children during the problem-solving process provides researchers and educators valuable insights in children's problem-solving competencies. This way, the use of PROSPER in future (intervention) studies and educational practice can help to deepen insights into the development of children's problem-solving competencies and how to optimally stimulate this development, enabling future generations to meet the challenges of vital STEM careers.

STUDY 4 | AN ARGUMENT-BASED APPROACH TO VALIDATION OF FORMATIVE PROBLEM-SOLVING ASSESSMENT IN PRESCHOOL

This chapter is currently under review as:

Van Elsen, J., Torbeyns, J., & De Maeyer, S., (2026). An argument-based approach to validation of formative problem-solving assessment in preschool [Manuscript submitted for publication].

Abstract

Background: Problem-solving competencies are complex and context-dependent, making them difficult to assess. Although formative preschool assessment is considered a viable approach, it has been criticised for lacking rigour. The argument-based approach to validation offers a systematic framework for addressing this criticism by mapping and evaluating all inferences and assumptions linking performance to the interpretations and uses of the assessment results.

Aims: This study aims to construct a validity argument for the interpretation and use of formative preschool assessment of problem solving.

Sample: Participants were four preschool teachers and 53 children aged 5–6 years.

Methods: Teachers assessed children's problem-solving competencies through the Observation-based Early Problem Solving (OEPS) assessment. Six inferences were identified as relevant to the Interpretation/Use Argument (IUA): performance, scoring, generalisation, explanation, extrapolation, and decision/use. Claims, warrants, and underlying assumptions were defined, and relevant backings and evidence were identified, collected, and analysed. Quantitative and qualitative data from 102 assessments were supplemented with data from teacher's feedback and from adherence observations.

Results: The validity argument showed strong support for the performance, explanation, extrapolation, and decision/use inferences. The scoring inference was fairly supported and support for the generalisation inference was moderate. Interrater reliability, across-task consistency, completing the training, and adherence to the protocols emerged as the primary points of concern. Teachers found the information they obtained useful for updating their understanding of children's competencies and for making instructional decisions.

Conclusion: Despite some concerns related to generalisability, the OEPS assessment provides preschool teachers with useful information to guide instructional decisions.

Introduction

Problem solving is a complex, context-dependent process involving varying sets of competencies (Funke et al., 2018; Jonassen, 2000). Problem-solving competencies can be developed through experience with appropriate tutoring (Vygotsky, 2017). Ideally, this process begins in early childhood (Burns et al., 2025; Diamond, 2018), with preschool teachers playing a pivotal role (Erol, 2024). To determine which problem-solving competencies each child has already mastered, to identify their needs for support, to inform instructional decisions, and to monitor the development of children's problem-solving competencies (Moss, 2003), educators require reliable, multifaceted, process-oriented assessment instruments that can be validly used in the classroom (Burns et al., 2025). However, in Study 2 we found that most assessment instruments used to investigate preschoolers' problem-solving competencies address isolated skills (e.g., thinking-up possible solutions) and result in general quantitative outcomes (e.g., binary, count, or scale scores). Except for two instruments (notably two scales in which children were individually asked to formulate multiple solutions to depicted problems), analysis of the instruments' psychometric properties was largely absent. In addition, we found that the assessments were typically conducted individually, in a separate room, by an external researcher. These assessments differ substantially from authentic classroom practice and provide teachers with limited information about the precise strengths and instructional needs of young children.

The multifaceted and layered nature of problem solving, combined with its embeddedness in specific contexts, makes problem-solving competencies difficult to assess. Standardised, norm-referenced tests fall short because they treat the rich and varied contexts in which the assessed skills occur as measurement error rather than as an integral component of the assessment (AERA et al., 2014) and provide limited information for improving performance in a real-world context (Archbald & Newmann, 1988). Authentic assessment is regarded as a viable alternative (Varier & Yun, 2023). Authentic assessment makes use of meaningful, real-world tasks performed in natural settings. This makes authentic assessment particularly suitable for identifying children's strengths and instructional needs in problem solving and for expressing expectations regarding real-world performance (Kane, 2013b). However, authentic assessment has been criticised for lacking methodological rigour (Bagnato, 2005). This criticism partly stems from the fact that methodological rigour is often evaluated against standards developed for standardised tests. The methodological quality of formative assessment instruments differs from that of standardised ones. The quality of formative assessment depends on the extent to which it supports teachers in making instructional decisions, rather than on its precision, generalisability across children or populations, or its capacity to measure performance against fixed standards (Kane & Wools, 2019).

The argument-based approach to validation provides a framework for identifying the critical aspects of validity by meticulously mapping and evaluating all inferences and assumptions underlying the interpretation and use of test scores, from performance to the decisions based on the outcomes (Kane, 2013a; Lane & Marion, 2025). In addition to conventional quality indicators of validity and reliability, such as construct validity, content validity, and inter-rater reliability, this framework places particular emphasis on the extent to which results are interpreted and used in accordance with their intended purpose.

The aim of this chapter is to build and evaluate an interpretation/use argument (IUA) for the Observation-based Early Problem Solving (OEPS) assessment. This authentic, formative assessment builds on the protocols and observation scheme from Study 3 and is designed to identify preschoolers' strengths and instructional needs. By developing such an argument, we aim to identify the strengths and limitations of the OEPS assessment and to determine the conditions under which valid interpretations and uses of the results are warranted (Kane, 2013b). The findings of this study further offer guidance and key considerations for the development and use of similar authentic, observation-based assessments for formative purposes in the preschool classroom.

The chapter is structured as follows. First, we present the conceptual and theoretical frameworks in which this study is grounded. Second, we describe the intended purpose and use and the main features of the OEPS assessment, including its materials and procedures. Third, we present the sources of validity evidence to support the validity argument and outline the data collection procedures, data types, and descriptive statistics. Fourth, we construct the IUA in which the relevant inferences, claims, warrants, assumptions, and sources of evidence are defined. Fifth, we present a validity argument in which the assumptions associated with each inference are evaluated to determine the extent to which the warrants hold. Sixth, we identify the strengths and points of concern in the argument and the conditions that must be met to make substantiated instructional decisions based on the assessment results. The chapter concludes with a discussion of the implications for research and practice.

Theoretical Framework

Early problem solving

Problem solving can be defined as "an individual's capacity to use cognitive processes to confront and resolve real, cross-disciplinary situations where the solution path is not immediately" (OECD, 2004, p. 156). According to this definition, problem solving is the ability to apply mental processes in cases where the individual does not exactly know what to do. However, real-world problems cannot be solved through thinking alone. Solving problems requires not only cognitive but also metacognitive, motor, social, affective, and volitional processes (Jonassen, 2000). The crux of problem solving lies in coordinating or orchestrating these processes to transform a problem situation into a goal situation (Kilpatrick, 1978; OECD, 2004). This orchestration itself is difficult to measure. Therefore, scholars have tried to operationalise problem solving by identifying relevant steps or phases in the problem-solving process (Rott et al., 2021). The resulting theoretical models show substantial similarities and typically include an orientation phase, a planning or preparation phase, an execution phase, and a reflection phase (e.g., Polya, 1945; Wilson et al., 1993; Zelazo et al., 1997).

These models typically focus on well-structured problems encountered by older students, presenting problem solving as a systematic, structured process in which planning plays a central role (Polya, 1945). However, preschoolers often struggle to plan in advance, as their planning abilities are still developing and they lack the prior experience and knowledge needed to predict the outcomes of their actions (Hudson et al., 1995; Lambert, 2001; Zelazo et al., 1997). To solve problems successfully, they must therefore rely on exploratory and creative processes and on monitoring their actions (Study 3; English, 1992; Evans, Todaro, et al., 2021). This makes problem solving in young children an iterative process that often begins with trial-and-error and progresses towards a solution with varying degrees of success (English, 1992).

A problem can be solved individually or in collaboration with others. Collaborative problem solving occurs when two or more children "attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2017, p. 134). Allowing children to collaborate while solving problems offers several advantages. Children can provide support to each other, they can learn from each other, and collaboration encourages children to articulate their ideas, enabling teachers to better understand their thinking. In addition, allowing children to collaborate during problem solving makes the assessment more authentic (Frey et al., 2012).

Formative preschool assessment

Although preschool assessment can serve multiple purposes, such as school readiness screening, identifying children at risk for learning difficulties, evaluating the effectiveness of instructional interventions and programmes, determining children's strengths and instructional needs, and informing educators' instructional planning, the goal of formative preschool assessment is primarily to improve learning (Brassard & Boehm, 2007). Formative assessment takes a central place in preschool practice (Pyle et al., 2020). It is defined as "an assessment process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning with the goal of improving students' achievement of intended instructional outcomes" (AERA et al., 2014, p. 219). Formative assessment aims to determine the distance between children's current level of learning and development and the intended goals, and to identify the support or actions needed to help them progress (Black & Wiliam, 2009). In preschool, formative assessment supports educators in identifying children's strengths and instructional needs in a given domain, thereby informing decisions to plan or modify instruction (Brassard & Boehm, 2007). This is illustrated by the definition of formative assessment provided by Black and Wiliam (2009):

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p.9)

Assessment in preschool settings poses several challenges. Young children develop in nonlinear, irregular, and idiosyncratic ways, vary in motivation and tolerance for frustration, have limited experience with formal testing, possess limited reading and writing skills, often struggle to express themselves verbally, and are sensitive to testing conditions such as familiarity with the test administrator, format, task characteristics, and materials (Brassard & Boehm, 2007; Mascolo & Fischer, 2015; Panter & Bracken, 2013). Moreover, preschool assessment must be authentic to adequately capture children's strengths and instructional needs (Bagnato, 2005; Grisham-Brown et al., 2006). This is particularly true for problem-solving assessment, as children may not make genuine attempts to engage with the problems if the presented problems are not authentic (Hardy et al., 2024). Authentic assessments are ideally performance-based, administered in realistic, group-based contexts, and involve cognitively complex tasks (Frey et al., 2012). Observing children in their natural environment constitutes an essential component of authentic preschool assessment (Burke et al., 2025). Observation-based assessment offers several advantages compared to other assessment forms: it focuses on what children actually do, enables linking behaviours to the context, does not rely heavily on verbal competencies, and allows consideration of social processes involved (Azevedo, 2009). As such, observation-based assessment can have high validity (Randel & Clark, 2013). Authentic, observation-based assessment is particularly appropriate for identifying preschoolers' strengths and instructional needs in the context of problem solving, as it yields richer information about the processes involved than test scores alone (Gullo, 2005).

In this study, we use the term *formative preschool assessment* to refer to authentic, observation-based assessments with a formative purpose in preschool. It is a form of preschool assessment situated at the intersection of classroom assessment (McMillan, 2013), formative assessment or assessment for learning (Black & Wiliam, 2009), authentic or play-based assessment (Pyle et al., 2020), and observation-based assessment (Brassard & Boehm, 2007).

The argument-based approach to validation

Validity is not a property of the test itself, but of the intended interpretations and uses of the test scores (AERA et al., 2014; Kane, 2013b; Messick, 1989). Therefore, validation cannot be limited to some psychometric properties of a test, but must account for the many assumptions involved in interpreting test scores (Kane, 2006). The argument-based approach to validation provides a framework to critically evaluate all inferences and assumptions made when interpreting and using test scores (Kane, 2013a). The framework is grounded in Toulmin's model of inference (Toulmin, 2003) and aligns with the *Standards for Educational and Psychological Testing* which defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11).

The framework consists of two types of arguments: an interpretation/use argument (IUA) and a validity argument. The IUA specifies how test scores are intended to be interpreted and used. It lays out the inferences and assumptions that support those interpretations and uses. The validity argument provides the theoretical and empirical evaluation of the proposed interpretations and uses of test scores as outlined in the IUA (Kane, 2013a; Lane & Marion, 2025).

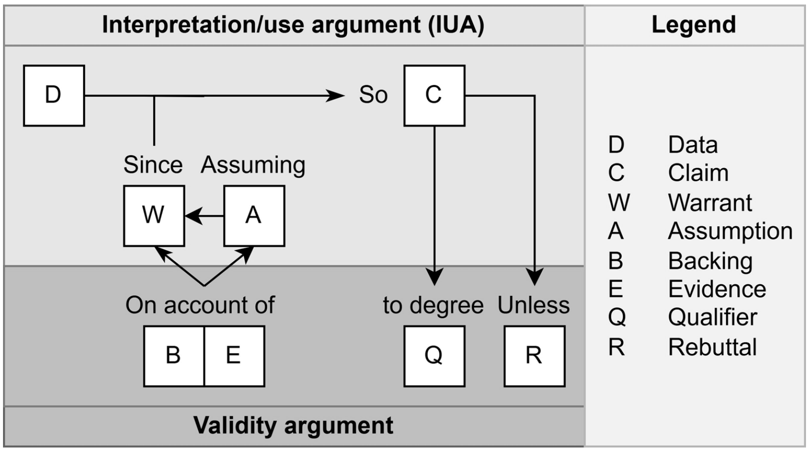
When teachers test or assess a particular skill in children, their intention is to gain some benefit from the results. This may involve identifying children with learning difficulties, evaluating the impact of an intervention, or providing information about children's strengths and areas for development. Therefore, the IUA starts from these intended uses and then works backwards to develop a chain of inferences and assumptions associated with them (Kane, 2021).

Validating the interpretation and use of test scores means examining the extent to which decisions based on these scores are supported by evidence. When using test scores, it is assumed that the tasks elicit the behaviours to be evaluated (*performance inference*), that the test administrator can accurately score these behaviours as intended (*scoring inference*), that the observed scores reflect expected performance on similar tasks (*generalisation inference*), that these scores are representative for the construct of interest (*explanation inference*), that the scores are indicators for performance in real-world situations (*extrapolation inference*), and that the test scores are interpreted and used as intended (*decision/use inference*) (Chapelle, 2021; Lane & Marion, 2025). Each inference must be supported by a *warrant* (Kane, 2013a). A warrant is a general hypothetical statement that justifies the inference. The warrants are supported by a backing, which provides the logical, theoretical, empirical, or methodological foundation for the warrants. In addition, warrants are supported by assumptions; context-specific conditions that must be true for the warrant to hold. All claims, warrants, backings, and assumptions together constitute the IUA.

The validity argument evaluates the IUA in terms of coherence, completeness, and its assumptions (Kane, 2013a). This requires evidence that supports or rejects the proposed interpretations and uses or suggests plausible rival interpretations (Kane, 2006). This evidence can be quantitative or qualitative (Cook et al., 2015) and can take various forms, including expert judgement, analysis of response processes, statistical data analysis, convergence with other tests or variables, and evidence concerning the consequences of the assessment (AERA et al., 2014; Chapelle, 2021). Based on the evaluation of the backings and the evidence, statements can be made regarding the strength of the warrant and the exceptional conditions under which the warrant does not hold. These are referred to as the *qualifiers* and the *rebuttals*, respectively. The use of qualifiers and rebuttals aligns with the interpretation of the *Standards* that validity is not true or false but a matter of degree (AERA et al., 2014).

Although the distinction between the IUA and the validity argument is conceptually clear as the theorem or proposition on the one hand and the proof on the other, different authors categorise the components of their argumentation in varying ways. For example, Kane (2021) places the backings and assumptions within the IUA, whereas Lane and Marion (2025) assign them to the validity argument. Chapelle (2021) does not distinguish between IUA and validity argument and treats interpretations and uses as components within the construction of a validity argument. As the IUA concerns *what must be demonstrated*, and the validity argument provides the *evidence* for this demonstration, we consider the claims, warrants, and assumptions as part of the IUA (consistent with Kane, 2021), and the backings and evidence as part of the validity argument (consistent with Lane & Marion, 2025). Figure 8 presents the Toulmin (2003) diagram for making inferences, adapted to the argument-based approach to validation.

Figure 8. Diagram for making inferences (adapted from Toulmin, 2003)



In research literature, a distinction is typically made between validity and reliability. Reliability is subsumed under validity and is considered a necessary but not sufficient condition for validity (Brennan, 1998; Kane, 2013b). Within the argument-based approach to validation, reliability functions as backing for the generalisation inference (Kane, 2013a). The assumption is that if test scores vary substantially across occasions, tasks, raters, or other potential sources of variation, it becomes difficult to generalise the interpretation and use of the results across situations (Kane, 2013b).

Validity in formative preschool assessment

As with any type of assessment, interpretation and uses of results from formative assessment should be valid (Stobart, 2012). Validation of formative assessment entails the specification, justification, and support of interpretations and uses of results from formative assessments (Gu, 2021). Since the main purpose of formative assessment is to improve learning, the validity of any formative assessment can be judged by the extent to which the collected evidence substantiates the decisions or actions taken to support children's learning or development in a specific domain (Hopster-den Otter et al., 2019; Kane & Wools, 2019). If the conclusions drawn from formative assessment are inaccurate, the rationale for adapting instruction is undermined (Bennett, 2011), potentially resulting in actions that are less effective than would have been achieved on the basis of valid conclusions.

Validation of formative assessment differs in several important respects from the validation processes used for standardised tests, large-scale assessments, or measurement instruments for scientific research. In standardised testing, for example, the aim is to ensure equal opportunities for all test-takers (fairness) and to enable comparisons across individuals or occasions requiring high levels of reliability and generalisability (AERA et al., 2014; Kane & Wools, 2019). Achieving high levels of generalisability requires controlling contextual influences as much as possible to minimise measurement error. However, the learning context plays a crucial role in formative assessment (Moss, 2003). Formative assessment is more likely to achieve a high degree of validity when it takes place in environments that closely resemble the real learning context, when its form and content reflect authentic learning situations, and when the assessment itself is meaningful and provides learning opportunities (Hargreaves, 2007). The critical concern in formative assessment is how well the decision and use of the test results support children's learning (Hopster-den Otter et al., 2019; Stobart, 2012), rather than the generalisability of the results (Crooks et al., 1996; Kane & Wools, 2019; Lane & Marion, 2025). Standardisation of procedures is therefore less central, as flexibility is valued for enabling teachers to respond to learners' abilities, experiences, and backgrounds (Lane & Marion, 2025).

Kane (2013b) argues that the standardisation of procedures improves reliability at the expense of validity. Although standardisation enhances fairness and generalisability, it does so by restricting the situations over which generalisation is possible. As a result, the authenticity of the test is reduced and so the usefulness of the results in the context of formative assessment. Authentic formative assessments generally demonstrate lower reliability but greater validity than standardised tests (Brennan, 1998). However, the challenge is to find an appropriate balance between validity and reliability (Kane, 2013b), as limited reliability in formative assessment may offer insufficient grounds for making instructional decisions.

The Observation-based Early Problem Solving (OEPS) assessment

Our aim is to develop a validity argument for the interpretation and use of the Observation-based Early Problem Solving (OEPS) assessment. The OEPS assessment is an observation-based instrument that enables teachers to assess young children's problem-solving competencies. It consists of a structured observation form, three problem tasks designed to elicit problem-solving behaviours in children aged 5 to 6 years, and an accompanying manual. This section discusses the intended purpose and use of the OEPS assessment, the indicators of problem solving, and the problem tasks.

The purpose and use

The OEPS assessment is primarily designed to inform preschool teachers about children's problem-solving competencies, allowing them to identify both strengths and instructional needs. It aims to broaden and deepen teachers' understanding of competencies that are difficult to assess with conventional measurement instruments or tests. The assessment is intended to give teachers, parents, and other stakeholders insights into children's strengths and instructional needs, both at the individual level and at the level of preschoolers as an age group, and to support instructional decisions. The assessment is not intended to be used for high-stakes decisions concerning school readiness, grading, academic progression (including promotion to the next year), or the identification of children with or at risk for learning difficulties.

Indicators of problem solving

The OEPS assessment is based on PROSPER, the observation scheme developed in Study 3. As discussed earlier, PROSPER is grounded in the PISA framework but omits representing, which proved to be almost impossible to observe. Given the collaborative context employed in the OEPS assessment, *collaboration* has been added to the four overarching problem-solving behaviours identified in Study 3 (commitment, perseverance, autonomy, and responsiveness). The OEPS assessment includes a total of 12 problem-solving indicators. Seven of these correspond to the steps of the PISA framework, while the remaining five correspond to overarching problem-solving behaviours.

The 12 indicators of problem solving constitute the backbone of the OEPS assessment. All indicators are scored on a five-point scale ranging from (1) 'The child does/shows this not at all, hardly, with difficulty, under guidance, after prompting' via (3) 'The child does/shows this occasionally, appropriately, and spontaneously, with limited guidance or support', to (5) 'The child does/shows this continuously, when necessary, fluently, thoroughly, on own initiative, independently.' An additional 'Difficult to say' (DTS) option is included for situations in which an indicator cannot be adequately assessed based on the observed behaviours. Finally, text boxes are provided for both general and indicator-specific comments. A sample observation form is provided in Appendix 3.

The problems

The OEPS assessment employs engineering design problems to elicit problem-solving behaviours in preschoolers. Engineering design problems are authentic, open-ended problems in which children are asked to craft or build a solution that meets specific criteria within given constraints (Arık & Topçu, 2022; Cunningham et al., 2018; Purzer & Douglas, 2018). Three engineering design problems were developed for the OEPS assessment. The Hook-a-Duck Problem requires participants to create a tool to get a ringed rubber duckling out of a transparent bottle. The Volcano Bridge Problem requires participants to build a bridge over a volcano. The Racetrack Problem requires children to build a racetrack suitable for toy cars. The first two problems are adapted from Van Elsen et al. (2025). The racetrack problem is developed specifically for this study. A full description of the problems, materials, setup, and examples of solutions is provided in Appendix 4.

To solve the Hook-a-Duck Problem and the Volcano Bridge Problem, children can use familiar, researcher-provided materials, including low-cost craft materials (e.g., cardboard, paper, toilet rolls, small containers), connecting materials (e.g., pipe cleaners, rubber bands, clothespins, hooks, modelling clay), tools (e.g., scissors, ruler, screwdriver), and kitchenware (e.g., children's fork and spoon, wooden spoon). For the Racetrack Problem, children are allowed to use any materials available in the classroom. The three problems are developed following the age appropriate design parameters provided by Cunningham et al. (2018), discussed with preschool teachers, extensively pilot tested ($N_{Children}$ = 61), and refined to ensure they were feasible yet challenging.

Data collection

Participants

Teachers were recruited through the first author's professional and personal network, a project website on problem solving in preschoolers (www.kleuterkronkels.be; in Dutch), and an invitation distributed through social media (LinkedIn and Facebook). Four teachers from four different schools participated in this study. Participating teachers were all women with 24–35 years of teaching experience in preschool education. All teachers taught in the third year of preschool (the year before the start of primary education) and were employed in publicly funded private schools located in rural municipalities.

Parents of all children in the participating classes were informed about the study in advance and were given the opportunity to exclude their child from participation. No parent requested to withdraw their child. A total of 68 children were enrolled across the participating classrooms. To maintain the feasibility of the study, teachers were given the option to not involve all children from their class. In case they preferred not to involve all children, teachers were asked to select a subsample of children with diverse profiles (e.g., in terms of language proficiency or creativity). In total, the assessment was administered to

53 children (30 boys and 23 girls) aged 5–6 years. Nine (17%) participants did not speak the assessment language (Dutch) at home.

The research was approved by the Ethical Advisory Committee for Social and Human Sciences (EASHW) of the University of Antwerp on 26 March 2025 (file SHW_2025_19). At no point were participants identifiable in the data to the researchers or to anyone other than the participants' own teacher.

Procedures

Prior to data collection, the participating teachers completed a training programme. The training was offered through the project website and comprised studying the manual, conducting a multiple-choice test, conducting a video-based training observation, and an individual consultation via Teams. Teachers were given the option to practice further with a second training video. A detailed step-by-step guide supported teachers throughout the training, preparation, and data-collection process.

Data were collected through direct, structured observations conducted by the teacher of the children following the protocol provided in the assessment manual. Observations took place during school hours, either in the child's own classroom or in a spare room at the child's school. Children worked in alternating pairs on two of the three engineering-design problems, resulting in 51 observations and 102 assessments, with a total duration of 1144 minutes (19.07 hours). Teachers noted their observations on paper or directly in a provided Excel file and took pictures of the setup and the solutions created by the children. All data were pseudonymised by the teachers using a class icon or number and uploaded to a secure folder hosted by the university.

Participating teachers were asked to document their experiences and provide feedback in an open format. Feedback submitted via email was also included. After data collection, a debriefing session was held to discuss issues related to materials, procedures, adherence to the protocol, reliability, feasibility, and use.

Finally, 13 observations (26 assessments) were co-observed by the first author (adherence observations). During these observations, data were collected to analyse interrater reliability and adherence to the protocols. In addition, the teacher's comments and reflections on acquired insights and the actual or planned use of the information obtained were recorded in field notes and used in accordance with the feedback.

Data types and analyses

The quantitative data collected to support the validity argument consisted of scores provided by the teachers ranging from 1 to 5, scores from a second observer for the analysis of interrater reliability, and the recorded duration of the sessions. Detailed information on the specific analyses is provided along with the assumption that each analysis is intended to support. All quantitative analyses were conducted in R Statistical Software (Version 4.4.2;

R Core Team, 2024) using the packages *tidyverse* (Version 2.0.0; Wickham et al., 2019), *psych* (Version 2.4.6.26; Revelle, 2024), *lavaan* (Version 0.6.20; Rosseel, 2012), and *lme4* (Version 1.1.35.5; Bates et al., 2015).

The qualitative data included text comments recorded by the teachers during the observations, feedback data, debriefing data, and fieldnotes from the adherence observations. The text comments from the observations were coded according to the inferences and assumptions they address. The comments were analysed for content and used to qualitatively support and illustrate the assumptions. The teachers' feedback and debriefing data were segmented into chunks, collated in a spreadsheet, coded, and analysed in a similar way. Finally, pictures were analysed to examine adherence to the protocols (setup, materials, context) and the diversity of solutions. Examples of solutions for each problem are provided in Appendix 4.

Descriptive statistics

Descriptive statistics on the number and duration of the observations are presented in Table 13 (by teacher) and Table 14 (by problem). The participating teachers conducted 51 observations, resulting in 102 assessments. The Hook-a-Duck Problem was used 16 times, the Volcano Bridge Problem 15 times, and the Racetrack Problem 20 times. Observations lasted on average 22.43 minutes (*SD* = 15.22 minutes), ranging from 2 to 60 minutes. However, substantial variation occurred between teachers and problems. The mean duration per observation per teacher ranged from 8.36 to 32.94 minutes. The Hook-a-Duck Problem lasted on average 13.62 minutes, the Volcano Bridge Problem 24.27 minutes, and the Racetrack Problem 28.10 minutes (Table 14).

Table 13. Descriptive statistics on the number and duration of observations by teacher

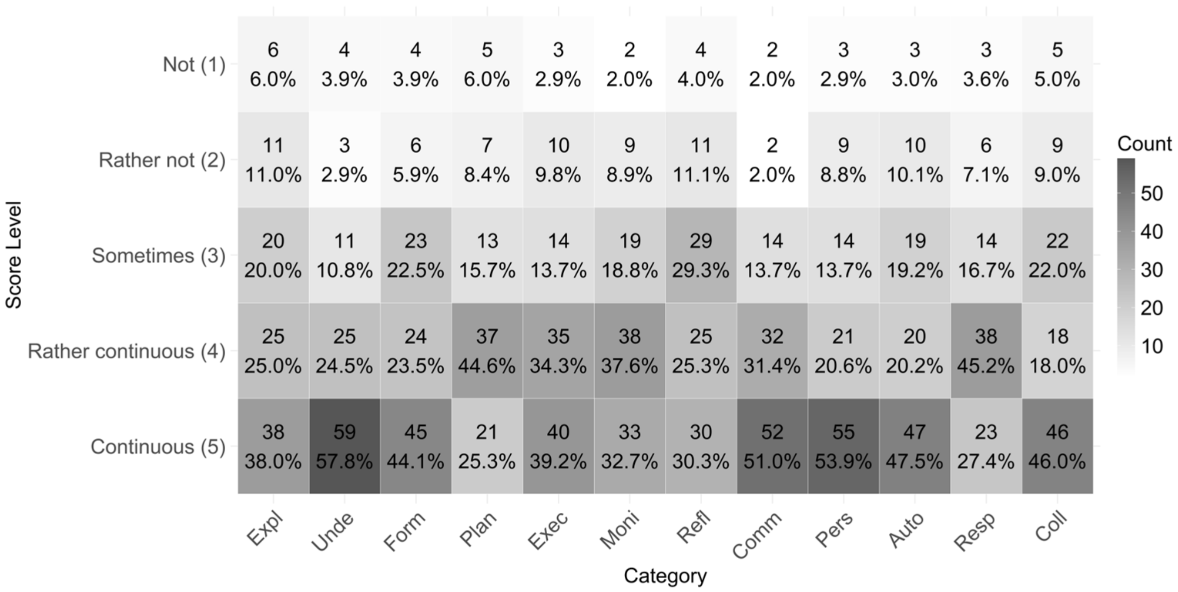
Metric	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Overall
Assessments (*N*)	34	20	22	26	102
Observations (*N*)	17	10	11	13	51
Duck (*N*)	6	2	5	3	16
Volcano (*N*)	5	3	0	7	15
Race (*N*)	6	5	6	3	20
Mean Duration (min)	32.94	15.50	8.36	25.92	22.43
Min Duration (min)	10	4	2	9	2
Max Duration (min)	60	40	20	50	60
SD Duration (min)	13.47	11.79	6.07	14.06	15.22
Total Duration (min)	560	155	92	337	1144

Table 14. Descriptive statistics on the number and duration of observations by problem

Metric	Hook-a-Duck	Racetrack	Volcano Bridge	Overall
Assessments (*N*)	31	40	31	102
Observations (*N*)	16	20	15	51
Mean Duration (min)	13.62	28.10	24.27	22.43
Min Duration (min)	3	2	6	2
Max Duration (min)	30	60	50	60
SD Duration (min)	9.72	17.23	13.61	15.22
Total Duration (min)	218	562	364	1144

Figure 9 summarises the score distribution for each problem-solving indicator, showing how often teachers assigned each score. The distribution of the scores is strongly skewed for all indicators. Scores were predominantly high (4–5), while the number of assessments resulting in a low score (1–2) was limited.

Figure 9. Heatmap of scores per problem-solving indicator ($N_{Observations}$ = 102)



Note. The numbers and percentages represent the scores for individual assessments across problems and children.

The interpretation/use argument (IUA)

As outlined in the theoretical framework on the argument-based approach to validation, validating the interpretations and uses of test scores involves two components: an interpretation/use argument (IUA) and a validity argument (Kane, 2013a). In this section, we present the IUA with its inferences, claims, warrants, and assumptions.

Table 15 presents the IUA for the interpretation and use of the results of the OEPS assessment for formative purposes. The IUA provides an overview of the inferences, claims, warrants, and assumptions that must hold to enable teachers to use the results to identify children's strengths and instructional needs in problem solving. The IUA is based on an extensive review of the scientific literature on formative assessment, authentic assessment, preschool assessment, observation-based assessment, problem solving, engineering design in early childhood education, and the argument-based approach to validation. The IUA is structured around six inferences: performance, scoring, generalisation, explanation, extrapolation, and decision/use (Chapelle, 2021; Lane & Marion, 2025). For each inference, a general claim and associated warrant were formulated, and the underlying assumptions were identified. Consistent with Toulmin (2003), the claims are formulated as general statements and the warrants as if-then statements. The assumptions are formulated as specific statements related to the problems (tasks), assessment, rubric, context, children, teacher, scores, and results.

The IUA, including the list of assumptions for each inference, was reviewed with the research team and a researcher familiar with both the framework and the early childhood education context. This process took place over three rounds. After each iteration, the IUA was refined: assumptions were added, split, merged, reformulated, relocated, or removed.

The validity argument

Sources of validity evidence

The validity argument seeks to evaluate the IUA by examining whether its underlying assumptions are supported by the evidence. To substantiate the validity argument for the interpretations and uses of the OEPS assessment, we draw on multiple sources of evidence. A first source of evidence consists of insights from theoretical and empirical research. A second source concerns the design of the assessment itself and its alignment with scientific theory and evidence. A third source comprises newly collected and analysed quantitative and qualitative data. For clarity, we present the argumentation for each assumption following this order, as far as the different types of arguments are relevant to substantiate the assumption.

Consistent with Lane and Marion (2025), we discuss backings and evidence together. Although Toulmin (2003) differentiates between backings and evidence as supporting warrants and assumptions respectively, both ultimately serve the same function; to support or refute the inference. Once the argumentation has been developed for each assumption underlying the inference, the strength of the evidence for that inference is evaluated by assigning a qualifier, including any conditions that must be met for the warrant to hold.

Table 15. Outline of the interpretation/use argument (IUA)

Inferences	Claims	Warrants	Assumptions
Performance inference (From task to performance)	Problem-solving behaviours can be elicited by asking children to solve problems	If designed according to age-appropriate design principles and presented in a familiar and supportive context, then engineering design tasks elicit observable problem-solving behaviours that reflect children's problem-solving competencies	1. The problems are authentic, meaningful, and inherently motivating for the children 2. The problems provide children with opportunities to display behaviours relevant to all indicators of problem solving 3. The problems are age appropriate 4. The children are appropriately scaffolded when truly stuck 5. The children do not have prior experience with the same or a similar problem that would inflate or distort performance 6. The context and materials are familiar and accessible to children 7. The context does not hinder children’s engagement or performance
Scoring inference (From performance to score)	Problem-solving competencies can be evaluated by scoring children’s observed problem-solving behaviours during task performance	If children engage meaningfully with the problems and teachers implement the assessment as intended, then they can score children’s problem-solving behaviours accurately	8. The rubric includes observable behaviours associated with the indicators of problem solving 9. The teachers are familiar with the assessment protocols and rubric and apply the procedures and scoring rules as intended 10. The teachers can observe children closely enough to capture relevant behaviours 11. The assessment is not overly demanding for the teacher
Generalisation inference	The scores representing children’s strengths and weaknesses in problem	If the tasks represent the problem-solving domain and scoring procedures are applied accurately and consistently across tasks,	12. The problems are representative of the broader domain of engineering-design problems

Inferences	Claims	Warrants	Assumptions
(From observed score to universe score)	solving are generalisable across tasks, raters, occasions, and classroom contexts	raters, occasions, and classroom contexts, then the resulting scores can be treated as indicative of performance on other problems within the domain	13. The scores are consistent across the different engineering design problems 14. The scores are consistent across observers 15 The assessment includes multiple occasions
Explanation inference (From scores to explanation)	The scores are representative for the construct as defined by the theory	If the test scores are not biased, then they are indicators for the construct of problem solving as theorised	16. The scores accurately represent the underlying construct of problem solving 17. The scores primarily reflect problem-solving competencies rather than other constructs
Extrapolation inference (From universe score to real-world performance)	The scores representing children's strengths and instructional needs in problem solving reflect their performance when solving real-world problems	If the tasks capture the essential cognitive and behavioural demands of real-world problem solving and children behave during assessment as they typically do in authentic contexts, then scores obtained from these tasks can be interpreted as indicators of their problem-solving performance in real-world situations	18. The assessment captures key cognitive and behavioural demands of real-world problem situations 19. The assessment context reflects real-world problem situations
Decision/use inference (From test results to decisions)	Teachers use results to update their understanding of children's problem-solving competencies and to adjust their instruction and support to better address children's learning needs	If teachers understand the meaning and limitations of the scores and have the capacity to act on them, then the results can inform teachers to take appropriate instructional decisions that support children's learning and development	20. The teachers use the results to update their understanding of children's capabilities, strengths, and needs 21. The teachers use the results to make instructional decisions 22. The teachers learn from the assessment

In this section, the sources of qualitative data are indicated between square brackets. "[FB]" refers to data from the teachers' feedback in their feedback documents, during debriefing, or verbal comments during adherence observations. "[Ass]" refers to the comments provided by the teachers on the observation forms. The letter after the hyphen indicates the teacher (A–D), and the number refers to the number of the assessment, observation, or chunk in the feedback data. For Example: "[FB-D175]" refers to feedback chunk number 175 from teacher D in which she states: "I loved doing it. I enjoy it."

Performance inference

Assumption 1 - The problems are authentic, meaningful, and inherently motivating for the children

To substantiate the assumption, we first draw on existing scientific evidence. To capture preschoolers' real-world competencies, assessments should include authentic (Bagnato, 2005), meaningful (Brassard & Boehm, 2007; Frey et al., 2012) activities that are motivating (Study 3; Snow & Van Hemel, 2008). The OEPS assessment makes use of engineering design problems to elicit problem-solving behaviours. Engineering design is a process directed at "making 'things' that work and making 'things' work better" (Lucas et al., 2014, p. 2). Engineering design problems require the children to create, construct, build, or craft a solution that meets specified needs under constraints (ABET, 2024). Engineering design problems are considered to be authentic and meaningful for young children because they include elements that reflect both the work of real engineers and daily craft and construction play in preschool classrooms (Isabelle et al., 2021; Tank et al., 2018). Engineering activities spark children's curiosity and motivate them to explore their environment (Moore et al., 2018). Drawing on this natural curiosity is what makes engineering design "extremely motivating" (Isabelle et al., 2021, p. 43) for preschoolers.

Besides arguments from the scientific literature, we can also substantiate the first assumption based on the assessment design. The problems used for the assessment were developed to be authentic, meaningful and motivating by aligning them with children's everyday engagement in craft activities and construction play. To further motivate the children, the problems were embedded in a story in which they were, for example, asked to help a toy animal (Arık & Topçu, 2022). The problems were extensively pilot tested by a research intern with 61 preschool children in three classes. Our observations during the pilot revealed considerable enthusiasm among the children, both during and between sessions, as they frequently continued talking about the problem after they had solved it and asked when they could participate again.

Finally, we can substantiate the assumption with the quantitative and qualitative data we collected. That problems are motivating for the children is reflected in the high scores teachers assigned on the indicator commitment, which refers to the child's engagement with the problem (see Figure 9). In 82.35% of the assessments, the teacher evaluated the child's commitment as strong (score 4–5). In only five (3.92%) assessments, commitment was insufficient (score 1–2). In their feedback, teachers reported that the children were "highly intrinsically motivated" [FB-C147], were curious, continued searching until they found a solution, and were excited when it was finally their turn. During the adherence observations, we also noted children's excitement when they finally found a working solution [Ass-D18].

Assumption 2 - The problems provide children with opportunities to display behaviours relevant to all indicators of problem solving

Engineering design problems, the type of problems used in the OEPS assessment, are essentially open-ended and ill-structured, allowing for multiple solutions and solution paths (Study 3; Arık & Topçu, 2022; Dubosarsky et al., 2018; Mangold & Robinson, 2013; Purzer & Douglas, 2018). Engineering design is basically a problem-solving process that parallels the phases and steps of problem-solving frameworks (Gold et al., 2021). To solve engineering design problems, children must go through all steps of the problem-solving process including identifying the problem, gathering information, generating ideas, planning and implementing a possible solution, and evaluating and improving it (Mangold & Robinson, 2013; Sung & Kelley, 2023; Tank et al., 2018).

The capacity of the Volcano Bridge Problem and the Hook-a-Duck Problem to elicit behaviours related to the indicators of problem solving was established in Study 3. The Racetrack Problem was not included in the previous study, but was developed according to the same design criteria (Cunningham et al., 2018). By allowing children to use a variety of familiar materials, or any materials available in the classroom, they have ample opportunities to explore, generate, and test ideas. The photographs of the solutions display a variety of solutions, illustrating that the problems can be solved in multiple ways (Appendix 4).

Finally, the fact that half of the participating teachers systematically rated all indicators and made no use of the DTS option when a judgement was not possible, supports the assumption that the problems elicited behaviours related to all problem-solving indicators.

Assumption 3 - The problems are age appropriate

The scientific literature confirms that engineering design problems, as those used in the OEPS assessment, are age appropriate. According to English (2018), preschool-aged children are capable of solving engineering design problems. They engage in engineering problem solving on a daily basis as they build constructions and use tools during play to achieve their goals (Raven et al., 2018). When engaged in engineering design, children investigate their environment for resources, generate ideas, plan, test, and improve solutions (Shechter et al., 2021).

To ensure the age-appropriateness of the problems used in the OEPS assessment, they were adapted from problems used in studies with children aged 4–6 years (Study 3; Evans, Todaro, et al., 2021; Shechter et al., 2021) and developed according to the age specific design parameters provided by Cunningham et al. (2018). This includes that the problems are presented in a short story, contain up to three success criteria, and can be solved by using a variety of familiar materials. The problems were extensively piloted with 61 preschoolers and refined to ensure that all 5–6-year-olds could start and get stuck, situating them within their zone of proximal development (Vygotsky, 2017).

Data analysis shows that the problems are suitable for children aged 5–6 years. In two out of three (67.7%) assessments, the teacher indicated that the children worked largely autonomously (see Figure 9), indicating that they are able to take ownership of the problem. A participating teacher also noted in the feedback that the tasks were feasible and sufficiently challenging for the children [FB-B85].

Assumption 4 - The children are appropriately scaffolded when truly stuck

As the problems were intentionally designed to be challenging children, children are expected to encounter an impasse at some point. When children cannot overcome the impasse on their own, scaffolds can be provided to motivate them (recruitment and direction maintenance), to avoid disengagement due to frustration (frustration control), and to ensure the can overcome the obstacle while receiving as little support as possible (reduction of degrees of freedom, marking critical features, demonstration; Wood et al., 1976). Scaffolds can be cognitive, affective, and technical (Yelland & Masters, 2007) and should be provided timely (Belland, 2014).

To ensure that children are scaffolded when truly stuck and avoid children losing engagement, becoming frustrated, or getting completely stuck, the assessment protocol specifies when, how, and what type of support may be offered. This support is organised in seven levels, progressing from motivational to task-focused, from vague to concrete, and from suggestive to more directive. Support is provided in the form of questions, supplemented with demonstration if necessary. To prevent support from being given too early (which would limit opportunities to display problem-solving behaviours) or too late (which may lead to frustration or disengagement), it is only offered when children show signs of disengagement or dysfunctional frustration, when they are genuinely stuck, or when they keep circling without making progress. For each new impasse, support is built up from the lowest level.

Although the protocol clearly indicates when and how support should be provided, its application in practice proved less straightforward for teachers. During the adherence observations, it was noted that the teacher regularly deviated from the scaffolding protocol. In more than half of the co-observed sessions (n = 7), support was immediately concrete and directive rather than being built up gradually from vague and suggestive to more concrete and directive. This takes the problem out of the children's hands and limits opportunities to elicit problem-solving behaviours. Although one teacher reported in the feedback that she had followed the support procedure closely [FB-A22], other teachers noted that it is difficult to remain silent and refrain from intervening when children are stuck or making little progress [FB-B99/D173]. The evidence suggests that the assumption is supported only to the extent that teachers adhere to the scaffolding protocol.

We also found that children were not only supported by the teacher but also by one another. Teachers were instructed to pair children who struggled to follow verbal instructions (e.g., children with limited language proficiency or second-language learners) with children who had stronger language proficiency. One teacher reported how a preschooler asked targeted questions to a peer who did not understand the assessment language, thereby effectively supporting the child's participation in solving the problem [FB-B96].

Assumption 5 - The children do not have prior experience with the same or a similar problem that would inflate or distort performance

A problem only exists when the solver has no obvious solution at hand (Mayer, 1994). When confronted with a problem, solvers may draw on similar previous experiences (Jonassen et al., 2006). However, if the solver recalls the solution from earlier experience or has learned a specific strategy, the situation is no longer a problem, but rather a task, meaning that problem-solving behaviours may not be elicited or observed.

To avoid confounding influences from children's prior experience, the problems were developed specifically for this assessment, making it unlikely that preschoolers had encountered them before. If children did have experience with one of the problems that might inhibit the elicitation of problem-solving behaviours, the other two problems could be administered instead.

Despite these precautions, the evidence showed that some children had experience with other, similar problems, which influenced the problem-solving process. One teacher reported the following regarding the Hook-a-Duck problem, indicating potential effects of prior experiences on the required time to solve a problem:

> When solving this, the children proceeded very quickly. Two weeks earlier, a vase had been placed in the reception area, and a whistle had fallen into it during a movement activity. A child then found a solution (turning the vase upside down), which led to a subsequent discussion about alternative approaches. [FB-B126]

If children were to transfer a solution directly from one problem to another, they might solve it without engaging in genuine problem-solving. However, in this situation such direct transfer was not possible. Turning the bottle upside down was not permitted, and children were instructed to design a tool to free the duck, rather than merely to retrieve it.

Children do not only learn from their own experiences but also by observing others. A teacher reported that children were curious and tried to watch what their peers were doing. In some cases, this led to imitation, which can hinder the elicitation of problem-solving behaviours [FB-A68]. More specific, imitation reduces the perceived need to generate new ideas.

Assumption 6 - The context and materials are familiar and accessible to children

To yield useful information about preschoolers' competencies, assessment ideally occurs in a context that reflects their real-life experiences (Bagnato, 2005; Brassard & Boehm, 2007). Moreover, children need to be familiar with the materials, including their possibilities and uses, before they can apply them to solve a problem (Isabelle et al., 2021).

To ensure the context is familiar for the children, the protocol specifies that the assessments should take place in the children's classroom or a spare room at the children's own school and is administered by the children's own (care)teacher. By design, the children could use familiar craft materials, construction materials, tools, and kitchenware to solve the problems.

All participating teachers reported conducting the assessments in their classroom or in a nearby spare room. The photographs of the set-up, materials, and outcomes show that the tasks were carried out using the prescribed materials. Adherence observations likewise revealed no deviations from the protocol at these points, supporting the assumption that the context and materials are familiar and accessible to the children.

Assumption 7 - The context does not hinder children's engagement or performance

Although the context plays an important role in problem solving (Jonassen, 1997; Kirsh, 2008) and preschool assessment requires the context to be authentic (Frey et al., 2012), contextual factors must not hinder children's performance. The context in which the assessment is administered is determined by the time, the place, and the social setting (Damşa et al., 2025). Constraints related to these factors can prevent children from demonstrating their problem-solving competencies. The first factor is time. It is essential that children are given sufficient time to solve the problem to allow them to explore the problem and the environment, and become familiar with the materials and their possibilities (Isabelle et al., 2021), to generate ideas (Beaty & Silvia, 2012), to share their ideas and plans (Ramani & Brownell, 2014; Sormunen et al., 2026), to try things out, to fail, to try again, or to try something different. When children must work within a restricted timeframe, they cannot fully demonstrate their repertoire of problem-solving competencies (Sireci & Botha, 2020).

To ensure that children could solve the problem at their own pace and to prevent time constraints to hinder children's performance, no time limits were imposed for the assessment. Analysis of the observation data indicates that children generally had sufficient time to solve the problem. The observations lasted on average 20.06 minutes (SD = 15.17, min = 2, max = 60). However, the shortest observation lasted only two minutes. Two minutes is very limited to complete the full problem-solving process and may indicate that the problem was already familiar or too easy. When the duration is broken down by teacher and problem, substantial variation emerges (see Tables 13-14). The mean observation time per teacher ranged from 8.36 minutes (*SD* = 6.07) to 32.94 minutes (*SD* = 13.47). The Volcano Bridge Problem took on average 24.27 minutes (*SD* = 13.61), the Hook-a-Duck Problem 13.62 minutes (*SD* = 9.72), and the Racetrack Problem 28.10 minutes (*SD* = 17.23). One teacher also noted in her feedback that some tasks inherently required more time than others [FB-A17]. Based on our experience in Study 3, the pilot study, and the adherence observations, we would argue that at least 10 minutes is required to allow children to demonstrate their problem-solving competencies.

The second contextual factor that may influence performance is the place where the assessment is conducted. Environmental distractions such as noise can disrupt the problem-solving process and assessment (Brassard & Boehm, 2007). To prevent environmental distractions from hindering the problem-solving process during the assessment, teachers were instructed to conduct the observations in a designated, quiet area in the classroom or, if organisationally feasible, in another nearby classroom.

The collected evidence indicates that finding a quiet environment was not always easy. One teacher reported that it initially required some effort to find a quiet location or an appropriate moment, for instance when an additional teacher was present in the classroom [FB-A31]. Teachers also reported that curious preschoolers came to watch and sometimes even offered suggestions, potentially hindering the process. Therefore, teachers preferred to administer the assessments in a separate room [FB-B110].

> It was popular, so there were always people standing around. Sometimes other preschoolers gave suggestions. [FB-A67/A53]

> The other preschoolers in the class were highly intrigued by what the two children were doing and found it difficult not to interfere [FB-C143]

During the adherence observations, which took place in a partially enclosed corner of the children's own classroom, we found that events in the classroom did not hinder children in solving the problem.

The specific location may influence performance, particularly in relation to the Racetrack Problem, as children can use any materials available in the classroom. When the assessment is administered in their own classroom, adequate materials are typically at hand and children know where to find them. However, one teacher reported that, due to unforeseen circumstances, she had to move to an almost empty room [FB-BX93]. She had hastily gathered some items and the materials box for the other problems. Although she encouraged the children to look for additional materials in the corridor, they continued to search for a solution using only the limited materials immediately available. The absence of materials hindered problem-solving performance, although this appears to be an exceptional case.

The third contextual factor is the social setting. The partner with whom the child collaborates may affect performance. If the partner has limited problem-solving, communication, or collaboration skills, children may be unable to demonstrate their full capabilities (Lane & Marion, 2025). To ensure social relationships were supportive rather than hindering performance, teachers were instructed to form heterogeneous pairs, combining verbally strong with less verbally strong children and creative with less creative children and not to pair children who had a difficult relationship. To compensate for peer influences, pairs were rotated for each session.

In their feedback, teachers acknowledged that the composition of the pairs influenced the children's performance [FP-CX141] and that children with lower language-proficiency benefited from the support of their more proficient peer [FB-BX97]. One teacher noted that collaboration proceeded more smoothly when children worked with a friend [FB-C151]. In the open comments of the observations, teachers noted that some children took the lead [e.g., Ass-C38/D21], while others adopted a more compliant role [Ass-A01/D06]. For example, a teacher noted on the observation form: "Is a follower. Copies what the other does." [Ass-D16]. This indicates that the social relationships and the role children assume influences their behaviour and performance.

The presented evidence, in combination with the high scores for commitment as presented in Figure 9, supports the assumption that the context does not hinder children's engagement or performance, at least if the children (are able to) spend sufficient time solving the problem.

Qualifier and potential rebuttals for the performance inference

Based on the backings and the evidence presented above, the performance inference is considered well supported. However, the performance inference only holds if the teacher adheres to the scaffolding protocol, the children cannot make direct transfer from similar problems they solved in the recent past or observed being solved by others, the problem-solving process lasts at least 10 minutes, and social relationships and peer interactions are supportive rather than hindering performance.

Scoring inference

Assumption 8 - The rubric includes observable behaviours associated with the indicators of problem solving.

Observation-based assessment involves scoring indicators on the basis of observable behaviours. Behaviour refers to any physical activity that can be observed externally (Bergner, 2011). A distinction can be made between overt behaviour, which is visible to others, and covert behaviour, which can only be experienced by the individual (Skinner, 1953). Observation-based assessment can only capture overt behaviour. Such behaviour may take the form of physical actions, facial expressions, body posture, or verbal expressions, including private speech, verbal interactions, or vocalisations of joy or frustration. For observation-based assessment, all behaviours linked to the indicators must be clearly defined and described (Tyler, 1934).

The rubric that is used for the OEPS assessment is based on the Problem-solving Observation Scheme for Playful Engineering and the Real-world (PROSPER) that was developed in Study 3. In this study, relevant and observable problem-solving behaviours were identified, assigned to the indicators of problem solving, and clarified through a description and an example. The behaviours linked to each indicator are discussed in detail in the OEPS assessment manual. Table 16 provides a brief description of de problem-solving indicators, along with the associated behaviours expressed as gerunds ("-ing" form) of active verbs. No separate behaviours were formulated for the overarching behaviours (commitment, perseverance, autonomy, responsiveness, collaboration) because they cannot be captured in a single action. An example of the observation form, including the rubric, is provided in Appendix 3.

Table 16. Overview of the problem-solving indicators with assigned observable behaviours

Indicator	Short description: The child …	Behaviours
Exploring	purposefully collects information	Observing, Naming, Describing, Comparing, Associating, Manipulating, Searching, Asking information
Understanding	understands the problem and shows insight	Listening, Confirming, Repeating, Complying
Formulating	generates and/or tries different possibilities	Thinking up, Considering, Asking ideas
Planning	plans and/or prepares actions	Preparing, Thinking ahead
Executing	acts purposefully towards a solution	Doing, Refining, Asking motor help
Monitoring	follows up on progress	Testing, Measuring, Determining, Modifying, Declaring, Asking confirmation
Reflecting	evaluates the process and the result	Recalling, Evaluating solution, Evaluating process
Commitment	shows engagement and remains focused on the task	
Perseverance	persists and does not give up when facing setbacks	
Autonomy	works independently or with minimal support	
Responsiveness	adjusts actions based on tips and suggestions	
Collaboration	shows willingness to cooperate	

Assumption 9 - The teachers are familiar with the assessment protocols and rubric and apply the procedures and scoring rules as intended

To enable teachers to administer the assessment as intended, they should be provided with appropriate documentation and training to familiarise them with the procedures and scoring rules (Snow & Van Hemel, 2008). This documentation, typically provided in user manuals, includes information on the nature of the assessment, its intended uses, scoring procedures, and protocols for administration (AERA et al., 2014). Adherence to the protocols and scoring rules enhances the quality of the scores and the interpretation of the results.

In line with the recommendation of *The Standards* (AERA et al., 2014), teachers were provided with an online training programme, including a comprehensive manual, a multiple-choice test, a video-based practice observation, and an individual meeting. The manual included information on the theoretical framework, a description of the problem-solving process, a description of the problems, and an implementation plan. The multiple-choice test was provided through Qualtrics (https://www.qualtrics.com) and comprised 33 descriptions of behaviours taken from Study 3, which respondents were asked to assign to the corresponding indicators of the problem-solving process. Two video recordings of 26 and 36 minutes from Study 3 were provided as training observations. In these recordings, children individually solved problems similar to those used in the OEPS assessment. Teachers were asked to score the children's behaviours using the observation form and to submit their responses. Finally, the protocols were discussed during an individual, one-hour session (online or in person). Teachers could ask further questions at any time by email.

Despite explicit and repeated requests to complete the full training, the data revealed that only two participating teachers completed the multiple-choice test. Both indicated that they were not fully satisfied with their performance and subsequently revisited the manual thoroughly. Only one teacher scored the practice videos and submitted the results. Failing to complete the training may result in teachers not being fully familiar with the rubric or scoring rules.

Regarding familiarity with the protocols, we already noted that some teachers experienced difficulties applying the scaffolding protocol as prescribed (see Assumption 4). During the adherence observations, we also noted deviations from the reflection protocol. This protocol provides guiding questions about the problem, the outcome, the process, and the children's feelings, with the questions increasing in complexity. In none of the co-observed sessions all components were addressed. The questions remained closed and superficial, follow-up questions were often lacking, and at times only one of the two children was addressed. In four sessions, the reflection phase was too short to allow an adequate assessment of children's reflective proficiency, potentially biasing the scores for reflection.

Analysis of the observation forms indicated that teachers completed the observation schemes as instructed. No scoring irregularities were identified during the adherence observations. However, we noted that most teachers avoided the DTS option in case of uncertainty. During both the pilot test and the adherence observations, it became evident that it was sometimes not possible to score all indicators accurately. For instance, planning behaviour is not always visible, responsiveness is difficult to assess when no prompts are provided, and collaboration can be challenging to assess when children work side by side, especially on the Hook-a-Duck problem. Two teachers provided a score for each indicator for every assessment, leaving nothing blank and without using the DTS option. One teacher left three boxes blank, and another left 45 boxes blank, mainly for planning (*n* = 19) and responsivity (*n* = 17). It is therefore possible that teachers assigned scores based on assumptions or 'gut feeling', rather than based on observed behaviours, which may introduce bias.

The collected evidence suggests that it cannot be assumed that teachers will complete the offered training upon request and that they experience difficulties in putting the scaffolding and reflection guidelines into practice. Moreover, questions arise as to whether teachers always base their scores on observed, rather than assumed, behaviour.

Assumption 10 - The teachers can observe children closely enough to capture relevant behaviours

Children's behaviours can only be observed and scored when the teacher can stay close to them for at least a substantial part of the process. The teacher must also be able to devote full attention to the children being assessed. When attention is divided between observation and classroom management, or when distractions occur, relevant behaviours may be missed, thereby potentially compromising the assessment (Cohen et al., 2018). In addition, background noise in the classroom may hinder capturing children's verbal interactions.

To minimise disruptions from other children and ensure full attention to the observations, the protocol specifies that assessments should take place in a quiet corner of the classroom or, if possible, in a spare room. When conducted in the regular classroom, the protocol suggests placing a low cupboard or table strategically to create a partial visual barrier around the workspace.

Two teachers conducted the assessments in their own classrooms [TC; TD]. To devote sufficient attention to the observations, the assessments were almost always conducted when a second teacher (e.g., a team-teaching colleague or care teacher) was present in the classroom. One teacher alternated between the classroom and a separate room [TA]. To facilitate focused observation, she aimed to conduct the assessments when another teacher was present or during moments without free play, although this was not always feasible [FB-A30/31]. The teacher reported that she was occasionally called away by other children but continued to observe and listen attentively. She also noted that it was possible to follow approximately 95% of the interaction and that, when something was missed, she sometimes asked the child or the partner to clarify what had happened [FB-A65/66].

The adherence observations took place in the children's classroom within a team-teaching context. A second teacher was always present during the observations. During the observations, the teacher consistently remained close to the children. Occasionally, another child approached her with a question, and at times she was called away briefly, but these interruptions were short and did not impair her observations or scoring. However, it was sometimes difficult to fully capture potentially relevant children's dialogues and private speech because of background noise from classmates playing in the room. Taken all together, the collected evidence supports the assumption that teachers are able to observe children closely enough to capture relevant behaviours.

Assumption 11 - The assessment is not overly demanding for the teacher

Administering the OEPS assessment requires teachers to simultaneously observe, score and support children if needed. To ensure feasibility, the observation form was purposefully kept concise (Martin, 1977; Popham, 1997). Scoring involves selecting a performance level for each of the 12 indicators of problem solving and optionally adding comments. Metadata (name, date, problem name, start time, and end time) are completed immediately before or after the observation. At the end of the session, a photograph of the solution was taken.

Teachers reported that simultaneously observing, assessing, and supporting children while managing incidents in the classroom is challenging. Therefore, the observation form was sometimes completed only after the observations [FB-A27]. This is consistent with our observations during the adherence observations. One teacher noted that assessing two children simultaneously in real-life settings is considerably more difficult than scoring one child at a time through video observation, as was the case with the training videos [FB-D168].

The demanding nature of the assessment is also reflected in missing metadata and photographs. Time registration is sometimes overlooked, leading teachers to rely on an estimated duration. During adherence observations, the teacher forgot to record the start and/or end time for four observations. For three observations, the total duration registered by the teacher deviated by more than five minutes from the time recorded by the researcher. At the start of the activity, the focus lies on engaging the children and introducing the problem, which can cause initial time registration to be forgotten. Towards the end of the session, the focus shifts to scoring the observed behaviours, which may explain missing time registration. This may also explain why teachers sometimes forgot to take a photograph of the solution, as noted by one of the teachers, who explained that taking photographs was not part of her routine [FB-A75].

Despite measures to facilitate scoring by limiting the observation form, the evidence indicates that the assessment places considerable demands on teachers. This workload results in delayed scoring, potentially introducing bias, as well as missing data. It is expected that, as teachers gain more experience with the assessment, they will become more adept at scoring and that the recording of metadata and the taking of photographs will become part of their routine.

Qualifier and potential rebuttals for the scoring inference

Based on the evidence and analysis, we consider the scoring inference to be fairly supported. Scoring is facilitated by associating observable behaviours to the indicators of problem solving, providing a comprehensive manual and teacher training, and keeping the observation form concise. However, the scoring inference only holds if teachers complete the training, adhere to the protocols, refrain from scoring based on assumed behaviours rather than observed behaviours, and become more familiar with the assessment through experience.

Generalisation inference

Assumption 12 - The problems are representative of the broader domain of engineering-design problems

As discussed in Assumption 1, the OEPS assessment makes use of engineering design problems. Engineering design is an iterative, creative decision-making process aimed at solving a problem by developing something while taking into account specific criteria and constraints (Asunda & Hill, 2007; Gold et al., 2021; Grubbs & Strimel, 2015). To ensure that the problems used in the assessment are representative of the broader domain of engineering design problems, they were designed in line with criteria derived from the scientific literature on engineering design in early childhood education:

- The problems are hands-on and require children to build or craft a solution (Cunningham et al., 2018; Gold et al., 2021; Lippard et al., 2018).
- The problems share similarities with real-life situations during children's craft and construction play (Shechter et al., 2021).
- The problems are open ended and allow for a variety of solutions (Cunningham et al., 2018; Dubosarsky et al., 2018).
- The problems should be solved within certain constraints (e.g., which materials can be used, where the problem should be solved) and contain two or three success criteria (e.g., concerning stability, dimensions, shape) (Cunningham et al., 2018).
- The problems allow for the application of scientific, technological, engineering, and mathematical concepts and skills (Mangold & Robinson, 2013).
- The problems can be solved with diverse, familiar materials (Isabelle et al., 2021).
- The problems can be embedded in a story (Arık & Topçu, 2022).

Additionally, we ensured that the three problems differed in their objectives (crafting, building, or constructing), criteria, and constraints to include various types of engineering design problems. As a result, the problems used in the OEPS assessment provide adequate coverage of engineering design problems.

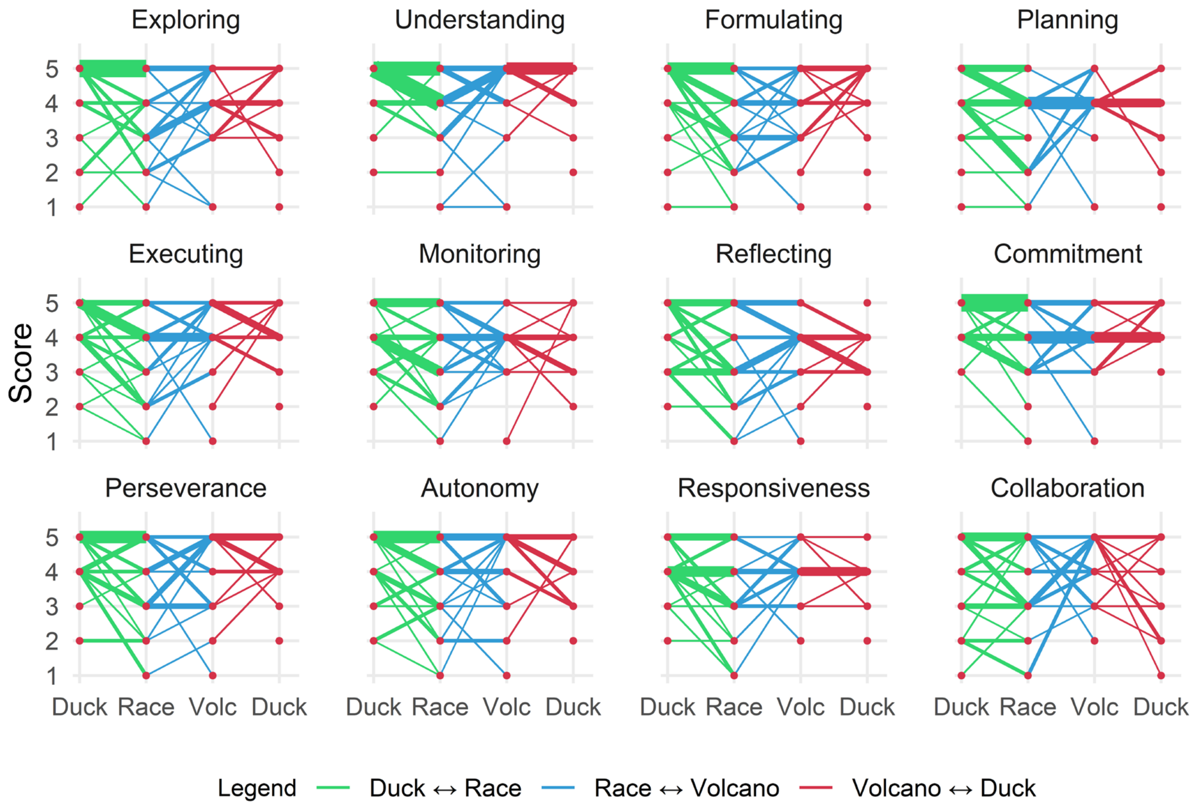
Assumption 13 - The scores are consistent across the different engineering design problems

Problems vary, and therefore the processes used to solve them also differ (Jonassen, 2010a). Different problems elicit different behaviours and thus reveal the indicators of problem solving in distinct ways, as evidenced by the results of Study 3. As the assessment aims to identify children's strengths and instructional needs in problem solving, it is essential to avoid drawing different conclusions merely because of variations between problems.

The three engineering design problems used in the OEPS assessment differ in their objectives (building, constructing, crafting), the success criteria, and the materials that can be used (provided materials versus all materials available in the classroom). The data also indicate that the problems differ in perceived difficulty. According to one of the teachers, children experienced the Hook-a-Duck Problem easier than the Racetrack problem [FB-A80]. This is reflected in the finding that the Hook-a-Duck Problem was, on average, solved in half the time it took to solve the Racetrack Problem (see Figure 9). Differences in difficulty may result in varying support needs and, consequently, in different scores.

Figure 10 presents the differences in scores between problems per child for each indicator. Each line represents the difference in score between two problems solved by a child, with colours indicating which pair of problems is compared. A horizontal line indicates equal scores, whereas steeper lines reflect larger score differences between the two problems. In addition to the horizontal lines that generally appear at the top of the graphs, indicating potential ceiling effects, some lines cross multiple points on the rating scale, reflecting substantial differences between the scores. The substantial variability in collaboration scores may result from children working with different peers on both problems. Only commitment seems to remain relatively stable across problems. However, this can be explained by potential ceiling effects.

Figure 10. Differences in scores between problems per indicator and child (Spaghetti plot)



Note. Each line represents the score difference between two problems. Colours indicate the problem pairs. Line thickness reflects the number of children with the same scores for both problems.

To examine the differences in scores between the three problems in greater depth, we used binomial multilevel analyses. Given the highly skewed distribution of the data (see Figure 9) and teachers' tendency to assign ratings at the upper end of the scale, a pronounced ceiling effect emerged. This limits the identification of exceptionally strong performances. Therefore, we focus on the extent to which the different problems enable differentiation between low-performing children on a given indicator (scores 1–2) and mid- or high-performing ones (scores 3–5). For this purpose, scores were dichotomised. Table 17 presents the frequencies of assessments in which the teacher scored the performance of the child as low, organised by problem and problem-solving indicator. Across problems and indicators, only a limited number of children are identified as low performing. For all indicators, the highest numbers of low-performing children were identified when solving the Racetrack Problem, with the exception of collaboration for which low performers were primarily identified in the Hook-a-Duck Problem.

Table 17. Frequencies of identified low performers by problem and problem-solving indicator

Indicator	Hook-a-Duck (N = 31)		Racetrack (N = 40)		Volcano Bridge (N = 31)	
	Mid or high performing	Low performing	Mid or high performing	Low performing	Mid or high performing	Low performing
Exploring	25 (80.6%)	6 (19.4%)	31 (77.5%)	9 (22.5%)	27 (93.1%)	2 (6.9%)
Understanding	30 (96.8%)	1 (3.2%)	36 (90.0%)	4 (10.0%)	29 (93.5%)	2 (6.5%)
Formulating	29 (93.5%)	2 (6.5%)	34 (85.0%)	6 (15.0%)	29 (93.5%)	2 (6.5%)
Planning	23 (92.0%)	2 (8.0%)	28 (77.8%)	8 (22.2%)	20 (90.9%)	2 (9.1%)
Executing	29 (93.5%)	2 (6.5%)	31 (77.5%)	9 (22.5%)	29 (93.5%)	2 (6.5%)
Monitoring	29 (93.5%)	2 (6.5%)	32 (80.0%)	8 (20.0%)	29 (96.7%)	1 (3.3%)
Reflecting	29 (93.5%)	2 (6.5%)	29 (74.4%)	10 (25.6%)	26 (89.7%)	3 (10.3%)
Commitment	31 (100%)	0 (0.0%)	37 (92.5%)	3 (7.5%)	30 (96.8%)	1 (3.2%)
Perseverance	29 (93.5%)	2 (6.5%)	32 (80.0%)	8 (20.0%)	29 (93.5%)	2 (6.5%)
Autonomy	27 (90.0%)	3 (10.0%)	33 (82.5%)	7 (17.5%)	26 (89.7%)	3 (10.3%)
Responsiveness	26 (96.3%)	1 (3.7%)	31 (81.6%)	7 (18.4%)	18 (94.7%)	1 (5.3%)
Collaboration	22 (73.3%)	8 (26.7%)	34 (87.2%)	5 (12.8%)	30 (96.8%)	1 (3.2%)

Because scores were nested within children, children were included as random effects in the binomial model. As children were also nested within classes, with their teachers as assessors, we additionally estimated a model with class as a third level. However, for none of the indicators did these models fit the data better than the simpler model with children as the only random effect. Table 18 presents the parameter estimates from the model with children as random effects and the problem as a fixed effect. The estimates indicate the probability that children were identified as low performing. For reflecting and commitment, the model could not be fitted adequately, resulting in confounded estimates. To avoid misinterpretation, these indicators were excluded from the overview.

The random intercept estimates for several indicators are extremely high (up to 435.66 for perseverance). This can be attributed to the very small number of children identified as low performing, which inflates the variance attributed to children and therefore does not accurately reflect child-level variability. In contrast, the variance attributed to children is zero for monitoring, indicating a singular fit. Consequently, the random effects estimates provide little meaningful information about the variance explained at the child level.

Table 18. Random and fixed effects of binomial mixed effects models for problem-solving indicators predicting the probability of low performance

Indicator	Random Effect	Var	SD	Fixed Effect	Est	SE	Z value
Exploring	KidID	2.03	1.43	Racetrack (Intercept)	-1.73*	0.75	-2.32
				Hook-a-Duck	-0.11	0.7	-0.16
				Volcano Bridge	-1.55	0.94	-1.65
Understanding	KidID	161.87	12.72	Racetrack (Intercept)	-9.11*	2.49	-3.65
				Hook-a-Duck	-1.2	3.49	-0.34
				Volcano Bridge	-1.98	2.29	-0.86
Formulating	KidID	54.21	7.36	Racetrack (Intercept)	-6.46*	1.93	-3.35
				Hook-a-Duck	-3.51	1.94	-1.81
				Volcano Bridge	-1.53	1.52	-1.01
Planning	KidID	185.14	13.61	Racetrack (Intercept)	-7.93*	2.05	-3.87
				Hook-a-Duck	-6.33*	2.55	-2.49
				Volcano Bridge	-6.33*	2.55	-2.48
Executing	KidID	0.70	0.83	Racetrack (Intercept)	-1.43*	0.66	-2.16
				Hook-a-Duck	-1.56	0.91	-1.72
				Volcano Bridge	-1.51	0.87	-1.74
Monitoring	KidID	0.00	0.00	Racetrack (Intercept)	-1.39*	0.4	-3.51
				Hook-a-Duck	-1.29	0.83	-1.55
				Volcano Bridge	-1.98	1.09	-1.82
Perseverance	KidID	435.66	20.87	Racetrack (Intercept)	-9.64*	2.23	-4.32
				Hook-a-Duck	-10.67*	4.27	-2.5
				Volcano Bridge	-2.16	2.19	-0.99
Autonomy	KidID	59.94	7.74	Racetrack (Intercept)	-6.79*	1.9	-3.57
				Hook-a-Duck	-1.34	1.33	-1
				Volcano Bridge	-0.46	1.29	-0.36
Responsiveness	KidID	274.13	16.56	Racetrack (Intercept)	-8.94*	2.14	-4.18
				Hook-a-Duck	-9.77*	2.33	-4.19
				Volcano Bridge	-2.8	2.66	-1.05
Collaboration	KidID	90.02	9.49	Racetrack (Intercept)	-8.5*	2.42	-3.51
				Hook-a-Duck	1.85	1.42	1.31
				Volcano Bridge	-4.27*	2.11	-2.02

Note. *Significant at z-value = 1.96; Var(0.00): Model showed a singular fit, indicating that the random-effect variance was estimated at zero. Reflecting and commitment were excluded from the overview due to model misfit.

The Racetrack Problem was used as the baseline (intercept) because it was associated with the highest observed prevalence of low performance among children (see Table 17). Negative estimates for the intercept indicate that the probability of a child being identified as low performing on the Racetrack Problem is below 50%. Negative estimates for the other problems indicate a lower likelihood of being identified as low performing relative to the Racetrack Problem (reference category). Estimates express the likelihood of low performance on the log-odds scale. For example, for the Racetrack Problem (intercept), the estimate (β) for exploring of -1.73 on the log-odds scale corresponds to a 15% probability ($p = \frac{e^{-1.73}}{1+e^{-1.73}}$) that a teacher identifies the child as low performing. This probability is statistically significantly different from zero (z = -2.32).

For most indicators, there is no significant difference between the problems in the likelihood that a teacher identifies a child as low performing. Although, a few indicators did show meaningful differences. Children are most likely to be identified as low performing in planning when working on the Racetrack Problem. Compared to the Racetrack Problem, the Hook-a-Duck Problem reduces the likelihood of children being identified as low performing for perseverance (β = -10.67, z = -2.50) and responsiveness (β = -9.77, z = -4.19), and the Volcano Bridge Problem reduces it for collaboration (β = -4.27, z = -2.02).

Due to the differences in difficulty between the problems as reported by the teachers, the substantial within-child differences in scores between problems (Figure 10), and the differences in the likelihood of identifying children as low performing for certain indicators, it cannot be assumed that the scores are consistent across engineering design problems. However, the limited number of data points and the highly skewed distribution of the data limit the precision of analyses and the clarity of interpretations.

Assumption 14 - The scores are consistent across observers

Although across-observer consistency or interrater reliability is not a primary concern in formative assessment (for a discussion on this topic, see Parkes, 2013), it remains relevant to determine how consistently teachers identify children as low performing. Analyses of interrater reliability can indicate the risk that certain instructional needs are detected by some teachers but go unnoticed by others. Failing to notice such needs may result in suboptimal information for making instructional decisions, leaving opportunities to support children's problem-solving development unexploited.

To ensure consistent scoring, all teachers received the same training (see Assumption 9). The assessment manual provides clear instructions on the scoring rules which were further clarified during an individual meeting. Across-observer consistency was assessed through 13 parallel observations (totalling 5.62 hours) conducted by the first author alongside and independently from a teacher, resulting in 26 individual assessments.

Because the data were highly skewed and the analysis focused on interobserver consistency in identifying children as low performing, scores were recoded as dichotomous variables. Table 19 summarises the absolute number of differences between raters and the percentage of agreement based on the dichotomised variables. The data indicate high interrater agreement for almost all problem-solving indicators. However, absolute differences may be confounded by chance. Commonly used indices for analysing inter-rater reliability that adjust for chance agreement, such as Cohen's kappa, are strongly penalised when raters almost always agree (Heyman et al., 2014). Gwet's first-order agreement coefficient (AC_1) compensates for this imbalance by estimating chance agreement only in situations where raters could realistically disagree, thereby providing a more stable and interpretable reliability estimate in cases with high inter-rater agreement (Gwet, 2008; Wongpakaran et al., 2013). However, Gwet's AC_1 is found to produce interrater estimates that are overly optimistic and critics warn against interpreting the estimates according to the Landis and Koch (1977) cut-offs for Cohen's Kappa (Vach & Gerke, 2023; Xu & Lorber, 2014). For transparency, Cohen's Kappa is presented along with Gwet's AC_1. The results indicate absolute agreement between the observers for planning, commitment, perseverance, and collaboration, with exploring and reflecting showing the lowest agreement across indices. These results provide moderate support for the assumption that scores are consistent across observers.

Table 19. Interrater reliability in identifying low performance per indicator of problem solving

Indicator	N	Differences (N)	Agreement (%)	Cohen's Kappa	Gwet's AC_1	se
Exploring	24	7	70.8	0	0.612	0.161
Understanding	26	2	92.3	0	0.917	0.062
Formulating	26	4	84.6	0.288	0.807	0.104
Planning	5	0	100	1	1	0
Executing	26	1	96.2	0.649	0.957	0.045
Monitoring	23	2	91.3	0	0.905	0.071
Reflecting	19	5	73.7	0.128	0.624	0.182
Commitment	26	0	100	Na	1	Na
Perseverance	24	0	100	Na	1	Na
Autonomy	23	1	95.6	0	0.955	0.047
Responsiveness	9	2	77.8	0	0.723	0.223
Collaboration	23	0	100	100	1	Na

Note. AC_1: Gwet's Agreement Coefficient; se: standard error for the AC_1.

Assumption 15 - The assessment includes multiple occasions

As performance and scores may be influenced by the context and the presented problem (Jonassen, 2010a), it is important that the assessment is conducted across multiple occasions with different partners and problems. According to Brassard and Boehm (2007), at least two observations per child are needed for instructional planning. To capture performance variability across occasions, all preschoolers solved two out of three engineering design problems. In each session, the children were paired with a different classmate to compensate for peer influences.

Teachers occasionally noted differences they observed during the second session on the observation form. For example, one teacher wrote, in relation to the Hook-a-Duck problem: “There was a strong contrast with the Racetrack Problem” [Ass-B12], indicating that she took task-related differences into account. Another teacher demonstrated awareness of potential peer influences by reporting differences related to the interaction between children compared with the first session, as illustrated by the following comments:

> Monitoring went better because she was paired with a different child who communicated more with her. [Ass-C06]
>
> Collaboration was much better this time because he was working with a friend. [Ass-C13]
>
> She quickly lost interest the second time. During the first session, she was carried along by the other child’s enthusiasm. [Ass-C22]

Figure 10 already demonstrated significant differences in the scores teachers assigned to a child across the two problems. If scores from two observations diverge substantially, additional observations are warranted to identify strengths and instructional needs across occasions.

Qualifier and potential rebuttals for the generalisation inference

Based on the evidence and analysis, we consider the generalisation inference to be moderately supported. By developing the problems according to specific criteria for engineering design problems, the problems used for the OEPS assessment are representative of the domain. We also found that different observers generally, though not always, agreed on whether the child’s performance was low for any of the indicators of problem solving. However, we found substantial differences in scores across problems. By conducting observations on two occasions, context- or task-related score differences can be identified. When scores differ substantially between occasions, additional observations are needed to make valid instructional decisions.

Explanation inference

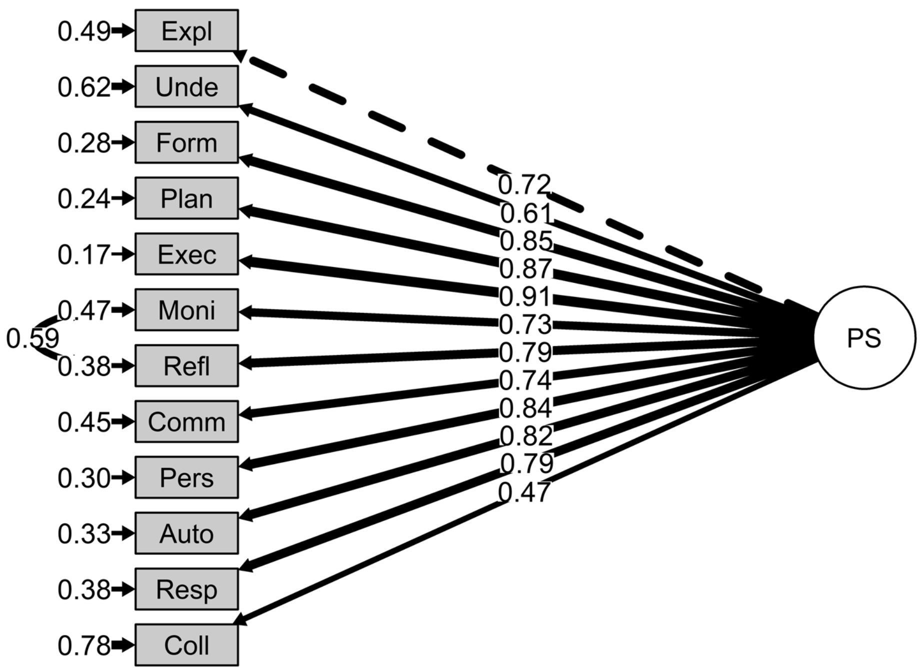
Assumption 16 - The scores accurately represent the underlying construct of problem solving

The observation instrument is grounded in the PISA problem-solving framework (OECD, 2013) which, in turn, is based on Polya’s (1945) four-phase model of problem solving. The instrument includes seven of the eight PISA indicators. Representing was omitted in Study 3 as it proved nearly unobservable, even under more controlled conditions. The observation scheme was supplemented by four generic indicators identified in Study 3: commitment, perseverance, autonomy, and responsiveness. Collaboration was added because children are allowed to work in pairs during the assessment.

Given the strong theoretical and empirical foundation, construct validity was examined through a confirmatory factor analysis (CFA) specifying problem solving as a single latent factor. The initial model with all 12 indicators showed acceptable fit for CFI (.868), TLI (.839), and SRMR (.068), but poor fit for RMSEA (.154). Inspection of modification indices indicated that monitoring and reflecting shared substantial residual variance beyond what was explained by the latent factor. This is theoretically plausible, as these two processes are paired within the PISA framework. Allowing their residuals to covary significantly improved model fit ($\Delta\chi^2(1) = 29.07$, $p < .001$), resulting in CFI = .907, TLI = .884, SRMR = .063, and RMSEA = .131. However, the TLI remains under the conventional .90 threshold for acceptable model fit and RMSEA exceeds the commonly recommended cutoff of .08 (Brown, 2015). The poor model fit may be explained by the relatively small sample size in relation to the number of degrees of freedom (Kline, 2016).

Although the model fit was suboptimal, all factor loadings were statistically significant and generally strong. As shown in Figure 11, 10 of the 12 indicators showed strong loadings ($\lambda \geq .70$). Formulating, planning, executing, perseverance, and autonomy demonstrated particularly high loadings ($\lambda > .80$), indicating that these indicators are strong representations of the latent construct. Understanding showed a more moderate loading ($\lambda = .61$) with comparatively higher residual variance (.62). Collaboration emerged as the weakest indicator ($\lambda = .47$), and its high residual variance (.78) suggests that it captures substantial variance not shared with the latent factor and may partly reflect a distinct construct. This can be attributed to the fact that collaboration is not included in the PISA model nor in the model proposed in Study 3. The relatively low factor loading for collaboration may contribute to the observed model misfit. These results provide support for the assumption that the scores represent the underlying construct of problem solving.

Figure 11. One-factor CFA model of problem solving with standardised factor loadings and residual variances



Assumption 17 - The scores primarily reflect problem-solving competencies rather than other constructs

Theoretically, problem solving is closely related to creativity, executive functions, and self-regulation. Guilford (1967) argued that the similarities between problem solving and creativity are so substantial that it is advisable to treat them as a single construct. Zelazo et al. (1997) link the different forms of executive functioning to the distinct phases of the problem-solving process, whereas Diamond (2013) identifies problem solving as a higher-level executive function grounded in inhibitory control, working memory, and cognitive flexibility. Zimmerman and Campillo (2003) argue that problem solving integrates metacognitive and social cognitive processes such as planning and self-efficacy. In addition, language proficiency (receptive and expressive language) plays an important role throughout the problem-solving process. Grubbs and Strimel (2015) identify engineering design as a problem-solving model that acts as “the great integrator” that integrates knowledge and skills across STEM domains.

In this study, we did not systematically assess discriminant validity. No data were collected that would allow for reliable examination of whether the scores correlate strongly with validated measures of creativity, executive functions, self-regulation, or language proficiency. However, based on the conceptualisation of the indicators, we expect formulating to relate to creativity, as it involves generating multiple ideas. Planning and monitoring are conceptually linked to executive functions and self-regulation. Receptive and expressive language proficiency can be linked to multiple indicators of the problem-solving process, including asking for additional information about the problem or criteria (exploring), understanding the verbally explained problem, goal, and success criteria (understanding), discussing ideas with a peer (formulating), sharing plans with the peer (planning), communicating observations from trials (monitoring), evaluating the result and the process (reflecting), understanding the teacher’s tips and suggestions (responsiveness), and collaborating with a peer (collaboration).

Drawing on the literature and the conceptualisation of the problem-solving indicators, some overlap between the indicators and related constructs is to be expected. This overlap is not problematic, given that these constructs are an integral part of the problem-solving process.

Qualifier and potential rebuttals for the explanation inference

Based on the evidence from the confirmatory factor analysis and theoretical considerations, we consider the explanation inference to be well supported. The CFA indicated that the 12 indicators loaded well on a single latent factor representing problem solving, except for collaboration. Theoretically, problem solving is closely linked to related constructs such as creativity, executive functions, self-regulation, and language proficiency. Further research incorporating validated measures of creativity, executive functions, self-regulation, and language proficiency may provide a clearer understanding of the relationships between the problem-solving indicators and related constructs.

Extrapolation inference

Assumption 18 - The assessment captures key cognitive and behavioural demands of real-world problem situations

The OEPS assessment makes use of engineering design problems, which bear strong similarities to the craft and construction play that young children naturally engage in (Fleer, 2022; Heroman, 2017; Shechter et al., 2021). During free play, preschoolers frequently engage in engineering activities that require them to solve problems (Bairaktarova et al., 2011; Gold & Elicker, 2020). Moreover, children's engineering play reflects the work processes of real engineers (Bairaktarova et al., 2011; Gold & Elicker, 2020; Isabelle et al., 2021). Consequently, the problems used in the OEPS assessment closely resemble real-world situations experienced by both preschool children and adults.

The observation scheme used for the assessment is based on the PISA problem-solving framework, which was originally intended to capture children's "overall capability to solve problems in real-life situations beyond the specific context of school subject areas" (OECD, 2004, p. 16). The PISA framework is a generic problem-solving framework that parallels well-known frameworks by Polya (1945) and Schoenfeld (1985) on mathematical problem solving; Treffinger (1995) and Osborn (1953) on creative problem solving; and Bransford and Stein (1993) on real-world problem solving. It can be used for both technology-oriented and social problems (OECD, 2013), for individual and collaborative problem solving (OECD, 2017a), and in digital environments (OECD, 2013) as well as real-life settings (Study 3). Consequently, the observation scheme captures the key elements of the real-world problem-solving process.

Assumption 19 - The assessment context reflects real-world problem situations

The assessment mirrors real-world contexts in terms of time, place, and social environment, as it is administered during normal school hours in the children's own classroom or a familiar classroom and administered by the children's teacher. The school constitutes an important living and learning space for children aged 5–6 years who attend preschool at least 290 half-days per year (https://www.vlaanderen.be/onderwijs-en-vorming/leerplicht/leerplicht-van-5-tot-18-jaar).

Regarding the place or physical environment, the observations generally took place in one of the learning centres in the children's classroom. Flemish preschool classrooms are typically organised into learning centres or play centres. These are clearly defined classroom areas dedicated to compatible activities such as block play, construction, dramatic play, or reading (Sanoff, 1995). The materials available in these centres enable and provoke play and learning activities aimed at achieving specific educational goals. One teacher reported administering the assessments in the loose parts centre [FB-B125], while the assessments observed by the researcher as part of the adherence observations were conducted in the construction centre. These centres constitute natural settings in which children engage in making and crafting activities.

Concerning the social environment, many problems encountered in the real world requires some kind of collaboration to be solved (OECD, 2017b). Similarly, children frequently engage in collaborative problem solving during play (Ramani & Brownell, 2014). Collaboration is inherent to problem solving and to children's natural play and should therefore be reflected in authentic assessment (Frey et al., 2012; Wiggins, 2011). Accordingly, the OEPS assessment requires preschool children to solve the problems in pairs.

In addition, the OEPS assessment was administered by the children's own teachers, and the protocols were designed to reflect typical teacher-child interactions during guided play and structured learning activities. This way, the assessment conditions closely reflect the real-world context of preschoolers.

Qualifier and potential rebuttals for the extrapolation inference

Based on the close alignment between the OEPS assessment and preschool children's everyday experiences, real-world problem solving, and the practices of professional engineers, we consider the extrapolation inference to be well supported.

Decision/use inference

Assumption 20 - The teachers use the results to update their understanding of children's capabilities, strengths, and needs

As illustrated in the heatmap in Figure 9, teachers rated children's performances predominantly very positively across all indicators. In 25.3% (planning) to 57.8% (understanding) of the assessments, the teacher assigned the maximum score (5). These high scores produce a ceiling effect, making it difficult to distinguish truly exceptional performances based on the quantitative scores alone. However, in their open-text comments teachers did identify notable strengths in children, as evidenced by the following excerpts from the observation forms:

> [The child] finds a solution very quickly. Works calmly but precisely. Good preparation. Also tells the other child what she is going to do and how. [Ass-A17]

> Very engaged. This is really his thing. His imagination sets the activity in motion. He quickly immerses himself in that world, allowing him to develop a beautiful creation. [Ass-A25]

> Very entrepreneurial. [Ass-B04]

> It is very striking how efficiently she acts immediately. Very accurate. [Ass-B05]

> [Planning] Very calculated and precise. [Ass-B05]

> [Commitment] Asks spontaneously whether he can/may solve more problems. [Ass-B12]

> [Collaboration] Is very helpful towards the other child. [Ass-A18]

> [Collaboration] Very diplomatic. [Ass-B13]

Weaker performances are more noticeable in the data than strong ones. Only 2.0% (commitment) to 6.0% (exploring and planning) of the assessments received the lowest score. In almost all cases, teachers explicitly elaborated on the low scores in their comments. These remarks provide clarification or interpretations of the scores, as illustrated by the following comments:

> Fear of failure. Does not get to the point of starting the task. [Ass-A05]

> Asks many questions. Very insecure. Often responds: I don't know, I can't … [Ass-A10]

> It is very striking that he undertakes little himself. He verbalises a lot of what he sees and thinks and asks the other child for permission noticeably often before doing anything. [Ass-B07]

> Works rigidly alongside the other child. Has an idea immediately and carries it out without many words or listening to the other's solution. [Ass-A21]

> He waits and imitates without understanding what he is doing. Often asks for help. [Ass-C27]

> She did not really work together; there was no communication at all, apart from going for the same materials. She was clearly the leader, the other child followed. They did not reach a solution and did not take the tips on board. [Ass-C11]

In addition to the identification of children's problem-solving strengths and instructional needs, the assessment enables teachers to refine their broader understanding of young children and their capabilities. Teachers reported that the observations helped them get to know their pupils better or that they were surprised by their performances as is illustrated with the following excerpts from the feedback:

> It is fascinating to observe how children arrive at a solution. You also get to know them better. [FB-C136]

> There were children who really surprised me. [FB-B99]

> I had not expected [anonymised] to be so capable. She is usually very insecure. [FB-D180]

Teachers also noted that the observations helped them understand children's social-emotional functioning.

> Children with fear of failure really stand out here. What you notice less in regular classroom practice becomes much more visible. [FB-A13]

> You can clearly recognise leadership or more compliant children. [FB-A5]

> You get to know children socially: patience, helpfulness, stubbornness, or act for personal profit. [FB-A7]

Teachers further indicated that the observations offered better insight into children's language skills and language use. The protocol instructed teachers to form heterogeneous pairs, for example by pairing verbally strong with less verbally strong children. Teachers observed that proficient native speakers sometimes supported newcomers with limited vocabulary:

> The children without language knowledge were paired with a verbally strong child. This worked well. The child actually learned a lot of vocabulary from the other. The other child asked very targeted questions. It was rich language interaction. Afterwards one child began giving instructions and the other followed them. [FB-B96]

> I have now seen a completely different child from what I normally see. She never uses so much language. She was really thinking along. I did not expect that. [FB-D171]

These examples from the teachers' feedback illustrate that the observations effectively helped them update their understanding of children's capabilities, their strengths and instructional needs.

Assumption 21 - The teachers use the results to make instructional decisions

Formative assessment primarily concerns changes in teachers' classroom practice that enable them to tailor instruction more effectively to children's learning needs (Black et al., 2003). According to Wiliam (2011), teachers use results from formative assessments to improve the quality of evidence on which their instructional decisions are based.

In their feedback and during the debriefing, teachers reported adjustments they had made or planned to make to their teaching and instruction. Based on their observations, teachers indicated that they intend to focus more on collaboration [FB-A15; FB-B88], on familiarising children with a wider range of materials [FB-A18; FB-B89; FB-D161], on the development of specific skills such as independent thinking and reflection [FB-B87], on (situated) dispositions such as self-confidence [FB-A52] and self-reliance [FB-B134], and on increasing differentiation in provided challenges [FB-B91]. They also expressed their intention to devote more time to support and observe children solving unexpected, daily problems [FB-B120].

> From that experimentation phase, we also identified the researchers. Who tries? Who tinkers? That is why we decided to create experimentation boxes, so that children can freely experiment with different materials. [FB-A18]

> The factor of self-confidence plays a very important role. We really need to stimulate it, because it can form a barrier. [FB-A52]

> You have to see it. You have to see that there is a problem, and you have to make time for it. Every moment is a learning moment. [FB-B120]

Teachers also reported sharing the results of their observations with other stakeholders, such as care coordinators [FB-D183], learning support staff [FB-A14], and parents [FB-B99; FB-D182] to discuss children's performance and development.

> We learned things from the preschoolers. We even used them during the parent-teacher meetings. [FB-B99]

> I will enter that immediately into the student monitoring system to use it during the parent-teacher meeting. [FB-D182]

These examples demonstrate that teachers effectively used the assessment results to inform instructional decisions aimed at supporting children's development, and to guide discussions with other stakeholders surrounding the child.

Assumption 22 - The teachers learn from the assessment

Black et al. (2003) argue that formative assessment changes teachers' thinking about their teaching. This implies that formative assessment benefits not only children's learning, but also teachers' professional development. The authors report that formative assessment helps teachers to reconsider their teaching, shifting from a curriculum-oriented approach to a child-centred one.

In their feedback, teachers reported that, through the observations, they learned more about how preschoolers (as an age group) solve problems [FB-A16], that they gained insights into other competencies such as collaboration and creativity [FB-D160], and that they developed their skills in observing young children [FB-A11].

> We do this far too little in early childhood education. We should invest more in problem-solving activities and executive functions. I think it's great. It gave me a lot of knowledge and ideas to build on. It provides insight. [FB-A34]

> I will definitely keep these problems to use again with preschoolers in the future, and to observe them as well. You can learn so much from them. [FB-C159]

The observations also prompted teachers to reflect on their own teaching practice and their role as teachers. One teacher noted that she often asks the same questions, which limits the progress she aims to achieve, and that it is sometimes difficult to come up with more in-depth questions [FB-A42]. Another teacher observed that too little attention is paid to process-oriented practice and that the teacher's role as a coach could be strengthened as is evidenced by the following quotes.

> I gained many insights from these observations during the STEM activities. They provided valuable material for reflection and further development. [FB-A1]

> As a teacher, guide rather than direct: ask open questions, give thinking time, and encourage children to persevere when facing difficulties. [FB-B92]

Qualifier and potential rebuttals for the decision/use inference

Given the strong alignment between the intended purpose and use of the assessment and the reported use by participating teachers, we consider the decision/use inference to be well supported. Teachers reported using the assessment results to identify children's strengths and instructional needs, to adapt instruction based on assessment results, to discuss outcomes with relevant stakeholders (e.g., care teachers, co-teachers, parents), to learn about children's problem-solving processes, and to reflect on their own teaching practices. However, we should note that support for the decision/use inference relies largely on teachers' self-reports. Further research is needed to investigate the actual use of the assessment data for instructional decision-making.

Evaluating the validity argument

Table 20 summarises the main outcomes of the validity argument, indicating the qualifiers and rebuttals for each inference. Based on the evaluation of the assumptions underlying the warrants for the inferences, the existing and newly collected theoretical and empirical evidence supports the performance, extrapolation, explanation, and decision/use inferences. However, the warrants for these inferences only hold if the children cannot make a direct transfer from earlier experiences, peers do not hinder performance, the problems are not solved too quickly, additional observations are conducted in case of discrepancies between scores from two observations, and quantitative results are complemented with qualitative comments on children's performance. In addition, further research is needed to investigate the relationships between problem-solving indicators and related constructs.

The scoring inference is fairly supported. Although we found that teachers were able to observe the children closely enough to capture their problem-solving behaviours, the assessment was found to be cognitively demanding, which makes it prone to error. In addition, we found that not all teachers completed the training. Failure to complete the training may result in deviations from the protocols, which could hinder the elicitation of problem-solving behaviours and compromise the accuracy of scoring.

The weakest link in the chain of inferences is the generalisation inference, which we consider only moderately supported. Although the presented problems are representative of the broader domain of engineering design tasks, we found meaningful differences between problems and observers for some problem-solving indicators.

Table 20. Summary of the validity argument

Inferences	Qualifier	Rebuttals (The inference holds if …)
Performance inference	Well-supported	The teachers adhere to the scaffolding protocol. The children cannot make direct transfer from previous problems or solutions they have observed from others The problems are not solved too quickly. The social relationships between collaborators are supportive.
Scoring inference	Fairly supported	The teachers complete the training. The teachers adhere to the scaffolding and reflection protocol. The scores are based on observed behaviours rather than assumed behaviours. The teachers have experience with administering the assessment. The scoring rules are followed.
Generalisation inference	Moderately supported	Additional observations are conducted in cases of discrepancies between observations.
Explanation inference	Well-supported	The relationship between problem-solving indicators and related constructs is established in follow-up studies.
Extrapolation inference	Well-supported	
Decision/use inference	Well-supported	Quantitative results are complemented with qualitative comments on children’s performance.

The finding that generalisability, and more specific consistency in scores, constitutes the weakest link in the chain of inferences can be situated within an ongoing scholarly debate on the role and importance of reliability in formative assessment. Reliability concerns the precision, accuracy, consistency, and stability of a measurement (Cohen et al., 2018). It is assumed that observed scores consist of a true score and random error (Cronbach et al., 1972; Ebel & Frisbie, 1991). To ensure the reliability of the scores, this random error should be minimised. However, the importance of reliability increases with the stakes associated with the decisions and interpretations drawn from the assessment.

Some scholars argue that reliability, traditionally conceptualised as consistency of scores across occasions and assessors (Lee & Harris, 2025), is a prerequisite for validity, and that reducing error should therefore be a key consideration in all forms of assessment (e.g, AERA et al., 2014; Brennan, 1998). Others argue that, in formative assessment, reliability is better approached as the reduction of errors and mistakes rather than as a statistical index of consistency and that consistency is less central than authenticity and the ability to capture meaningful learning (Parkes, 2013). In the context of formative assessment, limited reliability is not necessarily problematic (Crooks et al., 1996; Kane & Wools, 2019; Lane & Marion, 2025). The purpose of formative assessment is not to generalise, compare, or support high-stakes decision-making, but rather to generate timely information that can guide instructional decisions and support children's learning (Wiliam, 2011).

Nevertheless, achieving the highest possible reliability is also desirable in the context of formative assessment (Brennan, 1998). The reliability of the assessment could be increased by a higher degree of standardisation. However, this would simultaneously move the assessment further away from problem solving as the assessment's target domain. This is what Kane (2013b) referred to as the *reliability/validity paradox*. Standardisation enhances the generalisability of scores by narrowing down the situations over which generalisation is possible. Kane (2013b) points out that if performance varies across problems and contexts, and task formats are fixed, then standardisation introduces systematic error. Although standardisation may thus improve reliability, it can limit the extent to which assessments provide teachers with information to inform instructional decisions that support young children's learning in real-world contexts (Archbald & Newmann, 1988). It is precisely the flexibility of formative assessment that enables teachers to gather targeted information within *their* context and with *their* children (Lane & Marion, 2025).

In formative preschool assessment, both adherence to the protocols and flexibility towards children's individual needs and contextual circumstances are essential. This tension between adherence and flexibility is acknowledged in *The Standards* as it states:

> Each step toward greater flexibility in the assessment procedures enlarges the scope of the variations allowed in replications of the testing procedure, and therefore tends to increase the measurement error. However, some of these sacrifices in reliability/precision may reduce construct irrelevance or construct underrepresentation and thereby improve the validity of the intended interpretations of the scores. (AERA et al., 2014, p. 36)

Some degree of flexibility is required to respond to young children's volatile nature (Ellingsen, 2016), especially in performance-based assessments conducted in natural settings (Riley et al., 2016). Furthermore, flexibility is required to ensure fairness and offer all children, including those from disadvantaged backgrounds or with atypical development, equivalent opportunities to demonstrate their competencies. (AERA et al., 2014; Ercikan & Solano-Flores, 2025). In addition, flexibility is needed for dealing with the dynamic and often unpredictable environment of preschool classrooms.

In other words, the tension between reliability and validity can also be approached as a tension between standardisation and flexibility (Lane & Marion, 2025). By analogy with Kane's reliability/validity paradox, we could call this the *standardisation/flexibility paradox*. The standardisation/flexibility paradox requires careful balancing and continuous consideration of the need for adherence to the protocols versus the need for flexibility in response to individual differences between children and specific contextual factors. The higher the stakes, the more important reliability becomes (Parkes, 2013), and consequently the greater the need for standardisation to ensure consistency in scoring. When the emphasis shifts towards the timely collection of data to inform instructional decisions, flexibility gains importance to account for individual differences between children and varying contextual factors and to improve fairness and alignment of the assessment with the instructional goals (Lane & Marion, 2025).

Carefully considering the tension between standardisation and flexibility is particularly relevant in the context of formative preschool assessment aimed at supporting the development of problem-solving competencies. First, formative preschool assessment ideally takes place in an authentic environment (Grisham-Brown et al., 2006). Preschoolers are sensitive to changes in their context (e.g., distractors in the room, familiarity with the test administrator, or events preceding the assessment) and their behaviour can vary from day to day (Brassard & Boehm, 2007). Second, ill-structured problems such as engineering-design problems are essentially embedded in their context (Jonassen, 1997). Even more, Jonassen (2010a) argues that in ill-structured problems, "the context in which problems are embedded becomes a significant part of the problem and necessarily part of its solution" (p. 8).

The context in which preschool children's problem-solving competencies are elicited, observed, and assessed constitutes an integral component of the assessment itself. This context dependency of the assessment was also noted by the teachers who participated in the study. In response to the question "Which factors do you consider crucial for successful problem solving?", two teachers independently responded: "The context" [FB-A52/B109]. At the same time, this contextuality complicates the reliable assessment of problem-solving competencies. If problem solving is inherently context-bound, then generalisation over problems and occasions is, by definition, challenging. To enhance the reliability of problem-solving assessment for research purposes or high-stakes assessments, further research is needed into the specific contextual factors that influence children's problem-solving behaviours. These may include for example characteristics of the presented problems, the provided or available materials, the physical environment, relationships and collaboration with peers, the relationship with and support from adults, or the time of the day.

In light of the context-bound nature of problem solving and the aforementioned tension between standardisation and flexibility, it is advisable to consider in advance which aspects of the protocol can accommodate a degree of flexibility and which require strict adherence when developing and implementing formative preschool assessments. This approach allows for finding the ideal balance between reliability and validity. The following examples related to the OEPS assessment illustrate this point. When teachers adapt the narrative in which a problem is embedded to align more closely with a broader theme or children's interests, this may enhance children's motivation to engage in problem solving and thereby improve validity without undermining reliability. Similarly, adding or omitting certain prescribed materials may help to better align the problems with children's capabilities, for instance when a task proves to be too difficult or too easy. Although the materials provided may influence the range of possible solutions and the problem-solving process, such adaptations can be beneficial and are sometimes even necessary for assessment purposes, for example to avoid floor and ceiling effects. Conversely, deviations from the scaffolding protocols, for example by intervening too quickly or by providing overly specific guidance, may hinder the elicitation of problem-solving behaviours and compromise the outcomes. The same applies to deviations from the reflection protocol. Asking only superficial, closed questions, or moving too quickly on to more complex, open questions, may prevent children from demonstrating the full extent of their reflective abilities. When it comes to eliciting reflective thoughts, flexibility is less appropriate and standardisation is warranted.

Limitations and implications for research and practice

It should be acknowledged that it is impossible to make a conclusive validation argument on the basis of a single study (Chapelle & Lee, 2021). The aim of this study was therefore not to validate the OEPS assessment itself. This would contradict the core principle of the argument-based approach to validation that validity is not a property of the instrument but of its interpretation and use (Kane, 2013b). Rather, the aim was to identify the critical issues that must be considered when conducting formative preschool assessment of problem solving or other complex constructs. In the process, we identified several significant limitations of the present study.

First, we collected data from 102 assessments across 51 observations involving 53 children and 4 teachers. Depending on the assumption, the unit of analysis varied. We only collected data on the interpretation and use of the assessment results from four teachers. Collecting observational data in preschools is time consuming, for both teachers and researchers, and getting teachers on board was challenging. Some of the teachers who had initially agreed to participate in the study dropped out due to illness, an absent colleague, or time constraints. For this type of research, and within the available time and resources, the dataset is considerable. Nevertheless, it would be valuable to study the implementation and use of the instrument and its results with a larger sample of teachers.

Second, not all teachers could be observed while conducting the assessments to evaluate adherence to the protocols. We attempted to address this through collecting teachers' feedback, debriefings, and analysis of photographs of the set-up, materials, and results. However, this information largely relied on self-report. Adherence observations and inter-rater reliability analyses were conducted with only one teacher. Although we expect similar results of limited inter-rater reliability across teachers, caution is warranted when generalising these findings.

Third, we cannot draw strong conclusions about how teachers actually used the results. Systematic follow-up was not feasible within the timeframe of this study. Follow-up studies could focus on teachers' use of assessment results from this or similar instruments when making instructional decisions, grounded in theory of data-driven decision making.

Fourth, we observed that, although teacher training was limited to an online course with a step-by-step guide, a manual, a multiple-choice test and practice video's, not all teachers completed the full training programme. This may lead to deviations from the protocols or alternative interpretations of the problem-solving indicators which can result in suboptimal elicitation of problem-solving behaviours or interpretations that are not fully backed by theory or empirical evidence. Adequate training is necessary to ensure consistent observation and interpretation of children's behaviours (AERA et al., 2014; Parkes, 2013; Snow & Van Hemel, 2008). However, the limited time Flemish preschool teachers can devote to professional development hinders time-intensive training. Therefore, it may be advisable for teachers to conduct at least some observations jointly with a colleague to promote a shared understanding (Adie & Klenowski, 2016). This practice may also help to identify deviations from the protocol that gradually emerge in everyday practice.

Finally, only three problems have been developed for the OEPS assessment to date. To enable teachers to monitor the development of preschool children's problem-solving competencies over the course of a school year or across successive years, additional problems should be developed in accordance with the age-appropriate design criteria for engineering design problems. In addition, the OEPS assessment involves teacher-initiated problems. However, the observation form could also be applied to evaluate children's problem-solving behaviours in other contexts, such as free play, the playground, or the home environment. Follow-up studies could examine the extent to which results converge across different contexts.

Despite these limitations, the OEPS assessment offers valuable opportunities for preschool teachers to systematically collect and interpret information that refines their understanding of children's problem-solving competencies and supports instructional decision-making. However, instructional decisions should not be based on a single observation, as children's performance varies considerably across tasks (Volpe et al., 2005). Best practice in formative preschool assessment requires multiple occasions and the use of various information sources and assessment formats (Brassard & Boehm, 2007). Within this context, results from the OEPS assessment can serve as one information source to identify children's strengths and instructional needs. Further research could examine whether similar strengths and instructional needs emerge across other information sources, activities, and contexts, and whether they are also identified by other stakeholders (e.g., parents or support teachers).

The present study demonstrates that the OEPS assessment can support teachers in gaining insight into preschoolers' problem-solving competencies and that teachers are able and willing to use the collected data to inform instructional decisions. The OEPS assessment thus provides preschool teachers with a powerful instrument to design learning environments and activities that stimulate the development of young children's problem-solving competencies.

General discussion

This dissertation emerged from the conviction that problem-solving competencies are crucial for navigating ever changing 21st century societies. Problem solving is complex and, therefore, its development is best supported from an early age. STEM activities in the preschool classroom provide a particularly powerful context for developing early problem-solving competencies, as they offer rich opportunities for children to engage in meaningful, authentic, and hands on problems (Burns et al., 2025; Dubosarsky et al., 2018). To optimally support the development of preschoolers' problem-solving competencies, teachers require insight into children's strengths and support needs. Formative assessment is a suitable approach that enables teachers to systematically collect information and use it to update their understanding of children's problem-solving competencies and inform instructional decisions.

The overarching objective of this dissertation was to advance our knowledge on assessing problem-solving competencies in preschoolers. More specifically, we aimed to investigate critical aspects related to the validity, reliability, feasibility, and usability of formative problem-solving assessment in the preschool classroom. Four central questions guided this work: *What is early problem solving? How can preschoolers' problem-solving competencies be assessed? What observable behaviours do preschoolers demonstrate when solving problems? What are the critical factors that enable or hinder valid, reliable, feasible, and usable formative assessment of problem-solving competencies in preschoolers?*

The four studies included in this dissertation each address one of these questions. Study 1 maps the research literature on early problem solving across research fields and identifies its various manifestations and dimensions. Study 2 provides an overview of the methods used in the scientific research to assess preschoolers' competencies to solve ill-structured problems and synthesises the scope, characteristics, and psychometric properties of these methods. In Study 3, we identified observable problem-solving behaviours in preschoolers and developed the Problem-solving Observation Scheme for Playful Engineering and the Real World (PROSPER), grounded in the PISA problem-solving framework (OECD, 2013). Drawing on our findings from the previous studies, we developed the Observation-based Early Problem Solving (OEPS) assessment, an instrument for formative preschool assessment on problem solving. Finally, Study 4 presents a comprehensive validity argument on the interpretation and use of the OEPS assessment (Kane, 2013a), based on the observations and experiences of preschool teachers.

This general discussion is structured around four key findings, each addressing one of the central research questions. The findings integrate evidence from the four studies and highlight their contribution to educational research and practice. Each key finding is accompanied by a critical reflection on the most important limitations, forming the basis for suggestions for further research. We then broaden the discussion to the wider scholarly debate, concluding each section with a proposition. The propositions are formulated in a thought-provoking way and aim to challenge common ideas, theories, and practices. These propositions are not absolute. Rather, they aim at deepening and advancing scholarly debate by eliciting reflection and discussion.

This research was conducted and funded within the context of strategic basic research, which aims to generate both scientific and societal value. Accordingly, we conclude this dissertation with practical recommendations for policy and practice intended to support policymakers, educators, and other stakeholders in better preparing future generations for tomorrow's challenges.

Key findings

Key finding 1 – Early problem solving is a multi-dimensional construct that manifests in various ways

Across the four studies, we sought to offer a clear conceptualisation of problem solving in the context of early childhood education. In Study 1, we deliberately adopted the broadest possible definition of problem solving to avoid privileging one perspective over another. We only stated that problem solving involves transforming a problem state into a goal state when there is uncertainty about how that goal can be achieved (Duncker, 1945; Mayer, 1994). For the same reason, in Study 2 we defined problem solving as "the process of dealing with a problem for which no obvious solution method is available." Such general conceptualisations are useful for accommodating a diversity of perspectives in literature reviews, but they offer limited guidance for the operationalisation and the assessment of problem-solving competencies. Therefore, in Study 3 we defined problem solving more specifically as "the dynamic interplay between cognitive, metacognitive, physical, volitional, and social processes involved in resolving an unfamiliar task or situation." First, this definition makes explicit that problem solving is not solely a cognitive process but also encompasses regulatory, motor, motivational, and interpersonal aspects. Second, it emphasises that problem solving is not merely an aggregation of discrete skills or processes, but is concerned with the ways in which these competencies or processes are coordinated and integrated to resolve a problem (e.g., OECD, 2004). Finally, we characterised problem solving as dynamic, in the sense that the process can continuously change, thereby distinguishing it from routinised, linear, or predictable processes (e.g., Wilson et al., 1993). In Study 4, we added *affective* to the list of processes involved, as we gradually observed that emotions play an important role in young children's problem solving (see also Jonassen, 2000).

Across the four studies, we narrowed our conceptualisation of early problem solving, the core construct of this dissertation. This refinement was driven by the aims and focus of the studies. In Study 1, we aimed to map how and for what purposes the construct of problem solving is used to understand the diverse meanings and interpretations attributed to it by scholars across different research fields. In Study 2, we focused on problem solving in the context of ill-structured STEM problems. This was also the starting point for Studies 3 and 4, in which we further narrowed the focus to problem solving within engineering design, the *E* in STEM.

Conceptual clarity and alignment between a study's aims, scope, conceptualisation, operationalisation, measurement, analysis, and interpretation are essential to avoid mismatches between the constructs under investigation and the conclusions drawn from the findings (Bringmann et al., 2022; Podsakoff et al., 2016). However, this also implies that the conceptualisations differ across studies and research fields, which complicates cumulative knowledge building. Therefore, the initial aim of Study 1 was to develop a unified conceptual definition of problem solving in the context of early childhood education (early problem solving) through a conceptual systematic review (Schreiber & Cramer, 2024). However, it became evident that problem solving as a concept is so broadly used, both within and beyond research contexts, that striving for a unified definition would result in 'conceptual stretching' and inevitably lead to a mismatch between the conceptual definition and what is actually intended or measured in specific studies (Podsakoff et al., 2016). Instead of working towards a unified conceptualisation, we therefore chose to map the different research fields in which problem solving in early childhood education is studied and to identify the types and dimensions of early problem solving that serve as the building blocks for its conceptualisation. These building blocks provide a framework that researchers can use to delimit and conceptualise problem solving in alignment with their studies. Such a framework facilitates knowledge building around early problem solving, both within and across research fields.

The first building block for a clear conceptualisation of problem solving concerns the research field within which early problem solving is examined. This may involve one of the five major domains identified in Study 1 (Psychology, Mathematics, Social Sciences: Education, Computer Science, or Health Professions) as well as less frequently represented fields such as Engineering, Arts and Humanities, or Environmental Science. This building block enables researchers to situate their work within the broader scientific landscape. The other building blocks concern the dimensions of early problem solving. These dimensions refer to the problem type, problem topic, problem-solving skills, problem-solving processes, and problem-solving context and help further delineate the construct and the scope of the study.

The framework resulting form Study 1 can be applied to Studies 2, 3, and 4 (see Table 21). The three studies can be situated within the field of Social Sciences: Education. They draw on ill-structured problems related to STEM. In Study 3 and 4, this focus is further refined to engineering design. Solving the problems in study 3 and 4 requires primarily inquisitiveness (exploring), creativity (formulating), and self-regulation (planning and monitoring), and unfolds in an iterative manner. The problems are presented in an authentic, real-world context and are solved individually (Study 3) or collaboratively with a peer (Study 4). Although this overview does not take the form of a formal conceptual definition, it specifies how problem solving in these studies should be thought of by delineating the construct to an operationalisable scope. This delineation constitutes an important step towards the operationalisation and assessment of early problem solving and facilitates the comparison of findings from these studies with those of other studies that define the building blocks in a similar way.

Table 21. Conceptual delineation of early problem solving for Studies 2, 3, and 4

Building block	Study 2	Studies 3 and 4
Research field	Social Sciences: Education	Social Sciences: Education
Problem types	Ill-structured problems	Ill-structured problems
Problem topics	STEM	Engineering design
Problem-solving skills	/	Inquisitiveness, creativity, self-regulation
Problem-solving processes	/	Iterative
Problem-solving context	/	Authentic; individual (Study 3) or in pairs (Study 4)

Note. The publications included in Study 2 address a range of skills, processes, and contexts.

To achieve greater conceptual clarity and better alignment between conceptualisation and operationalisation, further conceptual work on early problem solving is needed. Several promising avenues for future research emerge. First, the database and categorisation of the research literature developed in Study 1 can be used to conduct conceptual systematic reviews (Schreiber & Cramer, 2024) of early problem solving within the different research fields. Such reviews could clarify the field-specific nature and characteristics of problem-solving. Subsequently, synthesising the results of such reviews could help to identify similarities and differences in the conceptualisation and operationalisation of early problem solving across research fields, thereby contributing to further theory development in early problem solving.

Second, it would be valuable to examine differences in the conceptualisation and operationalisation of problem solving in early childhood research compared with research on older students. This can be addressed through systematic reviews of the scientific literature, both within or across research fields, or by consulting experts from research and educational practice on how they conceptualise problem solving, for example within the context of a Delphi study.

Research on the conceptualisation and operationalisation of early problem solving can be situated within a long-standing debate on the nature and defining characteristics of problem solving as a scientific construct. Ambiguity in conceptualisation is a major threat to research, as it affects the entire research cycle and obscures the operationalisation, measurement, and interpretation of research results (Bringmann et al., 2022; Podsakoff et al., 2016). This issue compelled van Merriënboer (2013) to "sort out the chaos" (p. 153) by distinguishing the main perspectives on problem solving. van Merriënboer conceptualised problem solving as an educational goal, an educational method, and a skill. Similarly, Csapó and Funke (2017b), together with their co-authors, defined problem solving as a process, a skill, a set of skills, an ability, a competency, a capacity, a method, a context to learn or practice, a component of intelligence, and even an art. This list can be further extended with problem solving as behaviour (Skinner, 1953), an activity (Jonassen, 2010a), an approach (Wilson et al., 1993), an endeavour (English et al., 2008), a task (Bonotto & Santo, 2015), and thinking (Garton, 2004).

However, this is not where the discussion ends. Due to its nature, it is difficult to separate problem solving from other scientific concepts, such as executive functions (Zelazo et al., 1997), self-regulation (Zimmerman & Campillo, 2003), and metacognition (Mayer, 1998). In fact, prominent scholars have argued that problem solving is essentially synonymous with creativity (Guilford, 1967), with learning (Brownell, 1942), or with any goal-directed behaviour under conditions of uncertainty (Duncker, 1945).

It is beyond dispute that clear conceptualisations of problem solving help distinguish it from related constructs and support its operationalisation and measurement (Podsakoff et al., 2016). But consensus across research fields, topics, contexts, and problem types on the precise meaning and operationalisation of the construct is unlikely to be achieved. This is not inherently problematic, as long as authors clearly communicate how they define and operationalise the construct. However, this issue can also be approached from a different perspective. There may be little agreement on what problem solving *is*, but there may be consensus on what problem solving *does*; it integrates. Solving problems requires the integration of multiple processes, competencies, and recourses, whether cognitive or otherwise. According to Jonassen (2000), problem solving involves the integration of domain knowledge with different cognitive, affective, and conative processes, depending on the problem. Schoenfeld (2011) identifies a knowledge base, problem-solving strategies, monitoring and self-regulation, and beliefs as essential for successful problem solving, and Funke et al. (2018) identify problem solving as a “bundle of skills” (p. 41) including cognitive and non-cognitive components such as motivation, self-regulation, and social skills. The art of problem solving lies in adequately integrating these and other processes involved. We therefore position problem solving as “the great integrator”. The term is borrowed from Grubbs and Strimel (2015), who use it to refer to the potential of engineering design problems to connect different STEM disciplines. However, the integrative power of engineering design problems, and more broadly of ill-structured problems, extends far beyond the integration of knowledge and skills across domains. Based on the findings of the studies presented in this dissertation, we argue that problem solving has the potential to stimulate children to integrate everything they know, can do, want, believe, and even who they are in order to resolve a puzzling problem. From this consideration, we articulate our first proposition:

> **Proposition 1**
> Don’t just ask what problem solving *is*, but also what it *does*.

Key finding 2 – Early problem solving is an integrated and situated construct that is typically assessed through discrete components, outside its natural context.

The second key finding of this dissertation concerns the way problem solving is assessed in scientific research and comprises two components: the integrated nature of problem solving and its embeddedness within context. As discussed in Key finding 1, early problem solving requires the integration of multiple processes and competencies. By integration, we mean that different processes interact and can mutually reinforce one another. Early problem solving is therefore more than the sum of its parts. For example, having substantial knowledge, being skilled, and showing perseverance does not guarantee proficiency in problem solving. What matters is how children integrate and coordinate their available resources to achieve a desired goal.

Although problem solving requires an integrative approach involving cognitive, metacognitive, motivational, affective, motor, and social processes (Jonassen, 2000; OECD, 2004), Study 2 revealed that problem-solving competencies are typically measured as isolated components. For example, approximately half of the identified assessment instruments focus exclusively, or almost exclusively, on generating multiple solutions (Alexander et al., 1994; Leikin, 2013; Van Elsen et al., 2024). Generating possible solutions is an important aspect of any problem-solving process (Guilford, 1967; Osborn, 1953), but when assessed in isolation, it raises the question of whether creativity rather than problem solving is being measured. Moreover, the ability to generate solutions in structured creative production tasks may not reflect preschool children's ability to generate ideas when solving real-world problems. Our review identified only three instruments that covered most steps of the problem-solving process (Anggoro et al., 2021; Bartholomew et al., 2019; Shechter et al., 2021).

However, covering all steps of the problem-solving process is still different from capturing the extent to which children integrate and coordinate these steps and processes. In retrospect, it should be acknowledged that the OEPS assessment we developed in this dissertation also falls short in this respect. Although the assessment includes the different steps of the problem-solving process, supplemented by five overarching behaviours or dispositions, problem solving is nevertheless assessed through discrete indicators. Further research is needed to investigate methods for determining how, and how effectively, preschool children integrate and coordinate different processes and skills in their attempts to solve problems.

A first step could be to examine whether distinct behavioural patterns in problem solving can be identified. Most assessment methods focus on outcomes rather than processes. To date, little is known about how the problem-solving process unfolds in young children and how it relates to its effectiveness. In further research, the Problem-solving Observation Scheme for Playful Engineering and the Real-world (PROSPER) developed in Study 3 can be used to systematically map the sequence of actions. Subsequently, sequence analysis can be applied to reveal different patterns of problem solving (Tschisgale et al., 2025). Sequence analysis is a set of analytical techniques used to study the structure, timing, and patterns of processes (Abbott, 1995). Alternatively, latent class cluster analysis can be applied to identify profiles based on the observed behaviours of children during problem solving. Latent class cluster analysis enables the identification of hidden groups by detecting recurring patterns in categorical data and estimating the probability that individuals belong to each group based on their observed behaviours (DiStefano & Kamphaus, 2006; Vermunt & Magidson, 2002). Identifying recurring patterns of problem-solving behaviours can provide insight into the extent to which certain processes are interrelated and interact with one another. These insights may subsequently contribute to the development of methods for evaluating how, and to what extent, children integrate and coordinate their actions during problem solving.

The second component of this key finding concerns the contextual nature of problem solving in relation to its assessment. Real-world problems are inherently embedded in specific contexts, particularly ill-structured problems for which multiple solutions are possible (Jonassen, 1997; Kirsh, 2008). Nevertheless, Study 2 showed that competencies for solving ill-structured problems are most often assessed in controlled settings (see also Rhodes et al., 2024). Typically, assessments are administered individually by an external researcher in a separate room, removed from the dynamic classroom context that shapes preschoolers' everyday learning experiences. This practice threatens the validity of assessment outcomes (Kyllonen & Lee, 2005).

In this dissertation, we aimed to address these issues by developing an assessment method (OEPS) that can be applied in authentic classroom contexts. However, the classroom context represents only one of the many contexts in which preschool children can be found. A key distinction from many other assessment approaches is that the OEPS assessment explicitly acknowledges and embraces the context-bound nature of problem solving, as discussed extensively in Study 4. In doing so, this dissertation takes an important step forward in the assessment of preschool children's problem-solving competencies. Nevertheless, the question remains whether children display similar problem-solving behaviours outside the classroom, for example during free play in the playground or in the home environment. The rubric developed for the OEPS assessment can be used in future research to investigate and compare children's problem-solving behaviours across contexts. By treating context as an integral component of problem-solving assessment, researchers can gain insights into contextual differences in preschoolers' problem-solving behaviours and identify specific contextual factors that influence these behaviours. Findings from studies that take contextual factors into account can help ensure that learning activities designed to develop young children's problem-solving competencies are tailored to their needs.

Research into the interplay between problem solving and the context in which it is embedded is situated within a broader scientific debate on the exact role of context in problem solving and how it should be accounted for in assessment. The discussion on the contextuality of problem solving and its assessment can be framed within a situationalist approach to problem solving. For example, Kirsh (2008) challenges classical theories of problem solving by arguing that all cognition is embedded in its social, material, and cultural context. The way in which a problem is presented, and the environment in which this occurs, largely determines how the problem is perceived and which possibilities are recognised as valid strategies. Moreover, problem solving depends on the solvers' ability to draw on external resources (Pouw et al., 2014). This has implications for the measurement and assessment of problem-solving competencies (Hardy et al., 2024). Kirsh (2008) formulates the problem as follows:

> Examination of actual problem solving in ecologically natural contexts as opposed to white-room environments reveals a host of interactions with resources and cultural elements that figure in the many phases of problem solving, such as understanding the problem, exploring its scope and constraints, getting a sense of options, and developing a metric for evaluating progress toward a solution (p. 270)

As with problem solving, preschool assessment is best embedded in context (Bagnato et al., 2014; DeLuca et al., 2020; Grisham-Brown et al., 2006), especially in the context of observation-based formative assessment when the aim is to guide teacher's instructional decisions (Brassard & Boehm, 2007; Grisham-Brown et al., 2006). As Bagnato (2005) puts it:

> Testing of infants, toddlers, and pre-schoolers by strange adults, in unnatural places, at tables with flip cards and small, unmotivating toys does not make much sense. Infants and preschoolers are naturally on the go; they do not wait, share, sit still, take turns, or follow directions. (p. 17)

The tendency in scientific research to measure the components of problem solving in isolation and to strictly control environmental variables is questionable for several reasons. First, it overlooks the dynamic interactions between processes involved in solving ill-structured problems and the need to coordinate these processes. Second, by ignoring the context (Frensch & Funke 1995; Rhodes et al., 2024) or treating context variables as a source of error rather than an essential component of the construct, traditional measurement instruments overlook the dynamic interactions between the child, the problem, and the environment that form the core of the problem-solving process (Steffensen & Vallée-Tourangeau, 2018). Finally, the emphasis on isolated, quantifiable outcomes rather than on complex processes provides little useful information about how children solve real-world problems. It is precisely this type of information that teachers need to guide their instructional decisions (Archbald & Newmann, 1988). For research purposes, process-level information is essential to gain a deeper understanding of early problem solving. As Simon (1975) said, "we must avoid blending together in a statistical stew quite diverse problem solving behaviors whose real significance is lost in the averaging process" (p. 288).

To accurately measure or assess problem-solving competencies, the context in which problems are embedded must be taken into account. Context can be understood as the time, place, social, and cultural setting in which the assessment is conducted (Damşa et al., 2025; Rhodes et al., 2024). For teachers seeking information to inform and justify instructional decisions, the context is shaped by the classroom characterised by the presence of lively children and learning centres filled with a wide range of materials and colours. This is far removed from the relatively quiet, orderly classrooms filled with monochrome desks and chairs that are characteristic of primary and secondary schools. For researchers aiming to investigate preschoolers' abilities to solve real-world problems, this entails additional challenges, as they must find a balance between scientific rigour, characterised by standardisation and control, and authenticity, characterised by flexibility and children's agency. Although strict control of contextual variables may enhance measurement reliability, it inevitably undermines validity, thereby providing limited insight into what preschoolers are actually able to do in real-world settings (see also the discussion on the reliability/validity paradox in Study 4 and Kane, 2013b). Drawing on these considerations, we propose our second proposition.

> **Proposition 2**
> Problem solving is an integrated, contextual process that is best assessed 'in the wild'.

Key finding 3 – Preschoolers demonstrate a rich variety of problem-solving behaviours

Traditional problem-solving models, such as those proposed by Polya (1945) and OECD (2013), operationalise problem solving by decomposing it into a sequence of phases or steps. These steps are, in turn, latent constructs rather than manifest variables as they cannot be directly observed. One of the major contributions of this dissertation to scientific research into early problem solving is the further operationalisation of the steps through observable behaviours. In Study 3, we identified 33 distinct problem-solving behaviours, of which 29 were assigned to the steps of the PISA problem-solving framework (OECD, 2013), while four spanned the entire process. In study 4, collaboration was added as a fifth overarching behaviour since children were allowed to solve the problems in pairs.

The identification of commitment, perseverance, autonomy, responsiveness, and collaboration as overarching indicators of problem solving complements traditional phase models and constitutes an important theoretical contribution of this dissertation to the operationalisation of problem solving. These indicators serve as foundational prerequisites for the emergence of genuine problem-solving. If these conditions are not met, preschoolers are unlikely to fully mobilise their resources, resulting in suboptimal problem-solving processes and hindering accurate assessment of children's actual problem-solving abilities. To assess and support the development of problem-solving competencies, children must be encouraged to engage deeply with challenges, persevere despite setbacks, take ownership of the process, respond appropriately to environmental cues and provided support, and communicate and collaborate with others.

The precise role and contribution of these overarching problem-solving indicators have largely fallen outside the scope of research on early problem solving and warrant further investigation. For example, the importance of responsiveness in the problem-solving process was highlighted by both researchers and teachers involved in the video observations in Study 3. Responsiveness refers to how children react to stimuli from their environment, including questions or suggestions from teachers (McLeod et al., 2024), but also to comments or ideas from peers in collaborative contexts, as observed in Study 4. Adequate responsiveness to contextual cues is particularly relevant when solving embedded, ill-structured problems (Jonassen, 1997). From a social-constructivist perspective on learning, responding to prompts is also important for the development of problem-solving competencies (Vygotsky, 2017). However, to date, there is very limited research on preschoolers' responsiveness and its relation to problem solving (McLeod et al., 2024). Further research into the role of responsiveness in problem solving may provide educators with useful guidance for supporting the development of preschoolers' problem-solving competencies.

When looking at the behaviours assigned to the different steps of the PISA process, the greatest variety in observed behaviours was found for exploring. In Study 3 and 4 we found children spontaneously observing, naming, describing, comparing, and manipulating materials. They also searched for suitable materials, formed associations with familiar objects or prior experiences, and asked questions about unfamiliar materials and their functions. Schoenfeld (1985) refers to exploration as "the heuristic heart" (p. 110) of the problem-solving process. Moreover, children's tendency to explore is associated with successful problem solving in preschoolers (Evans, Todaro, et al., 2021). Exploration supports preschoolers' understanding of the properties and uses of objects and stimulates the generation of multiple possibilities (Cunningham et al., 2018; Evans, Todaro, et al., 2021). Through exploration, children gather information about the problem and the available options (OECD, 2013). Exploratory behaviour is driven by curiosity (Jirout & Klahr, 2012), which closely aligns with children's natural inquisitiveness (Engel, 2011; Fusaro & Smith, 2018). The wide range of exploratory behaviours observed in Study 3 reflects preschoolers' need for, and familiarity with, exploration as a means to understand problems and materials. Given both their natural inclination towards exploration and their limited prior knowledge base and experience with problem solving, it is important that children are given sufficient time to explore before and during problem solving (Gardiner et al., 2012). However, in several influential problem-solving models, exploration is absent or only implicitly addressed (Mayer, 1994; Polya, 1945; Zelazo et al., 1997). Based on the findings discussed in this dissertation, we argue that exploring deserves a prominent place in early problem-solving models.

Representing, by contrast, proved difficult to observe. Pretz et al. (2003) define problem representation as “the manner in which the information known about a problem is mentally organized” (p. 6). Although internal representations can be externalised through gestures, drawings, symbols, or language (Bat Or et al., 2022; Cartwright, 2024; Papandreou, 2009; Shiakalli & Zacharos, 2012), in Study 3 we were not able to observe behaviours that clearly expressed the extent to which preschoolers represented the problem, its goal, criteria, and constraints (Pretz et al., 2003). This absence of observable representing behaviours may be explained by preschoolers’ inadequate problem representation, their limited ability to externalise mental representations, or their lack of awareness of the relevance of making their mental representations explicit (Copple, 2003; Fleer, 2000).

Before representing behaviour in preschoolers can be observed, assessed, and supported, children must learn how to externalise their internal representations. This externalisation can occur verbally, but also through drawing a problem or a potential solution. For example, (Bartholomew et al., 2019) asked 5–6-year-old children to make a drawing of a problem presented in a children’s rhyme and their ideas for solving it. Because drawing skills at this age are not always sufficiently developed to produce recognisable representations, the researchers asked follow-up questions to clarify children’s intentions. However, this approach requires substantial guidance and support from an adult. If spontaneous representing behaviour is to be captured without interrupting or directing the natural problem-solving process, children must not only learn to express their thoughts but also come to understand the purpose and function of problem representation. In the context of ill-structured problems, this may be too challenging for preschoolers. Allowing children to collaborate, as in Study 4, is a promising approach because collaboration encourages children to express or externalise their thoughts about the problem and potential solutions. Further research could explore how young children can be supported in externalising their internal problem representations, enabling teachers and researchers to gain better insight into the extent to which preschool children form accurate representations of problems, their key elements, and possible solutions.

This dissertation has given us a better understanding of the behaviours that preschoolers exhibit when solving ill-structured problems. The development of the PROSPER observation scheme in Study 3 opens new avenues to systematically investigate the stability of problem-solving behaviours across different problems. Throughout the observation studies reported in this dissertation, we found that problem-solving behaviours varied between problems. However, due to ceiling effects and the limited number of observations per child, the evidence remains tentative. Further research into the consistency of children’s problem-solving behaviours across problems may provide better insight into which behaviours are relatively stable and which are more strongly influenced by problem characteristics. Such research could make an important contribution to the long-standing and ongoing scientific debate on the domain-specific (problem dependent) or domain-general (problem independent) nature of problem solving (Foster, 2023; Greiff et al., 2014; Sternberg, 1989;

Tricot & Sweller, 2014). If such research shows that preschool children consistently display or fail to display particular behaviours across problems, this may indicate domain-general problem-solving behaviour. Conversely, if specific behaviours vary markedly from one problem to another, they are more likely problem, domain, or context specific.

Another promising avenue for future research is the examination of developmental changes in children's problem-solving behaviours over time. For instance, research using a microgenetic approach can offer detailed insights into the development of problem-solving behaviours in children, as well as the ways in which this development varies between them. The microgenetic method was developed by Siegler and Crowley (1991) and enables the qualitative and quantitative analysis of cognitive change processes within and across individuals by observing children at short intervals over a specific period of time. For example, Guevara and Puche-Navarro (2015) used the microgenetic method to map intra-individual patterns in preschoolers' cognitive planning while solving a virtual planning problem. However, they did not examine these patterns in relation to other indicators of problem solving and focused exclusively on well-structured, digitally presented problems.

But one of the most important unanswered questions concerning children's problem-solving behaviours is directly related to the integrated nature of problem solving discussed in Key Findings 1 and 2. This question concerns the primary regulatory mechanisms children use to coordinate their actions. More specifically, it addresses the role of planning and monitoring in regulating young children's problem-solving processes (Schraw, 1998; Shin et al., 2003; Susanna et al., 2026; Veenman et al., 2006). Planning takes a central place in many influential problem-solving models. According to Newell and Simon (1972), planning guides and structures actions during problem solving. Bransford and Stein (1993) suggest that the IDEAL problem solver does not act until the problem has been identified, goals have been set, possible strategies explored, and consequences evaluated. The centrality of planning is strongly emphasised in Polya's four-phase model, in which the second phase concerns devising a plan and the third carrying it out (Polya, 1945). However, planning is a complex process involving multiple activities and skills (Rott et al., 2021) and requires "good ideas" grounded in prior knowledge and experience (Polya, 1945). Planning in advance is particularly difficult for preschoolers (Hudson et al., 1995). Planning skills develop markedly between the ages of 3 and 5 and continue to develop beyond the preschool years (Arfé et al., 2020; Warneken et al., 2014; Zelazo et al., 1997). In Studies 3 and 4 we observed that preschoolers often moved to action very quickly, sometimes even before the problem or goal was fully explained, leaving little opportunity for planning in advance. This is consistent with the findings of English (1992) that preschoolers often start with unplanned trial-and-error. Although trial-and-error can sometimes work (Funke et al., 2018), it is relatively ineffective and time-consuming (Stern & Hertel, 2022). Nevertheless, our observations revealed how preschoolers learned from their unplanned attempts, discovering what worked and what did not. They monitored outcomes and thereby compensated for their lack of planning in advance (see also English, 1992).

Previous research has shown that monitoring behaviours such as checking and optimising are associated with a higher likelihood of success, fewer errors, higher-quality solutions, but also longer time (English, 1992; Reuter & Leuchter, 2022). Although monitoring progress towards a solution is a complex skill involving observing, checking, and revising plans (English, 1992; OECD, 2013), Study 3 revealed a wide variety of monitoring behaviours in preschoolers. These findings are consistent with those of Bryce and Whitebread (2012) and Anggoro et al. (2021) who also identified multiple monitoring behaviours in this age group. Results from Study 4 further showed that preschoolers spontaneously monitored their progress, without prompting from adults. This aligns with the findings of Reuter and Leuchter (2022) that preschool-aged children spontaneously engage in testing and optimisation, which effectively improved outcomes.

The observation that preschool children display more monitoring behaviours than planning behaviours, that planning requires knowledge and experience which young children may not yet possess, and that monitoring appears to develop earlier than planning in advance (Bryce & Whitebread, 2012), raises the question of whether young children should always be expected to plan their actions in advance. The primacy of planning over monitoring in established problem-solving models can be attributed to their emphasis on older learners and well-structured problems (Newell & Simon, 1972; Polya, 1945). Compared with preschoolers, older students can rely on a broader knowledge base, more advanced self-regulation skills, and greater experience in solving problems. Furthermore, when solving ill-structured problems, planning in advance is inherently difficult. Ill-structured problems entail a greater degree of uncertainty than well-structured problems. This makes planning more challenging and increases reliance on monitoring during the process (Jonassen, 1997; Shin et al., 2003), particularly when the knowledge base is limited, as in young children (Schäfer et al., 2024). Planning in advance also appears to be less effective as problem complexity increases and exploration becomes necessary before meaningful plans can be formulated (Eichmann et al., 2019). Given developmental differences between preschoolers and older students, as well as differences in the processes involved in solving well-structured versus ill-structured problems, caution is warranted when applying traditional problem-solving models such as Polya's to authentic problem solving in early childhood. Viewing preschoolers' problem solving through such frameworks risks promoting a deficit perspective, emphasising limited planning abilities rather than their proficiency in monitoring progress.

Both planning and monitoring are beneficial for successful problem solving (English, 1992). The key distinction between them lies in the timing of regulation: planning occurs prior to action, whereas monitoring takes place during action. Cohen (1982) argues that plans enable monitoring and error detection, suggesting a complementary and sequential relationship in which planning precedes monitoring. In the context of preschoolers solving ill-structured problems, we argue that this sequence may need to be reversed. When children lack prior experience or knowledge, particularly in complex engineering design or real-world problems, they often need to act first in order to learn about the problem, materials, possibilities, problem space, and consequences of their actions (Evans, Todaro, et al., 2021). Preschoolers develop plans through action as young children need their hands to think (Ata-Aktürk, 2023). Consequently, for preschool children, action tends to precede planning and monitoring functions as “planning on the go”. We therefore suggest that, in the context of ill-structured STEM problems in preschool education, it may be appropriate to invert the adage *Plan–Do–Check* to *Do–Check–Plan*. This leads us to our third proposition.

> **Proposition 3**
> Preschoolers don’t always plan in advance; they plan on the go. And that is perfectly fine.

Key finding 4 – Formative assessment enables teachers to refine their understanding of children’s problem-solving competencies and tailor instruction to their needs

The previous key findings demonstrate that problem solving is a complex, multidimensional, and context-dependent construct, that assessing young children’s problem-solving competencies poses substantial challenges, and that preschoolers display a rich variety of problem-solving behaviours when engaging with ill-structured problems. But the primary aim of this dissertation was to investigate the validity, reliability, feasibility, and usability of formative problem-solving assessment in the preschool classroom.

To this end, the OEPS assessment was developed and implemented by teachers in their own classrooms. We found that, with only limited organisational adjustments, administering the OEPS assessment is feasible in everyday practice. In addition, the findings indicate that engineering design problems are motivating for children and elicit a broad range of problem-solving behaviours. Moreover, we demonstrated that the assessment supports teachers in refining their understanding of preschoolers’ problem-solving competencies and in making instructional decisions. Teachers’ positive responses in feedback and debriefing sessions further suggest that they perceive clear added value in the assessment and the structured observation of preschool children.

Although this dissertation demonstrates that teachers are able to implement the assessment in their classrooms and to interpret and use the outcomes to adjust their instruction, several avenues remain to further develop and strengthen the validity, reliability, feasibility, and usability of the OEPS assessment. First, we did not examine convergent validity of the assessment. It would be valuable to investigate the extent to which the indicators of problem solving included in the OEPS assessment correlate with outcomes of standardised instruments designed to measure related constructs such as executive functions, self-regulation, metacognition, creativity, and language proficiency. Theoretically, these constructs are expected to correlate with specific indicators of problem solving. For example, creativity is expected to correlate with formulating possible solutions and self-regulation with planning and monitoring.

Second, it would be valuable to establish normative patterns of expected behaviours for each indicator of early problem solving across different age groups. Such norms would give teachers clearer guidance for determining whether a child is excelling or requires additional support and may enable them to use the OEPS assessment to monitor the development of problem-solving competencies throughout early childhood education.

Third, the problems used in the OEPS assessment are tailored to the developmental level of children aged 5 to 6 years. To use the OEPS assessment to monitor the development of problem-solving competencies over time, there is a need for problems that are suitable for both younger and older children or problems that can easily be adjusted to provide varying levels of difficulty. Such problems can be developed based on the design criteria underlying the problems used in the OEPS assessment.

Fourth, this dissertation has not examined the actual use of the outcomes of the OEPS assessment. In order to make effective use of formative assessment, teachers require guidance on how to interpret assessment data and how to use these data to design or strengthen effective learning environments and instructional practices. This can, for instance, be situated within a learning trajectories approach (Clements & Sarama, 2014). Learning trajectories consist of three parts: a learning goal, a developmental path towards that goal, and instructional activities that support children along this path towards the goal. Formative assessment closely aligns with learning trajectories and supports teachers in optimally guiding children towards the intended goal by providing insight into children's strengths and instructional needs (Supovitz et al., 2024). Such learning trajectories can be developed for the different indicators of early problem solving. Follow-up research could examine the effectiveness of interventions that combine formative assessment with learning trajectories. Ideally, this would involve a longitudinal study in which the developmental effects are measured over the course of one or more school years.

Furthermore, the development of a comprehensive teacher training programme is essential to ensure that teachers implement, interpret, and use the OEPS assessment and its outcomes as intended. Given the limited time that Flemish preschool teachers can devote to professional development outside the classroom (e.g., due to limited staff to fill-in during absence; OECD, 2025), a hybrid training format is required in which teachers can become familiar with all components of the assessment. Alongside a manual and an introductory session outlining the purpose, the problem-solving framework, the assessment framework, the rubric, and the protocols, opportunities for practice and follow-up support are required to ensure long-term impact of the training programme (Schneider & Randel, 2010). External guidance can support this process, but teachers can also benefit from co-observing children and discussing their observations with colleagues to develop a shared understanding (Adie & Klenowski, 2016; Sancar et al., 2021).

But most importantly, effective research on formative assessment necessitates close collaboration between researchers and teachers. Studies on formative assessment, both in general and in relation to problem solving in preschool, benefit substantially from knowledge co-construction (Parsons, 2021), which is facilitated by the strong parallels between formative assessment as a pedagogical practice and scientific research (Torrance & Pryor, 2001).

In this dissertation, preschool teachers were involved in all phases of the research, from design to dissemination. Their contextual expertise enriched the research process (Sjölund et al., 2023). Teachers acted as critical friends in drafting the project proposal, developing tasks and protocols, and communicating findings to a wider audience. They organised data collection, informed parents and obtained consent, collected and processed data, contributed to the coding of recordings, implemented the assessment, and provided feedback.

This sustained collaboration between research and practice enabled knowledge co-construction in which diverse perspectives were not only brought together (Kelchtermans, 2021) but also generated new insights “at the intersection of research and practice” (Parsons, 2021, p. 1493). Such partnerships between research and practice offer numerous opportunities for both teachers and researchers (McGeown et al., 2023; Wyse et al., 2021). For teachers, they help bridge everyday practice and scientific evidence, stimulate reflection on teaching, support professional development, and result in findings that are practically relevant (Gilmore et al., 2025; Kelchtermans, 2021; Torrance & Pryor, 2001). For researchers, such partnerships align research with practical needs, help identify priorities, and provide access to the lived reality of preschool classrooms (Gilmore et al., 2025; Torrance & Pryor, 2001). Based on our experiences and backed by the scientific literature on participatory research, we put forward our fourth and final proposition.

> **Proposition 4**
> In educational research, teachers are not only informants, participants, or end-users, but active partners within the research team.

Recommendations for policy and practice

Problem solving deserves a prominent place in the preschool curriculum

Current education has the challenging task to prepare future generations for a world we don’t yet know with problems we never faced before. Equipping children with the tools to take on unknown challenges is a major concern in contemporary education. These may include knowledge, skills, and attitudes, but also the will and drive to tackle problems until they are solved.

The new Flemish preschool curriculum, conceived as minimum goals and implemented from the 2026–2027 school year onwards, explicitly emphasises knowledge acquisition as the primary cornerstone of high-quality education (Muijs et al., 2025). Problem solving is included in the learning objectives related to mathematics and science and technology (https://onderwijsdoelen.be/). Focussing on knowledge acquisition and teaching specific strategies within specific disciplines can help children learn to solve well-structured problems within specific subject areas, but contributes only to a limited extent to learning how to solve dynamic, ill-structured, real-world, or cross-curricular problems (Greiff et al., 2014). Solving ill-structured problems requires different cognitive processes and learning environments than well-structured, domain specific ones (Funke & Frensch, 2007; Jonassen, 2010a). A truly ambitious early childhood curriculum therefore focuses not only on domain-specific knowledge and skills, but also on complex, cross-domain competencies in which acquired knowledge and skills must be applied in an integrated way.

Within early childhood education, problem solving can be applied across a wide range of domains, in diverse ways, and for multiple purposes. Problem-solving activities can be integrated into preschool practice to:

- encourage children to discover new phenomena,
- allow children to gain experience in problem solving to learn from it,
- evaluate children's problem-solving competencies,
- motivate children to apply specific knowledge and skills,
- evaluate domain specific knowledge and skills,
- evaluate general skills.

As such, problem solving constitutes an indispensable component of any early childhood curriculum, functioning as an educational goal, an instructional method, and a competence (van Merriënboer, 2013). Moreover, Study 1 showed that problem solving provides significant educational opportunities for neurodiverse populations as problem solving can be used to provide additional support for children with learning difficulties or from disadvantaged backgrounds, as well as offering appropriate challenges for gifted children.

Where is the E in STEM?

Although problem solving can be addressed across the early childhood curriculum, it is particularly relevant in STEM education. Problem solving is an indispensable part of STEM education (Dubosarsky et al., 2018; Priemer et al., 2020). Problem-solving competencies are needed across all STEM domains and STEM offers plenty of opportunities to develop problem-solving competencies. This is especially true for engineering. Engineering is basically a problem-solving process (Gold et al., 2021; Malcok & Ceylan, 2022) and plays a unifying role within STEM (Tank et al., 2018).

However, across the globe the E for Engineering is underrepresented in early STEM education (Bustamante et al., 2018; English, 2018). This is also the case for the Flemish curriculum, which includes minimum goals for mathematics, science, and technology, but not for engineering. Yet engineering provides fertile ground for engaging young children in a wide range of problem-solving skills, especially when it is combined with design. The engineering design process is fundamentally a problem-solving process (Grubbs & Strimel, 2015; Isabelle et al., 2021), and concerns creating things that work or improving existing solutions (Lucas et al., 2014). It also aligns closely with what young children naturally do and enjoy, such as crafting and building (Isabelle et al., 2021; Tank et al., 2018). Moreover, engineering design acts as the interdisciplinary glue connecting the STEM domains (Tank et al., 2018), offering opportunities to explore or apply mathematical, scientific, and technological concepts and skills. In short, engineering design creates opportunities to integrate knowledge and skills across disciplines (Grubbs & Strimel, 2015). Unfortunately, early childhood education systems offer young boys and girls limited opportunities to thrive as engineers (Lucas et al., 2014).

And there is more. Integrating engineering design into the preschool curriculum offers opportunities to engage girls in STEM from an early age. Globally, girls remain underrepresented in STEM fields, both in education and in professional careers (Cheryan et al., 2025). This gender gap emerges early in childhood. Prior research shows that in preschool settings girls engage less frequently than boys in engineering play (Fleer, 2021). One commonly cited explanation for the gender gap is girls' comparatively lower self-confidence in STEM domains, which can be countered through early STEM education and positive experiences of success (Cheryan et al., 2025; Erol et al., 2025)

Research indicates that girls, like boys, can be engaged in engineering design when problems are embedded in an appealing narrative (Cunningham et al., 2018). A story can emotionally resonate with children by presenting socially meaningful problems that require an engineering solution (Fleer, 2021). Taking into account girls' preferences for socially relevant technologies (National Academy of Engineering & National Research Council, 2009), the problems for Study 3 were introduced by Mr Rabbit, a soft bunny hand puppet who asked for the children's help to support a friend in need. And this seemed to work well for both boys and girls, although we did not systematically investigate the "bunny effect".

Based on findings from the scientific literature and our own experiences during the studies reported in this dissertation, we argue that, with meaningful problems embedded in engaging stories and appropriate support, engineering design activities in preschool can strengthen girls' self-confidence and interest in STEM and contribute to reducing the persistent gender gap in STEM from an early age.

Provide problems with low floor, heigh ceiling, and wide walls

Engineering design problems are particularly well suited to challenge children in ways that align with their abilities and interests. The most engaging problems for young children are those that offer a low floor, high ceiling, and wide walls (English, 2017; Resnick & Silverman, 2005). A low floor ensures that all children can participate, meaning the problem should not introduce any barriers or obstacles that prevent or discourage children from engaging with it. A high ceiling ensures that every child, including gifted ones, can be sufficiently challenged. This can be realised by presenting sufficiently challenging problems or by calibrating complexity to the children's abilities. Wide walls highlight the importance of providing opportunities for children to explore multiple solution paths and to discuss their ideas openly with peers inside and outside the classroom.

The notions of low floor, high ceiling, wide walls can be further expanded with the concepts of agency and conceptual surprise (English, 2017; Gadanidis et al., 2017). Agency refers to children's ability to set their own additional goals, which can be encouraged by offering open-ended, ill-structured problems that provide opportunities for initiative and self-direction. Conceptual surprise occurs when children experience excitement and amazement upon discovering new phenomena. Such moments can be created by introducing meaningful, challenging problems that can be solved through mathematical, scientific, technological, or other principles that are still unfamiliar to them.

Allowing children to engage in problem solving and take up agency requires teachers to dare to relinquish some control over what happens in the classroom and how learning unfolds (Nadelson & Seifert, 2017). Children need opportunities to explore, experiment, and make sense of things on their own terms. It is therefore important not to define the problem too narrowly, not to intervene too quickly with guidance, and, above all, not to take over the problem from the child too soon.

Observe children while they solve problems to gain a better understanding of children's competencies and inform instructional decisions

When solving problems, children draw on everything they know, can do, believe, and want. The extent to which they use their available resources varies, offering insight into their abilities and preferences. This makes problem solving an ideal context for forming a holistic picture of young children. Closely observing children as they solve problems reveals their autonomy, responsiveness, leadership, willingness to collaborate, perseverance, creativity, flexibility, self-regulation, inhibition, frustration control, language proficiency, misconceptions, and many other capabilities and characteristics. Engaging preschoolers in problem solving allows them to fully demonstrate their strengths, makes visible areas for growth, and provides opportunities to get to know children from a different perspective.

Observing preschoolers as they solve problems can also help teachers make informed instructional decisions. It can contribute to identifying children who would benefit from being additionally challenged, children who may require extra support, or even potential learning and developmental barriers that warrant further examination. In this way, instruction can be better aligned with the educational and developmental needs of individual children or the specific classroom group.

Ten tips to teach children how to solve problems

From our experience of working with teachers, we observed that teachers not only benefit from robust theoretical frameworks and evidence-informed tools, but also from practical tips to help them improve their instruction. Therefore, this dissertation concludes with ten practical recommendations emerging from the four studies.

How to help children solve their own problems?

Don't solve all their problems yourself

You don't need to solve all children's problems to be a good teacher or parent. On the contrary, each problem presents learning opportunities. Children do not only learn to solve problems by being told how to do so, but by engaging in problem solving themselves. This does not mean that you cannot help them. Some problems are simply too difficult for them. Yet, with appropriate support, children can solve many problems on their own.

'What could you do?'

When children notice a problem and express it sometimes loudly ("Miss!!!"; "Dad!!!"), ask what happened and invite them to describe the issue. Then ask: 'What could you do?' These words often set cognitive gears in motion. Children frequently come up with one or more ideas straight away and dash off to try them out.

Encourage trying new approaches

The first idea is not always the best. Sometimes, alternative solutions or strategies are needed. Letting go of an idea and trying something new is not always easy. When children repeatedly apply the strategy without success, or stop trying when they realise their idea is ineffective, ask them: 'What else could you do?' Again, this question prompts the generation of new ideas. Because children do not yet have fixed schemas for common problems, they often come up with highly creative approaches. Value these alternative and unconventional solutions, even if they do not fully match your expectations.

Encourage perseverance

Problem solving requires perseverance. Things rarely work out on the first try, and this can be frustrating. Their limited experience makes it difficult for children to predict the outcomes of their actions. It is a process of trial-and-error. When their motivation seems to fade after several unsuccessful attempts, ask: 'What are you doing?' or 'What is not working?' This shows involvement without directing children's actions. Knowing that you are there for them often gives children renewed motivation and energy to continue.

Time and calibrate your help

Children learn most from problems that are just beyond what they can solve independently. Sometimes they will need support. The key is to provide support at the right moment. Providing support too quickly takes the problem out of their hands; providing support too late can lead to frustration or a sense of failure. When children encounter an obstacle they cannot overcome themselves, give them a small prompt that is just sufficient to help them move forward. Gradually increase the level of guidance: from vague suggestions to more specific, directive hints. Provide support in the form of a question to avoid taking over the problem. For example, after a child has unsuccessfully tried to retrieve a toy car from under a cupboard using their hands, a block, a pencil, and a soft toy, you might ask whether using something longer could help (vague suggestion). If that fails, you might ask whether the handle of a broom could work (specific hint).

Remove obstacles that are beyond their reach

However, some problems require skills that children do not yet fully master. In such cases, removing an obstacle may be necessary. You can do this by demonstrating an action and asking the child to imitate it. If that does not work, you can take over part of the problem so that the child can continue independently. In the toy-car example, you might show how to reach under the cupboard with a broom. But if the car is stuck against the wall at the very back, even this may not work. Moving the car slightly away from the wall offers just enough help without taking over the whole problem.

Recognise that problem solving takes time

Problems are not solved in the blink of an eye. Children must explore and understand the problem, generate and evaluate options, test ideas, determine what works and what does not, and adjust where needed. Many problems are not urgent. Give children time to experiment and take the time yourself to observe how they do so.

Let children collaborate

Children learn not only from adults but also from one another. It is remarkable to see how children arrive at solutions together. Collaboration encourages them to verbalise their ideas. Children who struggle may gain inspiration from their peers' solutions and then continue independently.

Look back and ahead

To truly learn from problems, children need opportunities to reflect on both the process and the outcome. After a problem is solved, ask some simple closed questions such as 'Did it work?' or 'Was it difficult?' Then move towards simple open questions like 'What was difficult?' If children can respond appropriately to these questions, you can introduce more complex why- and how-questions, such as 'Why did it not work?' and 'How did you solve it?'

Provide problems yourself

Young children enjoy solving problems. Make use of this natural inclination and create engaging problems. Encourage them to build bridges from blocks or towers from shoes. Invite them to make lids for jars filled with marbles. Ask them to retrieve polystyrene balls from bottles or objects from ice cubes. Or let them help you solve everyday household problems such as hanging streamers for their birthday party.

These recommendations were published as an EOS blog post (translated from Van Elsen, 2025)

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, *21*, 93–113. https://doi.org/10.1146/annurev.so.21.080195.000521
ABET. (2024). *Criteria for accrediting engineering programs: 2025 - 2026*. https://www.abet.org/wp-content/uploads/2024/11/2025-2026_EAC_Criteria.pdf
Adie, L., & Klenowski, V. (2016). Moderation and assessment. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1–6). Springer. https://doi.org/10.1007/978-981-287-532-7_393-1
AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
Aldemir, J., & Kermani, H. (2017). Integrated STEM curriculum: Improving educational outcomes for Head Start children. *Early Child Development and Care*, *187*(11), 1694–1706. https://doi.org/10.1080/03004430.2016.1185102
Alexander, P. A., Jetton, T. L., White, S. H., Parsons, J. L., Cotropia, K. K., Liu, H.-C., & Ackerman, C. M. (1994). Young children's creative solutions to realistic and fanciful story problems. *Journal of Creative Behavior*, *28*(2), 89–106. https://doi.org/10.1002/j.2162-6057.1994.tb00723.x
Anggoro, F. K., Dubosarsky, M., & Kabourek, S. (2021). Developing an observation tool to measure preschool children’s problem-solving skills. *Education Sciences*, *11*(12), Article 779. https://doi.org/10.3390/educsci11120779
Annevirta, T., & Vauras, M. (2006). Developmental changes of metacognitive skill in elementary school children. *Journal of Experimental Education*, *74*(3), 195–226. https://doi.org/10.3200/JEXE.74.3.195-226
Araiza-Alba, P., Keane, T., Chen, W. S., & Kaufman, J. (2021). Immersive virtual reality as a tool to learn problem-solving skills. *Computers and Education*, *164*. https://doi.org/10.1016/j.compedu.2020.104121
Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. National Center on Effective Secondary Schools. https://eric.ed.gov/?id=ED301587
Arfé, B., Vardanega, T., & Ronconi, L. (2020). The effects of coding on children's planning and inhibition skills. *Computers & Education*, *148*, Article 103807. https://doi.org/10.1016/j.compedu.2020.103807
Aria, M., Le, T., Cuccurullo, C., Belfiore, A., & Choe, J. (2024). openalexR: An R-tool for collecting bibliometric data from OpenAlex. *The R Journal*, *15*(4), 167–180. https://doi.org/10.32614/RJ-2023-089
Arık, M., & Topçu, M. S. (2022). Implementation of engineering design process in the K-12 science classrooms: Trends and issues. *Research in Science Education*, *52*(1), 21–43. https://doi.org/10.1007/s11165-019-09912-x
Aromataris, E., & Munn, Z. (2020). *JBI manual for evidence synthesis*. JBI. https://doi.org/10.46658/JBIMES-20-01
Asunda, P. A., & Hill, R. B. (2007). Critical features of engineering design in technology education. *Journal of Industrial Teacher Education*, *44*(1), 25–48. http://scholar.lib.vt.edu/ejournals/JITE/v44n1/pdf/asunda.pdf
Ata-Aktürk, A. (2023). “Teacher, I know how to do it”: An engineering design-based STEM activity on the concepts of forces and floating/sinking for young problem solvers. *Science Activities*, *60*(1), 12–24. https://doi.org/10.1080/00368121.2022.2128709
Aydin Gunbatar, S., Ekiz Kiran, B., Boz, Y., & Roehrig, G. H. (2022). A closer examination of the STEM characteristics of the STEM activities published in NSTA journals. *Research in Science*

& Technological Education, *42*(4), 952–977. https://doi.org/10.1080/02635143.2022.2121692

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning*, *4*(1), 87–95. https://doi.org/10.1007/s11409-009-9035-7

Azizan, M. F., Mohd Matore, M. E. E., & Omar, M. (2025). A bibliometric analysis of complex problem-solving approaches in engineering education. *Journal of Applied Science, Engineering, Technology, and Education*, *7*(1), 17–28. https://doi.org/10.35877/454RI.asci3846

Bagnato, S. J. (2005). The authentic alternative for assessment in early intervention: An emerging evidence-based practice. *Journal of Early Intervention*, *28*(1), 17–22. https://doi.org/10.1177/105381510502800102

Bagnato, S. J., Goins, D. D., Pretti-Frontczak, K., & Neisworth, J. T. (2014). Authentic assessment as "best practice" for early childhood intervention: National consumer social validity research. *Topics in Early Childhood Special Education*, *34*(2), 116–127. https://doi.org/10.1177/0271121414523652

Bairaktarova, D., Evangelou, D., Bagiati, A., & Brophy, S. (2011). Early engineering in young children's exploratory play with tangible materials. *Children, Youth and Environments*, *21*(2), 212–235. https://journals.uc.edu/index.php/cye/article/view/6602

Barghaus, K. M., Fantuzzo, J. W., Buek, K., & Gullo, D. F. (2022). Neglected validities: A diagnostic look at the state of early childhood assessment. *Early Childhood Research Quarterly*, *58*, 287–299. https://doi.org/10.1016/j.ecresq.2021.09.007

Bartholomew, S. R., Moon, C., Ruesch, E. Y., & Strimel, G. J. (2019). Kindergarten student's approaches to resolving open-ended design tasks. *Journal of Technology Education*, *30*(2), 90–115. https://doi.org/10.21061/jte.v30i2.a.6

Bat Or, M., Ben-Shoshan Chen, M., & Shalev, O. (2022). Problem-solving expressions in preschool children's "Person Picking an Apple from a Tree" drawings and verbal narratives. *The Arts in Psychotherapy*, *81*, Article 101964. https://doi.org/10.1016/j.aip.2022.101964

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

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309–319. https://doi.org/10.1037/a0029171

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 505–518). Springer New York. https://doi.org/10.1007/978-1-4614-3185-5_39

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bergner, R. M. (2011). What is behavior? And so what? *New Ideas in Psychology*, *29*(2), 147–155. https://doi.org/10.1016/j.newideapsych.2010.08.001

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for Learning: Putting it into practice*. Open University Press.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Bobrowicz, K., & Thibaut, J.-P. (2023). The development of flexible problem solving: An integrative approach. *Journal of Intelligence*, *11*(6), 119. https://doi.org/10.3390/jintelligence11060119

Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, *13*(1), Article 81. https://doi.org/10.1186/s13643-024-02502-7

Bolden, B., DeLuca, C., Kukkonen, T., Roy, S., & Wearing, J. (2020). Assessment of creativity in K-12 education: A scoping review. *Review of Education*, *8*(2), 343–376. https://doi.org/10.1002/rev3.3188

Bonner, S., Chen, P., Jones, K., & Milonovich, B. (2021). Formative assessment of computational thinking: Cognitive and metacognitive processes. *Applied Measurement in Education*, *34*(1), 27–45. https://doi.org/10.1080/08957347.2020.1835912

Bonotto, C., & Santo, L. D. (2015). On the relationship between problem posing, problem solving, and creativity in the primary school. In F. M. Singer, N. F. Ellerton, & J. Cai (Eds.), *Mathematical problem posing: From research to effective practice* (pp. 103–123). Springer. https://doi.org/10.1007/978-1-4614-6258-3_5

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685. https://doi.org/10.1002/asi.22990

Bransford, J. D., & Stein, B. S. (1993). *The ideal problem solver: A guide for improving thinking, learning, and creativity* (2nd ed.). W. H. Freeman.

Brassard, M. R., & Boehm, A. E. (2007). *Preschool assessment: Principles and practice*. Guilford.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, *17*(1), 5–9. https://doi.org/10.1111/j.1745-3992.1998.tb00615.x

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2 ed.). Guilford.

Brownell, W. A. (1942). Problem solving. *Teachers College Record*, *43*(10), 415–443. https://doi.org/10.1177/016146814204301012

Bryan, L., & Guzey, S. S. (2020). K-12 STEM education: An overview of perspectives and considerations. *Hellenic Journal of STEM Education*, *1*(1), 5–15. https://doi.org/10.51724/hjstemed.v1i1.5

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: Evidence from observational analysis of young children's behavior during problem-solving. *Metacognition and Learning*, *7*(3), 197–217. https://doi.org/10.1007/s11409-012-9091-2

Burke, K., Macy, M., & Edokhamhen, E. (2025). Authentic assessment practices for supporting child development during the preschool-to-kindergarten transition: A multiyear comparison of parental and teacher ratings. *Early Childhood Education Journal*, *53*(8), 2933–2947. https://doi.org/10.1007/s10643-025-01884-0

Burns, S., Saleem, S., McMullen, E., Falenchuk, O., White, L., Dhuey, E., & Perlman, M. (2025). A systematic review and meta-analysis of approaches to teaching problem-solving skills in early childhood education and care settings: A focus on science, technology, engineering and mathematics activities. *Review of Education*, *13*(2), Article e70079. https://doi.org/10.1002/rev3.70079

Bustamante, A. S., Greenfield, D. B., & Nayfeld, I. (2018). Early childhood science and engineering: Engaging platforms for fostering domain-general learning skills. *Education Sciences*, *8*(3), 144. https://www.mdpi.com/2227-7102/8/3/144

Cakir, R., Korkmaz, O., Idil, O., & Erdogmus, F. U. (2021). The effect of robotic coding education on preschoolers' problem solving and creative thinking skills. *Thinking Skills and Creativity*, *40*, Article 100812. https://doi.org/10.1016/j.tsc.2021.100812

Cartwright, K. (2024). Interpreting young children's multiplicative strategies through their drawn representations. *Mathematics Education Research Journal*, *36*, 367–397. https://doi.org/10.1007/s13394-023-00450-4

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE Publications, Inc. https://doi.org/10.4135/9781071878811

Chapelle, C. A., & Lee, H.-w. (2021). Understanding argument-based validity in language testing. In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 19–44). Cambridge University Press. https://doi.org/10.1017/9781108669849.004

Cheruiyot, L. G., & Molnár, G. (2026). Technology-based assessment of collaborative problem-solving skills: a bibliometric analysis and review. *Journal of Computers in Education*, *13*, 389–410. https://doi.org/10.1007/s40692-025-00361-2

Cheryan, S., Lombard, E. J., Hailu, F., Pham, L. N. H., & Weltzien, K. (2025). Global patterns of gender disparities in STEM and explanations for their persistence. *Nature Reviews Psychology*, *4*(1), 6–19. https://doi.org/10.1038/s44159-024-00380-3

Christou, E., Parmaxi, A., & Zaphiris, P. (2025). A systematic exploration of scoping and mapping literature reviews. *Universal Access in the Information Society*, *24*(1), 941–951. https://doi.org/10.1007/s10209-024-01120-3

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math. The learning trajectories approach* (2 ed.). Routledge.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Cohen, P. R. (1982). Planning and problem solving. In P. R. Cohen & E. A. Feigenbaum (Eds.), *The handbook of artificial intelligence* (Vol. III, pp. 513–562). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-86576-091-2.50010-1

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Copple, C. (2003). Fostering young children's representation, planning, and reflection: A focus in three current early childhood models. *Journal of Applied Developmental Psychology*, *24*(6), 763–771. https://doi.org/10.1016/j.appdev.2003.09.007

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. Wiley.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, *3*(3), 265–286. https://doi.org/10.1080/0969594960030302

Csapó, B., & Funke, J. (2017a). *The development and assessment of problem solving in 21st-century schools*. https://doi.org/10.1787/9789264273955-3-en

Csapó, B., & Funke, J. (Eds.). (2017b). *The nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing. https://doi.org/10.1787/9789264273955-en.

Culbert, J. H., Hobert, A., & Jahn, N. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, *130*, 2475–2492. https://doi.org/10.1007/s11192-025-05293-3

Cunningham, C. M., Lachapelle, C. P., & Davis, M. E. (2018). Engineering concepts, practices, and trajectories for early childhood education. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 135–174). Springer. https://doi.org/10.1007/978-981-10-8621-2_8

Cutumisu, M., Adams, C., & Lu, C. (2019). A scoping review of empirical research on recent computational thinking assessments. *Journal of Science Education and Technology*, *28*(6), 651–676. https://doi.org/10.1007/s10956-019-09799-3

Damşa, C., Hofmann, R., Ingerman, Å., Ligorio, M. B., & Myhill, D. (2025). Setting the scene for Learning in Context. *Learning in Context*, *2*(1), Article 100013. https://doi.org/10.1016/j.lecon.2025.100013

De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2019). Kappa coefficients for missing data. *Educational and Psychological Measurement*, *79*(3), 558–576. https://doi.org/10.1177/0013164418823249

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology*, *53*(6), 1024–1037. https://doi.org/10.1037/0022-3514.53.6.1024

Deluca, C., Pyle, A., Roy, S., Chalas, A., & Danniels, E. (2019). Perspectives on kindergarten assessment: Toward a common understanding. *Teachers College Record*, *121*(3), 1–58. https://doi.org/10.1177/016146811912100302

DeLuca, C., Pyle, A., Valiquette, A., & LaPointe-McEwan, D. (2020). New directions for kindergarten education: Embedding assessment in play-based learning. *The Elementary School Journal*, *120*(3), 455–479. https://doi.org/10.1086/707008

Demir, B. K., & Öztürk, B. (2024). Development of 21st century skills in the preschool period. In B. Aydınlı & B. D. Altunoğlu (Eds.), *Transforming early childhood education: Technology, sustainability, and foundational skills for the 21st century* (pp. 85–111). Özgür Yayınları. https://doi.org/10.58830/ozgur.pub534.c2211

Demirel, B., & Deretarla Gul, E. (2021). The effect of riddles on problem solving skills of 57–66 months old children. *Cukurova University Faculty of Education Journal*, *50*(2), 721–748. https://doi.org/10.14812/cufej.943800

Dewey, J. (1910). *How we think*. D. C. Heath & Co. https://www.gutenberg.org/files/37423/37423-h/37423-h.htm

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, L. L. (2018). Problem solving in the early years. *Intervention in School and Clinic*, *53*(4), 220–223. https://doi.org/10.1177/1053451217712957

DiStefano, C., & Kamphaus, R. W. (2006). Investigating subtypes of child development: A comparison of cluster analysis and latent class cluster analysis in typology creation. *Educational and Psychological Measurement*, *66*(5), 778–794. https://doi.org/10.1177/0013164405284033

Dochy, F. (2009). The edumetric quality of new modes of assessment: Some issues and prospects. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 85–114). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8905-3_6

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Dorie, B. L., Tranby, Z., Van Cleave, S. K., Cardella, M., & Svarovsky, G. N. (2013, June 23–26). *Using puppets to elicit talk during interviews on engineering with young children* [Paper presentation]. ASEE Annual Conference & Exposition, Atlanta, GA, United States. http://doi.org/10.18260/1-2--22719

Douglas, K. A., Gane, B. D., Neumann, K., & Pellegrino, J. W. (2020). Contemporary methods of assessing integrated STEM competencies. In C. C. Johnson, M. J. Mohr-Schroeder, T. J. Moore, & L. D. English (Eds.), *Handbook of research on STEM education* (pp. 234–254). Routledge.

Dubosarsky, M., John, M. S., Anggoro, F., Wunnava, S., & Celik, U. (2018). Seeds of STEM: The development of a problem-based STEM curriculum for early childhood classrooms. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 249–269). Springer. https://doi.org/10.1007/978-981-10-8621-2_12

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *58*(5), i–113. https://doi.org/10.1037/h0093599

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12. https://doi.org/10.1016/j.compedu.2018.08.004

Ellingsen, K. M. (2016). Standardized assessment of cognitive development: Instruments and issues. In A. Garro (Ed.), *Early childhood assessment in school and clinical child psychology* (pp. 25–49). Springer. https://doi.org/10.1007/978-1-4939-6349-2_2

Engel, S. (2011). Children's need to know: Curiosity in schools. *Harvard Educational Review*, *81*(4), 625–645. https://doi.org/10.17763/haer.81.4.h054131316473115

English, L. D. (1992). *Children's use of domain-specific knowledge and domain-general strategies in novel problem solving*. https://files.eric.ed.gov/fulltext/ED343952.pdf

English, L. D. (2016). STEM education K-12: Perspectives on integration. *International Journal of STEM Education*, *3*(1), Article 3. https://doi.org/10.1186/s40594-016-0036-1

English, L. D. (2017). Advancing elementary and middle school STEM education. *International Journal of Science and Mathematics Education*, *15*(1), 5–24. https://doi.org/10.1007/s10763-017-9802-x

English, L. D. (2018). Early engineering: An introduction to young children's potential. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 1–5). Springer. https://doi.org/10.1007/978-981-10-8621-2_1

English, L. D. (2023). Ways of thinking in STEM-based problem solving. *ZDM – Mathematics Education*, *55*, 1219–1230. https://doi.org/10.1007/s11858-023-01474-7

English, L. D., Lesh, R., & Fennewald, T. (2008, July 6–13). *Future directions and perspectives for problem solving research and curriculum development* [Paper presentation]. 11th International Congress on Mathematical Education, Monterrey, Mexico. https://www.researchgate.net/publication/38183931_Future_directions_and_perspectives_for_problem_solving_research_and_curriculum_development

Epstein, A. S. (2003). How planning and reflection develop young children's thinking skills. *Young Children*, *58*(5), 28–36. http://www.jstor.org/stable/42728978

Ercikan, K., & Solano-Flores, G. (2025). The sociocultural context of educational assessment. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (pp. 115–188). Oxford University Press. https://doi.org/10.1093/oso/9780197654965.003.0003

Erol, A. (2024). Failure analysis and continual improvement in the engineering design process: Teacher roles in children's problem-solving processes. *Education and Information Technologies*, *29*(12), 15229–15248. https://doi.org/10.1007/s10639-024-12489-2

Erol, A., Erol, M., & Canbeldek Erol, M. (2025). Supporting girls' engineering habits of mind with STEM education. *Psychology in the Schools*, *62*(7), 2192–2207. https://doi.org/10.1002/pits.23461

Escolano-Pérez, E., Herrero-Nivela, M. L., & Anguera, M. T. (2019). Preschool metacognitive skill assessment in order to promote educational sensitive response from mixed-methods approach: Complementarity of data analysis. *Frontiers in Psychology*, *10*, Article 1298. https://doi.org/10.3389/fpsyg.2019.01298

European Commission. (2019). *Key competences for lifelong learning*. Publications Office of the European Union. https://doi.org/10.2766/569540

Evans, N. S., Schlesinger, M. A., Hopkins, E. J., Jaeger, G. J., Golinkoff, R. M., & Hirsh-Pasek, K. (2021). Are preschoolers creative? A review of the literature. In S. W. Russ, J. D. Hoffmann, & J. C. Kaufman (Eds.), *The Cambridge handbook of lifespan development of creativity* (pp. 101–125). Cambridge University Press. https://doi.org/10.1017/9781108755726.008

Evans, N. S., Todaro, R. D., Schlesinger, M. A., Golinkoff, R. M., & Hirsh-Pasek, K. (2021). Examining the impact of children’s exploration behaviors on creativity. *Journal of Experimental Child Psychology*, *207*, Article 105091. https://doi.org/10.1016/j.jecp.2021.105091

Fleer, M. (2000). Working technologically: Investigations into how young children design and make during technology education. *International Journal of Technology and Design Education*, *10*(1), 43–59. https://doi.org/10.1023/A:1008923410441

Fleer, M. (2021). When preschool girls engineer: Future imaginings of being and becoming an engineer. *Learning, Culture and Social Interaction*, *30*, Article 100372. https://doi.org/10.1016/j.lcsi.2019.100372

Fleer, M. (2022). Engineering playworld: A model of practice to support children to collectively design, imagine and think using engineering concepts. *Research in Science Education*, *52*(2), 583–598. https://doi.org/10.1007/s11165-020-09970-6

Foster, C. (2023). Problem solving in the mathematics curriculum: From domain-general strategies to domain-specific tactics. *The Curriculum Journal*, *34*(4), 594–612. https://doi.org/10.1002/curj.213

Frensch, P., & Funke, J. (Eds.). (1995). *Complex problem solving: The European perspective*. Lawrence Erlbaum Associates.

Frensch, P. A., & Funke , J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. Frensch & J. Funke (Eds.), *Complex problem solving. The European perspective* (pp. 3–25). Lawrence Erlbaum Associates.

Frey, B. B., Schmitt, V. L., & Allen , J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research, and Evaluation*, *17*, Article 2. https://doi.org/10.7275/sxbs-0829

Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., Hamlett, C. L., Sones, E. M., & Hope, S. K. (2006). Teaching third graders about real-life mathematical problem solving: A randomized controlled study. *The Elementary School Journal*, *106*(4), 293–311. https://doi.org/10.1086/503633

Funke, J., Fischer, A., & Holt, D. V. (2018). Competencies for complexity: Problem solving in the twenty-first century. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills: Research and applications* (pp. 41–53). Springer International Publishing. https://doi.org/10.1007/978-3-319-65368-6_3

Funke, J., & Frensch, P. (2007). Complex problem solving: The European perspective - Ten years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). Lawrence Erlbaum. https://doi.org/10.4324/9781315091938-2

Fusaro, M., & Smith, M. C. (2018). Preschoolers' inquisitiveness and science-relevant problem solving. *Early Childhood Research Quarterly*, *42*, 119–127. https://doi.org/10.1016/j.ecresq.2017.09.002

Gadanidis, G., Hughes, J. M., Minniti, L., & White, B. J. G. (2017). Computational thinking, grade 1 students and the binomial theorem. *Digital Experiences in Mathematics Education*, *3*(2), 77–96. https://doi.org/10.1007/s40751-016-0019-3

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (Version 0.84.1) [R package]. https://CRAN.R-project.org/package=irr

Gardiner, A. K., Bjorklund, D. F., Greif, M. L., & Gray, S. K. (2012). Choosing and using tools: Prior experience and task difficulty influence preschoolers' tool-use strategies. *Cognitive Development*, *27*(3), 240–254. https://doi.org/10.1016/j.cogdev.2012.05.001

Garton, A. F. (2004). *Exploring cognitive development: The child as problem solver*. Blackwell. https://doi.org/10.1002/9780470773574

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15. https://doi.org/10.1177/00222194040370010201

Geist, E. (2025). Reinventing STEM. In E. Geist (Ed.), *Reinventing STEM in early childhood education: A constructivist approach* (pp. 1–31). Routledge. https://doi.org/10.4324/9781003488873-1

German, T. P., & Defeyter, M. A. (2000). Immunity to functional fixedness in young children. *Psychonomic Bulletin and Review*, *7*(4), 707–712. https://doi.org/10.3758/BF03213010

Gilmore, C., Bahnmueller, J., Baikie, A., Burke, S., Edginton, H., Hawkins, J., Rubini, A., Trundley, R., Tynemouth, A., & Xenidou-Dervou, I. (2025). Supporting children's inhibitory control in mathematics learning: A participatory research study in classroom contexts. *Learning in Context*, *2*(1), Article 100009. https://doi.org/10.1016/j.lecon.2025.100009

Ginsburg, H. P. (Ed.). (1983). *The development of mathematical thinking*. Academic Press.

Gold, Z. S., & Elicker, J. (2020). Engineering peer play: A new perspective on science, technology, engineering, and mathematics (STEM) early childhood education. In A. Ridgway, G. Quiñones, & L. Li (Eds.), *Peer play and relationships in early childhood: International research perspectives* (pp. 61–75). Springer. https://doi.org/10.1007/978-3-030-42331-5_5

Gold, Z. S., Elicker, J., Evich, C. D., Mishra, A. A., Howe, N., & Weil, A. E. (2021). Engineering play with blocks as an informal learning context for executive function and planning. *Journal of Engineering Education*, *110*(4), 803–818. https://doi.org/10.1002/jee.20421

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74–83. https://doi.org/10.1016/j.edurev.2014.10.002

Grisham-Brown, J., Hallam, R., & Brookshire, R. (2006). Using authentic assessment to evidence children's progress toward early learning standards. *Early Childhood Education Journal*, *34*(1), 45–51. https://doi.org/10.1007/s10643-006-0106-y

Grubbs, M., & Strimel, G. (2015). Engineering design: The great integrator. *Journal of STEM Teacher Education*, *50*(1), 77–90. https://doi.org/10.30707/JSTE50.1Grubbs

Gu, P. Y. (2021). An argument-based framework for validating formative assessment in the classroom. *Frontiers in Education*, *6*, Article 605999. https://doi.org/10.3389/feduc.2021.605999

Guevara, M., & Puche-Navarro, R. (2015). The emergence of cognitive short-term planning: Performance of preschoolers in a problem-solving task. *Acta Colombiana de Psicología*, *18*(2), 13–27. https://doi.org/10.14718/ACP.2015.18.2.2

Guevara, M., Rojas Ospina, T., & van Geert, P. (2020). Understanding transfer from a dynamic system approach: Two studies of children using problem-solving tasks. *Cogent Education*, *7*(1), Article 1814512. https://doi.org/10.1080/2331186X.2020.1814512
Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.
Gullo, D. F. (2005). *Understanding assessment and evaluation in early childhood education*. Teachers College Press.
Gunawan, I., Darhim, D., & Kusnandi, K. (2019). Exploration of the behavior of understanding mathematical concepts of junior high school students. *Journal of Physics: Conference Series*, *1157*(4), Article 042098. https://doi.org/10.1088/1742-6596/1157/4/042098
Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. https://doi.org/10.1348/000711006X126600
Hallinger, P. (2023). Bibliometric review methodology and state of the science bibliometric review of research on problem-based learning, 2017-2021. *The Interdisciplinary Journal of Problem-based Learning*, *17*(2). https://doi.org/10.14434/ijpbl.v17i2.35761
Hardy, J. K., Mere-Cook, Y., & Yang, H.-W. (2024). Critical issues in measuring and teaching social problem-solving in early childhood research. *Topics in Early Childhood Special Education*, *46*(1), 5–14. https://doi.org/10.1177/02711214241288214
Hargreaves, E. (2007). The validity of collaborative assessment for learning. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 185–199. https://doi.org/10.1080/09695940701478594
Harlen, W. (2007). Meanings, processes and properties of assessment. In W. Harlen (Ed.), *Assessment of Learning* (pp. 11–24). SAGE Publications. https://doi.org/10.4135/9781446214695.n2
Hayes, J. R. (1981). *The complete problem solver*. Routledge. https://doi.org/10.4324/9780203062715
Hendriani, M., Made Arnawa, I., & Zainil, M. (2025). Global research trends in mathematical problem-solving skills: A bibliometric review of 2020–2024 publications. *Seminars in Medical Writing and Education*, *4*, Article 453. https://doi.org/10.56294/mw2025453
Heroman, C. (2017). Preschool through grade 3: Making and tinkering: Bringing design challenges to the classroom. *Young Children*, *72*(2), 72–79. https://www.jstor.org/stable/90004128
Heyman, R. E., Lorber, M. F., Eddy, J. M., & West, T. V. (2014). Behavioral observation and coding. In C. M. Judd & H. T. Reis (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 345–372). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.018
Hong, E., & Milgram, R. M. (1991). Original thinking in preschool children: A validation of ideational fluency measures. *Creativity Research Journal*, *4*(3), 253–260. https://doi.org/10.1080/10400419109534397
Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement*, *56*(4), 715–732. https://doi.org/10.1111/jedm.12234
Hudson, J. A., Shapiro, L. R., & Sosa, B. B. (1995). Planning in the real world: Preschool children's scripts and plans for familiar events. *Child Development*, *66*(4), 984–998. https://doi.org/10.2307/1131793
Hudson, J. A., Sosa, B. B., & Shapiro, L. R. (1997). Scripts and plans: The development of preschool children's event knowledge and event planning. In S. L. Friedman & E. K. Scholnick (Eds.), *The developmental psychology of planning: Why, how, and when do we plan?* (pp. 77–102). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315806495

Isabelle, A. D., Russo, L., & Velazquez-Rojas, A. (2021). Using the engineering design process (EDP) to guide block play in the kindergarten classroom: exploring effects on learning outcomes. *International Journal of Play*, *10*(1), 43–62. https://doi.org/10.1080/21594937.2021.1878772

ITEEA. (2020). *Standards for technological and engineering literacy: The role of technology and engineering in STEM education*. https://www.iteea.org/downloadpurchase-stel

Jing, Y., Wang, C., Chen, Y., Wang, H., Yu, T., & Shadiev, R. (2024). Bibliometric mapping techniques in educational technology research: A systematic literature review. *Education and Information Technologies*, *29*(8), 9283–9311. https://doi.org/10.1007/s10639-023-12178-6

Jirout, J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental Review*, *32*(2), 125–160. https://doi.org/10.1016/j.dr.2012.04.002

Jirout, J., & Klahr, D. (2020). Questions – and some answers – about young children's questions. *Journal of Cognition and Development*, *21*(5), 729–753. https://doi.org/10.1080/15248372.2020.1832492

Jirout, J., & Zimmerman, C. (2015). Development of science process skills in the early childhood years. In K. Cabe Trundle & M. Saçkes (Eds.), *Research in early childhood science education* (pp. 143–165). Springer. https://doi.org/10.1007/978-94-017-9505-0_7

Jirout, J. J., Evans, N. S., & Son, L. K. (2024). Curiosity in children across ages and contexts. *Nature Reviews Psychology*, *3*(9), 622–635. https://doi.org/10.1038/s44159-024-00346-5

Johnson, C. C., Mohr-Schroeder, M. J., Moore, T. J., & English, L. D. (Eds.). (2020). *Handbook of research on STEM education*. Routledge.

Jonassen, D., Strobel, J., & Lee, C. B. (2006). Everyday problem solving in engineering: Lessons for engineering educators. *Journal of Engineering Education*, *95*(2), 139–151. https://doi.org/10.1002/j.2168-9830.2006.tb00885.x

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. https://doi.org/10.1007/BF02299613

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*(4), 63–85. https://doi.org/10.1007/BF02300500

Jonassen, D. H. (2010a). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge. https://doi.org/10.4324/9780203847527

Jonassen, D. H. (2010b, September 29 – October 1). *Research issues in problem solving* [Paper presentation]. The 11th International Conference on Education Research, Seoul, Korea. https://aect.org/docs/JonassenICER.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Praeger.

Kane, M. T. (2013a). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Kane, M. T. (2013b). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2021). Articulating a validity argument. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (2nd ed., pp. 32–47). Routledge. https://doi.org/10.4324/9781003220756-4

Kane, M. T., & Wools, S. (2019). Perspectives on the validity of classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement*. Routledge. https://doi.org/10.4324/9780429507533-2

Kangas, K., Sormunen, K., & Korhonen, T. (2022). Creative learning with technologies in young students' STEAM education. In S. Papadakis & M. Kalogiannakis (Eds.), *STEM, robotics, mobile apps in early childhood and primary education: Technology to promote teaching and learning* (pp. 157–179). Springer Nature. https://doi.org/10.1007/978-981-19-0568-1_9

Keen, R. (2011). The development of problem solving in young children: A critical cognitive skill. *Annual Review of Psychology*, *62*(1), 1–21. https://doi.org/10.1146/annurev.psych.031809.130730

Kelchtermans, G. (2021). Keeping educational research close to practice. *British Educational Research Journal*, *47*(6), 1504–1511. https://doi.org/10.1002/berj.3716

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(1), Article 11. https://doi.org/10.1186/s40594-016-0046-z

Kilpatrick, J. (1978). Variables and methodologies in research on problem solving. In L. L. Hatfield & D. A. Bradbard (Eds.), *Mathematical problem solving: Papers from a research workshop* (pp. 7–20). ERIC Information Analysis Center for Science, Mathematics, and Environmental Education. https://files.eric.ed.gov/fulltext/ED156446.pdf

Kim, E. J., & Koo, C. D. (2019). Nature activities in urban parks to encourage curiosity and scientific problem-solving ability in kindergarteners. *Journal of People, Plants, and Environment*, *22*(5), 515–524. https://doi.org/10.11628/ksppe.2019.22.5.515

Kirsh, D. (2008). Problem solving and situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 264–306). Cambridge University Press. https://doi.org/10.1017/CBO9780511816826.015

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology*, *62*(1), 1–18. https://doi.org/10.1002/asi.21444

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Kuo, C.-C., Maker, J., Su, F.-L., & Hu, C. (2010). Identifying young gifted children and cultivating problem solving abilities and multiple intelligences. *Learning and Individual Differences*, *20*(4), 365–379. https://doi.org/10.1016/j.lindif.2010.05.005

Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. W. Engle (Eds.), *Assessing problem solving in context* (pp. 11–26). SAGE Publications, Inc. https://doi.org/10.4135/9781452233529.n2

Lambert, E. B. (2001). Metacognitive problem solving in preschoolers. *Australasian Journal of Early Childhood*, *26*(3), 24–29. https://doi.org/10.1177/183693910102600306

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lane, S., & Marion, S. F. (2025). Validity and validation. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational Measurement* (5th ed., pp. 191–275). Oxford University Press.

Lee, W.-C., & Harris, D. J. (2025). Reliability in educational measurement. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 277–382). Oxford University Press. https://doi.org/10.1093/oso/9780197654965.003.0005

Leikin, M. (2013). The effect of bilingualism on creativity: Developmental and educational perspectives. *International Journal of Bilingualism*, *17*(4), 431–447. https://doi.org/10.1177/1367006912438300

Leikin, M., & Tovli, E. (2014). Bilingualism and creativity in early childhood. *Creativity Research Journal*, *26*(4), 411–417. https://doi.org/10.1080/10400419.2014.961779

Leikin, R. (2009). Exploring mathematical creativity using multiple solution tasks. In R. Leikin, A. Berman, & B. Koichu (Eds.), *Creativity in mathematics and the education of gifted students* (pp. 129–145). Sense Publishers.

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*(1), Article 69. https://doi.org/10.1186/1748-5908-5-69

Liljedahl, P. (2016). Building thinking classrooms: Conditions for problem-solving. In P. Felmer, E. Pehkonen, & J. Kilpatrick (Eds.), *Posing and solving mathematical problems: Advances and new perspectives* (pp. 361–386). Springer International Publishing. https://doi.org/10.1007/978-3-319-28023-3_21

Lin, X., Yang, W., Wu, L., Zhu, L., Wu, D., & Li, H. (2021). Using an inquiry-based science and engineering program to promote science knowledge, problem-solving skills and approaches to learning in preschool children. *Early Education and Development*, *32*(5), 695–713. https://doi.org/10.1080/10409289.2020.1795333

Lin, X. Y., Yang, W. P., Xie, W. L., & Li, H. (2023). The integrative role of parenting styles and parental involvement in young children's science problem-solving skills. *Frontiers in Psychology*, *14*, Article 1096846. https://doi.org/10.3389/fpsyg.2023.1096846

Lippard, C. N., Lamm, M. H., & Riley, K. L. (2017). Engineering thinking in prekindergarten children: A systematic literature review. *Journal of Engineering Education*, *106*(3), 454–474. https://doi.org/10.1002/jee.20174

Lippard, C. N., Lamm, M. H., Tank, K. M., & Choi, J. Y. (2019). Pre-engineering thinking and the engineering habits of mind in preschool classroom. *Early Childhood Education Journal*, *47*(2), 187–198. https://doi.org/10.1007/s10643-018-0898-6

Lippard, C. N., Riley, K. L., & Lamm, M. H. (2018). Encouraging the development of engineering habits of mind in prekindergarten learners. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 19–36). Springer. https://doi.org/10.1007/978-981-10-8621-2_3

Lombaers, P., de Bruin, J., & van de Schoot, R. (2024). Reproducibility and data storage for active learning-aided systematic reviews. *Applied Sciences*, *14*(9), Article 3842. https://www.mdpi.com/2076-3417/14/9/3842

Lopes, C. E., Grando, R. C., & D'Ambrosio, B. S. (2017). Experiences situating mathematical problem solving at the core of early childhood classrooms. *Early Childhood Education Journal*, *45*(2), 251–259. https://doi.org/10.1007/s10643-016-0775-0

Lu, C., Macdonald, R., Odell, B., Kokhan, V., Demmans Epp, C., & Cutumisu, M. (2022). A scoping review of computational thinking assessments in higher education. *Journal of Computing in Higher Education*, *34*(2), 416–461. https://doi.org/10.1007/s12528-021-09305-y

Lucas, B., Claxton, G., & Hanson, J. (2014). *Thinking like an engineer: Implications for the education system*. Royal Academy of Engineering. https://www.researchgate.net/publication/264950432_Thinking_like_an_engineer_implications_for_the_education_system

Luce, M. R., & Hsi, S. (2015). Science-relevant curiosity expression and interest in science: An exploratory study. *Science Education*, *99*(1), 70–97. https://doi.org/10.1002/sce.21144

MacDonald, B. L., Tofel-Grehl, C., & Searle, K. A. (2022). Play, problem-solving, STEM conceptions, and efficacy in STEM: An introduction to the STEM in early childhood education special issue. *Education Sciences*, *12*(5), 352. https://doi.org/10.3390/educsci12050352

Maker, C. J. (2001). DISCOVER: Assessing and developing problem solving. *Gifted Education International*, *15*(3), 232–251. https://doi.org/10.1177/026142940101500303

Malcok, B. A., & Ceylan, R. (2022). The effects of STEM activities on the problem-solving skills of 6-year-old preschool children. *European Early Childhood Education Research Journal*, *30*(3), 423–436. https://doi.org/10.1080/1350293x.2021.1965639

Mangold, J., & Robinson, S. (2013, June 23–26). *The engineering design process as a problem solving and learning tool in K-12 classrooms* [Paper presentation]. ASEE Annual Conference & Exposition, Atlanta, GA, United States. http://doi.org/10.18260/1-2--22581

Martín-Páez, T., Aguilera, D., Perales-Palacios, F. J., & Vílchez-González, J. M. (2019). What are we talking about when we talk about STEM education? A review of literature. *Science Education*, *103*(4), 799–822. https://doi.org/10.1002/sce.21522

Martin, J. (1977). The development and use of classroom observation instruments. *Canadian Journal of Education / Revue canadienne de l'éducation*, *2*(3), 43–54. https://doi.org/10.2307/1494419

Marulis, L. M., & Nelson, L. J. (2021). Metacognitive processes and associations to executive function and motivation during a problem-solving task in 3–5 year olds. *Metacognition and Learning*, *16*(1), 207–231. https://doi.org/10.1007/s11409-020-09244-6

Mascolo, M. F., & Fischer, K. W. (2015). Dynamic development of thinking, feeling, and acting. In R. M. Lerner (Ed.), *Handbook of child psychology and developmental science* (7th ed., Vol. 1, pp. 113–161). https://doi.org/10.1002/9781118963418.childpsy104

Mayer, R. E. (1994). Problem solving. In M. W. Eysenck (Ed.), *The Blackwell dictionary of cognitive psychology* (pp. 284–288). Blackwell.

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, *26*(1), 49–63. https://doi.org/10.1023/A:1003088013286

McGeown, S., Oxley, E., Practice Partners, L. t. R., Ricketts, J., & Shapiro, L. (2023). Working at the intersection of research and practice: The love to read project. *International Journal of Educational Research*, *117*, Article 102134. https://doi.org/10.1016/j.ijer.2022.102134

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282.

McLeod, B. D., Sutherland, K. S., Broda, M. D., Granger, K. L., Hollins, N., Frey, A. J., Markowicz, K. A., & Dear, E. (2024). Initial development of the child responsiveness scale for early childhood settings. *Journal of Emotional and Behavioral Disorders*, *33*(3), 159–174. https://doi.org/10.1177/10634266241271392

McMillan, J. (2013). *SAGE handbook of research on classroom assessment*. SAGE Publications. https://doi.org/10.4135/9781452218649

Mehsen, V., Morag, L., Chesta, S., Cleaton, K., & Burgos, H. (2022). Hot executive function assessment instruments in preschool children: A systematic review. *International Journal of Environmental Research and Public Health*, *19*(1), 95. https://www.mdpi.com/1660-4601/19/1/95

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Milburn, T. F., Lonigan, C. J., DeFlorio, L., & Klein, A. (2019). Dimensionality of preschoolers' informal mathematical abilities. *Early Childhood Research Quarterly*, *47*, 487–495. https://doi.org/10.1016/j.ecresq.2018.07.006

Mohamed, A., & Maker, C. J. (2011). Creative storytelling: Evaluating problem solving in children's invented stories. *Gifted Education International*, *27*(3), 327–348. https://doi.org/10.1177/026142941102700309

Mokkink, L. B., Prinsen, C. A., Bouter, L. M., Vet, H. C., & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther*, *20*(2), 105–113. https://doi.org/10.1590/bjpt-rbf.2014.0143

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity*, *9*, 35–45. https://doi.org/10.1016/j.tsc.2013.03.002

Monteira, S. F., & Jiménez-Aleixandre, M. P. (2016). The practice of using evidence in kindergarten: The role of purposeful observation. *Journal of Research in Science Teaching*, *53*(8), 1232–1258. https://doi.org/10.1002/tea.21259

Moore, T. J., Glancy, A. W., Tank, K. M., Kersten, J. A., Smith, K. A., & Stohlmann, M. S. (2014). A framework for quality K-12 engineering education: Research and development. *Journal of Pre-College Engineering Education Research*, *4*(1), Article 2. https://doi.org/10.7771/2157-9288.1069

Moore, T. J., Johnston, A. C., & Glancy, A. W. (2020). STEM integration: A synthesis of conceptual frameworks and definitions. In C. C. Johnson, M. J. Mohr-Schroeder, T. J. Moore, & L. D. English (Eds.), *Handbook of research on STEM education* (pp. 3–16). Routledge.

Moore, T. J., Tank, K. M., & English, L. (2018). Engineering in the early grades: Harnessing children's natural ways of thinking. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 9–18). Springer. https://doi.org/10.1007/978-981-10-8621-2_2

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, *22*(4), 13–25. https://doi.org/10.1111/j.1745-3992.2003.tb00140.x

Muijs, D., Bortels, H., Bosman, A., Counsell, C., Heynsens, E., Kellens, N., Oates, T., Spielman, A., Surma, T., Thyssen, K., & Vancoppenolle, A. (2025). *Kennisrijke minimumdoelen in het Vlaams basisonderwijs: Rapport van de Commissie Muijs [Knowledge-rich minimum attainment targets in Flemish primary education: Report of the Muijs Commission]* [Report]. AHOVOKS. https://assets.vlaanderen.be/image/upload/v1758804426/repositories-prd/DEF_Rapport_Vlaamse_kennisrijke_Minimumdoelen_PDF_lusrpp.pdf

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), Article 143. https://doi.org/10.1186/s12874-018-0611-x

Nadelson, L. S., & Seifert, A. L. (2017). Integrated STEM defined: Contexts, challenges, and the future. *The Journal of Educational Research*, *110*(3), 221–223. https://doi.org/10.1080/00220671.2017.1289775

Nam, K. W., Kim, H. J., & Lee, S. (2019). Connecting plans to action: The effects of a card-coded robotics curriculum and activities on Korean kindergartners. *Asia-Pacific Education Researcher*, *28*(5), 387–397. https://doi.org/10.1007/s40299-019-00438-4

National Academy of Engineering, & National Research Council. (2009). *Engineering in K-12 Education: Understanding the Status and Improving the Prospects*. The National Academies Press. https://doi.org/10.17226/12635

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

O'Reilly, C., Devitt, A., & Hayes, N. (2022). Critical thinking in the preschool classroom: A systematic literature review. *Thinking Skills and Creativity*, *46*, Article 101110. https://doi.org/10.1016/j.tsc.2022.101110

OECD. (2004). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. https://doi.org/10.1787/9789264101739-en

OECD. (2005). *Problem solving for tomorrow's world: First measures of cross-curricular competencies from PISA 2003*. https://doi.org/10.1787/9789264006430-en

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. https://doi.org/10.1787/9789264190511-en

OECD. (2014). *PISA 2012 results: Creative problem solving* (Vol. V). https://doi.org/10.1787/9789264208070-en

OECD. (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 results: Collaborative problem solving* (Vol. V). https://doi.org/10.1787/9789264285521-en

OECD. (2019). *PISA 2021 creative thinking framework (third draft)*. https://www.teachertoolkit.co.uk/wp-content/uploads/2024/09/PISA-2021-creative-thinking-framework.pdf

OECD. (2025). *Results from TALIS starting strong 2024: Strengthening early childhood education and care*. https://doi.org/10.1787/20af08c0-en

Oğuz, V., & Köksal- Akyol, A. (2015). Problem Çözme Becerisi Ölçeği (PÇBÖ) geçerlik ve güvenirlik çalışması [Problem-Solving Skills Scale (PSSS) validity and reliability]. *Cukurova University Faculty of Education Journal*, *44*(1), 105–122. https://doi.org/10.14812/cuefd.54362

OpenAlex. (2024a). *OpenAlex topic mapping table* [Taxonomy]. https://github.com/ourresearch/openalex-topic-classification/tree/main?tab=readme-ov-file

OpenAlex. (2024b). *OpenAlex: End-to-end process for topic classification*. https://github.com/ourresearch/openalex-topic-classification/tree/main

Osborn, A. F. (1953). *Applied imagination: Principles and procedures of creative thinking*. Scribners.

Otero, T. M., & Barker, L. A. (2014). The frontal lobes and executive functioning. In S. Goldstein & J. A. Naglieri (Eds.), *Handbook of executive functioning* (pp. 29–44). Springer. https://doi.org/10.1007/978-1-4614-8106-5_3

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article 71. https://doi.org/10.1136/bmj.n71

Panter, J. E., & Bracken, B. A. (2013). Preschool assessment. In *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.* (pp. 21–37). American Psychological Association. https://doi.org/10.1037/14049-002

Papandreou, M. (2009, July 19–24). Preschoolers semiotic activity: Additive problem-solving and the representation of quantity. In M. Tzekaki, M. Kaldrimidou, & H. Sakonidis (Eds.), *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 321–328). Thessaloniki, Greece.

Parkes, J. (2013). Reliability in classroom assessment. In J. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 107–123). SAGE. https://doi.org/10.4135/9781452218649.n7

Parsons, S. (2021). The importance of collaboration for knowledge co-construction in 'close-to-practice' research. *British Educational Research Journal*, *47*(6), 1490–1499. https://doi.org/10.1002/berj.3714

Parsonson, B. S., & Baer, D. M. (1978). Training generalized improvisation of tools by preschool children. *Journal of Applied Behavior Analysis*, *11*(3), 363–380. https://doi.org/10.1901/jaba.1978.11-363

Philpott-Robinson, K., Johnson, T., Evans, L., Wales, K., Leonard, C., & Lane, A. E. (2023). Measurement of self-regulation in preschool and elementary children: A scoping review. *Physical & Occupational Therapy In Pediatrics*, *43*(4), 403–429. https://doi.org/10.1080/01942638.2022.2158055

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, *19*(2), 159–203. https://doi.org/10.1177/1094428115624965

Pollock, D., Davies, E. L., Peters, M. D. J., Tricco, A. C., Alexander, L., McInerney, P., Godfrey, C. M., Khalil, H., & Munn, Z. (2021). Undertaking a scoping review: A practical guide for nursing and midwifery students, clinicians, researchers, and academics. *Journal of Advanced Nursing*, *77*(4), 2102–2113. https://doi.org/10.1111/jan.14743

Polya, G. (1945). *How we solve it: A new aspect of mathematical method*. Princeton.

Popham, J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership*, *55*(2), 72–75. https://www.ascd.org/el/articles/whats-wrong-and-whats-right-with-rubrics

Popham, W. J. (2017). *Classroom assessment: What teachers need to know* (8th ed.). Pearson.

Popper, K. (1999). *All life is problem solving*. Routledge.

Pouw, W. T. J. L., van Gog, T., & Paas, F. (2014). An embedded and embodied cognition review of instructional manipulatives. *Educational Psychology Review*, *26*(1), 51–72. https://doi.org/10.1007/s10648-014-9255-5

Pretz, J. E., Naples, A. J., & Sternberg, R. J. (2003). Recognizing, defining, and representing problems. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 3–30). Cambridge University Press. https://doi.org/10.1017/CBO9780511615771.002

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*. https://doi.org/10.48550/arXiv.2205.01833

Priemer, B., Eilerts, K., Filler, A., Pinkwart, N., Rösken-Winter, B., Tiemann, R., & Zu Belzen, A. U. (2020). A framework to foster problem-solving in STEM and computing education. *Research in Science & Technological Education*, *38*(1), 105–130. https://doi.org/10.1080/02635143.2019.1600490

Purzer, Ş., & Douglas, K. A. (2018). Assessing early engineering thinking and design competencies in the classroom. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 113–132). Springer. https://doi.org/10.1007/978-981-10-8621-2_7

Pyle, A., DeLuca, C., Danniels, E., & Wickstrom, H. (2020). A model for assessment in play-based kindergarten education. *American Educational Research Journal*, *57*(6), 2251–2292. https://doi.org/10.3102/0002831220908800

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramani, G. B., & Brownell, C. A. (2014). Preschoolers' cooperative problem solving: Integrating play and problem solving. *Journal of Early Childhood Research*, *12*(1), 92–108. https://doi.org/10.1177/1476718x13498337

Randel, B., & Clark, T. (2013). Measuring classroom assessment practices. In J. H. McMillan, B. Randel, & T. Clark (Eds.), *SAGE handbook of research on classroom assessment* (pp. 145–163). SAGE Publications. https://doi.org/10.4135/9781452218649.n9

Raven, S., Husseini, D. A., & Cevik, E. (2018). We are Engineers! *Science and Children*, *56*(1), 55–61. http://www.jstor.org/stable/44843260

Ray, W. S. (1955). Complex tasks for use in human problem-solving research. *Psychological Bulletin*, *52*(2), 134–149. https://doi.org/10.1037/h0044763

Reed, S. K. (2016). The structure of ill-structured (and well-structured) problems revisited. *Educational Psychology Review*, *28*(4), 691–716. https://doi.org/10.1007/s10648-015-9343-1

Reilly, S. E., Downer, J. T., & Grimm, K. J. (2022). Developmental trajectories of executive functions from preschool to kindergarten. *Developmental Science*, *25*(5), Article e13236. https://doi.org/10.1111/desc.13236

Resnick, M., & Silverman, B. (2005). *Some reflections on designing construction kits for kids.* Proceedings of the 2005 conference on Interaction design and children, Boulder, Colorado. https://doi.org/10.1145/1109540.1109556

Reuter, T., & Leuchter, M. (2022). Examining kindergarten children's testing and optimising in the context of a gear engineering task. *European Journal of STEM Education*, *7*(1), Article 04. https://doi.org/10.20897/ejsteme/11827

Revelle, W. (2024). *psych: Procedures for personality and psychological research* (Version 2.4.4.26) [R package]. Northwestern University. https://CRAN.R-project.org/package=psych

Rhodes, K. T., Richland, L. E., & Alcalá, L. (2024). Problem solving is embedded in context… so how do we measure it? *Frontiers in Psychology*, *15*, Article 1380178. https://doi.org/10.3389/fpsyg.2024.1380178

Riley, K., Miller, G. E., & Sorenson, C. (2016). Early childhood authentic and performance-based assessment. In A. Garro (Ed.), *Early childhood assessment in school and clinical child psychology* (pp. 95–117). Springer. https://doi.org/10.1007/978-1-4939-6349-2_5

Rogers, J. (2004). Autonomy and mathematical problem-solving: The early years. *Education 3-13*, *32*(3), 24–31. https://doi.org/10.1080/03004270485200291

Ros, R., Bjarnason, E., & Runeson, P. (2017, June 15–16). A machine learning approach for semi-automated search and selection in literature studies. In E. Mendes (Ed.) *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering* (pp. 118–127). Karlskrona, Sweden.

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

Rott, B., Specht, B., & Knipping, C. (2021). A descriptive phase model of problem-solving processes. *ZDM – Mathematics Education*, *53*(4), 737–752. https://doi.org/10.1007/s11858-021-01244-3

Said-Metwaly, S., Noortgate, W. V. d., & Kyndt, E. (2017). Methodological issues in measuring creativity: A systematic literature review. *Creativity. Theories–Research–Applications*, *4*(2), 276. https://doi.org/10.1515/ctra-2017-0014

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). SAGE Publications.

Salminen-Saari, J. F. A., Garcia Moreno-Esteva, E., Haataja, E., Toivanen, M., Hannula, M. S., & Laine, A. (2021). Phases of collaborative mathematical problem solving and joint attention: a case study utilizing mobile gaze tracking. *ZDM – Mathematics Education*, *53*(4), 771–784. https://doi.org/10.1007/s11858-021-01280-z

Sancar, R., Atal, D., & Deryakulu, D. (2021). A new framework for teachers' professional development. *Teaching and Teacher Education*, *101*, Article 103305. https://doi.org/10.1016/j.tate.2021.103305

Sanoff, H. (1995). *Creating environments for young children*. North Carolina State University, School of Design. https://eric.ed.gov/?id=ED394640

Schäfer, J., Reuter, T., Karbach, J., & Leuchter, M. (2024). Domain-specific knowledge and domain-general abilities in children's science problem-solving. *British Journal of Educational Psychology*, *94*(2), 346–366. https://doi.org/10.1111/bjep.12649

Scherer, R., & Beckmann, J. F. (2014). The acquisition of problem solving competence: evidence from 41 countries that math and science education matters. *Large-scale Assessments in Education*, *2*(1), 10. https://doi.org/10.1186/s40536-014-0010-7

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, Article 101602. https://doi.org/10.1016/j.ijer.2020.101602

Schneider, C. M., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 251–276). Routledge. https://doi.org/10.4324/9780203874851

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Schoenfeld, A. H. (2011). *How we think: A theory of goal-oriented decision making and its educational applications*. Routledge.

Schoenfeld, A. H. (2016). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (reprint). *Journal of Education*, *196*(2), 1–38. https://doi.org/10.1177/002205741619600202

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*(1), 113–125. https://doi.org/10.1023/A:1003044231033

Schreiber, F., & Cramer, C. (2024). Towards a conceptual systematic review: proposing a methodological framework. *Educational Review*, *76*(6), 1458–1479. https://doi.org/10.1080/00131911.2022.2116561

Seitz, H., & Bartholomew, C. (2008). Powerful portfolios for young children. *Early Childhood Education Journal*, *36*(1), 63–68. https://doi.org/10.1007/s10643-008-0242-7

Shechter, T., Eden, S., & Spektor-Levy, O. (2021). Preschoolers' nascent engineering thinking during a construction task. *Journal of Cognitive Education and Psychology*, *20*(2), 83–111. https://doi.org/10.1891/JCEP-D-20-00010

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of researchon teaching* (4th ed., pp. 1066–1101). AERA.

Shiakalli, M. A., & Zacharos, K. (2012). The contribution of external representations in pre-school mathematical problem solving. *International Journal of Early Years Education*, *20*(4), 315–331. https://doi.org/10.1080/09669760.2012.714992

Shin, N., Jonassen, D. H., & McGee, S. (2003). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of Research in Science Teaching*, *40*(1), 6–33. https://doi.org/10.1002/tea.10058

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005

Shure, M. B., & Spivack, G. (1980). Interpersonal problem solving as a mediator of behavioral adjustment in preschool and kindergarten children. *Journal of Applied Developmental Psychology*, *1*(1), 29–44. https://doi.org/10.1016/0193-3973(80)90060-X

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, *46*(6), 606–620. https://doi.org/10.1037//0003-066x.46.6.606

Silva, C., Sousa-Gomes, V., Fávero, M., Oliveira-Lopes, S., Merendeiro, C. S., Oliveira, J., & Moreira, D. (2022). Assessment of preschool-age executive functions: A systematic review. *Clinical Psychology & Psychotherapy*, *29*(4), 1374–1391. https://doi.org/10.1002/cpp.2718

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, *4*(3), 181–201. https://doi.org/10.1016/0004-3702(73)90011-8

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, *7*(2), 268–288. https://doi.org/10.1016/0010-0285(75)90012-2

Simon, H. A. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), *Handbook of learning & cognitive processes: V. Human information* (pp. 271–295). Lawrence Erlbaum.

Sireci, S. G., & Botha, S. M. (2020). Timing considerations in test development and administration. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 32–46). Routledge. https://doi.org/10.4324/9781351064781-3

Sjölund, S., Lindvall, J., Larsson, M., & Ryve, A. (2023). Mapping roles in research-practice partnerships: A systematic literature review. *Educational Review*, *75*(7), 1490–1518. https://doi.org/10.1080/00131911.2021.2023103

Skinner, B. F. (1953). *Science and human behavior*. Free Press.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, *22*(4), 26–33. https://doi.org/10.1111/j.1745-3992.2003.tb00141.x

Snow, C. E., & Van Hemel, S. B. (Eds.). (2008). *Early childhood assessment: Why, what, and how*. The National Academies Press.

Sormunen, K., Yliverronen, V., & Rönkkö, M.-L. (2026). Unveiling young children's collaborative communication in hands-on problem-solving. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-025-10049-4

Spoon, R., Rubenstein, L. D., & Terwillegar, S. R. (2021). Team effectiveness in creative problem solving: Examining the role of students' motivational beliefs and task analyses in team performance [Article]. *Thinking Skills and Creativity*, *40*, Article 100792. https://doi.org/10.1016/j.tsc.2021.100792

Steffensen, S. V., & Vallée-Tourangeau, F. (2018). An ecological perspective on insight problem solving. In F. Vallée-Tourangeau (Ed.), *Insight: On the origins of new ideas* (pp. 169–190). Routledge. https://doi.org/10.4324/9781315268118-9

Stern, M., & Hertel, S. (2022). Relationship between maternal scaffolding and preschooler's metacognitive strategies in a problem-solving situation. *Learning and Instruction*, *80*, Article 101631. https://doi.org/10.1016/j.learninstruc.2022.101631

Sternberg, R. J. (1989). Domain-generality versus domain-specificity: The life and impending death of a false dichotomy. *Merrill-Palmer Quarterly*, *35*(1), 115–130. http://www.jstor.org.kuleuven.e-bronnen.be/stable/23086427

Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 233–242). SAGE. https://doi.org/10.4135/9781446250808.n14

Stoll, J., Hamilton, A., Oxley, E., Eastman, A. M., & Brent, R. (2012). Young thinkers in motion: Problem solving and physics in preschool. *Young Children*, *67*(2), 20–26. http://www.jstor.org/stable/42731148

Sucilestari, R., Ramdani, A., Susilawati, Sukarso, A., Rokhmat, J., & Arizona, K. (2025). Exploring the global trends of creative problem solving (CPS): A systematic literature review and bibliometric analysis. *Multidisciplinary Reviews*, *9*(5), Article e2026228. https://doi.org/10.31893/multirev.2026228

Suherman, S., & Vidákovich, T. (2022). Assessment of mathematical creative thinking: A systematic review. *Thinking Skills and Creativity*, *44*, Article 101019. https://doi.org/10.1016/j.tsc.2022.101019

Sung, E., & Kelley, T. R. (2023). Elementary students' engineering design process: How young students solve engineering problems. *International Journal of Science and Mathematics Education*, *21*(5), 1615–1638. https://doi.org/10.1007/s10763-022-10317-y

Supovitz, J. A., Ebby, C. B., & Collins, G. (2024). Pathways to performance: The experimental impacts of learning trajectory-oriented formative assessment in mathematics. *American Journal of Education*, *130*(4), 621–651. https://doi.org/10.1086/730994

Susanna, T., Lasse, E., Henrik, N. J., & Sari, H.-N. (2026). From problem-solving to reflection: Activating diverse metacognitive skills in mathematics. *International Journal of Science and Mathematics Education*, *24*(2), 10. https://doi.org/10.1007/s10763-025-10643-x

Suseelan, M., Chew, C., & Chin, H. (2022). Research on mathematics problem solving in elementary education conducted from 1969 to 2021: A bibliometric review. *International Journal of Education in Mathematics, Science and Technology*, *10*(4), 1003–1029. https://doi.org/10.46328/ijemst.2198

Suto, I. (2013). 21st century skills: Ancient, ubiquitous, enigmatic? *Research Matters*, *15*. https://doi.org/10.17863/CAM.100282

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, *148*, Article 103798. https://doi.org/10.1016/j.compedu.2019.103798

Tank, K. M., Moore, T. J., Dorie, B. L., Gajdzik, E., Terri Sanger, M., Rynearson, A. M., & Mann, E. F. (2018). Engineering in early elementary classrooms through the integration of high-quality literature, design, and STEM+C content. In L. English & T. Moore (Eds.), *Early engineering learning* (pp. 175–201). Springer. https://doi.org/10.1007/978-981-10-8621-2_9

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research*, *27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Thornberg, R., & Charmaz, K. (2014). Grounded theory and theoretical coding. In U. Flick (Ed.), *Grounded theory and theoretical coding* (pp. 153–169). SAGE Publications. https://doi.org/10.4135/9781446282243.n11

Thunder, K., Almrode, J., & Hattie, J. (2022). *Visible learning in early childhood*. Corwin.

Torrance, E. P. (1974). *Torrance tests of creative thinking: Directions manual and scoring guide*. Personnel Press.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, *27*(5), 615–631. https://doi.org/10.1080/01411920120095780

Toulmin, S. E. (2003). *The uses of argument* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Treffinger, D. J. (1995). Creative problem solving: Overview and educational implications. *Educational Psychology Review*, *7*(3), 301–312. http://www.jstor.org/stable/23359352

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., . . . Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of internal medicine*, *169*(7), 467–473. https://doi.org/10.7326/m18-0850

Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, *26*(2), 265–283. https://doi.org/10.1007/s10648-013-9243-1

Tsamir, P., Tirosh, D., Tabach, M., & Levenson, E. (2010). Multiple solution methods and multiple outcomes: Is it a task for kindergarten children? *Educational Studies in Mathematics*, *73*(3), 217–231. https://doi.org/10.1007/s10649-009-9215-z

Tschisgale, P., Kubsch, M., Wulff, P., Petersen, S., & Neumann, K. (2025). Exploring the sequential structure of students' physics problem-solving approaches using process mining and sequence analysis. *Physical Review Physics Education Research*, *21*(1), Article 010111. https://doi.org/10.1103/PhysRevPhysEducRes.21.010111

Tyler, R. W. (1934). Techniques for evaluating behavior. *Educational Research Bulletin*, *13*(1), 1–11. http://www.jstor.org/stable/1470737

Tytler, R. (2020). STEM education for the twenty-first century. In J. Anderson & Y. Li (Eds.), *Integrated approaches to stem education: An international perspective* (pp. 21–43). Springer International Publishing. https://doi.org/10.1007/978-3-030-52229-2_3

Uher, J. (2016). What is behaviour? And (when) is language behaviour? A metatheoretical definition. *Journal for the Theory of Social Behaviour*, *46*(4), 475–501. https://doi.org/10.1111/jtsb.12104

Ukobizaba, F., Nizeyimana, G., & Mukuka, A. (2021). Assessment strategies for enhancing students' mathematical problem-solving skills: A review of literature. *EURASIA Journal of Mathematics, Science and Technology Education*, *17*(3), Article em1945. https://doi.org/10.29333/ejmste/9728

Unal, M., & Aral, N. (2014a). Development of the problem solving scale in science education (PSSSE): The reliability and validity study. *Egitim ve Bilim [Education and Science]*, *39*(176), 267–278. https://doi.org/10.15390/EB.2014.3585

Unal, M., & Aral, N. (2014b). An investigation on the effects of experiment based education program on six years olds' problem solving skills. *Egitim ve Bilim [Education and Science]*, *39*(176), 279–291. https://doi.org/10.15390/EB.2014.3592

Unal, M., & Saglam, M. (2018). Examination of the effect of the GEMS program on problem solving and science process skills of 6 years old children. *European Journal of Educational Research*, *7*(3), 567–581. https://doi.org/10.12973/eu-jer.7.3.567

UNESCO Institute for Statistics. (2012). *International standard classification of education (ISCED) 2011*. https://unesdoc.unesco.org/ark:/48223/pf0000219109

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa - A comparison of basic properties. *MethodsX*, *10*, Article 102212. https://doi.org/10.1016/j.mex.2023.102212

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Van Eck, N. J., & Waltman, L. (2024, Januari 24). An open approach for classifying research publications. *Leiden Madtrics*. https://doi.org/10.59350/qc0px-76778

Van Elsen, J. (2025, September 23). Zo leer je kinderen zelf hun problemen oplossen [How to help children solve their own problems?]. *EOS Blogs*. https://www.eoswetenschap.eu/psyche-brein/zo-leer-je-kinderen-zelf-hun-problemen-oplossen

Van Elsen, J., Buckers, L., Van Tricht, C., Torbeyns, J., & De Maeyer, S. (2025). PROSPER: A comprehensive, valid, and reliable instrument to observe problem-solving behaviours in preschoolers. *Thinking Skills and Creativity*, *58*, Article 101940. https://doi.org/10.1016/j.tsc.2025.101940

Van Elsen, J., Catrysse, L., & De Maeyer, S. (2024). The effect of interactive picturebook reading on problem-solving skills in preschool: A quasi-experiment. *Early Childhood Education Journal*, *52*(7), 1471–1485. https://doi.org/10.1007/s10643-023-01542-3

van Merriënboer, J. J. G. (2013). Perspectives on problem solving and instruction. *Computers & Education*, *64*, 153–160. https://doi.org/10.1016/j.compedu.2012.11.025

van Raan, A. (2019). Measuring science: Basic principles and application of advanced bibliometrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 237–280). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_10

Varghese, V. V. V., & Renumol, V. G. (2021, February 11–12). *Assessment methods and interventions to develop computational thinking: A literature review* [Paper presentation]. 2021 International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India. https://doi.org/10.1109/ICITIIT51526.2021.9399606

Varier, D., & Yun, S. (2023). A review of research on assessment use: implications for preparing early childhood educators for data-informed decision-making. *Journal of Early Childhood Teacher Education*, *44*(4), 1067–1092. https://doi.org/10.1080/10901027.2023.2257157

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.004

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: a survey. *ZDM – Mathematics Education*, *52*(1), 1–16. https://doi.org/10.1007/s11858-020-01130-4

Vessonen, T., Hellstrand, H., Aunio, P., & Laine, A. (2024). Individual differences in mathematical problem-solving skills among 3- to 5-year-old preschoolers. *International Journal of Early Childhood*, *56*, 339–357. https://doi.org/10.1007/s13158-023-00361-2

Vlachou, A., & Stavroussi, P. (2016). Promoting social inclusion: A structured intervention for enhancing interpersonal problem-solving skills in children with mild intellectual disabilities. *Support for Learning*, *31*(1), 27–45. https://doi.org/10.1111/1467-9604.12112

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing Students in Classroom Settings: A Review of Seven Coding Schemes. *School Psychology Review*, *34*(4), 454–474. https://doi.org/10.1080/02796015.2005.12088009

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Harvard University Press.

Vygotsky, L. S. (2017). The problem of teaching and mental development at school age [Problema obuchenija i umstvennogo razvitija v shkol’nom vozraste]. *Changing English*, *24*(4), 359–371. https://doi.org/10.1080/1358684X.2017.1395165 (Original work published 1935)

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Wan, Z. H., Jiang, Y., & Zhan, Y. (2021). STEM education in early childhood: A review of empirical studies. *Early Education and Development*, *32*(7), 940–962. https://doi.org/10.1080/10409289.2020.1814986

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PLOS ONE*, *15*(1), Article e0227742. https://doi.org/10.1371/journal.pone.0227742

Ward, S. D. (1993). *Developmental versus academic mathematics education: effects on problem solving performance and attitudes toward mathematics in kindergarten children [Unpublished doctoral dissertation]*. Vanderbilt University.

Warneken, F., Steinwender, J., Hamann, K., & Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. *Cognitive Development*, *31*, 48–58. https://doi.org/10.1016/j.cogdev.2014.02.003

Weiser, L. E. (2022). Young children's free play in nature: An essential foundation for STEM learning in Germany. In S. D. Tunnicliffe & T. J. Kennedy (Eds.), *Play and STEM education in the early years: International policies and practices* (pp. 85–103). Springer. https://doi.org/10.1007/978-3-030-99830-1_4

Wheatley, C. L., & Wheatley, G. H. (1984). Problem solving in the primary grades. *The Arithmetic Teacher*, *31*(8), 22–25. http://www.jstor.org/stable/41192375

Wicaksono, A. G. C., & Korom, E. (2022). Review of problem-solving measurement: an assessment developed in the Indonesian context. *Participatory Educational Research*, *9*(1), 116–136. https://doi.org/10.17275/per.22.7.9.1

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *43*(4), Article 1686. https://doi.org/10.21105/joss.01686

Wiggins, G. (2011). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, *92*(7), 81–93. https://doi.org/10.1177/003172171109200721

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wilson, J. W., Fernandez, M. L., & Hadaway, N. (1993). Mathematical problem solving. In P. S. Wilson (Ed.), *Research ideas for the classroom: High school mathematics* (pp. 57–77). MacMillan.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Wong, M., Choi, K., Barak, L., Lapidow, E., Austin, J., Shafto, P., & Bonawitz, E. (2024). Young children's directed question asking in preschool classrooms. *Behavioral Sciences*, *14*(9), Article 754. https://doi.org/10.3390/bs14090754

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), Article 61. https://doi.org/10.1186/1471-2288-13-61

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wyse, D., Brown, C., Oliver, S., & Poblete, X. (2021). Education research and educational practice: The qualities of a close relationship. *British Educational Research Journal*, *47*(6), 1466–1489. https://doi.org/10.1002/berj.3626

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219–1227. https://doi.org/10.1037/a0037489

Yalçın, V., & Erden, Ş. (2021). The effect of STEM activities prepared according to the design thinking model on preschool children's creativity and problem-solving skills. *Thinking Skills and Creativity*, *41*, Article 100864. https://doi.org/10.1016/j.tsc.2021.100864

Yelland, N., & Masters, J. (2007). Rethinking scaffolding in the information age. *Computers & Education*, *48*(3), 362–382. https://doi.org/10.1016/j.compedu.2005.01.010

Yuan, Y., Zeng, W., Kloos, H., Brown, R., & Carr, V. (2024). Preschool engineering play on nature playscapes. *Early Childhood Education Journal*, *53*, 2207–2216. https://doi.org/10.1007/s10643-024-01743-4

Yuhana, Y., & Fajari, L. E. W. (2025). Cross-disciplinary effects of problem-based learning on problem-solving skills in mathematics and beyond: A comprehensive meta-analysis and bibliometric. *EURASIA Journal of Mathematics, Science and Technology Education*, *21*(11), Article em2732. https://doi.org/10.29333/ejmste/17345

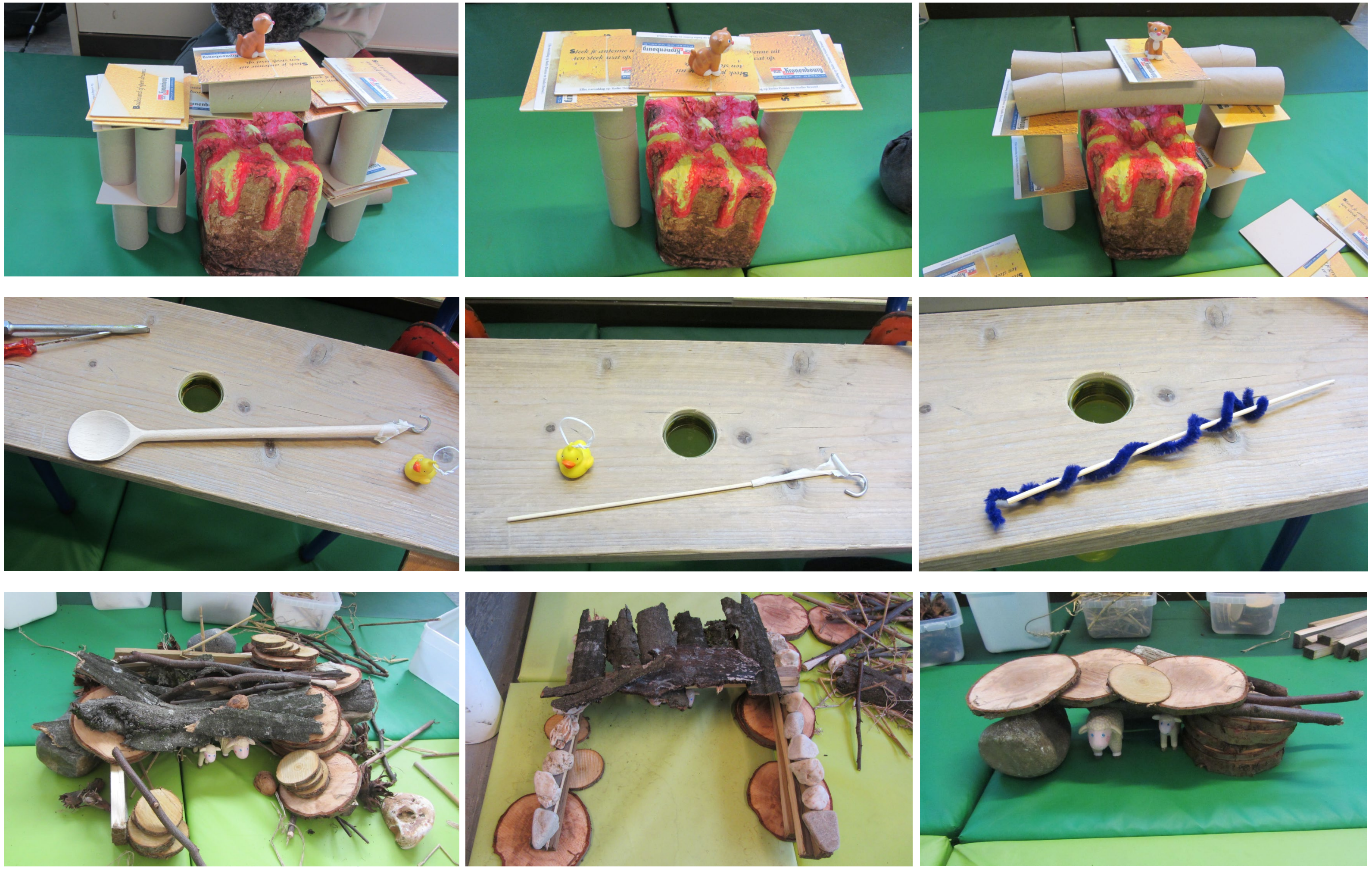
Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology*, *1*(2), 198–226. https://doi.org/10.1037/1089-2680.1.2.198

Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 233–262). Cambridge University Press. https://doi.org/10.1017/CBO9780511615771.009

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 299–315). Routledge.

APPENDICES

Appendix 1: Solutions engineering design problems (Study 3)



Appendix 2: The Problem-solving Observation Scheme for Playful Engineering and the Real-world (PROSPER)

Category	Behaviour	Description	Examples of identified behaviours (“The child …”)
Exploring		Deliberately gathering information about the problem, materials, and possibilities by observation and interaction with the problem and environment	
	Observing	Looks at the available materials or inspects the problem setup	Takes a closer look at the set-up and the duck in the bottle.
	Naming	Identifies materials by naming them	Points to the materials as she names them.
	Describing	Specifies characteristics or functions of materials or tools	Describes the volcano and what it is made of.
	Comparing	Examines or groups materials based on similarities or differences	Compares the sheep as he holds them: “this one is a baby, and this one is bigger.”
	Associating	Articulates links between objects or situations and other objects or situations	Says: “What is this? The cow likes it. It is from the farm,” as he looks at the straw.
	Manipulating	Handles objects by touching them, moving them and/or demonstrating which actions can be performed with them	Takes the ballpoint pen out of the toolbox, presses the button, finds that it does not work and says: “it’s off.”
	Searching	Rummages through the materials or looks around in search for things that can potentially be used	Rummages in the toolbox in search for what he can use.
	Asking information	Asks questions about the available materials, the problem, or the criteria	Takes a screw from the toolbox and asks: “What would this be? I don't know what this is”.
Understanding		Demonstrating understanding of the problem, its key elements, constraints and success criteria	
	Listening	Pays attention to the problem description or clarifications of the goal and success criteria	Listens and watches carefully at first, but wants to start before Mr Rabbit has finished explaining the problem and criteria.

Category	Behaviour	Description	Examples of identified behaviours (“The child …”)
	Confirming	Shows understanding by nodding or verbal expression	Nods to confirm when Mr Rabbit asks if she can make a bridge.
	Repeating	Articulates or demonstrates the problem, goal state, or success criteria	Demonstrates what the kitten wants to do by going over the volcano with the kitten in the air.
	Complying	Acts in accordance with the goal or success criteria	Continues to place items on and against the volcano, even after repeated reminders.
Representing		Creating a mental, verbal, visual and/or physical representation of the problem and/or a goal situation	
Formulating		Coming up with possible ways to solve the problem and evaluating alternatives	
	Thinking up	Articulates or shows a (new) idea or (other) possible solution	Comes up with different structures: 2 rolls on top of each other, 2 rolls next to each other, a ramp, a staircase, a third pillar...
	Considering	Evaluates alternative options or substantiates the choice for a specific solution or strategy	Holds a [popsicle] stick to a pipe cleaner and says: “Then we can stick these sticks together like this. But I can't paste this (...) because we don't have any glue.”
	Asking ideas	Aks for help in coming up with ideas	Turns to the supervisor and asks: “How did the others do it? Aren't you allowed to tell? But I want you to tell me.”
Planning		Determining (sub)goals, action plans, intermediate steps, and next steps	
	Preparing	Prepares work by collecting and arranging materials, tidying up ...	Says: “Now I'm going to take this big stone,” while taking the biggest stones from the box and putting them next to it.

Category	Behaviour	Description	Examples of identified behaviours (“The child …”)
	Thinking ahead	Anticipates next step(s) or event(s)	Says: “First, I must make a wall. And then with a stick ...”
Executing		Carrying out a plan, working towards a solution or trying to do so	
	Doing	Works towards a solution through physical actions	Works on through trial and error, gradually improving the construction.
	Refining	Finishes the result with care by adding extra elements or details	Turns all cards white side up throughout the process.
	Asking motor help	Asks for specific, motor assistance	Fails to cut the tape and asks: "Can you hold this for a moment?” while handing the roll to the supervisor.
Monitoring		Monitoring progress, identifying problems, interim check of success criteria, and adjusting plans or actions if necessary	
	Testing	Checks if the success criteria are being met	Tests with hairdryer and says: “Here comes the storm again.”
	Measuring	Fits, measures, weighs, counts things to check whether they are the correct size, height, weight ...	Measures the height of the stacked wooden discs against the stones and the sheep with his hand and says: “This must be as high as the stones (…) Must be above the sheep.”
	Determining	Indicates whether something is working or not, or if there is a problem	After accidently dropping a spoon into the bottle, he says: “First I have to get the spoon out.”
	Modifying	Adjusts elements of the solution or own actions to improve it	Reinforces the construction by adding a roll at each level.
	Declaring	Indicates that a solution has been reached	Fishes the duckling out of the bottle, shouts "Yeah!!!" and immediately calls Mr Rabbit.

Category	Behaviour	Description	Examples of identified behaviours ("The child …")
	Asking confirmation	Asks for approval or agreement from the supervisor	Looks at the supervisor and asks: "Should it be like this?"
Reflecting		Looking back at the problem and process and critically evaluating the outcome to learn from it	
	Recalling	Describes or demonstrates key elements from the previous assignment	Tells upon request what she had to do in the previous task, who she had to do it for, and names some materials she was allowed to use.
	Evaluating solution	Describes or demonstrates the solution, demonstrates how it works, appraises the solution based on the criteria or suggests possible improvements	Spontaneously demonstrates to Mr Rabbit that the stable can withstand the storm by aiming the hairdryer at it.
	Evaluating process	Describes or demonstrates trials, concrete events, situations, difficulties or emotions during the problem-solving process	Recalls several failed trials and says the pencil was too short and the spoon had fallen into the bottle.
Generic		Overarching problem-solving behaviours	
	Commitment	Is willing to solve the problem and is focused on the task	Starts playing/throwing the kitten.
	Perseverance	Keeps going and does not give up when faced with setbacks or impasse	Works patiently despite repeated collapses.
	Autonomy	Works independently with only limited need for support	Regularly asks for help, then waits; does not come up with a good plan of action himself.
	Responsiveness	Adjusts actions based on tips and suggestions from the supervisor	Does not initially follow the supervisor's suggestions. Follows only when the supervisor gives very specific instructions and helps physically.

Appendix 3: OEPS observation form (translated)

KidID ____________________

Date ____________________

Obs nr.

Problem ❑ Racetrack ❑ Volcano ❑ Duck

OEPS

Observing Early Problem Solving

Score	1	2	3	4	5
The child does / shows this ...	not at all, hardly, with difficulty, under guidance, after prompting	*in between*	occasionally, appropriately, spontaneously, with limited guidance or support	*in between*	continuously, when necessary, fluently, thoroughly, on own initiative, independently

Indicators	The child ...	1	2	3	4	5	Comments
Exploring	purposefully collects information						
Understanding	understands the problem and shows insight						
Formulating	generates and/or tries different possibilities						
Planning	plans and/or prepares actions						
Executing	acts purposefully towards a solution						
Monitoring	follows up on progress						
Reflecting	evaluates the process and the result						

Generic	The child ...	1	2	3	4	5	Comments
Committment	shows engagement and remains focused on the task						
Perseverance	persists and does not give up when facing setbacks						
Autonomy	works independently or with minimal support						
Responsivity	adjusts actions based on tips and suggestions						
Collaboration	shows willingness to cooperate						

General comments	Time	
	Start	
	Stop	
	Duration	minutes

Appendix 4: Engineering design problems (Study 4)

List of materials

Building materials	Dimensions	Quantity	Notes
Cardboard	5 x 20 cm	6	Sturdy cardboard
	2.5 x 20 cm	6	Sturdy cardboard
	10 x 10 cm	6	Sturdy cardboard
	5 x 10 cm	6	Sturdy cardboard
Paper	A4	2	Normal white or coloured
	A5	2	Normal white or coloured
Toilet rolls	Ø 4.5 x 9.5 cm	12	
Cups	ca. Ø 8 x 8 cm	4	Paper/cardboard or reusable plastic
Yogurt cups	Small	4	Empty, disposable yogurt cups
Skewers	ca. 25 cm	6	
Toothpicks	ca. 6.5 cm	6	
Popsicle sticks	ca. 11 cm	12	
Straws	ca. 15–21 cm	6	Paper/cardboard
Fasteners	**Dimensions**	**Quantity**	**Notes**
Adhesive tape		1	Painter's tape can be easily torn off
Pipe cleaners	ca. 30 cm	6	In different colours
Rubber band	ca. Ø 8 cm	6	Ordinary, small rubber bands
Paper clips		6	
Clothespins		6	
Hooks	Ø eye 1.5–2 cm	4	Use screw hooks or make your own hooks
Nails and screws	Short and long	4	Blunt the points slightly
Ball of modelling clay	Ø 2 cm	1	Place the plasticine or clay in a jar
Magnets		6	Small, strong magnets (OPTIONAL)
Tools	**Dimensions**	**Quantity**	
Scissors		1	
Ruler	30 cm	1	
Coloured pencils	Big and small	6	
Screwdrivers	Big and small	2	
Children's forks	Big and small	2	
Children's spoons	Big and small	2	
Wooden spoons	Big and small	2	

The Volcano Bridge Problem

Description

Build a bridge over the volcano so that the cat can safely get to the other side.

Criteria

- The cat can cross the bridge from one side to the other.
- The bridge remains stable while the cat stands at its centre.
- No part of the bridge comes into contact with the volcano.

Materials

- Volcano
- Box of material (see list of materials)

IMPORTANT: Remove the adhesive tape.

Make a volcano out of a cardboard box by painting it (or having it painted by the children), by placing a red cloth on it, or make one yourself using papier-mâché. The dimensions of the volcano are:

- Width: ca. 16 cm (14–20 cm)
- Height: ca. 16 cm (16–20 cm)
- Length: doesn't matter much

Setup

Minimum space requirement: 2 x 2 meters.

Place the volcano in the centre of the workplace. Place the box of materials at the edge of the workspace.


Extra challenge

- Build a bridge with other materials.
- Make a bridge without skewers / ruler / …

The Hook-a-Duck Problem

Description

Make something (a tool) to free the duckling from the well (a bottle).

Criteria

- You can free the duckling with the tool.
- The tool is easy to use.
- The tool is robust enough to be used several times.

Materials

- Bottle (neck Ø ca. 4 cm, height ca. 20–25 cm)
- Ducks or another object with a ring (Ø ring ca. 2 cm)
- Box of material (see material list)

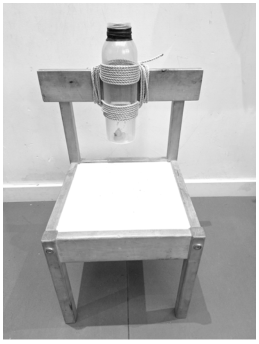
Look for a drinking bottle with an opening of ca. Ø 4 cm and a height of ca. 23 cm.

Make a ring of about Ø 2 cm on a duck or other toy or object.

Setup

Minimum space requirement: 2 x 2 meters.

Fix a drinking bottle so that it cannot be moved. This can be done with rope, with strong tape or in any other way. Place the ringed duckling in the bottle. Place a second duckling (test duck) next to it. Place the box with material at the edge of the workplace.


Extra challenge

- Make a tool with other materials.
- Drop a piece of Styrofoam of ca. Ø 2–3 cm into the bottle and ask the preschoolers to take it out. You can also use a scoop of plasticine, 'filling chips' or PlayMais (or anything else you can poke into to get it out of the bottle).

The Racetrack Problem

Description

Create a racetrack that allows cars to get from the mountain (chair) to the parking space.

Criteria

- The racetrack starts on the seat.
- The cars do not fall off the racetrack or tip over.
- The cars stop on the parking space.

Materials

- Low chair or bench (31–35 cm high)
- A-3 parking symbol

For this assignment, the preschoolers can use any materials available in the classroom. No materials need to be provided. The racetrack can be made of cardboard, shelves, tubes, books, etc.

Setup

Minimum space requirement: 2 x 3 meters.

Place a low chair or bench (31–35 cm high) against a wall or cupboard. This is the starting point of the racetrack.

Place an A-3 sheet (or two A-4 sheets taped together) with a large, white letter P on a blue background or an image of a parking sign about 1.5 meters away from the chair or bench. Secure the sheet to the floor with adhesive tape.

Make sure that there is at least 20 cm. free space behind the sheet.

Extra challenge

- Make the racetrack sturdier/safer.
- Create a racetrack from a (children's) table.
- Create a racetrack for larger cars.

Some solutions

AUTHOR CONTRIBUTIONS (CREDIT)

Study 1 – Mapping the landscape of research on early problem solving: A bibliometric review

Joris Van Elsen	Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Project administration, Funding acquisition.
Lieven Verschaffel	Conceptualization, Methodology, Writing - review & editing, Supervision.
Sven De Maeyer	Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Funding acquisition.
Joke Torbeyns	Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Funding acquisition.

Study 2 – Methods for assessing problem-solving competencies in preschoolers: A scoping review

Joris Van Elsen	Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data curation, Writing – Original Draft, Visualization, Project administration, Funding Acquisition.
Sven De Maeyer	Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision, Funding Acquisition.
Joke Torbeyns	Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision, Funding Acquisition.

Study 3 – PROSPER: A comprehensive, valid, and reliable instrument to observe problem-solving behaviours in preschoolers

Joris Van Elsen	Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Visualisation, Project administration, Funding acquisition
Lisanne Buckers	Validation, Formal Analysis
Charlotte Van Tricht	Validation
Joke Torbeyns	Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision, Funding acquisition
Sven De Maeyer	Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision, Funding acquisition

Study 4 – An argument-based approach to validation of formative problem-solving assessment in preschool

Joris Van Elsen	Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Project administration, Funding acquisition.
Sven De Maeyer	Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition.
Joke Torbeyns	Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition.

Dissertation submitted to obtain the joint degree of

Doctor of Education Sciences | University of Antwerp
Doctor of Educational Sciences | KU Leuven

Supervisors:

Prof. dr. Sven De Maeyer | University of Antwerp
Prof. dr. Joke Torbeyns | KU Leuven

Antwerp, 26 August 2026

University of Antwerp
Faculty of Social Sciences
Training and Education Sciences

KU LEUVEN
Faculty of Psychology and Educational Sciences
Centre for Instructional Psychology and Technology

fwo
Fonds Wetenschappelijk Onderzoek
Research Foundation – Flanders
[Grant Number 1S03525N]