COGNITIVE SYNERGY: EXPLORING THE TRANSFORMATIVE INTERSECTION OF HUMAN INTELLIGENCE AND ARTIFICIAL INTELLIGENCE IN DESIGNING EQUITABLE NEXT GENERATION SCIENCE ASSESSMENTS

By

Tingting Li

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Educational Psychology and Educational Technology – Doctor of Philosophy

2024

ABSTRACT

This study explores the intersection of human intelligence and Artificial Intelligence (AI) to design knowledge-in-use science assessments for supporting students' deep science learning. In the context of evolving educational paradigms, it seeks to harness AI tools (GPT), to enhance knowledge-in-use assessment design, ensuring equitable opportunities for diverse learners. Anchored in the Next Generation Science Assessment and an evidence-centered design, this study aspires to harmonize AI's computational strengths with human expertise in assessment design. Drawing from an array of theoretical frameworks—Hybrid Intelligence System, Distributed Cognition, and Self-Regulated Learning Theory—the study underscores the multi-faceted and dynamic nature of knowledge-in-use and the symbiotic integration of human and AI.

Employing a Design-Based Research approach, the study proceeds in three stages: (1) Iteratively training GPT models for effective designing knowledge-in-use assessments; (2) Gathering multidisciplinary expert feedback on AI-co-designed assessments; and (3) developing a domain-specific GPT-model for tailored assessment design that capture knowledge-in-use and address diverse student needs. Diverse data analysis techniques, encompassing thematic analysis, and descriptive statistics, such as heat map and scatter plot, are leveraged. Anticipated results spotlight an exploratory GPT model adept at creating tailored assessments resonating with diverse learning needs while emphasizing equity, adaptability, and inclusivity. This study holds the potential to significantly enhance the educational landscape by advocating a balanced approach where AI complements human expertise, paving the way for a progressive and inclusive future in education.

ACKNOWLEDGEMENTS

It is hard to believe that I am writing what I consider the most daunting section of my dissertation—the acknowledgment. Reflecting back to August 31, 2018, when I first ventured into this completely new world beyond my imagination, I see how different this moment feels compared to when I wrote my first dissertation acknowledgment. Now, I am calmer and deeply appreciative of everyone who has supported me on this incredible journey.

First and foremost, I am profoundly grateful to my co-advisors, Dr. Joseph Krajcik and Dr. Rand Spiro. I often reflect on my fortune to have had the opportunity to learn from such wise, experienced advisors. You have mentored me not only in the art of research and scholarship but also in becoming a more thoughtful human being. Your open-mindedness, dedication, diligence, and support have profoundly shaped me. I am eager to pass on your spirited dedication to my students and will always hold dear your fervent enthusiasm for education—immersing in theory, bridging it with practice, consistently focusing on students, and pondering the profound implications of educational research. I aspire to embody the ideals of being a lifelong learner, collaborator, and supporter. I am committed to living up to your expectations.

I extend my heartfelt thanks to my committee members: Dr. Jennifer Schmidt from the Educational Psychology and Educational Technology program; Dr. Jiliang Tang from the Computer Science Department; and Dr. Kevin Haudek from the Natural Science Department. Your expertise has not only enriched my dissertation but also my personal and professional growth. I am particularly indebted to Jennifer for your enduring guidance and unwavering support throughout this journey.

Special thanks go to Dr. Christine Greenhow for her guidance, and to Drs. Emily Adah Miller, Christina Schwarz, Amelia Gotwals, and David Stroupe for supporting both Peng's and my academic and professional paths.

To my CREATE for STEM colleagues—Drs. Bob Geier, Namsoo Shin, Cory Miller, Consuelo Morales, Emil Edin, Selin Akgun, and Jonathan Bowers—and to Ligita, Mary, Sue, Renee, Colter, Angie, and Alison: thank you for being part of the significant milestones in our lives. To my EPET friends—

Samuela, Anne Drew, Larissa, and John Kean—thank you for the mutual support and camaraderie. A special shout-out to Sharon Hammond, who has always been ready to assist whenever I had questions about the program or my progress.

I am grateful to the mentors and colleagues with whom I've collaborated on various research projects (Dr. Barbara Schneider, Dr. I-Chien Chen, Debroah Peek-Brown, Sue Codere, Kayla Bartz, and Lydia Bradford). My appreciation also extends to the teachers, district coordinators, and school principals I've worked with. Your insights have been invaluable. A special acknowledgment to the expert panels who reviewed the products of my dissertation—without your thoughtful feedback, this work would not have been possible.

Lastly, I owe a profound debt of gratitude to my family. To my parents, Xifeng Li and Xihong Yu, thank you for your encouragement and support in pursuing my dreams. To my elder sisters, Haiyan Li and Mengmeng Li, thank you for caring for our parents, allowing me to focus on my studies. To my husband, Dr. Peng He, thank you for your encouragement, solid support, and the powerful backing, helping me to trust and surpass myself. Thank you for being an incredible partner and father, filling our lives with love, care, and adventure. And to my daughter, Jinni (何祎然）, thank you for being a wonderful source of inspiration and joy—I wish you all the happiness and health in the world and hope you pursue whatever dreams you have and become the person you aspire to be. Lastly, I thank myself for remaining persistent and strong. Go Green!

Thank you, everyone! I wish you all the very best!

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Rationale

<u>1.1.1 The Evolution of Educational Assessments in Science Education</u>

Historically, educators and researchers largely perceived assessments as static milestones marking students' progress (Bloom, 1968; Sadler, 1989). Recent pedagogical advancements, such as supporting students' higher-order thinking like problem-solving skills (Kang et al., 2014; Pellegrino & Hilton, 2012) or transferable knowledge (Shepard et al., 2019), have transformed these evaluations from mere benchmarks to dynamic tools that actively support instructional practices. Educators need to develop localized assessments that emphasize "assessment for learning" rather than mere evaluation (Black & Wiliam, 1998; Shepard, 2000; Stiggins, 2014). This transformative perspective positions assessment as an ongoing dialogue within the classroom, steering pedagogical strategies (DiCerbo, 2020; Wiggins, 1998).

In science education, the notion of knowledge-in-use is central to this transformation. It refers to students' ability to apply their acquired knowledge to real-world scenarios or complex problems (Li et al., 2024; NGSS Lead States, 2013; NRC, 2012). To develop knowledge-in-use, students should actively engage with the three dimensions of scientific knowledge (3D learning): disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs) to make sense of compelling phenomena and design solutions to challenging problems (NGSS Lead States, 2013; NRC, 2012). This ambitious vision for science education is emphasized in foundational documents such as A Framework for K-12 Science Education (NRC, 2012) and the Next Generation Science Standards (NGSS Lead States, 2013). Subsequent policy documents, including Science and Engineering for Grades 6-12: Investigation and Design at the Center (National Academies of Sciences, Engineering, and Medicine, 2019) and Science and Engineering in Preschool Through Elementary Grades: The Brilliance of Children and the Strengths of Educators (National Academies of Sciences, Engineering, and Medicine, 2022), also focus on developing students' knowledge-in-use through 3D learning.

The transition to knowledge-in-use raises a vital challenge for the science education community: how can we effectively collect evidence to understand if students have developed knowledge-in-use?

(Furtak, 2017, 2023; Pellegrino, 2013; Penuel & Smolek, 2019). Addressing this challenge requires exploring the development of suitable assessment tasks that can capture students' knowledge-in-use (Li et al., 2024; NGSS Lead States, 2013). Such assessments prioritize learners' ability to leverage acquired knowledge in real-world scenarios, which, while vital, presents challenges in design and implementation (Bertenthal & Wilson, 2006; NGSS Lead States, 2013). Moreover, students must find these assessments sufficiently compelling and engaging to motivate them in their learning processes (Li, He, & Peng, 2023; Stiggins, 2014). Additionally, the notion of "assessment for learning" emphasizes the important role of formative assessment in supporting students' learning (Li et al., 2024).

Formative knowledge-in-use assessments are essential to effective science instruction (Harris et al., 2019). These high-quality assessments, which align with NGSS standards, offer crucial formative insights for educators (NRC, 2014; Shepard et al., 2018). They depict the progression of students' learning over time. Teachers often need to develop assessment tasks for their students based on their instruction and students' needs (Heritage, 2010). Yet, many teachers do not feel prepared to develop NGSS-aligned assessments or use them formatively (Furtak, 2017). Due to the complex and varied nature of local classrooms, science teachers need the capability to intentionally design assessment tasks that align with the NGSS and are easily integrable into their real-time, interactive classroom activities (Pellegrino, 2013). Designing these assessments demands adaptability and inclusivity for a diverse spectrum of learners, especially from minoritized and marginalized racial and ethnic groups (Darling-Hammond & Snyder, 2000). The process of developing such assessments requires collaboration between assessment experts, science education experts, and teachers, which is time-consuming and labor-intensive (Furtak & Lee, 2023). Moreover, it necessitates professional knowledge of assessment design, science content, teaching, and student knowledge (Brookhart, 2010).

1.1.2 The Confluence of Human and Artificial Intelligence in Assessment Design

Teachers need support to transition from assessment for evaluation to assessment for learning (Li et al., 2024). To provide feedback to support student learning, teachers need to design effective assessments that attend to different needs of learners and that provide evidence of student learning (Harris et al., 2019; Hattie & Timperley, 2007; Shute, 2008). Customization is essential to align with the diverse backgrounds and dynamic classroom scenarios, adding to the complexity and time demands on teachers (Brookhart, 2010; Darling-Hammond et al., 2020). This new vision brings both enlightenment and challenges, such as design intricacies, adaptability concerns, and the overarching quest for equity, especially for learners from minoritized and marginalized groups navigating varied pathways (Furtak & Lee, 2023).

Amid the ever-evolving educational landscape, this pressing situation emerges, compelling us to reckon with the promise and pitfalls of technological integration in academic evaluation (Luckin et al., 2017). The solution may lie at the intersection of human intelligence and artificial intelligence. This exploration is driven by technological innovation and a profound aspiration for equity. AI tools have the potential to democratize the process of knowledge-in-use assessments, making them more accessible and equitable for all students (Luckin et al.,2017). Emergent generative AI technologies, notably Large Language Models (LLMs) by OpenAI, offer a glimpse of this potential (Greengard, 2022). Yet, to fully harness AI's potential, educators need a foundational understanding of machine intelligence’s underlying principles and professional expertise in corresponding fields (Ifenthaler et al., 2024; Williams, 2023; Zawacki-Richter et al., 2019). Without adequate expertise to evaluate the outputs of tools like ChatGPT, there's a risk of misguided decisions, which could further erode trust in AI. Harmoniously melding human expertise with AI's capabilities is key.

A pressing concern is that many educators, especially those working on knowledge-in-use assessments, are increasingly relying on AI without fully grasping its nuances (Brown et al., 2020). Bridging this gap calls for innovations like domain-specific algorithms designed with educational paradigms in mind. Such tools can guide educators in integrating the strengths of AI with human insights for assessment design (Khosravi et al., 2022; Owan et al., 2023). This collaboration can redefine the role of assessments, ensuring they remain steadfast guiding lights in the learning voyage (Pellegrino, Chudowsky, & Glaser, 2001; Nguyen et al., 2021). This emphasis on human-machine collaboration in education reflects broader shifts in the AI era. As we delve deeper, technology's role in education becomes increasingly intricate, moving beyond basic computer-aided lessons to sophisticated, intelligent educational systems (Miller, 2023; Reiser, 2001 a, b). While this integration brings opportunities, it also demands a balance between machine-driven innovations and human-centric pedagogy (Halverson & Collins, 2009; Roberts, 2021). Using generative AI exemplifies the promise and potential pitfalls of this alliance, especially in the domain of knowledge-in-use assessments (Greengard, 2022).

This study stands at the crossroads of human intelligence and AI (Greengard, 2022; Johnson et al., 2022). To harness this potential, a balanced approach is essential: educators must synergize their diverse expertise with AI's prowess, ensuring the tools developed are rooted in human values, bias-free, and tailored to the diverse needs of education. Prioritizing collective human intelligence, this study integrates insights from experts across various disciplines, promoting a well-rounded approach to AI's role in education. Central to this endeavor is the belief that AI should augment, not replace, human expertise.

The heart of this exploration lies in the development and validation of knowledge-in-use assessments. Beyond mere evaluation, these assessments are pivotal for enhancing student learning. A core tenet of this study posits that these assessments should be dynamic, allowing educators the autonomy to design, adapt, and align them with their students' unique needs, fostering equity, and championing culturally relevant teaching. Equity and inclusion are critical elements of this research that weave throughout the research and development process, beginning with the initial domain analysis of performance-based learning goals and continuing through the development of tasks and rubrics, recruitment of teacher participants from diverse classroom settings for broad access and participation, and data analyses for validation.

This study aims to chart a course where human expertise and AI innovation converge, offering transformative insights for educational assessment. By delving into the nuances of AI-human collaboration (Dellermann et al, 2021; Fui-Hoon Nah et al., 2023; Johnson et al., 2022), I aim to refine the discourse on knowledge-in-use assessments. The goal is to seamlessly meld technology with methods that prioritize human values, adaptability, and equity. Key objectives include iteratively training a large language model for effective design of knowledge-in-use assessments, gathering multidisciplinary expert feedback on AI-generated assessments, and exploring how to incorporate collective human experts' intelligence to develop a domain-specific algorithm encompassing refined AI processes for tailored assessment design. At its core, this research seeks to foster a symbiotic relationship between AI and human agency, particularly in the realm of developing knowledge-in-use assessments. It endeavors to ensure that AI tools amplify human capabilities rather than replace them. The outcome will be a comprehensive guide to the potential, challenges, and effective practices for integrating AI into educational assessments.

Drawing from the theory of distributed intelligence, I envision a harmonious integration of AI and human intelligence in educational settings (Pea, 1993; Salomon, 1997). Upholding principle like "Human in the Loop" (Mosqueira-Rey et al., 2023), this synergy promises to revolutionize educational assessments. The collaborative dynamic between human and AI in the realm of education is an emerging area of research. As the relationship between AI and education garners attention, it's pivotal to view AI as extensions of human cognitive abilities, not mere adjuncts (Pea, 1993). With the potential to reshape cognitive functions, as postulated by Pea and Kurland (1987), understanding this relationship becomes crucial when aiming to broaden educational assessment horizons. While tools like generative AI offer valuable assistance in assessment design and interpretation, educators must be equipped to harness their capabilities effectively. This research underscores the need for a holistic approach, emphasizing that AI's influence transcends technical aspects, with human agency's emotional, cognitive, and ethical dimensions remaining central (De Cremer & Narayanan, 2023; Sundar, 2020).

This study is to harmonize AI with human experts' knowledge in designing knowledge-in-use assessments. Facing the challenges of designing assessments that reflect nuanced knowledge applications and cater to diverse student needs, this study explores how AI can be integrated with human expertise to yield innovative solutions. The study delves beyond just the technical, emphasizing the ethical implications of ensuring that AI enhances rather than overshadows human agency. The goal is to strike a balance between AI's prowess and human-driven pedagogy, anchoring the research in principles of equitable and meaningful education. Ultimately, this study aims to offer a nuanced, evidence-based framework for envisioning and developing knowledge-in-use assessments in an evolving, AI-influenced educational landscape.

1.2 Research Questions (RQs)

This study explores three major questions:

RQ1. How can generative AI models be effectively and iteratively trained to design knowledge-in-use assessments?

RQ2. How do human experts across different disciplines evaluate the AI-generated knowledge-in-use assessments, and what refinements do they suggest?

RQ3. What is the process of refining AI-designed knowledge-in-use assessments based on the feedback provided by human experts?

CHAPTER 2: LITERATURE REVIEW AND THEORETICAL FRAMEWORK

To better understand the landscape and key construct of this study, I reviewed the relevant studies about "knowledge-in-use", "AI for education", "Human-AI collaboration" and their respective theoretical foundations to set up a better understanding of the current landscape of these research fields and how my study fills the gap of leveraging AI in science education to develop knowledge-in-use assessment. This section also presents a theoretical framework about how humans and the AI machine can collaborate with each other to augment human intelligence in designing knowledge-in-use assessment.

2.1 Meaning of Knowledge-In-Use Proficiency

In the context of modern challenges such as food scarcity, pandemics, and climate change, it is essential for citizens to possess scientific knowledge to make informed decisions, support policy changes, and understand the consequences of inaction (Anderson et al., 2020; NRC, 2012; NRC, 2011; OECD, 2019). To develop a science-literate citizenry, educators need to focus on what students should ultimately know (big ideas) and be able to do (scientific practices), and create learning environments that support this integrated proficiency (NRC, 2012). Consequently, the goals of science education worldwide have shifted towards knowledge-in-use learning objectives (Kulgemeyer & Schecker, 2014; NRC, 2012; People's Republic of China Ministry of Education, 2017). Knowledge-in-use demands that students apply their knowledge by making sense of real-world phenomena, solving complex problems, and making informed decisions (NRC, 2012; NASEM, 2019; Pellegrino & Hilton, 2012). The concept of knowledge-in-use reflects a growing awareness among learning scientists, science educators, and policymakers about the skills required for global citizens in the 21st century (OECD, 2019). It suggests that knowledge is a product of the activities, context, and culture in which it is developed and used (Brown et al., 1989; Bonwell & Eison, 1991), and posits that individuals actively participate in the creation of their own knowledge (Schreiber & Valle, 2013). Instead of merely acquiring knowledge from teachers or textbooks, knowledge-in-use emphasizes the application of scientific knowledge to make sense of natural phenomena and solve complex, authentic problems, as promoted in The Framework (NRC, 2012). This approach allows students to explain new real-world phenomena or solve complex problems by applying their learning (NRC, 2012; NGSS Lead States, 2013).

The development of knowledge-in-use has been a significant focus in cognitive science and science education. From a cognitive science perspective, there is a strong link between knowledge-in-use and adaptive skills (Li et al., 2023; Ward et al., 2018), highlighting the need for cognitive abilities to be adaptive, flexible, and context-sensitive. Knowledge-in-use primarily involves applying previously mastered knowledge and skills to new situations (Bransford & Schwartz, 1999). It is akin to adaptive skills, which emphasize the learning process and the continual adjustment of one's approach in varied contexts (Hatano & Oura, 2003). Unlike transferable knowledge, which depends on contextual similarities for both far and near transfers (Ruiz-Primo et al., 2002), adaptive skills equip learners to handle unknown situations even without directly relevant prior knowledge. This aligns with the goals of knowledge-in-use, such as problem-solving and decision-making, fostering specific adaptive skills like flexibility, resilience, and metacognition (Spiro et al., 2017; Sternberg & Kaufman, 1998).

2.2 Supporting Knowledge-In-Use

One of the most effective strategies for helping learners adapt to novel situations is equipping them with the appropriate knowledge and skills to tackle and solve complex real-life problems (Brown & Duguid, 1993). This can be achieved by designing learning environments that provide authentic activities for novices to experience expert performance, offer scaffolding at crucial moments, support cooperative knowledge building, and include monitoring features (Herrington & Oliver, 1995, 2000). This approach is exemplified when learners apply their scientific understanding to make sense of phenomena or solve intricate problems, encapsulating adaptive skills, transferable knowledge, and cognitive flexibility (Mensah & Chen, 2022; Spiro et al., 2018; Ward et al., 2018). However, developing such proficiency is a gradual process that requires continuous exposure to disciplinary experiences involving open-ended, unresolved problems (Esposito & Bauer, 2017). The *Framework* and the NGSS advocate a three-dimensional (3D) learning approach to explain relevant phenomena and provide solutions to complex problems, proposing performance goals that develop knowledge-in-use across Grades K-12 (NASEM, 2022; NASEM, 2019; NRC, 2012). Despite the recognized value of 3D learning, it presents operational challenges for teachers (Penuel et al., 2015). Teachers must adapt their teaching and assessment practices, conceptualize learning as a trajectory toward generative ideas, and nurture the use of scientific practices.

Situated learning, one of the tactical foundations of 3D learning, posits that knowledge is partly a product of the activities, context, and culture in which it is developed and used (Brown, Collins, & Duguid, 1989).Situated learning theory suggests that learning environments should foster learners' participation in inquiry and support the development of their personal identities as capable and confident learners and knowers. Curricula should be designed to sequence learning activities with attention to students' progress in various disciplinary practices of discourse and representation. Learning activities should focus on meaningful, problematic situations that resonate with students' experiences and show how concepts and methods of subject-matter disciplines are embedded. This also requires the knowledge-in-use assessment to be designed to capture the situated and application nature of 3D learning. Research in the learning sciences (Krajcik et al., 2023; NRC, 2012) has shown that the most effective learning occurs when it is situated in authentic contexts.

2.3 Measuring Knowledge-In-Use

Given the complex cognitive nature of knowledge-in-use, measuring it presents significant challenges. However, understanding students' performance on knowledge-in-use activities and tasks is crucial for effective teaching and learning. Assessment, as a component of any educational system, plays a vital role in diagnosing, monitoring, and promoting students' development of knowledge-in-use in science learning (NRC, 2014). The intricate nature of knowledge-in-use constructs makes assessment design and validation particularly challenging (NRC, 2012; 2014). To address these challenges, the National Research Council's (2014) report, "Developing Assessment for the Next Generation Science Standards," recommended evidence-centered design (ECD) as the cognitive foundation for developing knowledge-in-use assessments (NRC, 2001). Several prominent research groups have made significant efforts to design classroom assessments for knowledge-in-use using principled design approaches (Harris, He et al., under review; Krajcik, & Pellegrino, 2023; Osborne & Wertheim, 2019; Penuel et al., 2019), such as the ECD approach (Mislevy & Haertel, 2006) and the construct-modeling approach (Wilson et al., 2005). For example, the Next Generation Science Assessment (NGSA) project applies a modified ECD design process to articulate a systematic design approach (Harris et al., 2019; 2023). The design and validation of assessments involve ensuring their reliability, validity, and fairness. ECD emphasizes aligning assessment tasks with the knowledge and skills to be measured, ensuring that the assessment provides valid evidence of student learning (Pellegrino, 2014). Validity frameworks, including content validity, construct validity, and criterion-related validity, ensure that assessments accurately measure what they are intended to measure and are fair to all students (Coffey, Black, & Atkin, 2001). Mark Wilson's work on the development and application of assessment frameworks has significantly contributed to this field. Wilson (2005) emphasizes the importance of construct modeling and the BEAR Assessment System, which provides a structured approach to designing, implementing, and validating assessments. In the study, I emphasize assessment for learning to underline the critical role of using assessment formatively to support teaching and learning.

2.4 Designing Knowledge-in-use Assessments

Designing effective knowledge-in-use assessments presents several challenges. One of the main challenges is ensuring that the tasks are accessible and engaging for all students, regardless of their background or prior knowledge. This requires careful consideration of language appropriateness, cultural relevance, and the inclusion of compelling and relatable phenomena (National Research Council, 2012). Additionally, equity considerations must be addressed to ensure that all students have an equal opportunity to demonstrate their understanding and abilities. The National Research Council (2012) highlights the importance of designing assessments that are inclusive and culturally responsive.

2.4.1 Assessment Frameworks and Theoretical Foundations

There are different perspectives on developing formative assessment for knowledge-in-use, including *sociocognitive* and *sociocultural* perspectives. Sociocognitive approaches assess students' understanding and skills as they engage in increasingly sophisticated practices typical of disciplinary experts. This method is grounded in the belief that thinking and learning are inherently social activities, and thus, assessments are based on "local instructional theories" of learning. These theories involve creating sequences of instructional activities tailored to support a specific group of students in developing proficiency (Bakker & Gravemeijer, 2004). In both scenarios, assessment materials are designed for specific content areas, considering the typical challenges students face and common strategies to help them progress. Sociocognitive strategies not only measure students' content knowledge but also aim to help them adopt the dispositions and identities of their field of study. This approach favors assessment practices such as collaborative inquiry, expertly facilitated questioning, discussion, and qualitative feedback, allowing teachers to observe how students are acting, thinking, and reasoning in disciplinary ways (Penuel et al., 2017; Smith et al., 2006).

The strengths of the sociocognitive approach lie in its discipline-specific learning goals and a well-defined learning theory (Penuel & Shepard, 2016). Instead of merely reporting the number of correct answers, this approach reveals how students think about and solve specific problems, providing teachers with valuable insights into students' knowledge, confusions, and immediate learning needs. However, developing such fine-grained, subject-specific assessment tools requires significant expertise and resources, making them less accessible to many schools and districts, especially for certain topics and grade levels. Additionally, these tools focus on assessing students' understanding of specific subject matter without considering their diverse values, experiences, and personal goals. This limitation raises concerns about whether these tools benefit all students equitably, particularly regarding their racial, ethnic, or gender identities, emerging bilingualism, conditions of poverty, and other critical aspects of their lives.

Sociocultural perspectives share many foundational ideas with the sociocognitive approach, such as the social nature of learning and the importance of student engagement in disciplinary practices. However, they differ significantly in how they address student diversity. Sociocultural theories of learning emphasize that students bring valuable knowledge and interests from their personal and community backgrounds into the classroom. Instead of disregarding these experiences, teachers should help students reflect on how the school's ways of knowing, doing, and being relate to those valued in their families and communities (Bang & Medin, 2010). This broad understanding and acceptance are essential for implementing and sustaining equitable assessment practices in diverse educational settings.

2.4.2 Critical Aspects for Designing Knowledge-In-Use Assessment

Designing assessments in science education involves several critical factors to ensure that assessments are effective, equitable, and capable of measuring students' higher-order thinking skills. To support equitable learning and teaching, assessments must be designed with inclusivity and accessibility in mind. This involves developing tasks that are not only challenging and engaging but also accessible to students from diverse backgrounds. Equity considerations include ensuring that the assessment tasks do not disadvantage any group of students and that they provide adequate support for all learners to understand and engage with the phenomena being assessed (Pellegrino & Hilton, 2012). Pellegrino and Hilton (2012) suggest that equitable assessments should accommodate diverse learning styles and provide multiple means of demonstrating understanding. Thus, *equity* is a paramount consideration in assessment design, ensuring that all students, regardless of their background, have equal opportunities to demonstrate their knowledge and skills. Equitable assessments address potential biases and barriers that might disadvantage certain groups of students. This includes considerations for cultural relevance, socioeconomic status, and varying levels of prior knowledge (Pellegrino & Hilton, 2012). *Language* is a critical factor, particularly for English language learners (ELLs). Assessments must ensure that language complexity does not impede students' ability to demonstrate their understanding of scientific concepts. This includes using clear, concise language and providing support such as glossaries or translated materials when necessary (Lee, Quinn, & Valdés, 2013). *Engagement* refers to the extent to which assessment tasks are interesting and relevant to students. Engaging assessments motivate students to perform their best and provide more accurate measures of their abilities. Engaging tasks often involve real-world problems and scenarios that are meaningful to students (Hidi & Renninger, 2006). *Accessibility* ensures that all students, including those with disabilities, can participate meaningfully in the assessment process.

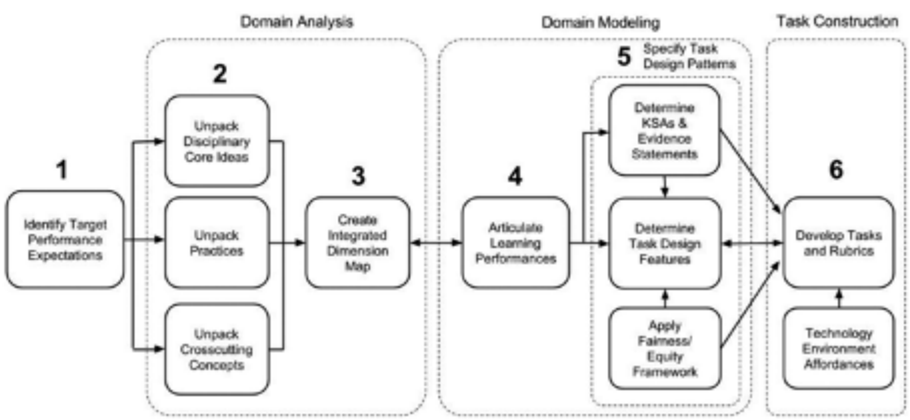
The National Research Council's (2014) report, "Developing Assessment for the Next Generation

Science Standards," recommended ECD as the cognitive foundation for developing knowledge-in-use assessments (NRC, 2001). Several prominent research groups have made significant efforts to design classroom assessments for knowledge-in-use using principled design approach (He et al., under review; Osborne & Wertheim, 2019; Penuel et al., 2019), such as the ECD approach (Mislevy & Haertel, 2006) and the construct-modeling approach (Wilson et al., 2005). For example, the NGSA project applies ECD design principles to articulate a systematic design approach (Harris et al., 2023).

2.4.3 The NGSA Approach to Design Knowledge-In-Use Assessment

In pursuit of the objective to design assessments for learning, given the systematic approach to design knowledge-in-use assessment, the NGSA design process (Harris et al., 2019; Figure 2-1) is employed, initiating with the deconstruction of the NGSS PEs. This process encompasses three primary phases.

Figure 2-1. Overview of the assessment design process (from Harris et al., 2019)



Within the Next Generation Science Standards (NGSS), there are three distinct and equally important dimensions to learning science. These dimensions are combined to form each standard—or performance expectation (PE)—and each dimension works with the other two to help students build a cohesive understanding of science over time. The Domain Analysis phase concentrates on breaking down the broad PEs into manageable components that facilitate the creation of more detailed learning performances. PEs are comprehensive statements outlining the knowledge and skills students should possess at the end of a grade or grade band. They originate from the Framework for K-12 Science Education (enhanced forth, the Framework, National Research Council, 2012), embodying a vast scope. Each PE is structured as a singular statement encapsulating competencies at a large grain size, without delving into the underlying specifics. PEs are inherently three-dimensional, always incorporating a Disciplinary Core Idea (DCI), a Science and Engineering Practice (SEP), and a Crosscutting Concept (CCC) as outlined in the Framework. The elements of each dimension, elaborated further in the NGSS, vary across different PEs within a grade level or band, growing in complexity as students advance through K-12. The integration of the dimensions in PEs depicts the application of DCIs and CCCs through engagement in SEPs for understanding phenomena or solving problems, not meant to be isolated from each other.

I use a middle school physical science PE as an example to explain its' structure and why it needs elaboration. MS-PS1-2: Matter and Its Interactions[1]. The PE is "**Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred.** [Clarification Statement: Examples of reactions could include burning sugar or steel wool, fat reacting with sodium hydroxide, and mixing zinc with hydrogen chloride.] [*Assessment boundary: Assessment is limited to analysis of the following properties: density, melting point, boiling point, solubility, flammability, and odor.*]" MS stands for middle school level. PS refers to "Physical Science," which is one of the four major domains [Life Science (LS), Earth and Space Science (ES), and Engineering] included in the NGSS. The numbers "1-2" indicate that this is the second performance expectation within the first major topic of Physical Science.

A concise PE statement, like MS-PS1-2, encompasses numerous concepts; it is essential not to overlook the intricacies of each dimension within a PE. The apparent simplicity of a PE, such as MS-PS1-2, belies its depth – for instance, the DCI aspect necessitates applying knowledge about substance properties for identifying chemical reactions. The CCC element, while not explicit in the PE statement, is

[1] https://www.nextgenscience.org/pe/ms-ps1-2-matter-and-its-interactions

crucial for pattern recognition and reasoning. The SEP explicitly focuses on data analysis for distinguishing similarities and differences before and after substance interactions. Unpacking these dimensions reveals the extensive range encompassed within a single PE statement, indicating the need for comprehensive and sequential learning experiences and assessments to progress towards achieving these multifaceted PEs. The NGSA design process is instrumental in dissecting and identifying the significant components of PEs suitable for classroom-based assessment, focusing on constructing detailed learning performances from these components.

2.4.4 Evidence-Centered Design in Understanding Student Learning

The process of drawing conclusions from assessments fundamentally relies on evidence-centered design. Pioneered by assessment experts like Robert Mislevy and colleagues (Mislevy, Steinberg & Almond, 2003; Mislevy & Haertel, 2006), ECD prioritizes establishing learning objectives and identifying the necessary evidence to judge student performance against these objectives, subsequently defining task features to elicit this evidence. Central to ECD is the goal of substantiating claims about students' knowledge and abilities with collected evidence, typically manifested through student responses to assessment tasks.

Following its introduction two decades ago, ECD has garnered significant attention in education for its principled approach to assessment design. Notably, post-NGSS release, the National Research Council (2014) advocated for ECD-aligned assessment designs to accurately measure three-dimensional learning. ECD's argumentative reasoning is integral to NGSA's design process, particularly in developing tasks that provide evidence of students' three-dimensional performance and route to meeting PEs.

The NGSA design process, depicted in Figure 2-1, guides the utilization of PEs as the basis for developing three-dimensional assessment tasks for classroom application, enhancing NGSS teaching and learning. The subsequent sections detail each step of the NGSA design process, elucidating the methodologies for selecting performance expectations, unpacking NGSS dimensions, mapping dimensions, articulating learning performances, specifying design blueprints, and constructing tasks and rubrics. These steps collectively ensure the creation of effective, equitable, and inclusive assessment tools aligned with NGSS standards.

Step 1: Selecting Performance Expectations

The initial step involves choosing a target PE or a coherent bundle of PEs suitable for classroom instruction. The selection should align with instructional content, enabling students to progressively build the necessary knowledge and skills required by the PE or bundle. A strong correlation between the PEs and instructional activities ensures their appropriateness as focal points for developing three-dimensional assessment tasks.

Unpacking entails a thorough exploration of each dimension's proficiency aspects, highlighting intersections between dimensions and considering additional SEPs and CCCs that might productively contribute to achieving the PE or bundle. This step also involves considering students' prior knowledge, potential challenges with the dimensions, equity and inclusion issues, and identifying relevant phenomena and realistic scenarios that can motivate and engage students. For example, unpacking MS-PS1-2 would include delving into the SEP of Analyzing and Interpreting Data, DCI elements related to the Structure and Properties of Matter, and the CCC of Patterns. This comprehensive unpacking process establishes a clear understanding of the required knowledge depth and scope for each dimension at the given grade level.

Step 2: Unpacking the NGSS Dimensions

A critical aspect in assessment design is understanding the depth of a PE, which often extends beyond its one-sentence statement. Unpacking is essential for revealing all components involved in a PE or PE bundle. This step is invaluable as it allows designers to detail the specifics of the three dimensions and the required student proficiencies in each. Documenting the results of unpacking provides a reliable reference throughout the design process, ensuring key design decisions are supported and verified.

Step 3: Mapping the Dimensions

The third step in the NGSA design process uses the detailed information from unpacking to create what is termed an "integrated dimension map." This map visually represents the key relationships among the DCI, SEP, and CCC, synthesizing the unpacked information to illustrate the most significant and productive intersections. The mapping process is akin to constructing a concept map, depicting the key sub-ideas and their interrelations, thereby forming a comprehensive visual guide for achieving the target PE or bundle.

Step 4: Crafting Inclusive Learning Performances

In this step, learning performances are articulated, drawing from the integrated dimension map. These performances are crafted as specific knowledge-in-use statements that are narrower in scope, covering distinct areas of the map. The learning performances are the claims in the ECD argument. Each learning performance, like a PE, is structured to be three-dimensional, ensuring that students apply their knowledge in practical contexts. The process involves integrating various SEPs and CCCs with DCI elements, offering students diverse ways to engage with the content and demonstrate their understanding. The aim is to cover the entire scope of a PE or bundle through a set of complementary learning performances.

Step 5: Developing Design Blueprints

Design blueprints are utilized to guide the development of assessment tasks aligned with each learning performance. These blueprints document the essential and variable task features, as well as equity and inclusion considerations, ensuring that tasks are both comprehensive and accessible. Characteristic task features describe the attributes that are common across all the tasks for a learning performance. Variable task features describe the features that can vary across tasks, such as the level of scaffolding to vary task difficulty. Both types of task features include equity/fairness considerations to help ensure that our tasks are accessible and fair to students of diverse cultural, linguistic, and socioeconomic backgrounds. The blueprints also include evidence statements that articulate the observable features of student performance that can provide evidence of a high-level demonstration of the learning performance, and we use these to inform the development of both tasks and scoring rubrics. The blueprints answer critical design questions, such as what students should know and be able to do, the evidence needed to demonstrate this knowledge, and how to construct tasks that are inclusive and fair.

This step is vital for creating a diverse range of tasks that are consistent in quality and aligned with the learning objectives.

Step 6: Constructing Tasks and Rubrics

The final step involves the actual construction of assessment tasks and rubrics based on the design blueprints. This process includes selecting phenomena or problems that are relevant and engaging, creating scenarios that prompt students to apply their knowledge, and writing task prompts that elicit integrated three-dimensional responses. The development of rubrics is an integral part of this step, providing a framework for evaluating student responses and ensuring that they reflect the multidimensional nature of the learning performances.

Throughout these steps, there is a constant emphasis on considering the diverse backgrounds and experiences of students. This includes using language and scenarios that are relatable and accessible, reducing bias in task content, and providing scaffolds where necessary to support all students in demonstrating their knowledge and skills. By incorporating these principles, the NGSA design process ensures that assessment tasks are not only effective in measuring student understanding but also inclusive and equitable, catering to the needs of a diverse student population.

In all, the NGSA design process, as outlined, represents a comprehensive approach to developing assessments in science education. It emphasizes the integration of knowledge dimensions, evidence-based design, and a strong commitment to equity and inclusion. This process ensures that assessments are not only aligned with educational standards but also responsive to the diverse needs and abilities of all students, fostering an inclusive and equitable learning environment.

2.4.5 Challenges in Measuring Knowledge-In-Use Proficiency

The global education paradigm is shifting from traditional rote learning to a focus on fostering adaptive thinking and championing knowledge-in-use (NRC, 2012; Pellegrino & Hilton, 2012). Consequently, the community is obligated to explore pioneering strategies for creating appropriate assessment tasks that capture students' knowledge-in-use and, importantly, determine methods to utilize these tasks to enhance deep science learning (Li et al., 2024). The design of performance-based tasks presents significant challenges, which requires students to apply their knowledge and experience to solving novel context problems or explaining real-world phenomena, presents significant challenges (He et al., 2023). Additionally, it is often laborious and time-intensive for educators to analyze students' constructed responses to these tasks (Li et al., 2023a). Furthermore, due to the constructed and formative nature of these assessments, they frequently illuminate students' diverse learning trajectories and needs. Moreover, they require students to apply their knowledge in new scenarios. Although critical to developing knowledge, analyzing students' 3D responses becomes intricate and time-consuming. These assessments reveal the various paths students take in their learning, which demands that teachers reconceptualize assessments to cater to diverse student backgrounds (particularly from minoritized and marginalized racial and ethnic groups), underscoring the critical need to empower teachers with robust assessment design skills. Designing assessments to capture the complex cognitive construct of knowledge-in-use remains challenging for the field (He et al., 2023).

2.5 Artificial Intelligence and Assessment

2.5.1 The Origins and Evolution of AI

The advent of big data, cloud computing, artificial neural networks, and machine learning has enabled the development of machines capable of mimicking human intelligence. These technologies underpin the creation of systems that can perceive, recognize, learn, respond, and solve problems, collectively known as artificial intelligence (AI) (Kumar & Thakur, 2012; Spector, Polson, & Muraida, 1993). These advanced technologies are set to revolutionize future workplaces (Lawler & Rushby, 2013). AI, with its capability to interact with and assist humans in performing complex tasks, is being recognized as a major disruptive innovation (Seldon & Abidoye, 2018). Often seen as a critical component of the fourth industrial revolution, AI also has the potential to initiate a significant transformation in the educational sector. Integrating AI into school curricula has already begun (Dai et al., 2020; Knox, 2020). However, similar to how television and computers were initially perceived as groundbreaking for education, AI's role is likely to enhance information accessibility without fundamentally altering the core educational practices. AI is defined as the capability of digital machines to perform tasks that typically require human intelligence. These technologies span various fields, including computer vision, speech recognition, machine learning, big data, and natural language processing (Chiu, 2021; Chiu et al., 2022; Xia et al., 2022). The rapid expansion of AI is profoundly altering how people interact, communicate, live, learn, and work (Chiu, 2021; Chiu et al., 2022; Xia et al., 2022; Pedró et al., 2019). In the context of education, AI in education (AIEd) refers to the application of AI technologies such as intelligent tutoring systems, chatbots, robots, and automated assessments to support and enhance educational processes. AIEd holds significant promise for improving learning, teaching, assessment, and educational administration by providing personalized and adaptive learning experiences, enhancing teachers' understanding of student learning processes, and enabling anywhere, anytime machine-supported queries and immediate feedback. Consequently, AIEd is driving an evolution in teaching practices and program development, making it a crucial area for educational research.

2.5.2 AI and Science Assessments

In science education, AI has been used primarily for the automated assessment of student-written text data, which is common because science educators often use open-ended items to assess students' explanations of phenomena (Liu et al., 2016; Shin & Shim, 2021). Initial studies demonstrated the feasibility of using machine learning models to assess student responses in large-scale classroom assessments, detached from teaching and learning contexts. For instance, automated assessments have been used to evaluate student concepts of natural selection (Ha & Nehm, 2016), climate change (Zhu et al., 2017, 2020), and acid-base reactions (Haudek et al., 2012). Recent research has also explored automated assessment of student-generated hand drawings and written responses about the particulate nature of matter (Lee et al., 2023). Recent advancements in automated assessment in science education have expanded to focus on students' application of knowledge in scientific practices. These advancements aim to provide individualized feedback and support learning through appropriate instructional interventions (Ha et al., 2019; Zhu et al., 2020). Although there are concerns about the socio-cultural and linguistic sensitivity of AI assessments (Li et al., 2023), the practice of integrating AI with formative assessment is becoming widespread (Li et al., 2023, 2024). This research indicates that AI integration in classroom assessment is significantly impacting science teaching and learning. Since the release of ChatGPT, research has highlighted increasing opportunities to use AI in science learning beyond assessment. For example, reviews also suggest that AI chatbots offer opportunities for learners to interact with AI to gain knowledge (Kuhail et al., 2023). This evidence suggests that AI can collaborate with humans to play a critical role in science learning and teaching.

2.5.3 Addressing Gaps and Harnessing AI's Potential in Assessments

Despite the significant advancements in AI and learning science, a systematic approach to using AI technologies in developing and implementing knowledge-in-use assessment tasks in science education remains elusive (NRC, 2012b; Pellegrino & Hilton, 2012). Formative assessments, especially those that require automation and optimization, are complex. They need a deep understanding of how students think and learn (cognitive processes) and how they plan, monitor, and assess their understanding and performance (metacognitive processes). AI systems need to be sophisticated enough to understand these intricate aspects of learning to be truly effective in educational settings. Therefore, a principal challenge resides in the successful translation of AI advancements into pedagogically sound practices for creating, interpreting, or analyzing tasks to provide feedback, and utilizing assessments that support and evaluate knowledge-in-use proficiency. It's crucial to shift the focus of assessment from merely evaluating students to facilitating their learning. To achieve this, teachers need robust support in crafting tailored materials that address the unique needs of each learner. This transition not only requires a change in the way assessments are designed but also underscores the need for teachers to have resources and guidance to effectively adapt to diverse learning styles and challenges. Herein lies the potential of AI. This tool can substantially aid teachers in designing, interpreting, and leveraging assessments that enhance student learning. However, a challenge persists: teachers may lack the requisite knowledge to efficiently utilize AI to provide support tailored to their specific needs. Thus, this paper delves into the iterative process of training generative AI to design, analyze, and utilize performance-based knowledge-in-use assessments as a lever for students' deep science learning. By doing so, I aim to pave the way for more inclusive and effective pedagogical practices, harnessing the power of AI in augmenting human intelligence and fostering students' proficiency in knowledge-in-use.

Presently, most AI solutions today focus on automating processes, often overlooking their potential role in educational models (Sanusi et al., 2024). This overlooks the need for a deep understanding of pedagogy and insight into learners' cognitive processes, especially when automating tasks. The challenge is translating AI advancements into pedagogically sound practices for creating and using knowledge-in-use assessments. With the shift towards assessments that focus on learning rather than evaluation, educators seek tools and guidance tailored to diverse student needs. AI emerges as a promising tool to aid educators in designing and interpreting these assessments. I presented the complicated design process of designing these types of assessment tasks. However, there is a hurdle: educators might not be familiar with using AI effectively for their specific needs. This study aims to hone generative AI's capabilities in designing assessments to enhance deep science learning. The goal is to utilize AI's potential to supplement human intelligence, fostering more inclusive and effective educational strategies and enhancing knowledge-in-use.

2.6 Human-AI Collaboration in Education

Human-AI collaboration presents opportunities to address the challenges. Despite promoting advancements in teaching and learning, AI technologies should primarily aim to enhance human capacities rather than merely replace human tasks (Hwang et al., 2020; Pedró et al., 2019). While AI excels in logical decision-making, it cannot emulate human perceptions, emotions, and cognitions (Yang et al., 2024). Thus, integrating human intelligence with machine intelligence may aid in transitioning towards a human-centered AI involves perceiving AI from a human perspective and acknowledging the multifaceted attributes and contexts of humans.

Human-Computer Interaction (HCI) has been a foundational area of research for decades. Berg (2000) notes that traditional HCI studies emphasized human factors, usability, and interface design, highlighting the computer primarily as a medium. This paradigm shifted significantly with the advent of AI, which has broadened the scope of interaction to include human-AI interaction (HAI) . The psychological aspects of HCI were significantly developed by Card et al. (1983), who conceptualized the human mind as an information-processing system. This view laid the groundwork for understanding how users interact with computers and, by extension, AI systems. With the rise of AI, the literature has increasingly focused on HAI, reflecting a growing interest in how AI can augment human decision-making processes. Hybrid intelligence, where human-AI collaboration leverages the complementary strengths of both, is crucial for effective teaming. In educational contexts, researchers have explored human-AI collaboration to promote student-centered learning (Kim, 2023).

2.7 Theoretical Underpinnings of This Study

In this section, based on the review above, I propose an adapted theoretical framework of human-AI collaboration to design knowledge-in-use assessment. To introduce the framework, I first discuss the differences between human intelligence and machine intelligence , then I define what kind of AI I use in my study with the definition of AI in my work. Finally, I introduce the theoretical framework of this study that guides the research design, data analysis and presentation.

2.7.1 Human Intelligence and Cognition

Human intelligence is a multifaceted cognitive ability that encompasses various mental capacities, including reasoning, problem-solving, planning, abstract thinking, comprehension, and learning from experience. It involves both cognitive processes, such as working memory and long-term memory, and the ability to manage cognitive load effectively (Baddeley, 2000; Sweller, 1988). Human intelligence is characterized by its complexity and adaptability, enabling individuals to handle ill-defined problems requiring flexibility and creativity (Sternberg, 1985).

Human intelligence and AI differ fundamentally in their nature and functioning. Human intelligence is characterized by its flexibility, adaptability, and emotional depth. It encompasses not only cognitive abilities but also emotional and social intelligence, enabling humans to navigate complex social interactions and emotional landscapes. In contrast, artificial intelligence is a product of human design and programming, aimed at replicating specific cognitive tasks. AI operates based on algorithms and data processing, excelling in tasks that require pattern recognition, data analysis, and computational efficiency.

AI systems are not inherently capable of emotional understanding or subjective experiences. They rely on large datasets and computational power to learn and improve, lacking the innate curiosity and creativity that drive human learning. While AI can surpass human performance in certain tasks, it lacks the holistic understanding and consciousness that characterize human intelligence.

2.7.2 Artificial Machine Intelligence and Relational Epistemology

The branch of intelligence focused on machines is referred to as AI. This encompasses systems that execute "activities that we associate with human thinking, activities such as decision-making, problem solving, learning" (Bellman, 1978). Despite the various definitions of AI, the overarching concept involves creating machines capable of achieving complex objectives. These objectives include natural language processing, object recognition, knowledge storage and application for problem-solving, and the ability to adapt and act within their environment through machine learning (Russell & Norvig, 2016). At its core, AI seeks to simulate human intelligence through computational methods. Alan Turing suggested that machines could perform tasks requiring human intelligence by automating calculations, a process that machines can execute much faster than humans (Turing, 1950). The famous Turing 'imitation game' posits that AI is achieved when distinguishing between a conversation with a human and a machine becomes impossible. Although the notion of the "Turing machine" has been critiqued (Searle, 1980), the core idea proposed by Turing remains compelling. Turing emphasized that the significance lies not in the inherent nature of the computer but in what a person perceives the computer to be.

Inspired by Turing's notion, this study adopts a relational epistemology (Bearman & Ajjawi, 2022), conceptualizing AI based on human-technology interactions rather than the computational approach. AI is defined not by its technological features but by the context-bound relationship between humans and computational artifacts during specific interactions. This perspective emphasizes the sociomaterial production of knowledge, focusing on what technologies do rather than their intrinsic properties. This dynamic conceptualization of AI interactions depends heavily on the circumstances of their use.

The relational epistemology proposes that knowledge exists between actors, meaning it is contextualized within specific relationships between people, things, and spaces. This idea aligns with connectivism, which recognizes the interconnectedness of all entities (Siemens, 2005). A sociomaterial perspective further appreciates human and non-human actions and knowledge as entangled in systemic webs (Fenwick, 2010). Non-humans are seen as active participants, following Latour's (2007) view that actors are defined by their actions and impacts rather than their human qualities. For instance, Latour (1999) suggests that a 'speed bump' is an actor whose agency is expressed through its effect on traffic, causing drivers to slow down. Similarly, we consider AIs to be agentic but not sentient, understanding knowledge and knowing as products of the social dynamics involving objects and spaces (Foucault, 1963). This view aligns with Johnson and Verdicchio's (2017) conceptualization of AI systems as 'sociotechnical ensembles ... combinations of artifacts, human behavior, social arrangements, and meaning.'

AI encompasses several distinctive features that enable it to perform cognitive tasks traditionally associated with human intelligence. One of AI's key characteristics is its ability to perform calculations at a speed and scale beyond human capacity. "*They (AI) are much less than human intelligence—they can only calculate. And they are much more—they can calculate larger numbers and faster than humans*." (Cope et al., 2021). Claude Shannon's development of binary calculation using relay circuits laid the groundwork for modern computing, allowing AI to process vast amounts of data efficiently (Shannon, 1938). AI systems often employ machine learning and deep learning techniques to analyze and interpret data. Machine learning involves algorithms that identify patterns within data, while deep learning uses multilayered neural networks to recognize intricate patterns, requiring substantial data and computational power (Krizhevsky, Sutskever, & Hinton, 2012).

Another essential feature of AI is its capability to name and categorize extensive datasets. This process involves representing real-world objects and concepts in binary form, enabling machines to recognize and process these entities more quickly than humans can (Cope & Kalantzis, 2020). AI's calculability allows it to handle large quantities of data swiftly, which is particularly useful in fields such as natural language processing and statistical modeling (Cope & Kalantzis, 2020).

AI systems also possess the ability to measure and interpret data through various sensors and data collection methods. This capability enables AI to gather real-time data and generate insights. For instance, in educational environments, AI can track student interactions to provide personalized feedback and adaptive learning pathways (Cope, Kalantzis, & Searsmith, 2021). Additionally, AI can represent information in multiple formats, such as text, images, sound, and videos, facilitating effective communication and data processing.

Despite these advanced capabilities, AI significantly differs from human cognitive processes. Human intelligence involves context, understanding, and experiential learning, which AI lacks. AI's power lies in its ability to perform detailed and extensive calculations and process vast datasets, rather than understanding or experiencing the world as humans do (Cope, Kalantzis, & Searsmith, 2021). Human intelligence is based on biological neural networks, while AI operates on silicon-based digital systems, resulting in distinct operating principles and capabilities (Korteling et al., 2021).

AI systems can process information at speeds far beyond human capabilities. Human nerve signals travel at most 120 m/s, whereas AI systems can operate at nearly the speed of light (Tegmark, 2018). Furthermore, human learning is influenced by biological and environmental factors, often requiring significant time and effort, while AI can rapidly learn from vast datasets and adapt through continuous training (Russell & Norvig, 2014). AI systems, when designed and validated appropriately, can mitigate human cognitive biases, providing more objective analyses (Korteling et al., 2021). However, humans possess emotional and social intelligence, allowing for nuanced interpersonal interactions, whereas AI, despite advancements in natural language processing and affective computing, lacks genuine emotional understanding and social intelligence (Shneiderman, 2020). Additionally, human decisions are often explainable through introspection and communication, while AI decisions, particularly those made by deep learning models, can be less transparent, necessitating efforts to improve explainability (Cope, Kalantzis, & Searsmith, 2021; Shneiderman, 2020).

2.7.3 Hybrid Intelligence System and Human-AI Collaboration

Human intelligence and artificial intelligence can complement each other in various ways, creating synergistic effects that enhance capabilities in numerous fields. Human intelligence brings creativity, intuition, and emotional understanding to the table, which are areas where AI currently falls short. Humans can excel at making sense of ambiguous and novel situations, understanding context, and applying ethical considerations to decision-making. On the other hand, AI can process and analyze vast amounts of data at unprecedented speeds, identify patterns that might elude human analysts, and perform repetitive tasks with high precision and consistency. By leveraging AI, humans can enhance their decision-making processes, gain insights from complex data sets, and automate mundane tasks, freeing up time and cognitive resources for more strategic and creative endeavors. For instance, in healthcare, AI can assist in diagnosing diseases by analyzing medical images and patient data, while human doctors provide the necessary context, empathy, and ethical judgment in patient care, and more importantly, doctors judge the ambiguous cases that AI has challenges to judge. In education, AI can personalize learning experiences by adapting to individual student's needs, while teachers guide and mentor students, fostering critical thinking and emotional development. The collaboration between human intelligence and artificial intelligence holds the potential to revolutionize various sectors, driving innovation and improving efficiency.

Rather than limiting human involvement to specific parts or times during the creation of machine learning models, real-world problem-solving applications require a continuous socio-technological collaboration between humans and machines. This approach contrasts with earlier research on decision support and expert systems (Gregor, 2001; Holzinger, 2016). Dellermann et al. (2021) argue that the most likely paradigm for the future division of labor between humans and machines is hybrid intelligence. This concept leverages the complementary strengths of human intelligence and AI, enabling them to function more intelligently together than separately (Kamar, 2016). The fundamental rationale is to merge the complementary strengths of heterogeneous intelligences (i.e., human and artificial agents) into a socio-technological ensemble.

Hybrid intelligence systems (HIS) are envisioned as those capable of achieving complex goals by combining human and artificial intelligence to collectively achieve superior results than either could independently, continuously improving through mutual learning (Dellermann et al., 2021). Tasks are performed *collectively*, meaning that while the activities conducted by each part are interdependent, they are not necessarily always aligned to achieve a common goal, such as teaching an AI adversarial tasks like playing games. The system achieves a performance level that none of the involved actors could have achieved alone (*superior results*). The goal is to make the outcome, such as a prediction, more efficient and effective at the socio-technical system level by achieving goals that were previously unattainable. Over time, the socio-technological system improves as a whole, with each component (i.e., humans and machines) learning from each other's experiences, thus enhancing performance in specific tasks (*continuous learning*). The performance of such systems is measured not only by the superior outcomes of the entire system but also by the learning progress of the human and machine agents within the socio-technical system. The concept of hybrid intelligence systems thus envisions socio-technical ensembles where human and AI components co-evolve to improve over time.

The HIS perspective reflects the idea of human–computer interaction (HCI). While extensive research has been conducted on general HCI aspects such as human factors, usability, and interface design, educational HCI studies have traditionally emphasized the computer as a medium (Berg, 2000). Card et al. (1983) laid the groundwork for the psychology of HCI by conceptualizing the human mind as an information-processing system. With the advent of AI technology, research attention has shifted toward human-AI interaction or human-centered AI (HAI) (Stanford HAI, 2020). Lai et al. (2021) reviewed over 80 empirical studies on human-AI decision-making across various fields, including education, and noted a substantial increase in publications on human-AI interaction and decision-making post-2010. The number of relevant papers surged from fewer than 100 every two years before 2016 to over 1000 per topic by 2020. Decision tasks such as predicting student performance, admissions, dropouts, and answering law school admission test questions have been particularly prevalent.

HAI can be interpreted from two perspectives: AI under human control and AI on the human condition. Shneiderman (2020) discusses AI under human control, where AI systems are judged based on the degree of human oversight. At one end of this spectrum is AI that operates entirely under human control, merely assisting with automation. At the other end is AI that operates autonomously, making decisions independently. Human-controlled AI leverages the collaboration between human oversight and AI automation to enhance human productivity, ensuring high levels of reliability, safety, and trust (Shneiderman, 2020). The second perspective, AI on the human condition, is discussed by Stanford HAI (2020). This approach reflects on the design of AI algorithms with humanity as the central consideration. AI on the human condition emphasizes the importance of creating AI systems that are explainable and interpretable, ensuring that their computational and judgment processes can be understood by humans. Additionally, these systems must continuously adjust their algorithms based on human context and societal phenomena. The goal is to augment human intelligence using machine intelligence, ultimately enhancing human welfare (Stanford HAI, 2020). I take the second perspective in this study about HAI.

2.7.4 Distributed Cognition Theory and HAI

The HIS and HAI also reflect the theory of Distributed Cognition (Hutchins, 2000; Pea, 1993), which asserts that cognitive processes are shared and shaped between humans and their tools, highlighting a collaborative cognitive dynamic. It offers a framework to understand the symbiotic dynamics between educators and AI tools. It steers the research methodologies and interpretation, especially in recognizing how ChatGPT can function as an active participant in the cognitive ecosystem of educational assessments. Within the context of developing a domain-specific AI algorithm, Distributed Cognition emphasizes AI's active and integral role in shaping its design, testing, and optimization beyond mere computational augmentation.

Hutchins' Distributed Cognition Theory (1995) posits that cognitive processes aren't singularly anchored but resonate across collective entities, both human and non-human. In conjunction, Roy Pea (1985, 1993) underscores the transformative role of digital tools as cognitive amplifiers that not only extend but also reshape human thinking and collaboration in educational contexts. At its core, Distributed Cognition serves as the theoretical bedrock guiding the structure and trajectory of this research. By adopting this lens, the study is explicitly oriented to capture the fluid interplay between diverse human experts and the sophisticated AI capabilities of ChatGPT. This perspective drives the research methodologies: from the design of experimental setups that foster seamless collaboration to the selection of evaluative metrics that capture both individual and collective cognitive contributions. In practical terms, when studying the process of creating, evaluating, and refining knowledge-in-use assessments, the research actively looks for evidence of distributed cognitive dynamics. For instance, it doesn't just observe an educator’s individual input but examines how that input morphs when interfaced with AI suggestions or when juxtaposed with insights from another domain expert. The interventions, iterative refinements, and validations conducted in the study are all set up to capture these dynamic cognitive exchanges. Furthermore, the research's emphasis on ChatGPT is not just a tool, but a 'cognitive partner' finds its roots in Pea's observations. The AI’s role is conceived not merely as a passive repository or a computational enhancer but as an active agent in the cognitive matrix, shaping and being shaped in turn.

By intertwining the precepts of Distributed Cognition and the insights from Roy Pea, this research champions a groundbreaking approach to understanding AI-human collaboration in educational settings. It strives for a nuanced appreciation of the cognitive orchestra that emerges when human expertise, in all its diverse richness, collaborates with the computational prowess of AI, promising a richer, more holistic outcome that transcends individual capabilities. This paradigm not only informs the study's foundational logic but also steers its empirical pursuits and interpretative analyses, setting a benchmark for future explorations in the realm of distributed cognitive research.

2.7.5 Interdisciplinary Collaborative Learning and HIS

Hybrid intelligence systems often necessitate varying levels of expertise from the humans providing input. Traditionally, both research and practical applications have emphasized the importance of input from machine learning (ML) experts, requiring deep expertise in AI (Attenberg et al., 2015; Chakarov et al., 2016; Kulesza et al., 2010; Patel et al., 2019). Additionally, end users can contribute to product recommendations and e-commerce, or human non-experts can provide input through crowd work platforms (Chang et al., 2018; Nushi et al., 2017). More recent efforts focus on integrating domain experts into hybrid intelligence architectures. These experts use their deep understanding of the semantics of a problem domain to teach machines without needing extensive ML expertise (Dellermann et al., 2019;

Simard et al., 2017). The quantity of human input can range from individual contributions to aggregated input from multiple individuals. Individual input is often used in recommender systems for personalization or cost efficiency (Li et al., 2017). Conversely, collective human input aggregates the contributions of several individuals through mechanisms of human computation (Dellermann et al., 2019; Quinn & Bederson, 2011). This method helps reduce errors and biases inherent in individual inputs and aggregates diverse knowledge (Cheng et al., 2023; Dellermann et al., 2019). Aggregation can be tailored to individual characteristics (Dawid & Skene, 1979; Kamar et al., 2012; Kim & Ghahramani, 2012) or adjusted based on the teaching task (Kosinski et al., 2014; Raykar et al., 2010; Whitehill et al., 2009). This approach informs the design of studies involving expert panels with diverse expertise to collaboratively provide feedback on hybrid intelligence systems' products.

2.7.6 Self-Regulated Learning and HAI

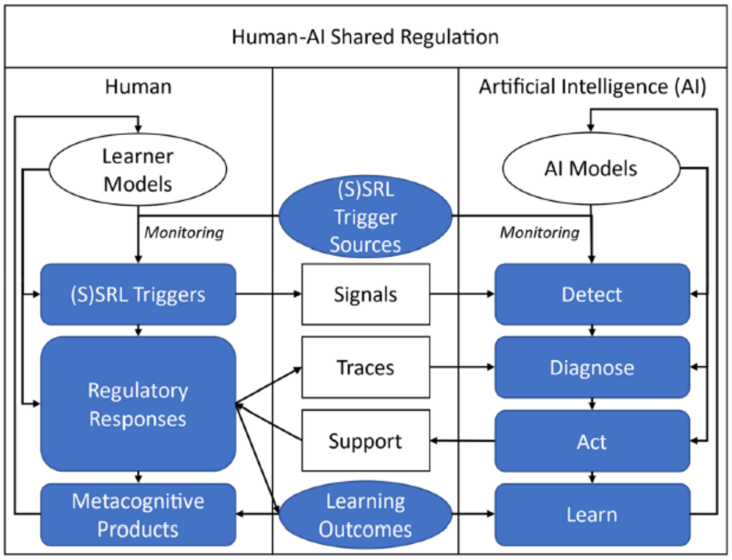
In addressing complex and novel problems while maintaining system efficiency, it is crucial to emphasize the significant role of humans in the HAI process. Consequently, self-regulated learning (SRL) serves as an essential theoretical framework for understanding and enhancing the human learning process and actions within this context. SRL is defined as a goal-oriented process where learners make conscious decisions to achieve their learning objectives (Azevedo, 2015; Winne, 2018). Self-regulated learners utilize cognitive processes such as summarizing, rereading, and elaboration, and metacognitive processes like orientation, planning, monitoring, and evaluation to control their learning and motivate themselves (Greene & Azevedo, 2007). Research on SRL has shown that self-regulated learners are adaptive, engaging metacognitively, motivationally, and behaviorally in their learning (Schunk & Greene, 2018). These learners implement appropriate learning strategies, monitor their progress towards goals, and adjust their strategies and learning conditions when progress is insufficient (Winne & Hadwin, 1998).

Effective self-regulating learners set learning goals to plan their activities and adjust strategies as needed to achieve these goals (Winne, 2017). They continuously monitor whether their actions are aiding progress towards their learning objectives (Azevedo, 2009). Zimmerman (2000) identified three phases in the self-regulated learning process: Forethought, Performance, and Self-reflection. In the forethought phase, learners analyze tasks, set specific goals, and plan strategies. During the performance phase, they implement these strategies, monitor their progress, and receive feedback. In the self-reflection phase, learners evaluate the effectiveness of their strategies and make necessary adjustments.

In my study, I incorporated Zimmerman (2000)'s three phase model and the COPES model (Winne & Hadwin, 1998) to understand the human cognitive conditions in the process of collaborating with AI to design knowledge-in-use tasks. COPES is an acronym representing conditions, operations, products, evaluations, and standards within a task completion framework. Conditions include the available resources and any constraints affecting the task, while standards are profiles of desired attributes refined through planning. Operations involve cognitive processes in working memory that transform information, ranging from innate, simple processes to more complex, acquired strategies. These operations generate products, which are evaluated against standards. Monitoring these comparisons is crucial and if discrepancies arise, it may lead to adjustments in the task, conditions, goals, and standards, or even to abandoning the task. Thus, the COPES model functions as a recursive, adaptable system in task management and learning.

Järvelä, Nguyen, and Hadwin (2023) introduced a framework to operationalize human-AI collaboration, proposing a hybrid human-AI shared regulation in learning (HASRL) model (Figure 2-2). This model positions human and AI collaboration for socially shared regulation (SSRL) in learning, highlighting the synergy between humans and AI to improve learning regulation. Through empirical examples, they demonstrate how hybrid intelligence can enhance learning sciences research, arguing that combining human and AI strengths is vital for advancing this field.

Figure 2-2. Human-AI shared regulation in learning (HASRL) model from Järvelä et al. (2023)

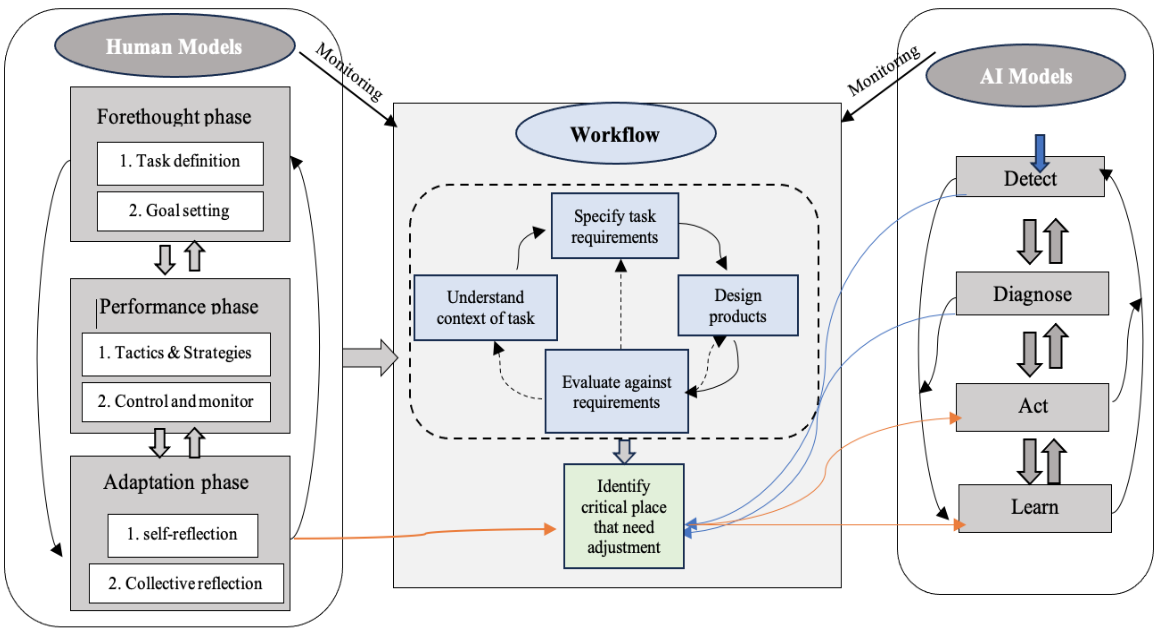


In their study, the HASRL framework is adapted to explore the collaborative potential of hybrid intelligence, leveraging the capabilities of humans and machines to design knowledge-in-use assessments. Human learners bring creative, flexible thinking, and long-term goal orientation to the process, while SRL provides a theoretical foundation for understanding the human-machine interaction in designing assessments. The framework (Figure 2-3) illustrates the interplay between human and AI components in a hybrid intelligent system.

On the human side, during the Forethought phase, humans set the context, define the scope, purpose, and goals, including background information on NGSS, PE, DCI, CCC, and SEP. They design tasks for collaboration with AI. In the Performance phase, humans monitor learning progress and guide AI to reflect on task completion. During the Adaptation phase, humans reflect on goals and requirements, evaluate products, and decide on necessary adjustments, incorporating interdisciplinary feedback from experts throughout the collaboration process.

The AI component, informed by Molenaar (2022), follows the detect-diagnose-act framework. In the detect phase, AI collects learning process data. In the diagnosis phase, AI assesses the current state and predicts future development of assessment tasks. In the act phase, AI implements plausible changes based on the diagnosis, while also adjusting AI models to support human cognitive development. This ensures that AI systems not only respond to critical needs but also scaffold human cognitive competencies. The HHACI model provides a conceptual architecture for integrating technology developers and science educators to create AI-enabled solutions for designing knowledge-in-use assessment tasks.

Figure 2-3. Hybrid human-AI collaborative model (HHACI) in complex task design



There are several noteworthy aspects of this model. First, it is an iterative training model. Central to this is the idea that assessment design is not a linear process; rather, it is a complex tapestry woven together with numerous variables, including students' cognitive states, social-emotional needs, language competencies, and diverse cultural backgrounds. Integrating GPT into this complex environment does not merely add another variable but acts as a catalyst, potentially fostering innovative patterns of interaction and pedagogical strategies (Johnson, 2001). This synergy between educators, students, and GPT forms what Complexity Theory designates as a "complex adaptive system." In this dynamic setup, the principles of Complexity Theory are prominent, emphasizing the adaptability and fluidity required for effective educational outcomes (Byrne, 1998).

Informed by this, my research recognizes that shaping GPT for assessment is not only a multifaceted task, influenced by evolving student needs and educational contexts, but also one that strives for equity in assessment. This ensures that all students, irrespective of their backgrounds, have fair opportunities. To comprehensively address these complexities, an interdisciplinary panel of expert reviewers will be assembled. Additionally, central to the project's methodology is the commitment to iterative training, adaptation, and refinement of ChatGPT, with the goal of achieving both optimal and equitable educational outcomes.

Second, due to the exploratory nature of this study, to better understand the black box, I emphasize human's ability to intentionally influence their functioning and life circumstances. Within my research, this theory illuminates how human experts actively shape AI's role in education rather than merely absorbing its outputs. Their feedback merges AI's potential with their deliberate cognitive tactics. Thus, this research perspective emphasizes the proactive collaboration between educators and students with AI, creating an environment where human intentionality coexists and flourishes with AI-enhanced capabilities.

Informed by Bandura's Human Agency Theory (1989), this research underlines the salience of human capacity to shape one's circumstances and functions, a perspective that becomes paramount when exploring the GPT model's potential to amplify human cognitive faculties. Delving into the theory's core tenets — intentionality, forethought, self-reactiveness, and self-reflectiveness — offers a nuanced lens to understand the multifaceted human-AI interplay in designing knowledge-in-use assessments. The principle of intentionality emerges prominently in the research as educators proactively harness ChatGPT, showcasing a conscious choice rather than passive acquiescence. Central to the research's premise, this principle aligns with humans' purposeful engagement with ChatGPT. Their proactive involvement suggests a conscious decision to harness AI, rather than a passive acceptance, underscoring the act of choosing specific paths and outcomes in AI-mediated educational settings. Forethought, meanwhile, is exemplified in the study's forward-looking approach, moving beyond immediate requirements to anticipate the future trajectories of educational AI. Beyond mere immediacy, the research adopts a visionary stance, enabled by the educator's strategic foresight. Guided by this principle, the study not only focuses on current pedagogical necessities but also aspires to anticipate and prepare for the evolving contours of educational AI.

Lastly, given the knowledge-in-use assessment features, this study also is informed by the Cognitive Flexibility Theory (CFT, Spiro et al., 1992), emphasizing adaptive cognition in ill-structured domains, suggesting that true understanding necessitates multiple viewpoints. This study uses CFT to analyze interdisciplinary expert feedback on AI-designed knowledge-in-use assessments. In this research, the CFT offers an indispensable lens through which the complex construct of knowledge-in-use can be understood and analyzed. By leveraging the principles of CFT, this study endeavors to collaborate with AI, specifically in crafting assessment tasks that can aptly measure such a nuanced domain. The amalgamation of AI capabilities and the insights from CFT holds the promise of generating more refined, context-sensitive, and adaptive assessment tools that can capture the dynamism and depth of knowledge-in-use.

CHAPTER 3: STUDY DESIGN AND METHODOLOGY

3.1 Positionality and the Assessment Development Framework

My involvement with the Next Generation Science Assessment (NGSA) project provided experience in designing knowledge-in-use assessments. I adopt this principled approach (Harris et al., 2019) designed using an evidence-centered design approach (ECD; Mislevy & Haertel, 2006) to guide the GPT-4 model to design assessment tasks. My foundational knowledge not only informs but also guides the GPT-4 model in creating assessments with precise prompts. This expertise is crucial for evaluating the quality of the outputs. I direct the GPT model to design assessment tasks to capture knowledge-in-use.

3.2 Focal Performance Expectations

This study focuses on two elementary school level performance expectations from the NGSS. The performance expectations focus on two major scientific and engineering practices of *developing models* and *constructing scientific explanations,* fundamental to students' knowledge-in-use (Krajcik et al., 2023; Schneider et al., 2022). The two PEs are both for 3rd grade level, one PE focuses on "Physical Sciences" and the other PE is from the "Life Sciences." The two PEs and their associated information are presented below.

3-PS2-1. Plan and conduct an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object.

Figure 3-1. Snapshot of PE 3-PS2-1 from NGSS Online Resources

Students who demonstrate understanding can:

3-PS2-1. **Plan and conduct an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object.** [Clarification Statement: Examples could include an unbalanced force on one side of a ball can make it start moving; and, balanced forces pushing on a box from both sides will not produce any motion at all.] [*Assessment Boundary: Assessment is limited to one variable at a time: number, size, or direction of forces. Assessment does not include quantitative force size, only qualitative and relative. Assessment is limited to gravity being addressed as a force that pulls objects down.*]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Planning and Carrying Out Investigations Planning and carrying out investigations to answer questions or test solutions to problems in 3–5 builds on K–2 experiences and progresses to include investigations that control variables and provide evidence to support explanations or design solutions. • Plan and conduct an investigation collaboratively to produce data to serve as the basis for evidence, using fair tests in which variables are controlled and the number of trials considered. - - - - - - - - - - - - - ***Connections to Nature of Science*** **Scientific Investigations Use a Variety of Methods** • Science investigations use a variety of methods, tools, and techniques.	**PS2.A: Forces and Motion** • Each force acts on one particular object and has both strength and a direction. An object at rest typically has multiple forces acting on it, but they add to give zero net force on the object. Forces that do not sum to zero can cause changes in the object's speed or direction of motion. (Boundary: Qualitative and conceptual, but not quantitative addition of forces are used at this level.) **PS2.B: Types of Interactions** • Objects in contact exert forces on each other.	**Cause and Effect** • Cause and effect relationships are routinely identified.

Connections to other DCIs in third grade: N/A

Articulation of DCIs across grade-levels:
K.PS2.A ; **K.PS2.B** ; **K.PS3.C** ; **5.PS2.B** ; **MS.PS2.A** ; **MS.ESS1.B** ; **MS.ESS2.C**

Common Core State Standards Connections:

ELA/Literacy -

RI.3.1	Ask and answer questions to demonstrate understanding of a text, referring explicitly to the text as the basis for the answers. *(3-PS2-1)*
W.3.7	Conduct short research projects that build knowledge about a topic. (3-PS2-1)
W.3.8	Recall information from experiences or gather information from print and digital sources; take brief notes on sources and sort evidence into provided categories. (3-PS2-1)

Mathematics -

MP.2	Reason abstractly and quantitatively. *(3-PS2-1)*
MP.5	Use appropriate tools strategically. *(3-PS2-1)*
3.MD.A.2	Measure and estimate liquid volumes and masses of objects using standard units of grams (g), kilograms (kg), and liters (l). Add, subtract, multiply, or divide to solve one-step word problems involving masses or volumes that are given in the same units, e.g., by using drawings (such as a beaker with a measurement scale) to represent the problem. *(3-PS2-1)*

3-LS4-3 Construct an argument with evidence that in a particular habitat some organisms can survive well, some survive less well, and some cannot survive at all.

Figure 3-2. Snapshot of PE 3-LS4-3 from NGSS Online Resources

Students who demonstrate understanding can:

3-LS4-3. **Construct an argument with evidence that in a particular habitat some organisms can survive well, some survive less well, and some cannot survive at all.** [Clarification Statement: Examples of evidence could include needs and characteristics of the organisms and habitats involved. The organisms and their habitat make up a system in which the parts depend on each other.]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Engaging in Argument from Evidence Engaging in argument from evidence in 3–5 builds on K–2 experiences and progresses to critiquing the scientific explanations or solutions proposed by peers by citing relevant evidence about the natural and designed world(s). • Construct an argument with evidence.	**LS4.C: Adaptation** • For any particular environment, some kinds of organisms survive well, some survive less well, and some cannot survive at all.	**Cause and Effect** • Cause and effect relationships are routinely identified and used to explain change.

Connections to other DCIs in third grade:
3.ESS2.D

Articulation of DCIs across grade-levels:
K.ESS3.A ; **2.LS2.A** ; **2.LS4.D** ; **MS.LS2.A** ; **MS.LS4.B** ; **MS.LS4.C** ; **MS.ESS1.C**

Common Core State Standards Connections:

ELA/Literacy —

RI.3.1	Ask and answer questions to demonstrate understanding of a text, referring explicitly to the text as the basis for the answers. *(3-LS4-3)*
RI.3.2	Determine the main idea of a text; recount the key details and explain how they support the main idea. *(3-LS4-3)*
RI.3.3	Describe the relationship between a series of historical events, scientific ideas or concepts, or steps in technical procedures in a text, using language that pertains to time, sequence, and cause/effect. *(3-LS4-3)*
W.3.1	Write opinion pieces on topics or texts, supporting a point of view with reasons. (3-LS4-3)
W.3.2	Write informative/explanatory texts to examine a topic and convey ideas and information clearly. *(3-LS4-3)*
SL.3.4	Report on a topic or text, tell a story, or recount an experience with appropriate facts and relevant, descriptive details, speaking clearly at an understandable pace. *(3-LS4-3)*

Mathematics —

MP.2	Reason abstractly and quantitatively. *(3-LS4-3)*
3.MD.B.3	Draw a scaled picture graph and a scaled bar graph to represent a data set with several categories. Solve one- and two-step "how many more" and "how many less" problems using information presented in scaled bar graphs. *(3-LS4-3)*

The PEs were unpacked through the unpacking process, which resulted in several learning performances for each PE. Notably, these learning performances should be able to cover the entire PE when used together. The next section details how to design assessments by leveraging the GPT-4 model through unpacking the research methodology and research process.

3.3 Study Design

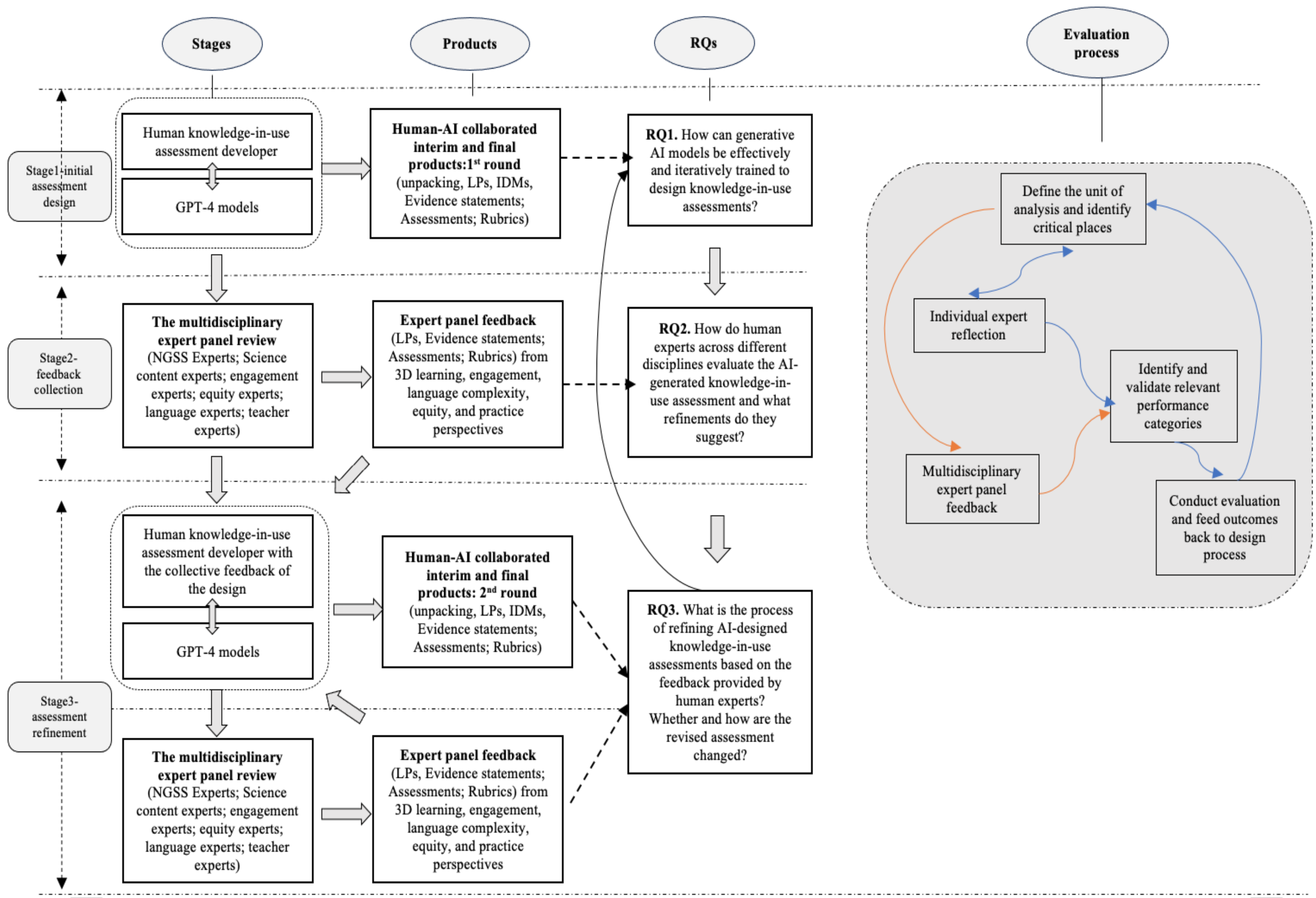
Anchored in the principles of Design-Based Research (DBR) (Barab & Squire, 2004; Collins, Joseph, & Bielaczyc, 2004), this study endeavors to extract the intricate interplay between AI and human intelligence within the realm of knowledge-in-use assessment design. DBR, recognized for its systematic and iterative approach, facilitates a profound exploration that seamlessly marries theoretical understanding with empirical applications.

In the pursuit to address the overarching questions concerning the potential collaboration between human intelligence and artificial intelligence in knowledge-in-use assessment design, the research is structured into three distinct yet interlinked stages with corresponding research questions (see Table 3-1).

These stages span from the initial training and design capabilities of the GPT-4 model, to the critical examination by human experts across disciplines, and finally, to the evolution and optimization of GPT-4 generated assessments. Each stage, while providing depth in its area, collectively contribute to a holistic understanding of the synergy between AI and human intelligence, setting the stage for a transformative leap in educational assessment practices. Through this DBR-driven approach, the study promises depth in exploration and breadth in application, paving the way for innovative strides in the landscape of educational assessment.

Given the features of DBR, for each stage of my study and data analysis, I follow the socio-technical evaluation strategies proposed by Waschull and Emmanouilidis (2023) to analyze the human-AI collaborative assessment system. Being specific, I use the implementation workflow and evaluation methodology presented in Figure 3-3 to evaluate the human-AI collaborative knowledge-in-use assessment design system across the three stages of my study. It is worth to note that this study is an exploratory study to investigate the possibility of AI and human intelligence interaction. Generalizability is beyond the scope of this study.

Figure 3-3. Implementation flow and evaluation methodology



The model involves a structured workflow comprising three stages: initial assessment design, feedback collection, and assessment refinement. In the initial assessment design stage, human knowledge-in-use assessment developers, with the aid of GPT-4 models, create the first round of human-AI collaborated interim and final products, including unpacking, learning performances (LPs), integrated dimension maps (IDMs), evidence statements, assessments, and rubrics. This stage addresses the RQ1, "*How can generative AI models be effectively and iteratively trained to design knowledge-in-use assessments?*" Next, the multidisciplinary expert panel review stage collects feedback on the interim products. This panel comprises NGSS experts, science content experts, engagement experts, equity experts, language experts, and teacher experts. The panel provides feedback on the LPs, evidence statements, assessments, and rubrics, focusing on 3D learning, engagement, language complexity, equity, and practice perspectives. This feedback process addresses the RQ2, "*How do human experts across different disciplines evaluate the AI-generated knowledge-in-use assessments, and what refinements do they suggest?*" The final stage is assessment refinement, where human knowledge-in-use assessment developers, integrating collective feedback, collaborate with GPT-4 models to produce the second round of interim and final products. The same multidisciplinary expert panel reviews these refined products, providing further feedback. This stage responds to the RQ3, "*What is the process of refining AI-designed knowledge-in-use assessments based on the feedback provided by human experts? Whether and how are the revised assessments changed?*"

The evaluation process involves defining the unit of analysis and identifying critical areas, individual expert reflection, and collecting and validating relevant performance categories. This process aims to conduct evaluations and feed outcomes back into the design process, ensuring continuous improvement and alignment with educational objectives and standards. This iterative cycle of design, feedback, and refinement ensures that the assessments developed are robust, context-sensitive, and pedagogically sound, leveraging the strengths of both human and AI intelligence. In the subsequent sections, the specific research design for each stage will be meticulously detailed.

Table 3-1. Data collection and analysis overview

Data Source	Data	Method of Analysis	Intended Inference
Research Question 1: *Iterative* T*raining and Initial GPT-4 Model-Based Assessment Design (Stage 1)* How can generative AI models be effectively and iteratively trained to design knowledge-in-use assessments?			
Self-reflection of the GPT-4 model generated outputs and key themes of high-quality prompt design	Analyze GPT-4 model's outputs after each training to pinpoint essential prompt features. By refining prompts iteratively, I will further probe the quality of generated content.	Thematic analysis & Reflective records	- Capture trends and deeper themes in GPT-4 model developed outputs, including in-the process outcomes and final assessments. - Assess the potential aptitude of the GPT-4 model in the domain of assessment design. - Set the groundwork for a comprehensive framework that nuances the integration of AI in specialized educational contexts or knowledge-in-use assessment design.
Research *Question 2: Human Expert Review and Feedback Collection (Stage 2)* How do human experts across different disciplines evaluate the AI-generated knowledge-in-use assessments, and what refinements do they suggest?			
Expert panel review of learning performances	Likert ratings	Descriptive statistics (Heatmap)	Confirm adequacy of the set of learning performances with respect to representing the domain
	Responses to open ended review questions	Thematic analysis	
Expert panel review of tasks and rubrics	Likert ratings	Descriptive statistics (Heatmap)	Confirm cognitive appropriateness of each task including task complexity and equity issues
	Responses to open ended review questions	Thematic analysis	
Expert panel review of tasks and rubrics regarding equity	Likert ratings	Descriptive statistics (Heatmap)	Confirm the adequacy of equitable opportunity for diverse students' needs
	Responses to open ended review questions	Thematic analysis	
Expert panel review of tasks and rubrics regarding engagement	Likert ratings	Descriptive statistics (Heatmap)	Confirm the adequacy of the assessment for students' engagement.
	Responses to open ended review questions	Thematic analysis	
Teacher cognitive interviews	Teacher reflections on task and overall reactions	Creswell (2003) hierarchical coding procedure	Understand the trends of AI-designed assessments in supporting diverse students' three-dimensional learning in the classroom.

Table 3-1 (cont'd)

Research Question 3: Assessment Refinement (Stage 3) What is the process of refining AI-designed knowledge-in-use assessments based on the feedback provided by human experts?			
Expert panel review of tasks and rubrics	Experts reflections on task and overall reactions	Descriptive statistics (scatter plot) and thematic analysis	Glean deeper insights into the extent of improvement of the refined assessments
Documented revision/refinement process by reflections	Self-reflections on the refinement process	Thematic analysis	

<u>3.3.1 Stage 1: Initial Iterative Training and Preliminary GPT-4 Assessment Design</u>

Stage 1 aims to respond to the research question 1: *"How can the GPT-4 model be effectively and iteratively trained to design knowledge-in-use assessments*?." I adopted the NGSA approach to establish a training blueprint for the GPT-4 model. This approach was explicitly present above in Chapter 2 section 2.4.2. There are two reasons why this study adopts the NGSA approach to design the knowledge-in-use assessment. One is based on the comprehensive analysis of the current approaches to design performance-based assessments to understand students' knowledge and skills of solving complex problems or explain real-world phenomena. This NGSA approach can ensure the designed assessments can capture both the scope and depth of the ideas and abilities that are embedded in certain PEs. The evidence-centered design process also allows the designed assessments to elicit and collect evidence to capture students' understandings. This approach also can ensure the designed assessments to align with the NGSS PEs (Harris et al., 2019; Li et al., 2024). The second reason is, as a person who has extensive assessment design experience enable me to serve as the critical person to use the design criteria to train the GPT-4 model to design the assessments and can judge the outputs of GPT-4's generations to make effective reflections on the outputs to give iterative feedback to the GPT-4 model for further improvement or adjustment.

This stage involves feeding GPT with background data, emphasizing design principles, and introducing domain analysis and modeling processes, following the workflow proposed in the HHAIC model in Figure 2-3. After each training session, I analyzed GPT's outputs to identify prominent prompt features through thematic analysis and reflection based on the assessment design approach and human cognitive functions when doing self-regulated learning (specify the task; set up goals and plans; enact the plans; monitor the learning process; finally reflect on the entire process of the human and AI interactive process). Through iterative refinement, I judge if the outputs against or meet the requirements/goals I set up before. For a more holistic analysis, I deploy thematic analysis complemented by reflective insights, aiming to understand both the explicit patterns and the underlying themes of the AI-generated assessments. The significance of this stage is twofold. Primarily, it seeks to evaluate the potential of the GPT-4 model, in the realm of assessment design. Subsequently, it aspires to craft an initial framework, detailing the nuances of molding AI for specialized educational applications.

3.3.1.1 Design-Based Research with Reflective Practice Highlight

In my work, I used the iterative design and feedback loops provided by my reflections to explore the research questions. In this process, I did not just observe the outcomes but actively engage in refining the assessment designs based on observations and reflections. In the process of exploring the research question 1, where I engaged in in-depth reflection to summarize effective strategies and identify future improvements. This process allows for deep insights into the iterative design process,

3.3.1.2 Data Collection and Analysis

Data collection

I input essential background information for the GPT-4 model to enable it to equip the basic understandings of NGSS, knowledge-in-use, and NGSA design procedures and criteria. Then, I gave detailed instructions about each step of the assessment design process and provided some examples that I want the GPT model to learn from. I collected the outputs generated by the GPT-4 model after each prompt for each step of the design process. I also collected the prompts and corresponding outputs of the training process.

Training process and environment setting: Set up my training with GPT-4 Turbo

In setting up my training process, I utilized OpenAI's Application Programming Interface (API) to interact with the GPT-4 Turbo model for generating structured responses and guiding me through the design process for knowledge-in-use science assessments. The API, a set of rules and protocols, allows different software applications to communicate with each other, enabling me to send requests to OpenAI servers and receive responses generated by their language models. I selected the “gpt-4-turbo-preview” model for its advanced capabilities, as it is designed to provide high-quality and efficient responses, particularly suited for tasks requiring detailed understanding and text generation. This makes it ideal for guiding me through structured processes. To ensure secure and authenticated interaction with the OpenAI API, I used an API key, a secret token that grants access to OpenAI services. I configured the request headers to include the content type as JSON and the authorization token. The payload for each request comprised the chosen model, gpt-4-turbo-preview, and a sequence of messages defining the conversation's structure, including roles such as "system" for setting the context and "user" for input prompts (see Figure 3-4). Additionally, I set a maximum token limit of 1500 to allow for comprehensive responses.

Figure 3-4. Screenshot for the training environment setup

```
import json
import requests

# Replace 'xxx' with your actual OpenAI API key
api_key = [redacted]

headers = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {api_key}"
}

# Define the payload with the structured conversation
payload = {
    "model": "gpt-4-turbo-preview",
    "messages": [
        {
            "role": "system",
            "content": "You are an assistant specialized in guiding users through a detailed and structured design
        },
        {
            "role": "user",
            "content": "Provide an overview of the design process for creating equitable science assessments tailo
        },
        {
            "role": "assistant",
            "content": "To start, let's define the main steps in the design process: 1) Identifying the Performanc
        },
```

POST requests were made to the API endpoint with these headers and payload, and the responses were processed by verifying successful status codes, parsing the JSON data, and saving it to a local file. This setup facilitated automated text generation, which was then formatted into a document recording the entire conversation, ensuring comprehensive logging of both user inputs and assistant responses for further analysis.

Setting up training for the assessment design

I began the training by initializing the OpenAI API interaction framework, where the AI is provided with a system message that sets the context (Refer to Figure 3-4). The AI is tasked with understanding complex instructions, breaking down tasks into smaller steps, and generating intermediate products at each stage. This foundational setup ensures that the AI comprehends its role within the broader framework of the design process. The training process unfolds through a sequence of interactions where the AI and the human user engage in a detailed and structured conversation. Initially, the AI is equipped with a set of instructions that define its role and objectives. Following this, the human user provides a series of prompts designed to guide the AI through various aspects of the science assessment design process. Each user prompt is crafted to elicit detailed, context-specific responses from the AI, ensuring that the output aligns with educational standards and equity goals.

I first set up the system to identify a role for the GPT-4 system as an assistant by giving the prompt: " You are an assistant specialized in guiding users through a detailed and structured design process for science assessments. Your role includes understanding complex instructions, breaking down tasks into smaller steps, and generating intermediate products at each stage. You need to communicate clearly, structuring your responses in a way that aligns with the users' design framework, and refer back to previous steps or information as needed." I then provided an overview for the system to understand the design task by providing the process of the NGSA approach. Here is the prompt I gave the GPT model, " To start, let's define the main steps in the design process: 1) Identifying the Performance Expectation; 2) Unpacking the Performance Expectation; 3) Mapping the Dimensions; 4) Designing Learning Performances; 5) Developing Assessment Tasks; 6) Creating Rubrics; 7) Iterative Review and Revision.

We will go through each step one by one, ensuring clarity and focus on equity and usability.” I then started the design process with one of the focal PEs, 3-PS2-1.

Data analysis

I analyzed the collected data using thematic analysis to extract the common themes that GPT-4 model may fail in understanding human prompts and common strategies that are efficient for supporting GPT-4 model to understand the design purposes and goals. I also maintained a close reflection throughout the process to document my observations, thoughts, and feelings about the GPT's outputs and the iterative training process. Analyze these reflections to identify patterns in my responses to the AI's outputs and how the information that is generated by the GPT-4 model may add to humans, in this case, my understanding or ideas of designing knowledge-in-use assessment.

Anticipated outcomes

Throughout this process, I anticipate achieving three key outcomes. First, I expect to obtain initial outputs from the GPT-4 model for each crucial step in the assessment design process. These steps include unpacking, learning performance generation, evidence statement generation, essential characteristics design, and varied characteristic design. Secondly, I aim to produce preliminary design assessment tasks. It is anticipated, based on the expertise of human assessment experts, that each PE yields at least three learning performances following the unpacking process. For each learning performance, I work with GPT-4 to generate two assessment tasks. These tasks are intended to evaluate or further probe the model's comprehension of various task features. The third expected outcome is the creation of a corresponding rubric for each assessment task. This rubric is designed to evaluate the tasks generated by GPT-4, ensuring they meet the established criteria for analyzing student understanding effectively.

3.3.2 Stage 2: Interdisciplinary Expert Review and Refinement

Stage 2 aims to address the research question 2: *"How do human experts across different disciplines evaluate the AI-generated knowledge-in-use assessments, and what refinements do they suggest?*). For this stage, I randomly select one Learning Performance (LP) from each PE that is generated at Stage 1. This selection forms the basis of the documentation prepared for review by human experts in this stage. To enable the collection of various feedback that focuses on different areas of expertise, Stage 2 requires assembling an interdisciplinary panel of experts. This expert panel (Table 3-2) will include science content experts in physical science and/or life science domain, experts in knowledge-in-use assessment design, experts who has deep understanding of next-generation science standards, experienced elementary science teachers, science education experts who have different focal research areas, and experts who have expertise in motivation, engagement and/or cognitive processes.

Table 3-2 shows how different experts who serve on the panel provide different focused feedback. Their feedback and comments will be sought on both the initial processing seminal products generated by GPT-4 and the preliminary assessment tasks and rubrics designed. This expert review is essential for refining the assessment tools and ensuring their alignment with educational objectives and standards.

Table 3-2. Expert panel and their feedback expertise

Panel members	Feedback expertise
Experts who have strong science content background in physical and/or life sciences	Provide content validity
Experts who have expertise in knowledge-in-use assessment design	Provide feedback on the assessment design process and interim products, such as unpacking documents, etc.
Expert who has deep understanding of the NGSS	Provide feedback on the interim and final products of designed assessments to ensure the coherent and aligned understanding of 3D and knowledge-in-use.
Experienced elementary science teachers	Provide feedback on the assessment tasks to ensure the tasks can be used for elementary, specifically 3rd grade students.
Experts who are science education researchers with different research focuses and/or expertise: two experts focus on literacy and language and two experts focus on equity and inclusion	Provide feedback from different perspectives, such as if the assessment language is appropriated for all students; if the scenarios or contexts in the assessment tasks are accessible to all students regardless of their backgrounds, etc.
Experts who have expertise in motivation, engagement, and cognitive process	Provide feedback about the designed assessment tasks on if the task phenomena are compelling enough to cognitively engage students in the task, etc.

3.3.2.1 Expert Panels' Composition and Background

To comprehensively evaluate the AI-co-designed knowledge-in-use assessments, two expert panels were assembled to review the LPs and related assessments for two distinct PEs: 3-PS2-1 and 3-LS4-3. Each panel comprises multidisciplinary experts, ensuring robust and comprehensive feedback from diverse perspectives. The panels consist of individuals with extensive backgrounds in their respective fields, offering a rich blend of perspectives and insights essential for a thorough evaluation of AI-generated assessments. For both PEs, the panels include NGSS experts who have significant experience in science education, curriculum development, and state-level curriculum frameworks and policy advising. Assessment design experts contribute deep knowledge of three-dimensional teaching and learning approaches and scalable methods for NGSS-aligned teaching and learning. Science content experts specialize in physical sciences and life sciences, providing detailed insights into the subject matter. Science education researchers focus on equity and language, ensuring that the assessments address diverse student needs and are inclusive and accessible. Engagement experts bring valuable perspectives on student motivation, cognitive engagement, and innovative teaching strategies. Elementary science teacher experts, with practical classroom experience, offer a grounded view of the assessments' applicability and implementation in real-world teaching scenarios. These panels bring together a comprehensive set of skills and knowledge, providing a holistic review of the AI-generated assessments. Table 3-3 details the composition and background of each expert panel member, highlighting their interdisciplinary expertise and the robust feedback they can provide. This interdisciplinary composition ensures that the feedback provided by the panels is comprehensive and robust, addressing various aspects of the AI-generated assessments from multiple perspectives.

Table 3-3. Expert panel and their backgrounds

Group	Expertise Area	N	Background
Group 1: PE 3-PS2-1	NGSS Expert	2	*T* has a robust background in physical science education, with a BS degree in Earth, Atmospheric, and Planetary Sciences. He has extensive experience with the NGSS, having served as the in-house expert at the National Science Teaching Association (NSTA) for eight years. Additionally, T has significant experience in curriculum development and standards-based education reform.
			E has over two decades of experience teaching preschool, elementary, and middle school, including ten years as a science specialist and ESL/Bilingual teacher. She was a co-writer for the NGSS and contributed to the NGSS Diversity and Equity Team's Appendix D. After earning a PhD, she became an assistant professor specializing in elementary science education.
	Assessment Design Experts	2	*C* is an expert in science education and assessment design, known for pioneering innovative approaches to support three-dimensional teaching and learning. With over a decade of experience, C has developed scalable methods to address the NGSS through curricula, assessments, and professional learning models. His work focuses on creating engaging, interactive, equitable, and accessible learning experiences for students and supporting teachers in implementing these strategies effectively.
			P is an expert in assessment design with a strong background in chemistry and education. Holding a BS in chemistry, a master's degree in chemistry education, and a PhD in curriculum and instruction (chemistry education), P conducts research on NGSS curriculum, assessment, and professional learning at middle and high schools. With over five years of experience in NGSS curriculum and assessment design, P identifies as a science educator, science teacher educator, and an international science education research scholar.
	Science Content Experts (Physical Science)	5	*J* is a science content expert specializing in physical science. Holding a bachelor's degree in physics and a doctorate in education, J is currently a postdoctoral researcher in science education. With previous experience in physics education research, J focuses on designing instructional environments using Project-Based LearninG to foster the development of students' knowledge-in-use and understanding of the nature of science.
			S holds a doctorate in physics education and a bachelor's degree in physics. Her research interests focus on pre-service teacher professional development and project-based learning.

Table 3-3 (cont’d)

			P is an expert in assessment design with a strong background in chemistry and education. Holding a BS in chemistry, a master’s degree in chemistry education, and a PhD in curriculum and instruction (chemistry education), P conducts research on NGSS curriculum, assessment, and professional learning at middle and high schools.
			T has a robust background in physical science education, with a BS degree in Earth, Atmospheric, and Planetary Sciences. He has extensive experience with the NGSS, having served as the in-house expert at the National Science Teaching Association (NSTA) for eight years. Additionally, T has significant experience in curriculum development and standards-based education reform.
			J has a background in microbiology and holds a PhD in Science Education. He specializes in supporting students in building and revising computational models. J has taught college-level science courses and science teaching methods for secondary pre-service teachers. He has also conducted professional development sessions for in-service teachers and has experience designing assessment tasks for standardized science exams.
	Science Education Researchers (equity)	2	E has over two decades of experience teaching preschool, elementary, and middle school, including ten years as a science specialist and ESL/Bilingual teacher. She was a co-writer for the NGSS and contributed to the NGSS Diversity and Equity Team’s Appendix D. After earning a PhD, she became an assistant professor specializing in elementary science education.
			Co is an assistant professor of Teacher Education. She teaches science methods courses in the science education department and also facilitates professional learning initiatives focused on urban school districts. Co has worked on major research projects related to project-based learning and has extensive teaching experience, primarily in pre-K through 7th-grade science, as well as teaching all subjects in a self-contained 3rd-grade classroom.
	Science Education Researchers (language)	2	E has over two decades of experience teaching preschool, elementary, and middle school, including ten years as a science specialist and ESL/Bilingual teacher. She was a co-writer for the NGSS and contributed to the NGSS Diversity and Equity Team’s Appendix D. After earning a PhD, she became an assistant professor specializing in elementary science education.
			Su has an extensive background in educational standards and curriculum development. She served as the Lead State Representative on the NGSS development team and led standards development in her previous role. Currently, she oversees an eight-year elementary science PBL project, focusing on literacy integration and standards implementation. With 28 years of experience across various content areas, Su brings a wealth of knowledge and expertise to her role.

Table 3-3 (cont'd)

	Engagement experts	3	Sa is a fourth-year PhD candidate specializing in motivation, engagement, and critical race theories, with additional experience in literacy. She has been working on an NSF project supporting middle school students' motivation in science learning.
			Q is a third-year PhD student specializing in cognitive flexibility, student engagement, game-based learning, and virtual learning. With an undergraduate background in psychology and business and a master's degree in cognitive science, she brings a unique interdisciplinary perspective.
			H is a fifth-year PhD candidate specializing in student engagement, language and literacy assessment, and special education. She has extensive experience supporting student science assessments through a linguistic perspective.
	Elementary science teacher experts	2	Le has over 30 years of teaching experience, specializing in science education for intermediate school students. She particularly enjoys working with sixth graders, leveraging their energy and curiosity to promote scientific inquiry. Le has facilitated local, state, and national workshops to advance science education and has led multiple community education initiatives.
			B has 30 years of teaching experience in a rural public school. Her teaching background includes 3rd grade, 6th grade science, and predominantly 5th grade. For the past five years, B has served as a K-5 STEM teacher utilizing a project-based learning curriculum. She has collaborated closely with a research team to provide feedback and observational data while teaching these units.
Group 1: PE 3-LS4-3	NGSS Expert	2	M is an expert in NGSS with extensive experience in science education. She has contributed to state-level curriculum frameworks and advises on science education policy. Her research focuses on teacher learning, professional development, and adapting pedagogies to support multilingual students.
			E has over two decades of experience teaching preschool, elementary, and middle school, including ten years as a science specialist and ESL/Bilingual teacher. She was a co-writer for the NGSS and contributed to the NGSS Diversity and Equity Team's Appendix D. After earning a PhD, she became an assistant professor specializing in elementary science education.

Table 3-3 (cont'd)

	Assessment Design Experts	3	*C* is an expert in science education and assessment design, known for pioneering innovative approaches to support three-dimensional teaching and learning. With over a decade of experience, C has developed scalable methods to address the NGSS through curricula, assessments, and professional learning models. His work focuses on creating engaging, interactive, equitable, and accessible learning experiences for students and supporting teachers in implementing these strategies effectively.
			Sm has a background as a middle and high school science teacher, primarily working with students underrepresented in STEM. Currently, Sm is a tenure-track professor at a research-intensive institution, focusing on the design of science education interventions for large-scale use, including curriculum, assessments, and professional development.
			P is an expert in assessment design with a strong background in chemistry and education. Holding a BS in chemistry, a master's degree in chemistry education, and a PhD in curriculum and instruction (chemistry education), P conducts research on NGSS curriculum, assessment, and professional learning at middle and high schools.
	Science Content Experts (Life Science)	5	J has a background in microbiology and holds a PhD in Science Education. He specializes in supporting students in building and revising computational models. J has taught college-level science courses and science teaching methods for secondary pre-service teachers. He has also conducted professional development sessions for in-service teachers and has experience designing assessment tasks for standardized science exams.
			Cn is a bilingual Latina with a rich background in science education. She is currently focused on research in her role as an Academic Specialist. With extensive experience in developing 3D PBL curriculum and assessments, she also brings a wealth of knowledge from her time as a middle and high school science teacher. Cn holds degrees in plant biology, education, and public health, and has a PhD in secondary science education. She is also well-versed in teacher professional learning.
			L is a third-year PhD student specializing in curriculum, instruction, and teacher education with a focus on science and urban education. She has four years of experience as a high school science teacher, where she taught biology, environmental science, and chemistry. Her research interests include noticing, classroom discourse, and group work, aiming to create more equitable and just science classrooms, especially for marginalized students.
			Sm has a background as a middle and high school science teacher, primarily working with students underrepresented in STEM. Currently, Sm is a tenure-track professor at a research-intensive institution, focusing on the design of science education interventions for large-scale use, including curriculum, assessments, and professional development.

Table 3-3 (cont'd)

			H is a PhD candidate in science education, who has extensive experience in both science and gifted education. She holds degrees in Biology and Biotechnology and has a master's degree in healthcare administration. Initially intending to pursue medicine, she shifted her focus to education, driven by a passion for teaching. Her work includes significant experience in Artificial Intelligence in Education (AIED) and supporting teachers in integrating new technologies into their classrooms.
	Equity and language experts	3	E has over two decades of experience teaching preschool, elementary, and middle school, including ten years as a science specialist and ESL/Bilingual teacher. She was a co-writer for the NGSS and contributed to the NGSS Diversity and Equity Team's Appendix D. After earning a PhD, she became an assistant professor specializing in elementary science education.
			Co is an assistant professor of Teacher Education. She teaches science methods courses in the science education department and also facilitates professional learning initiatives focused on urban school districts. Co has worked on major research projects related to project-based learning and has extensive teaching experience, primarily in pre-K through 7th-grade science, as well as teaching all subjects in a self-contained 3rd-grade classroom.
			Su has an extensive background in educational standards and curriculum development. She served as the Lead State Representative on the NGSS development team and led standards development in her previous role.
	Engagement experts	3	Sa is a fourth-year PhD candidate specializing in motivation, engagement, and critical race theories, with additional experience in literacy. She has been working on an NSF project supporting middle school students' motivation in science learning.
			Q is a third-year PhD student specializing in cognitive flexibility, student engagement, game-based learning, and virtual learning.
			H is a fifth-year PhD candidate specializing in student engagement, language and literacy assessment, and special education. She has extensive experience supporting student science assessments through a linguistic perspective.
			H is a fifth-year PhD candidate specializing in student engagement, language and literacy assessment, and special education. She has extensive experience supporting student science assessments through a linguistic perspective.

Table 3-3 (cont'd)

	Elementary science teacher experts	2	Le has over 30 years of teaching experience, specializing in science education for intermediate school students. She particularly enjoys working with sixth graders, leveraging their energy and curiosity to promote scientific inquiry. Le has facilitated local, state, and national workshops to advance science education and has led multiple community education initiatives. She holds a B.S. in elementary education and an M.A. in teacher development and educational technology.
			B has 30 years of teaching experience in a rural public school. Her teaching background includes 3rd grade, 6th grade science, and predominantly 5th grade. For the past five years, B has served as a K-5 STEM teacher utilizing a project-based learning curriculum. She has collaborated closely with a research team to provide feedback and observational data while teaching these units. The curriculum has proven effective, significantly improving test scores. Additionally, B incorporates a digital platform to deliver lessons and enhance student interaction.

3.3.2.2 Data Collection

I collected data to examine the cognitive, inferential, and instructional validity (Pellegrino et al., 2016) of GPT generated assessments produced in the first stage. Each panel assessed two GPT-4 designed assessment tasks from either of the two PEs. They offered both quantitative evaluations and qualitative feedback encompassing strengths, areas of concern, and potential improvements.

Instruments

I developed and used different instruments for experts with different expertise on the panel. The panel used a protocol to independently determine the appropriateness of designated LP and the adequacy of the set of LPs with respect to representing the domain (instructional validity). They reviewed the tasks designed to align with each LP and the scoring rubrics (inferential validity). During these reviews, they attended to cognitive validity issues, including ethnic and cultural bias, cognitive complexity, and task performance demands.

Feedback collection instruments were tailored for different expert groups, including NGSS experts, assessment design experts, science content experts, and science education researchers with a focus on equity and language. These groups received protocols designed to elicit detailed feedback on

Learning Performances (LPs) and Evidence Statements (Table 3-4), as well as two AI-co-designed assessment tasks (Table 3-5). Engagement experts and teacher experts were given protocols specifically designed to gather insights from their unique perspectives on the AI-co-designed tasks.

Three types of expert feedback collection instruments were developed. The first instrument is *science-focused feedback instrument* (see Table 3-4 & 3-5). which are about the designed LPs, evidence statements, and corresponding assessment tasks. The second type of instrument is *engagement and language-focused feedback instrument* (see Table 3-6). For the science-focused instrument, it was designed to capture the experts' feedback on the GPT-designed assessments and interim seminal products to collect their feedback if the designed interim products and assessments 1) captured the three-dimensions of science knowledge and skills; 2) align with the PEs/LPs, 3) elicit students' knowledge-in-use performance. Table 3-3 presents the instrument that is used for collecting the feedback from science content and knowledge-in-use assessment design experts about the designed learning performances by generative AI, which is a critical part of the assessment design process. Experts provided Likert scale ratings and open-ended feedback of the tasks. Table 3-5 presents the instrument for capturing science-focused feedback with respect to the designed assessments. These questions solicit detailed and actionable feedback from experts regarding the quality of the designed knowledge-in-use assessments. Science content, knowledge-in-use assessment design, and NGSS experts, and experienced teachers used the instrument. The questions on the instrument presented in Table 3-5 should elicit in-depth feedback on the quality of the AI-generated assessment tasks. Their collective human feedback ensures the designed assessment is robust and resonates with the pedagogical tenets, maintaining assessment validity and reliability.

Data were analyzed qualitatively using a thematic analysis of the open-ended responses. For the Likert scale ratings, I checked for consistency across reviews, provided descriptive analysis, and determined a set of revisions to the tasks. Descriptive statistics dissect the quantitative feedback, while thematic analysis delves into the qualitative feedback, unraveling emergent patterns and insights. I conducted a pilot test to ensure the instruments are accessible and understandable.

Table 3-4. Expert panel review protocol for AI-designed learning performance

<table>
<tr><th colspan="8">Feedback on Learning Performances and Evidence Statements for PE</th></tr>
<tr><td colspan="8">This instrument is designed to systematically assess the effectiveness and quality of learning performances. It aims to gauge various aspects including the integration of knowledge, the essentiality of the learning performance for achieving performance expectations, the sufficiency of evidence statements, collective representation of proficiencies, as well as identification of any gaps or overreach. Respondents are requested to provide both quantitative ratings using a Likert scale and qualitative feedback through rationales for their responses. This comprehensive approach ensures a nuanced understanding of the learning performances.</td></tr>
<tr><td>Dimension</td><td>No.</td><td>Question</td><td colspan="5">Rating Scale (1-5)</td></tr>
<tr><td colspan="3"></td><td>1- Not at all</td><td>2- Slightly</td><td>3- Moderately</td><td>4- Very much</td><td>5- Comp letely</td></tr>
<tr><td rowspan="2">Collective Representati on of Proficiencies</td><td>1</td><td>To what extent does the set of learning performances collectively represent the proficiencies that are necessary for attaining the performance expectation?</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>1a</td><td colspan="6">Rationale for your response:</td></tr>
<tr><td rowspan="2">Essentiality of the PEs</td><td>2</td><td>To what extent does LP4 comprise an essential part of what is needed to achieve the performance expectation?</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>2a</td><td colspan="6">Rationale for your response:</td></tr>
<tr><td rowspan="2">Sufficiency of Evidence</td><td>3</td><td>To what extent do the evidence statements of LP4 reflect obtainable pieces which, taken together, are sufficient for supporting a claim of student proficiency in this learning performance?</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>3a</td><td colspan="6">Rationale for your response:</td></tr>
<tr><td rowspan="2">Integration of Knowledge</td><td>4</td><td>To what extent is LP4 an integrated 3-dimensional statement of knowledge-in-use?</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>4a</td><td colspan="6">Rationale for your response:</td></tr>
</table>

Table 3-4 (cont'd)

Identification of Gaps	5	What gaps, if any, do you see in the set of learning performances (i.e., proficiencies required by the performance expectation that are not represented in the set of learning performances)?
	5a	Rationale for your response:
Identification of Overreach	6	What overreach occurs, if any, in the set of learning performances (i.e., proficiencies that ARE NOT required by the performance expectation but that ARE required in the set of learning performances)?
	6a	Rationale for your response:
Additional Comments	7	Please provide any further comments you would like to make about the set of learning performances relative to the performance expectation. Include both positive and negative comments:

Table 3-5. Expert panel feedback instrument for AI-designed assessments

Assessment Task

Each assessment task draft has a general format consisting of three parts:

(1) Stem includes the phenomena (aligned to the LP) embedded in a scenario.

(2) Prompt is the question we ask students that they respond to. All prompts need to be 3D. You may choose to have scaffolds in prompts to provide support for students in integrating the 3D. Within a given task, we may have many prompts in order to address different Integrated Proficiencies or to break down a complex task into manageable chunks.

(3) An exemplar response that an ideal student would respond with, if given the task.

Accessibility to a task includes comprehension of the stem (the description of the scenario) and prompt (the question for students to answer) and how students are to express what they know. The criteria are based on research from assessment design for English Language Learners, assessment bias, Universal Design for Learning, and validity of assessments. In addition, literature on culturally relevant learning opportunities was consulted.

ITEM STEM

Criteria	Review Questions	Reviewer's Comments (If you have concerns, explain the concerns)				
		1 - Not at all	2 – Slightly	3 – Moderately	4 - Very much	5 – Completely
1. Phenomena	(a) To what extent does the stem set forth a phenomenon that is compelling and comprehensible for students to make sense of?	Rating: Comment:				
Comprehension	(a) Prior Learning: What prior learning or background experiences should students have when engaged with the task?	List prior learning and/or background experience.				
	(b) Information coherence: To what extent the task requires students to leverage three-	Rating: Comment:				

Table 3-5 (cont'd)

	understand and answer the prompt? (e.g., construct an explanation, analyze patterns in data to answer...)	
	(b) Information coherence: To what extent does the prompt align with what is being described in the scenario?	Rating: Comment:
	(c) Consistency: To what extent are the terms used in the prompt consistent with how they are used in the stem?	Rating: Comment:
3. Language Complexity	(a) Sentence structure: To what extent are the sentences clear and direct?	Rating: Comment:
	(b) Other vocabulary: To what extent are all vocabulary words used (e.g., "cite"; "gradually") appropriate for grade level?	Rating: Comment:
	(c) Domain Specific Vocabulary: To what extent are there domain specific vocabulary (i.e. science vocabulary, terms) being used? If so, are students familiar with this vocabulary? (e.g., write a claim, make a prediction).	Rating: Comment:
Scaffolds	(a) If scaffolds are used, then to what extent are they presented in a way that they can help students break down the complexity of the task?	Rating: Comment:

RUBRIC

Criteria	Review Questions	Reviewer's Comments (If you have concerns, explain the concerns)
	(a) To what extent does the rubric capture ALL necessary evidence statements?	Rating: Comment:

Table 3-5 (cont'd)

	understand and answer the prompt? (e.g., construct an explanation, analyze patterns in data to answer...)	
	(b) Information coherence: To what extent does the prompt align with what is being described in the scenario?	Rating: Comment:
	(c) Consistency: To what extent are the terms used in the prompt consistent with how they are used in the stem?	Rating: Comment:
3. Language Complexity	(a) Sentence structure: To what extent are the sentences clear and direct?	Rating: Comment:
	(b) Other vocabulary: To what extent are all vocabulary words used (e.g., "cite"; "gradually") appropriate for grade level?	Rating: Comment:
	(c) Domain Specific Vocabulary: To what extent are there domain specific vocabulary (i.e. science vocabulary, terms) being used? If so, are students familiar with this vocabulary? (e.g., write a claim, make a prediction).	Rating: Comment:
Scaffolds	(a) If scaffolds are used, then to what extent are they presented in a way that they can help students break down the complexity of the task?	Rating: Comment:

RUBRIC

Criteria	Review Questions	Reviewer's Comments (If you have concerns, explain the concerns)
	(a) To what extent does the rubric capture ALL necessary evidence statements?	Rating: Comment:

Table 3-5 (cont'd)

	(b) To what extent is it possible for students to provide responses presented in the rubric	Rating: Comment:
	(c) To what extent is the rubric that is appropriate for the grade level?	Rating: Comment:

Engagement and language-focused feedback instrument. To facilitate a comprehensive understanding of the assessment tasks, the panel responded to prompts that examine the cognitive validity, equity, language appropriateness, and engagement evidence (see Table 3-5). All experts on the panel used the *Engagement and language-focused feedback instrument.* Qualitative data were analyzed through thematic analysis of the open-ended responses. For Likert scale ratings, I assessed consistency across reviews, performed descriptive statistical analysis, and identified necessary revisions to the tasks. I conducted pilot tests to ensure the instruments are accessible and understandable before formally applying them in the study.

Table 3-6. Expert panel review protocol (engagement) for AI-designed assessment tasks

Assessment Task

Each assessment task draft has a general format consisting of three parts:

(1) Stem includes the phenomena (aligned to the LP) embedded in a scenario.

(2) Prompt is the question we ask students that they respond to. All prompts need to be 3D. You may choose to have scaffolds in prompts to provide support for students in integrating the 3D. Within a given task, we may have many prompts in order to address different Integrated Proficiencies or to break down a complex task into manageable chunks.

(3) An exemplar response that an ideal student would respond with, if given the task.

Accessibility to a task includes comprehension of the *stem* (the description of the scenario) and *prompt* (the question for students to answer) and how students are to express what they know. The criteria are based on research from assessment design for English Language Learners, assessment bias, Universal Design for Learning, and validity of assessments. In addition, literature on culturally relevant learning opportunities was consulted.

ITEM STEM

Criteria	Review Questions	Reviewer's Comments (If you have concerns, explain the concerns)				
		1 - Not at all	2 - Slightly	3 - Moderately	4 - Very much	5 - Completely

Table 3-6 (cont'd)

1. Phenomena	(a) To what extent does the stem set forth a phenomenon that is compelling and comprehensible for students to make sense of?	Rating: Comment:
2. Comprehension	(a) Prior Learning: What prior learning or background experiences should students have when engaged with the task?	List prior learning and/or background experience.
	(b) Information coherence: To what extent the task requires students to leverage three-dimensional knowledge to respond to the questions?	Rating: Comment:
	(c) Information processing: To what extent is the information provided in a sequential order?	Scale: Comment:
	(d) Consistency: To what extent are the terms being used consistently?	Rating: Comment:
	(e) Visuals: To what extent the task includes helpful images/videos to help with comprehension? Is a caption necessary for students' understanding?	Rating: Comment: Caption: Yes No Comment:
3. Language Complexity	(a) Sentence structure: To what extent the sentences are clear and direct?	Rating: Comment (Suggest revisions for sentences):
	(b) Other vocabulary: To what extent all the vocabulary words used (e.g., "cite"; "gradually") are appropriate for grade level?	Rating: Comment:
	(c) Context-specific vocabulary (words that are used in daily life, but also in science with specific meanings): To what extent are the context-specific vocabulary clearly	Rating: Comment:

Table 3-6 (cont'd)

	interpreted for specific meaning in context?	
	(d) Reading level: To what extent is the reading level and load reasonable for students? (flag places where vocabulary can be simplified)	Rating: Comment: Identity places where vocabulary can be simplified:
	(e) Domain-specific Vocabulary: To what extent are there domain-specific vocabulary (i.e. science vocabulary, terms) being used? If so, are students familiar with this vocabulary?	Rating: Comment: List Domain-specific Vocabulary and familiarity for the grade level.
4. Engagement	(a) Interest to students: To what extent are students going to find the scenario interesting?	Rating: Comment:
	(b) Relevance to 3D Learning: To what extent are students going to find the scenario relevant to their lives?	Rating: Comment:
	(b) Skills of 3D Learning: To what extent does this scenario provide essential information that allows students to engage with the 3 dimensions of the LP?	Rating: Comment:
5. Cultural sensitivity	(a) To what extent are there groups for whom the task is <u>not inclusive</u> (e.g. SES, gender, culture, region in the country)?	Rating: Comment: List groups for whom the task is not inclusive:
	(b) To what extent is the scenario familiar and/or relevant (authentic to real life) to students?	Rating: Comment:

ITEM PROMPT

Table 3-6 (cont'd)

Criteria	Review Questions	Reviewer's Comments (If you have concerns, explain the concerns)				
		1 - Not at all	2 - Slightly	3 - Moderately	4 - Very much	5 - Com plete ly
1. 3D Prompt	(a) To what extent are the prompt(s) 3-dimensional and aligned to the integrated proficiencies?	Rating: Comment:				
	(b) To what extent are the questions elicited by the prompt(s) clearly motivated by the scenario described in the stem?	Rating: Comment:				
	(c) To what extent can students who are still developing relevant proficiency demonstrate their novice understandings when responding to the prompt?	Rating: Comment:				
2. Comprehension	(a) Prior Learning: What prior knowledge do students have the required prior learning to understand and answer the prompt? (e.g., construct an explanation, analyze patterns in data to answer...)	List prior learning assumed in the prompt:				
	(b) Information coherence: To what extent does the prompt align with what is being described in the scenario?	Rating: Comment:				
	(c) Consistency: To what extent are the terms used in the prompt consistent with how they are used in the stem?	Rating: Comment:				
3. Language Complexity	(a) Sentence structure: To what extent are the sentences clear and direct?	Rating: Comment:				
	(b) Other vocabulary: To what extent are all vocabulary words used (e.g., "cite";	Rating: Comment:				

Table 3-6 (cont'd)

	"gradually") appropriate for grade level?	
	(c) Domain Specific Vocabulary: To what extent are there domain specific vocabulary (i.e. science vocabulary, terms) being used? If so, are students familiar with this vocabulary? (e.g., write a claim, make a prediction).	Rating: Comment:
Scaffolds	(a) If scaffolds are used, then to what extent are they presented in a way that they can help students break down the complexity of the task?	Rating: Comment:

EXEMPLAR RESPONSE

Criteria	Review Questions	Reviewer's Comments (If you have concerns, explain the concerns)				
		1 - Not at all	2 - Slightly	3 - Modera tely	4 - Very much	5 - Completel y
1. Exemplar Student Response	(a) To what extent does the exemplar response(s) capture ALL necessary evidence statements?	Rating: Comment:				
	(b) To what extent is it possible for students to provide accurate responses without attending to the Integrated Proficiencies entailed by an LP?	Rating: Comment:				
	(c) To what extent is the response written in language that is appropriate for the grade level?	Rating: Comment:				

Teacher cognitive interview protocol. A semi-structured experienced teacher interview protocol was designed to conduct interviews with experienced teachers, especially when there are unique concerns or points raised in their questionnaire. This interview was conducted after teachers filled out the survey. Interview is to further detect or understand each teacher's perceptions on the GPT-designed assessments, their concerns or suggestions on the designed assessments. I use thematic analysis to analyze the interview to further capture the suggestions or feedback from teachers. Thematic analysis will synthesize the feedback, distilling salient patterns and enlightening nuances. The insights are anticipated to offer tangible directions for refining the assessments.

Table 3-7 presents the teacher interview protocol. Some prompts include: "How do you see the tasks providing appropriate opportunities for your students to demonstrate their proficiencies with 3-dimensional aspects of the NGSS PEs? *(cognitive validity)*" "What strengths, if any, do the AI-generated task contain?" "Which areas necessitate refinement or enhancement in the AI-generated task?" "In which ways, if any, do the specific AI-designed tasks fall short of your expectations? Could you detail the areas of deficit?" "How well do the tasks cater to students with diverse backgrounds, ensuring equitable opportunities for all to demonstrate their understanding? (*equity*)" "How do the tasks actively engage learners, prompting interest and sustain attention throughout the assessment? (*engagement*)" "In what ways do the tasks facilitate your students' ability to approach problems from multiple perspectives? (*knowledge-in-use*)."

Table 3-7. Teacher interview protocol

Directions: ● Record Teacher's Name ● Interviews could be done via video conference or phone. ● Confirm with the teacher that this interview is voluntary, and they do not have to answer questions they don't feel comfortable with. ● State that this is confidential ● Ask for permission to record the interview and take notes. ● Make the interview conversational in tone. ○ Ask the initial question, then ask teachers follow-up questions to probe deeper but keep it like a conversation. Use probes such as: Tell me more about that. Can you give me an example? Can you tell me what you mean by… ● Be careful not to lead the teacher. They should be doing 90% of the talking. 1. How do you see the tasks provide appropriate opportunities for your students to demonstrate their proficiencies with 3-dimensional aspects of the NGSS PEs? 2. What strengths, if any, strengths do the AI-generated task contain 3. Which areas necessitate refinement or enhancement in the AI-generated task? 4. In which ways, if any, do the specific AI-designed tasks fall short of your anticipations? Could you detail the areas of deficit? 5. How well do the tasks cater to your students from diverse backgrounds, ensuring equitable opportunities for all to demonstrate their understanding? 6. How do the tasks actively engage learners, prompting interest and sustained attention throughout the assessment? 7. In what ways do the tasks facilitate your students' ability to approach problems from multiple perspectives of the three dimensions (i.e., DCIs, SEPs, CCCs)?

3.3.2.3 Data Analysis

The feedback analysis was organized into three main sections: LPs and Evidence Statements, Task 1, and Task 2, corresponding to the evaluations of the Performance Expectations (PEs). Both qualitative and quantitative methods were employed in the data analysis. Each section of the report starts with an overview of the quantitative analysis, followed by an in-depth qualitative analysis that highlights key feedback and recommendations from the experts.

For the quantitative data analysis stage, I use heatmaps as both an analytical and representational tool to organize and interpret the expert feedback data. Heatmaps serve as graphical representations that employ color coding to illustrate complex data matrices. This visual method facilitates the immediate recognition of patterns and correlations across multiple dimensions, which is essential for the preliminary analysis (Wilkinson & Friendly, 2009). In educational research, heatmaps effectively depict variations and trends across different evaluative criteria, making them an invaluable tool for understanding assessments (Borkin et al., 2013). The color gradients in a heatmap range from lighter to darker hues, representing the spectrum of scores or feedback intensity. Typically, cooler colors (e.g., blues) indicate lower scores or less favorable feedback, while warmer colors (e.g., reds) denote higher scores or more positive evaluations. This color-coding aids in quickly identifying areas of concern where expert feedback suggests a need for improvement, as well as strengths where feedback is generally positive. The decision to employ heatmaps for data analysis in this context is strategic. They provide a clear, concise way to compare large volumes of data across multiple evaluative criteria and expert groups. This is particularly valuable where multifaceted feedback must be synthesized to guide revisions and improvements in learning performances and assessment tasks. Heatmaps enable stakeholders to visually digest complex information, promoting easier interpretation and facilitating more informed decision-making.

For the qualitative analysis, a dual approach was used, incorporating both a priori and thematic analysis methods. The a priori method, chosen for its relevance to the structured assessment design, involves using predefined themes or codes established from prior research and theoretical frameworks (Brooks et al., 2015). These codes, which include dimensions such as 3D learning, engagement, language, accessibility, and equity, provided a structured lens for the initial data examination and were detailed in Section 2.4.1. This structured approach allows for focused analysis while accommodating necessary adjustments as the analysis progresses (Crabtree & Miller, 1999). Following the a priori coding, thematic analysis was conducted to identify emergent themes, major concerns, and suggestions not initially anticipated. This stage involved a systematic review of the qualitative data to detect patterns that extend beyond the predefined codes (Braun & Clarke, 2006). This comprehensive approach ensures that the analysis captures both anticipated and emerging insights from the expert feedback.

3.3.3 Stage 3: GPT-Designed Assessment Refinement

Stage 3 aims to respond to the research question 3: *"What is the process of refining GPT-designed knowledge-in-use assessments based on the feedback provided by human experts?*" Drawing from insights gathered in earlier stages, I integrated human experts' feedback into the assessment refinement process. This stage focuses on the iterative refinement cycle, where the feedback from the interdisciplinary expert panel and the insights gained from the cognitive interviews with teachers are utilized to enhance the GPT-4 generated assessments. The outcome of this phase is a customized, domain-specific script that harmonizes the AI's functionalities with the adaptability required by educators to meet the varied needs of students and their teaching objectives. Central to this stage is exploring how to incorporate human experts' feedback with AI to refine the initial designed knowledge-in-use assessment.

3.3.3.1 Data Utilization and Refinement Process

I commence this stage by meticulously reviewing all the feedback obtained from the expert panel in Stage 2, which includes both their numerical ratings and detailed comments. I consider each piece of feedback to determine the most effective way to refine the assessments generated by GPT-4. I also leverage the themes I gained from my thematic analysis to refine the assessments. These adjustments are not merely superficial; they delve into the content, format, and rubrics to ensure that each assessment's integrity and pedagogical goals are maintained, if not enhanced.

3.3.3.2 Expert Re-Evaluation

Once I refined the assessments, I brought back the same interdisciplinary panel from Stage 2 for a re-evaluation. I used the same instruments presented in Stage 2 to collect new feedback. These instruments include Tables 3-3, 3-4 and 3-5. The reason why I decided to use the same evaluation tools that I used before is to ensure I can see clearly if the changes I've made have resolved the concerns the panel originally had. I kept a comprehensive record of every adjustment made to the assessments, including the original feedback that prompted the change and the reasoning behind each decision. This practice is not just for the sake of organization—it's a commitment to transparency and accountability. I want to provide a clear justification for every modification based on the expert input I received. Sticking with the same evaluation tools for this second review allows me to understand the true impact of the refinements. The panel's familiarity with these tools streamlines the process and reinforce the validity of the adjustments made to the assessments.

This phase also included a significant change: the introduction of a new group of experts who were unaware that the assessments and interim products had been co-designed with AI. The decision to form this new expert group was inspired by an interview with a teacher from the initial expert panel, who mentioned, "*I think the tasks are fine, but knowing they are AI-designed, I tend to be more critical compared to tasks designed by humans. It feels like there's less pressure to provide feedback.*" This revelation prompted me to consider the core purpose of the review— to ensure that the evaluations focused solely on the quality of the tasks, rather than the nature of their design process, which could introduce bias. Consequently, assembling a new group of experts who were not informed about the AI involvement was aimed at potentially reducing such bias. Details on the composition of this new group of experts and their backgrounds are provided in Table 3-8.

Table 3-8. New expert panel (blinded) and their backgrounds

Expert	**Expertise**
H	The framework for K-12 education writer; scientist; science content
O	NGSS writing team member; integrating science, language, and computational thinking with a focus on multilingual learners; equity, justice
D	Science assessment
A	Science assessment, teacher education
T	NGSS-aligned science assessment
M	Chemistry education; 3D learning

After this second round of evaluation, I compared the feedback about the first round assessment tasks and the second assessment tasks based on the experts feedback. I also took any additional feedback and made further refinements. This cycle is key to designing high-quality knowledge-in-use assessments. It's a careful iterative process of revision and refinement by incorporating experts' feedback. Ultimately, this stage ensures that the assessments GPT-4 helps us create are not just innovative but also practically useful and pedagogically sound. By marrying the capabilities of AI with the insights of human experts, I aim to create assessments that truly measure what students know and can do.

3.3.3.3 Data Analysis

The analysis of feedback provided by experts was methodically arranged into three primary sections: LPs and Evidence Statements, Task 1, and Task 2. These sections align with evaluations specific to distinct PEs. A blend of qualitative and quantitative methodologies was employed to analyze the data comprehensively.

For the quantitative data analysis stage, I used scatter plot to show the comparison between the first-round review and second round review across multiple expert groups on multiple dimensions. A scatter plot is a type of data visualization used to display the values of typically two variables for a set of data. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis (Cleveland & McGill, 1984). This type of visualization is particularly useful for identifying relationships, trends, or distributions within data, and is widely used in both scientific and business applications to explore potential correlations between variables. In the dissertation, scatter plots are used to display feedback scores across different feedback dimensions between two feedback rounds. Points on the plot show specific scores for a criterion at a particular round, with the vertical axis indicating the score and the horizontal axis listing different criteria. The use of different colors for points, such as red and blue, can differentiate scores from various rounds or groups, facilitating an easy comparison of score changes over time or between groups. Patterns observed, such as clustering of points or vertical dispersion within a single criterion, can suggest consensus among evaluators or significant changes in perceptions of quality over time, respectively. Scatter plots are chosen for their ability to clearly depict relationships and changes between two variables—feedback dimensions and feedback scores across rounds in this case. This visualization is effective for examining data where comparisons over time or between groups are essential for discerning underlying trends and patterns in feedback (Tufte, 2001). Scatter plots also assist in identifying outliers or anomalies, providing a basis for targeted investigation in subsequent qualitative analyses. For the qualitative analysis, I used a similar approach presented above in the section 3.3.2.3.

CHAPTER 4: FINDINGS AND DISCUSSIONS

In this chapter, I present a comprehensive analysis of the findings related to the three research questions, systematically derived from the data analysis process. This section aims to provide a detailed examination of the iterative training of generative AI models, the evaluation of AI-generated assessments by human experts, and the subsequent refinement processes, offering critical insights into the effectiveness and challenges of integrating AI in educational assessment design.

4.1 RQ1. How Can Generative AI Models Be Effectively and Iteratively Trained to Design Knowledge-In-Use Assessments?

To respond to this research question, I used thematic analysis to analyze interactions between the individual human knowledge-in-use assessment design expert and the GPT-4 model. I first present the design process by showing how the interim assessment products and final assessments were created, and then I present the themes identified based on the transcripts to discuss how to iteratively work with GPT-4 to design knowledge-in-use assessments and what kinds of challenges and opportunities this approach brought. In this section, I use PE 3-PS2-1 as an example to show the design process.

This section delineates the strategic exploration of leveraging the GPT to design knowledge-in-use assessments that align with the NGSS. This process was iterative, involving a synergy between human input and AI capabilities to shape the overall assessment design. I started the design process with one of the focal PEs, 3-PS2-1. I opened up each step closely by providing specific guidelines and goals. The following sections present the brief co-designing process for each step.

4.1.1 Unpacking Performance Expectations from the NGSS

The initial phase involved unpacking the three dimensions of the PE to gain a comprehensive understanding of the core concepts embodied in this concise statement. In the preliminary stage of the design process, I imparted foundational information to GPT-4 models about the unpacking process. This covered the goal of domain analysis within the ECD framework, the methodology to unpack the PE's three dimensions, related resources, targeted grade level, and the specific DCI element to be unpacked.

To ensure the unpacking meets the explicit and specific requirements and does not miss any critical sub-ideas of the DCIs, SEPs, and CCCs, I prompted the GPT-4 models by introducing each dimension separately. I first introduced the DCIs in the PE 3-PS2-1. Moreover, I further explicitly pointed out the major sub-disciplinary core ideas for the DCI dimension of the PE. By doing that, I hoped the GPT-4 models could cover comprehensive disciplinary core ideas. In the case of 3-PS2-1, I gave the prompts of two specific DCIs in the PE, which are "*DCI1 in the PE: PS2.A: Forces and Motion - Each force acts on one particular object and has both strength and a direction. An object at rest typically has multiple forces acting on it, but they add to give zero net force on the object. Forces that do not sum to zero can cause changes in the object's speed or direction of motion. (Boundary: Qualitative and conceptual, but not quantitative addition of forces are used at this level.)*" and "*DCI2 in the PE: PS2.B: Types of Interactions - Objects in contact exert forces on each other.*"

After providing the explicit DCIs the PE focuses on, I then gave the guideline and task of how to unpack the PE. Based on the previous unpacking experiences and literature, I provided explicit guidelines for the steps of how to unpack DCIs. For instance, for the DCI1, I gave the prompts of the unpacking process shown in Figure 4-1.

Figure 4-1. Prompts of DCI unpacking

```
{
            "role": "user",
            "content": "Please unpack the DCI1 from the aspects:1. Elaborating the DCI Elements and Sub-ideas, including the sub-ideas that are key to the element(s) of the DCI and the sub-ideas related to each other and to the PE or bundle."
        },
{
            "role": "user",
            "content": "Please unpack the DCI1 from the aspects:2. Defining Boundaries, including the areas of the DCI's element(s) lie outside the scope of learning at this grade band."
        },
{
            "role": "user",
            "content": "Please unpack the DCI1 from the aspects:3. Describing Prerequisite Knowledge, which is knowledge and capabilities (both from this topic and from other topics) do students need in order to achieve an understanding of the element(s) of the DCI."
        },
{
            "role": "user",
            "content": "Please unpack the DCI1 from the aspects:4. Identifying Relevant Phenomena that provide widely accessible examples or applications of the DCI elements and sub-ideas and relatable real-world situations could contextualize the phenomena and create a need for making sense of it."
        },
{
            "role": "user",
            "content": "Please unpack the DCI1 from the aspects:5. Identifying Student Strengths and Challenges, including the everyday experiences have students had that would be relevant for this DCI? And the commonly held student ideas differ in important ways from the scientifically accepted understanding of the DCI"
        },
```

4.1.1.1 DCI Unpacking

I began the unpacking by setting a 500 tokens' length response. However, the response is too concise and cannot provide explicit ideas. Thus, I gradually increased the length of the outputs from 500 tokens, 1000 tokens, to 1500 tokens. And I finally found the 1500 tokens is adequate for offering adequate information that meet the design requirements. I then kept the rest of my conversations with 1500 tokens. Table 4-1 presents the DCIs of the PE 3-PS2-1 unpacking with 1500 tokens. Following the same approach, I prompted the GPT-4 model to unpack the DCI2 of the PE.

My first-round iteration ends up with this version of DCIs' unpacking. Those DCI unpackings follow the specific unpacking procedures. But the degree of unpacking may need to be further addressed. For instance, for the "key sub-ideas", the current unpacking version does not define the idea of "net force," as well as the meaning of how to judge movement is not clearly pointed out. Then, I further provided extra information for the GPT model to guide its unpacking. Being specific, I reiterated the importance of identifying the critical sub-ideas of each DCI, provided extra information for the GPT models to gain enough information about the DCIs, and highlighted the importance of the grade boundary. I provided DCIs' information from the corresponding information of NGSS and *The Framework*. Figure 4-2 presents the prompts I provided to support further unpacking.

Figure 4-2. Prompts for 2nd iteration of DCI unpacking


```
{
            "role": "user",
            "content": "I will provide you: 1. Explicit ideas and understandings of the DCIs; 2. DCI unpacking process and aspects. Please ensure you're unpacking using the unpacking process/ framework I provided."
        },
{
            "role": "user",
            "content": "1.Explicit ideas and understandings of the DCIs: First DCI:PS2.A: Forces and Motion. The major expectation of this DCI in this PE focuses on 'Each force acts on one particular object and has both strength and a direction. An object at rest typically has multiple forces acting on it, but they add to give zero net force on the object. Forces that do not sum to zero can cause changes in the object's speed or direction of motion. (Boundary: Qualitative and conceptual, but not quantitative addition of forces are used at this level.)' Specific understandings of this DCI includes: How can one predict an object's continued motion, changes in motion, or stability? Interactions of an object with another object can be explained and predicted using the concept of forces, which can cause a change in motion of one or both of the interacting objects. An individual force acts on one particular object and is described by its strength and direction. The strengths of forces can be measured and their values compared. What happens when a force is applied to an object depends not only on that force but also on all the other forces acting on that object. A static object typically has multiple forces acting on it, but they sum to zero. If the total (vector sum) force on an object is not zero, however, its motion will change. Sometimes forces on an object can also change its shape or orientation. For any pair of interacting objects, the force exerted by the first object on the second object is equal in strength to the force that the second object exerts on the first but in the opposite direction (Newton's third law). At the macroscale, the motion of an object subject to forces is governed by Newton's second law of motion. Under everyday circumstances, the mathematical expression of this law in the form F = ma (total force = mass times acceleration) accurately predicts changes in the motion of a single macroscopic object of a given mass due to the total force on it. But at speeds close to the speed of light, the second law is not applicable without modification. Nor does it apply to objects at the molecular, atomic, and subatomic scales, or to an object whose mass is changing at the same time as its speed. An understanding of the forces between objects is important for describing how their motions change, as well as for predicting stability or instability in systems at any scale. For speeds that are small compared with the speed of light, the momentum of an object is defined as its mass times its velocity. For any system of interacting objects, the total momentum within the system changes only due to transfer of momentum into or out of the system, either because of external forces acting on the system or because of matter flows. Within an isolated system of interacting objects, any change in momentum of one object is balanced by an equal and oppositely directed change in the total momentum of the other objects. Thus total momentum is a conserved quantity."
        },
{
            "role": "user",
            "content": "Grade Band Endpoints for PS2.A: By the end of grade 2. Objects pull or push each other when they collide or are connected. Pushes and pulls can have different strengths and directions. Pushing or pulling on an object can change the speed or direction of its motion and can start or stop it. An object sliding on a surface or sitting on a slope experiences a pull due to friction on the object due to the surface that opposes the object's motion. By the end of grade 5. Each force acts on one particular object and has both a strength and a direction. An object at rest typically has multiple forces acting on it, but they add to give zero net force on the object. Forces that do not sum to zero can cause changes in the object's speed or direction of motion. (Boundary: Qualitative and conceptual, but not quantitative addition of forces are used at this level.) The patterns of an object's motion in various situations can be observed and measured; when past motion exhibits a regular pattern, future motion can be predicted from it. (Boundary: Technical terms, such as magnitude, velocity, momentum, and vector quantity, are not introduced at this level, but the concept that some quantities need both size and direction to be described is developed.) By the end of grade 8. For any pair of interacting objects, the force exerted by the first object on the second object is equal in strength to the force that the second object exerts on the first but in the opposite direction (Newton's third law). The motion of an object is determined by the sum of the forces acting on it; if the total force on the object is not zero, its motion will change. The greater the mass of the object, the greater the force needed to achieve the same change in motion. For any given object, a larger force causes a larger change in motion. Forces on an object can also change its shape or orientation. All positions of objects and the directions of forces and motions must be described in an arbitrarily chosen reference frame and arbitrarily chosen units of size. In order to share information with other people, these choices must also be shared. By the end of grade 12. Newton's second law accurately predicts changes in the motion of macroscopic objects, but it requires revision for subatomic scales or for speeds close to the speed of light. (Boundary: No details of quantum physics or relativity are included at this grade level.) Momentum is defined for a particular frame of reference; it is the mass times the velocity of the object. In any system, total momentum is always conserved. If a system interacts with objects outside itself, the total momentum of the system can change; however, any such change is balanced by changes in the momentum of objects outside the system."
        },
```

Table 4-1. Iterations of DCIs of PE 3-PS2-1 unpacking

DCI unpacking	
DCI1_PS2.A: Forces and Motion	DCI2_PS2.B: Types of Interactions
Iteration 1	
1. Elaborating the DCI Elements and Sub-ideas: DCI1, PS2.A, focuses on understanding that forces act on objects, these forces have both strength and direction, and how these forces affect an object's motion. Key sub-ideas include: -The concept that multiple forces can act on an object at rest, adding up to zero net force, keeping the object stationary. - The idea that forces which do not sum to zero affect the object\'s speed or direction of motion. 2. Defining Boundaries At this grade band, the focus is on qualitative and conceptual understanding of forces and their effects on motion. Quantitative aspects, such as calculating the exact sum of forces or using formulas to predict motion, are outside the learning scope here. The instructional emphasis is on observable phenomena and conceptual reasoning rather than numerical computation. 3. Describing Prerequisite Knowledge - Students need a foundational understanding of: - Basic motion concepts (e.g., speed, direction). - The idea of "force" as a push or a pull. - Simple understanding of interaction between objects (e.g., contact forces). This knowledge often stems from everyday experiences, like playing on a slide or kicking a ball, and from previous learning segments, possibly in earlier grades, focusing on basic physical concepts. 4. Identifying Relevant Phenomena - Relevant phenomena that could anchor the DCI include:	Unpacking DCI 2 (PS2.B: Types of Interactions) 1. Elaborating the DCI Elements and Sub-ideas: **Key Elements**: The core of PS2.B focuses on the interactions between objects that are in contact, highlighting that these interactions result in exertion of forces on each other. **Relationship with PE or Bundle**: This DCI is directly linked to the focal Performance Expectation (3-PS2-1) as understanding the types of interactions between objects is crucial for planning and conducting investigations to observe the effects of balanced and unbalanced forces. 2. Defining Boundaries: **Scope of Learning for This Grade Band**: At this level, the understanding of interactions is qualitative. Students are expected to recognize and describe the effects of interactions without the necessity for quantitative force calculations. **Outside the Scope**: Detailed mathematical modeling of forces (i.e., using formulas to calculate force magnitudes or directions) is beyond the expected learning outcomes for elementary students. 3. Describing Prerequisite Knowledge: **From This Topic**: Familiarity with general concepts of forces (as introduced possibly in earlier grades) – that they can push or pull objects. **From Other Topics**: Basic observational and reasoning skills to infer that objects in contact can influence each other's motion. Understanding of simple motion (moving, stopping) is also necessary. 4. Identifying Relevant Phenomena: **Accessible Examples**: - Playing on a seesaw: How the weight and position of one child affect the motion of

Table 4-1 (cont'd)

- A book remaining stationary on a table demonstrates balanced forces (gravity pulling down is balanced by the table pushing up). - Pushing a toy car to make it move illustrates how an unbalanced force (the push) changes the object\'s motion. - Observing leaves fluttering to the ground versus a rock falling quickly can spark discussion on how forces act on different objects. These phenomena are relatable and observable in students' everyday environments, making the abstract concepts of forces and motion more tangible. 5. Identifying Student Strengths and Challenges **Strengths**: Students often have intuitive understandings of motion and force from their daily experiences, such as playing sports or moving objects. **Challenges**: - Students may have misconceptions such as believing that only moving objects exert forces or that stationary objects are not subject to forces. - Understanding that forces are vectors, having both magnitude and direction, can be challenging at this level, particularly without grounding in actual experiences.	another. - Pushing a toy car and observing how it interacts with obstacles. - The stopping of a rolling ball when it comes into contact with a wall. **Real-world Situations**: Examples like the friction experienced when sliding down a slide or the way a soccer ball stops when it hits a goalpost help contextualize the phenomena of interactions between objects. 5. Identifying Student Strengths and Challenges: **Relevant Experiences**: Children often have firsthand experiences with the effects of force through play, such as pushing and pulling toys, feeling the resistance of water while swimming, or the effort needed to move heavier versus lighter objects. **Common Challenges**: A prevalent misunderstanding among students may be the notion that only moving objects exert forces, disregarding the forces present in static situations (e.g., a book resting on a table). Another challenge is conceptualizing that force involves an interaction between two objects, not an inherent property of a single object.
Iteration 2	
'Given the provided context and requirements for unpacking the first DCI (PS2.A: Forces and Motion) for the focal performance expectation 3-PS2-1, here's a detailed unpacking tailored for grades 3-5: Aspect 1: Elaborating the DCI Elements and Sub-ideas **DCI PS2.A: Forces and Motion focuses on understanding how forces affect the motion of objects. This includes:** - **Force Characteristics:** Forces are push or pull actions that can cause an object to start moving, stop moving, or change its motion's direction. Each force acts on a particular object with a specific strength (magnitude) and direction.	Based on the provided explicit ideas and understandings of the DCIs for PS2.B: Types of Interactions, here is the unpacking for the Second DCI: PS2.B in 3-PS2-1 according to the requested aspects: 1. Elaborating the DCI Elements and Sub-ideas **DCI Elements and Sub-ideas:** The key elements of PS2.B focus on the nature of the interactions between objects. Primarily, it deals with the concept that objects in contact exert forces on each other. This covers a broad range of forces, including friction, elastic pushes and pulls, and the forces exerted over a distance, such as gravitational, electric, and magnetic forces. The sub-ideas include: - **Friction:** The resistance that one surface or object encounters when moving over

Table 4-1 (cont'd)

- **Interaction of Forces:** Multiple forces can act on an object simultaneously. When the forces are balanced (sum to zero), the object's motion doesn't change; it remains at rest or continues moving at a constant speed in a straight line. Unbalanced forces (those that do not sum to zero) cause changes in the object\'s motion, either accelerating it or altering its direction. -**Newton's Third Law:** The mutual interactions between two objects can be described in terms of equal and opposite forces. When one object exerts a force on a second object, the second object simultaneously exerts a force equal in magnitude and opposite in direction on the first object. Aspect 2: Defining Boundaries **At this grade level, the focus is on:** - Qualitative and conceptual explanations of how balanced and unbalanced forces affect motion, without delving into quantitative analysis or calculations of force magnitudes. - Observations and descriptions of motion changes due to applied forces, without introducing technical terms like "magnitude," "velocity," or "vector." Aspect 3: Describing Prerequisite Knowledge **Students entering grade 3 should already have foundational knowledge, including:** - Basic understanding of motion — knowing that objects move in various ways, such as straight, zigzag, round and round, back and forth, fast and slow. - Familiarity with pushes and pulls as forces that can move objects. Aspect 4: Identifying Relevant Phenomena **Examples of relevant, real-world phenomena related to forces and motion include:** - Playing on a seesaw demonstrates how applying force (pushing down) on one end causes the other end to move up, illustrating action and reaction forces. - Observing how kicking a soccer ball	another. - **Elastic Forces:** The forces exhibited by objects that can be stretched or compressed and return to their original shape. - **Gravitational Forces:** The attraction between objects with mass. Although typically weak on a human scale, it's significant for massive objects like planets. - **Magnetic and Electric Forces:** These forces can be either attractive or repulsive and do not require the objects to be in contact. These sub-ideas are intrinsically linked to each other and to the broader phenomenon of motion and stability, as outlined in 3-PS2-1. Understanding the types and nature of these interactions provides a foundation for predicting and explaining motion. 2. Defining Boundaries **Grade Band Boundaries:** For grades 3-5, the focus is on observable forces and interactions, including friction, elastic forces, and the basic introduction to forces that act at a distance like gravity, magnetism, and static electricity. The boundaries at this level exclude detailed mathematical modeling of these forces (e.g., using Newton's law of universal gravitation or Coulomb's law) and the in-depth exploration of electromagnetic, strong and weak nuclear interactions. 3. Describing Prerequisite Knowledge Students need a basic understanding of motion, including the concepts of pushing and pulling, and that motion can change in speed or direction. Knowledge from earlier grades that objects can affect each other's motion through physical contact. An introductory understanding of magnets and static electricity, as well as the intuitive idea that the Earth pulls objects towards it. 4. Identifying Relevant Phenomena Playing with magnets to feel the attraction and repulsion forces without direct contact. - Observing objects rolling down different surfaces to introduce friction as a force that slows motion.

Table 4-1 (cont'd)

(applying a force) changes its motion depending on the strength and direction of the kick. Aspect 5: Identifying Student Strengths and Challenges **Strengths:** - Many students have practical experience with pushing and pulling objects, which can be leveraged to introduce the concepts of force and motion. **Challenges:** - Students may believe that constant motion requires a constant force application, conflating force with motion. - Some students might think that a larger object always moves slower than a smaller one, regardless of the forces applied, due to misconceptions about mass and force. By exploring and addressing these various aspects, educators can tailor their teaching strategies to effectively convey the foundational principles of forces and motion to elementary students, setting a robust groundwork for more advanced concepts in later grades.	- Using rubber bands to demonstrate how elastic forces work. - Dropping objects to discuss and observe gravity's effect on all objects, regardless of their size. 5. Identifying Student Strengths and Challenges - **Strengths:** Students often have direct, everyday experiences with the types of interactions described in PS2.B, such as playing with magnets, using elastic bands, and observing objects falling. These provide an intuitive grasp of the forces involved. - **Challenges:** Students may have misconceptions about forces, particularly about gravitational and magnetic forces acting at a distance. They might not easily accept that invisible forces can exist without direct contact between objects. Further, the concept that forces can be both push and pull, and the idea that objects can exert forces on each other simultaneously (as in Newton's third law), can be abstract and challenging to understand fully without concrete examples and guided exploration.

After two rounds of iterations, the unpacking of the DCIs became much more explicit compared to the initial generation. For instance, the unpacked sub-ideas for PS2.A evolved significantly. Initially, the unpacking provided a basic understanding: "*The concept that multiple forces can act on an object at rest, adding up to zero net force, keeping the object stationary," and "The idea that forces which do not sum to zero affect the object's speed or direction of motion." However, through iterative refinement, the unpacking became more detailed: "DCI PS2.A: Forces and Motion focuses on understanding how forces affect the motion of objects. This includes Force Characteristics: Forces are push or pull actions that can cause an object to start moving, stop moving, or change its motion's direction. Each force acts on a particular object with a specific strength (magnitude) and direction. Interaction of Forces: Multiple forces can act on an object simultaneously. When the forces are balanced (sum to zero), the object's motion doesn't change; it remains at rest or continues moving at a constant speed in a straight line.*

Unbalanced forces (those that do not sum to zero) cause changes in the object's motion, either accelerating it or altering its direction. Newton's Third Law: The mutual interactions between two objects can be described in terms of equal and opposite forces. When one object exerts a force on a second object, the second object simultaneously exerts a force equal in magnitude and opposite in direction on the first object."

Reflecting on the co-design process of unpacking, it is crucial to emphasize the importance of providing explicit outputs by setting specific task requirements and goals. Moreover, it is essential to equip the GPT models with sufficient domain-specific information to enable accurate analysis and completion of tasks. While GPT models can access general information, they lack depth in domain-specific knowledge unless explicitly provided. Consequently, the depth and appropriateness of the unpacking are limited by the scope of information the GPT model can analyze. Human experts, with their domain content knowledge and unpacking experience, play a critical role in identifying whether the outputs are appropriate or require further revisions. Their judgment ensures that the outputs meet the task requirements. Human experts' reflections are also vital for the iterative training process. For instance, the second iteration of DCI unpacking occurred after generating the initial DCI map, which revealed the unpacking level was insufficient. Subsequently, I re-unpacked the DCIs with more explicit prompts and additional scientific knowledge. This iterative refinement process may vary with different AI models, each having unique capabilities for acquiring training information.

4.1.1.2 SEPs and CCCs Unpacking

To unpack the SEPs and CCCs for the PE, I employed a similar approach as with the DCIs. I began by identifying the specific SEP related to the PE and provided foundational information, including definitions, theoretical foundations, the framework of the SEP, and the grade boundary for the SEP (refer to Figure 4-3). Following this, I outlined specific SEP unpacking principles to guide the GPT-4 models in their analysis (refer to Figure 4-4. This structured approach ensured a comprehensive and detailed unpacking of the SEP and CCC elements, aligning them with the educational standards and goals of the PE.

Figure 4-3. Prompts for SEP (planning and carry out investigations) fundamental information

```
{
            "role": "user",
            "content": "SEP in the PE: PS2.A: Planning and Carrying Out Investigations. Specifically, Planning and carrying out investigations to answer questions or test solutions to problems in 3–5 builds on K–2 experiences and progresses to include investigations that control variables and provide evidence to support explanations or design solutions. Explicit ideas for this SEP in this PE focus on Plan and conduct an investigation collaboratively to produce data to serve as the basis for evidence, using fair tests in which variables are controlled and the number of trials considered.)"
        },
{
            "role": "user",
            "content": "When unpacking the SEPs, an important aim is to identify and define the aspects of performance that comprise each practice and determine the appropriate ways students will engage in the practice at a particular grade or grade band. By “aspect” we mean a necessary part of an SEP that students would need to perform anytime the SEP was put into use. For example, the practice of modeling can involve students in developing a model and/or using a model. An essential aspect of developing a model includes representing the appropriate and necessary elements of a system or phenomenon to help predict the “how” or explain the “why.” Another essential aspect of developing a model includes representing the relationships or interactions among the model elements. Beyond identifying the aspects, we also need to consider how students will be expected to engage with the SEP."
        },
{
            "role": "user",
            "content": "When unpacking the SEP, you also want to consider grade boundary. For example, students in the elementary grades are expected to engage with this practice in ways different from middle school and high school students. For example, in grades K–2, students are expected to develop and use models that take the form of diagrams or physical replicas to represent concrete events or design solutions. In the upper elementary grades, the modeling practice expands to include evaluating and revising simple models to explain or predict phenomena. The expectations become more sophisticated in the secondary grades when students develop, use, and revise more complex models to explain, test, and predict phenomena. These grade band differences are important to document so that the assessment tasks that are developed include the SEP aspects that are appropriate for students and calibrated to their respective grade bands."
        },
{
            "role": "user",
            "content": "When unpacking the SEP, another important aim for unpacking the SEPs is to identify the productive intersections between the target SEPs in a bundle and other complimentary SEPs and then describe how they relate to one another. The value in doing this is that it enables you to earmark a set of SEPs that could potentially be used across a range of tasks for assessing progress toward achieving the PE bundle. Similar to instruction that might engage students with multiple SEPs as they build proficiency with a PE bundle, assessment should also consider the role of other complimentary SEPs for assessing students’ developing proficiencies."
        },
{
            "role": "user",
            "content": "For plan and carry out investigations, scientists and engineers investigate and observe the world with essentially two goals: (1) to systematically describe the world and (2) to develop and test theories and explanations of how the world works. In the first, careful observation and description often lead to identification of features that need to be explained or questions that need to be explored. The second goal requires investigations to test explanatory models of the world and their predictions and whether the inferences suggested by these models are supported by data. Planning and designing such investigations require the ability to design experimental or observational inquiries that are appropriate to answering the question being asked or testing a hypothesis that has been formed. This process begins by identifying the relevant variables and considering how they might be observed, measured, and controlled (constrained by the experimental design to take particular values."
        },
{
            "role": "user",
            "content": "For plan and carry out investigations, planning for controls is an important part of the design of an investigation. In laboratory experiments, it is critical to decide which variables are to be treated as results or outputs and thus left to vary at will and which are to be treated as input conditions and hence controlled. In many cases, particularly in the case of field observations, such planning involves deciding what can be controlled and how to collect different samples of data under different conditions, even though not all conditions are under the direct control of the investigator. Decisions must also be made about what measurements should be taken, the level of accuracy required, and the kinds of instrumentation best suited to making such measurements. As in other forms of inquiry, the key issue is one of precision—the goal is to measure the variable as accurately as possible and reduce sources of error. The investigator must therefore decide what constitutes a sufficient level of precision and what techniques can be used to reduce both random and systematic error."
        },
{
            "role": "user",
            "content": "For the goals of the plan and carry out investigations, by grade 12, students should be able to 1. Formulate a question that can be investigated within the scope of the classroom, school laboratory, or field with available resources and, when appropriate, frame a hypothesis (that is, a possible explanation that predicts a particular and stable outcome) based on a model or theory; 2. Decide what data are to be gathered, what tools are needed to do the gathering, and how measurements will be recorded; 3. Decide how much data are needed to produce reliable measurements and consider any limitations on the precision of the data; 4. Plan experimental or field-research procedures, identifying relevant independent and dependent variables and, when appropriate, the need for controls; 5. Consider possible confounding variables or effects and ensure that the investigation’s design has controlled for them."
        },
{
            "role": "user",
            "content": "For plan and carry out explanation, in the elementary years, students’ experiences should be structured to help them learn to define the features to be investigated, such as patterns that suggest causal relationships (e.g., What features of a ramp affect the speed of a given ball as it leaves the ramp?). The plan of the investigation, what trials to make and how to record information about them, then needs to be refined iteratively as students recognize from their experiences the limitations of their original plan. These investigations can be enriched and extended by linking them to engineering design projects—for example, how can students apply what they have learned about ramps to design a track that makes a ball travel a given distance, go around a loop, or stop on an uphill slope. From the earliest grades, students should have opportunities to carry out careful and systematic investigations, with appropriately supported prior experiences that develop their ability to observe and measure and to record data using appropriate tools and instruments. Students should have opportunities to plan and carry out several different kinds of investigations during their K–12 years. At all levels, they should engage in investigations that range from those structured by the teacher—in order to expose an issue or question that they would be unlikely to explore on their own (e.g., measuring specific properties of materials)—to those that emerge from students’ own questions. As they become more sophisticated, students also should have opportunities not only to identify questions to be researched but also to decide what data are to be gathered, what variables should be controlled, what tools or instruments are needed to gather and record data in an appropriate format, and eventually to consider how to incorporate measurement error in analyzing data."
        },
,
```

Figure 4-4. Prompts of SEP unpacking

```
{
          "role": "user",
          "content": "Please unpack the SEP from the aspects:1. describing the SEP and its essential aspects of performance, including a clear grade-appropriate definition of the SEP and the essential aspects of the SEP that students are to perform."
      },
{
          "role": "user",
          "content": "Please unpack the SEP from the aspects:2. Identifying productive intersections between the SEP and other SEPs that are relevant for building toward the PE or PE bundle, including the productive intersections between this SEP and other SEPs and the related SEPs could students engage with in concert with the DCI and CCC of the PE."
      },
{
          "role": "user",
          "content": "Please unpack the SEP from the aspects:3. specifying the evidence that will be needed to demonstrate the SEP, including the evidence required for students to demonstrate a high level of proficiency with the SEP."
      },
{
          "role": "user",
          "content": "Please unpack the SEP from the aspects:4. describing students' prerequisite knowledge, including the knowledge and capabilities do students need in order to demonstrate the SEP."
      },
{
          "role": "user",
          "content": "Please unpack the SEP from the aspects:5. Identifying Student Strengths and Challenges, including the everyday ways of knowing and doing relate to this SEP and how they are similar and different from the SEP and the common challenges might students encounter as they are developing sophistication in their use of the SEP."
      },
```

Following the initial exploration, I redirected the GPT-4 model to focus on unpacking two specific SEPs: developing and using models, and constructing scientific explanations, rather than the previously explored SEP of planning investigations. I employed similar strategies to prompt the model, beginning by specifying the task goal of unpacking the SEPs and providing fundamental knowledge about each SEP, including definitions, expectations, and grade boundaries. After establishing this foundational understanding, I introduced the SEP unpacking principles outlined to guide the GPT model in meeting the task requirements. The output of the initial unpacking of developing and using models reflected the general information provided in the prompts. Table 4-2 presents the iterations and the resulting unpacking, demonstrating the refinement process and the increasing specificity and depth of the unpacked SEPs.

Table 4-2. The iterations and adjustment of unpacking SEP of Developing and Using Models

<table>
<tr><th>Input information for the task</th><th>Unpacking</th><th>Adjustments for task requirements</th></tr>
<tr><td colspan="3">Iteration 1</td></tr>
<tr><td>Input of the meaning of developing and using models, the applications of scientific models, the grade boundary of developing and using scientific models, and the progressive perspectives of scientific models.</td><td>'Given the depth and breadth of the Science and Engineering Practice (SEP) of "Developing and Using Models," let's unpack this SEP according to the requested aspects:

1. Describing the SEP and Its Essential Aspects of Performance
Grade-Appropriate Definition: Developing and Using Models is about students constructing, analyzing, and working with representations that symbolize natural or designed systems. These models help make predictions, provide explanations, and propose solutions to scientific or engineering problems.
Essential Aspects of the SEP:
Constructing Models to represent physical, conceptual, or mathematical aspects of phenomena.
Analyzing and Interpreting Models to explore and predict behaviors or investigate the mechanisms within the phenomena.
Using Models to Test Solutions or explanations in a simulated, conceptual environment.

2. Identifying Productive Intersections between the SEP and Other SEPs
This SEP can productively intersect with:
Planning and Carrying Out Investigations: Models can be used to design experiments or predict outcomes, guiding investigative processes.
Analyzing and Interpreting Data: Models help in understanding data patterns and making sense of experimental results.
Using Mathematics and Computational Thinking: Developing mathematical or computational models to simulate phenomena.
Engaging in Argument from Evidence: Using models to support or refute arguments based on empirical evidence.
These intersections are crucial for developing a systematic understanding of phenomena in line with the DCI and CCC of the PE.

3. Specifying the Evidence That Will Be Needed to Demonstrate the SEP
Evidence of proficiency includes:
Ability to Build Models: Students create appropriate models of phenomena or systems that they study.
Model Analysis: Students can analyze a model to</td><td>~ The unpacking framework of developing and using models is not explicit enough, including constructing models, analyzing and interpreting models, and using models. However, they're all general larger grain-size of proficiency requirements of scientific modeling practices. When it comes to specific dimensions mentioned here, there should be a more explicit framework to guide students to construct models, or judge models. That is., how to construct a model, what kind of basic criteria that the scientific models should have to enable it to have the explanation function instead of only description. Also, when judging models, what should be the critical aspects that students want to consider in terms of the quality of explanatory ability of the models.

In addition, the unpacking</td></tr>
</table>

Table 4-2 (cont'd)

	explain phenomena, predict outcomes, or test solutions. **Model Revision:** They can iteratively refine their models based on new evidence or understanding. 4. Describing Students' Prerequisite Knowledge Students should have: **Understanding of the Phenomena:** Basic knowledge about the system or phenomena being modeled. **Skills in Observation and Data Collection:** Ability to gather and interpret data that informs the model. **Mathematical Reasoning:** Depending on the model, basic to advanced mathematical skills. 5. Identifying Student Strengths and Challenges **Strengths:** **Creativity in Model Construction:** Students often bring innovative approaches to representing complex ideas. **Intuitive Understanding of Basic Systems:** From their everyday experiences, students often have an intuitive sense of how things work, which can be a good foundation. **Challenges:** **Abstract Thinking:** Moving from concrete representations to abstract models can be challenging. **Complex Systems Understanding:** Grasping the intricacies of complex systems and accurately representing them in models requires development. **Connection with Everyday Knowing and Doing:** Everyday experiences, like using a map or building simple objects, relate to modeling but on a more intuitive level. Educators can build on these experiences to develop a deeper, more systematic approach to modeling scientific phenomena.	dimensions (4, 5, and challenges) are too general without providing explicit information for each aspect of modeling.
	Iteration 2	
To ensure the SEP unpacking meet the requirements of the task that can be explicit enough to guide the evidence statement design letter, I further provided information on the	Adjusted Unpacking for Developing Models (Grades 3-5) **Components of the Practice:** **Model Elements:** Students should focus on identifying basic observable elements (e.g., objects, organisms) and their noticeable attributes to include in their models, emphasizing the inclusion of components that are directly observable or can be researched at an elementary level. **Relationships Among Elements:** At this grade level, emphasis should be on illustrating simple and direct relationships between elements. For example, showing a predator-prey relationship in a food chain or the effect of sunlight on plant growth. **Sequence of Events:** Models by students should demonstrate a basic linear sequence of events that illustrates cause-and-effect relationships relevant to the phenomenon being explored, such as the stages of the water cycle.	The adjusted output aligns with the unpacking principles and framework of developing models, the integrated proficiencies of each sub aspect of the unpacked SEP, and

Table 4-2 (cont'd)

framework of developing models, including identifying the essential components in the system that can affect, mapping out the relationships among components of the model, and leveraging the components and their relationships to explain the corresponding phenomena and solve complex problems. I also provided some exemplar unpacking that were done by human experts. In addition, I asked the unpacking to focus on developing model aspects instead of the other two aspects,	**Correspondence:** Students should be encouraged to make connections between their model elements and real-world observations, using familiar and concrete examples wherever possible. **Limitations:** It's important for students to recognize that their models may not cover all aspects of a phenomenon, focusing on discussing what their model shows well and what it doesn't. **Explanation/Prediction:** Encourage students to use their models to explain simple phenomena or predict outcomes in familiar contexts, like predicting the growth of a plant under different conditions. **Knowledge, Skills, and Abilities:** **Model Elements:** Understanding that models represent real-world phenomena with specific, observable features that can be seen or measured. **Relationships Among Elements:** Ability to connect elements in a model to show their interactions (e.g., through arrows or spatial placement). **Sequence of Events:** Recognizing how one event can lead to another in a straightforward manner. **Correspondence:** Matching elements of the model with real-world counterparts or data they have collected or researched. **Limitations:** Awareness that models are simplifications and cannot capture all details of a phenomenon. **Explanation/Prediction:** Ability to use the model to explain observations or predict simple outcomes based on their understanding of the model. **Evidence for Each Component:** **Model Elements:** Model focuses on the necessary elements needed to explain a target phenomenon at a basic level, related to observable or researched information appropriate for Grades 3-5. **Relationships Among Elements:** Model clearly represents basic relationships or interactions needed to explain the phenomenon, avoiding unnecessary complexity. **Sequence of Events:** Illustrates a clear and correct sequence of events, with each step logically following the previous step in a manner appropriate for the understanding of a student at this grade level. **Correspondence:** Demonstrates a clear connection between model elements and the real-world phenomenon, using examples or data that are accessible to students in Grades 3-5. **Limitations:** Student can articulate basic limitations of the model in explaining the phenomenon, showing awareness of the model's simplicity. **Explanation/Prediction:** Uses the model to construct understandable explanations or predictions relevant to the	challenges and prerequisite knowledge that students may want to equip in order to meet the requirements of the grade boundary of developing model ability. More impressively, the challenges of prerequisite knowledge of developing models were also unpacked at a smaller grain size based on the framework I provided. Reflections: specify the task requirements and provide adequate and appropriate knowledge for the GPT-4 models. Providing exemplar

Table 4-2 (cont'd)

since in my study, I decided to measure students' ability in developing models to explain real-world phenomena.	phenomenon, appropriate for their developmental level. Adjusting the unpacking components for developing, using, and evaluating models to align with the grade boundary of Grades 3-5 involves simplifying the complexity of the concepts, focusing on observable and directly relational aspects of the phenomena, and ensuring that the language and examples are accessible and relevant to the experiences of students in these grades. This adjustment also emphasizes the development of foundational skills in modeling practices that students will build upon in later grades.	examples and specifying the difference between the examples and the previous generation are important for the outputs.

The iterative process of unpacking the SEP of Developing and Using Models involved refining the AI's outputs through multiple rounds of input and feedback, significantly enhancing the quality and detail of the unpacking over time. Initially, the task involved providing the GPT-4 model with fundamental knowledge about the SEP, including its definition, applications, and grade boundaries. The first iteration produced general insights but lacked specificity in guiding students on how to construct and evaluate scientific models. To address these shortcomings, subsequent iterations included more detailed and explicit guidelines, emphasizing the essential components of developing models. This involved clarifying the elements to be included in models, such as identifying observable components, mapping relationships among these components, and understanding the limitations of models. The adjusted approach also focused on developing students' abilities to use models for explanation and prediction, relevant to real-world phenomena.

Human experts played a critical role throughout this iterative process. Their feedback helped identify areas where the AI outputs were too general or did not meet educational standards. By integrating their insights and providing more domain-specific information, the unpacking became more detailed and aligned with educational goals. The iterative nature of the process, coupled with reflective practice, ensured continuous improvement and refinement of the AI-generated outputs.

To enhance the specificity of the unpacking, I provided the system with a framework of the subdimensions of developing models. This framework included the components of the system that need to be modeled, relationships among the components, and leveraging these relationships to explain relevant phenomena or solve complex problems. Additionally, I provided the grade boundaries for this SEP. Iteration 2 in Table 4-2 shows the adjusted unpacking of developing models. The adjusted output aligns with the unpacking principles and framework of developing models, integrating the proficiencies of each sub-aspect of the unpacked SEP and addressing the challenges and prerequisite knowledge students need to meet the grade boundary requirements. The challenges of prerequisite knowledge were unpacked in more detail based on the provided framework.

Reflecting on the process, it is clear that specifying task requirements and providing adequate and appropriate knowledge for the GPT-4 model were crucial. Providing exemplar examples and highlighting the differences between iterations significantly improved the quality of the outputs. This iterative and reflective approach ensured that the AI-generated unpacking were both comprehensive and aligned with educational standards, setting the stage for unpacking the CCCs in the next section.

Following a similar approach, I also prompted the GPT-4 model to unpack the other SEP focused on in this study, which is constructing scientific explanations and argumentations. Additionally, two CCCs were unpacked using the same method. Table 4-3 presents the final versions of the unpacking for both the SEPs and CCCs.

Table 4-3. The final output of unpacking of constructing scientific explanations and CCCs.

SEP/CCC	**Unpacking**
Constructing scientific explanations	**Components of Scientific Explanation:** 1. **Claim:** In the context of Grades 3-5, a claim would be a student\'s concise answer to a question about a scientific phenomenon or design solution, articulated in a manner understandable by their peers. For example, a student might claim, "Plants grow faster in sunlight than in the dark." 2. **Evidence:** At this level, evidence should be based on direct observations, simple measurements, or patterns discovered in data. Students might collect evidence about plant growth in sunlight versus darkness through regular measurement of plant height. Visual aids, such as photographs of the plants at various stages, might also be considered as supporting evidence. 3. **Reasoning:** Reasoning for students in Grades 3-5 involves linking the evidence to the claim in a straightforward manner, perhaps with the assistance of

Table 4-3 (cont’d)

	basic scientific principles (like photosynthesis). An example of reasoning at this level might be, "Plants need sunlight to perform photosynthesis, which helps them grow. Our plants in the sun grew taller, which supports our claim." **Knowledge, Skills, and Abilities for Constructing Explanations:** - Knowledge of the basic structure of a scientific explanation (claim, evidence, reasoning). - Skills in observing, measuring, and recording data accurately. - Ability to make straightforward observations and connect these observations to broader scientific concepts. - Ability to articulate a simple claim based on observations or findings. - Ability to use simple scientific vocabulary to explain how their evidence supports their claim. **Evidence for Each Component of the Practice:** **For Claims:** Students articulate a clear and testable claim that answers a question about a phenomenon or solves a problem. **For Evidence:** Students present specific observations, measurements, or identified patterns as evidence. This could include simple data tables, charts, or descriptive observations. **For Reasoning:** Students logically explain why the evidence they have gathered supports their claim, possibly linking to broader scientific ideas appropriate for their grade level. In adjusting these components for Grades 3-5 students, emphasis is placed on the clarity of expression, the directness of the observed evidence, and the basic logic linking evidence to claims. The complexity of the scientific principles involved is scaled to be age-appropriate, with a focus on fostering students\' ability to think critically and reason scientifically at an elementary level.
Evaluate scientific explanations	**Claim:** **Grade 3-5 Focus**: Determine if the claim provides a direct answer to the question based on what was observed or found in the investigation. **Evidence Requirement**: Checks if the claim makes sense given the observations or data collected during class activities or simple experiments. **Evidence:** **Grade 3-5 Focus**: Evaluate if the evidence includes observations or data that directly supports the claim. This could include things seen, counted, measured, or read about in trusted sources. **Evidence Requirement**: Verifies that students are not just listing observations but are connecting them explicitly to the claim as supportive evidence. **Reasoning:****Grade 3-5 Focus**: Assess whether the explanation includes basic scientific concepts or ideas that connect the claim and evidence. This might involve simple cause-and-effect relationships or observations of change. **Evidence Requirement**: Looks for statements where students explain "why" their evidence supports their claim, using age-appropriate scientific language and concepts.

Table 4-3 (cont’d)

	For elementary students, evaluating explanations often revolves around their ability to: **Understand and apply basic scientific vocabulary and concepts** related to the topic at hand. **Make clear connections** between what they observed (evidence) and what they claim those observations mean. **Use simple reasoning** to articulate why their evidence supports their claim, drawing on foundational scientific principles they\'ve learned. The focus is on fostering foundational skills in scientific inquiry, such as making observations, drawing conclusions from data, and beginning to articulate the rationale behind these conclusions with basic scientific reasoning. This adjusted unpacking aims to scaffold these practices in ways that are accessible and meaningful for Grades 3-5 students, preparing them for more complex scientific thinking in later grades.'
Cause and effect	Causes: **Correctly identifies** or describes basic cause(s) that lead to an observable effect(s), often in a simple and direct relationship. Effects: **Correctly identifies** simple and observable effect(s) that result from a specific cause(s), understanding that the same cause may not always lead to the same effect due to different conditions. Conditions: **Correctly recognizes** or describes, in basic terms, the conditions under which certain cause-and-effect relationships hold or don't hold, acknowledging that different outcomes can arise from changes in conditions. Mechanism/Intermediate Events:**Able to provide a basic explanation** of how one event leads to another, using simple scientific concepts where applicable. Identifies basic chains of events or actions that link a cause to its effect without necessarily needing to use specific scientific terminology. Evidence:**Describes or provides** simple observations or data that they have gathered or been given as evidence to support a causal claim. Understands that evidence is needed to support the link between a cause and its effect. Predictions/Theories:**Makes simple predictions** based on observed cause-and-effect relationships, understanding that if we know the cause, we can predict the effect or vice versa. Begins to see how these predictions can connect to broader scientific ideas or theories, even at a rudimentary level. Solutions:**Designs basic solutions** to simple problems by applying an understanding of cause and effect. For instance, if plants are not growing, recognizing that water (cause) might be necessary to affect growth (effect), and proposing watering the plant as a solution.

4.1.2 Generating Integrated Dimension Map for the PE

After unpacking the dimensions, the next step is guiding the GPT-4 model to generate an Integrated Dimension Map (IDM) to facilitate the subsequent learning performance design. The IDM should integrate the three dimensions (DCIs, SEPs, and CCCs) based on the unpacking, using the DCI as the main framework. This involves mapping out the major ideas and then adding the unpacked aspects of the SEPs and CCCs to link the different DCI components together, which will guide the learning performance design.

To illustrate this process, I use PE 3-PS2-1 as an example, demonstrating how I worked iteratively with the GPT-4 model to design the IDM. Table 4-4 presents the iterative process of co-designing the IDM for PE 3-PS2-1.

Table 4-4. Iterative process of co-designing IDM for PE 3-PS2-1

Input	**Output**	**Reflections**
	Iteration 1	
- provide rationale and purpose of IDM; - provide design principles of designing IDM; - provide explicit design process of generating IDM	To lay out and make the connections between the DCI sub-ideas for Performance Expectation (PE) 3-PS2-1, we will focus on the Disciplinary Core Ideas (DCIs) PS2.A: Forces and Motion, and PS2.B: Types of Interactions. The goal is to create a map that demonstrates how these sub-ideas relate within the context of balanced and unbalanced forces affecting the motion of an object. Step 1: Lay Out the DCI Sub-ideas Based on previous unpacking: 1. Forces and Motion (PS2.A): Describes how each force acts on a particular object and has both strength and a direction. This DCI also touches upon how balanced forces result in no change in motion, whereas unbalanced forces cause a change in speed or direction of motion. 2. Types of Interactions (PS2.B): Explains that objects in contact exert forces on each other. Diagram Representation: In the diagram:- Text Box 1:Forces and Motion (PS2.A): Each force acts on an object with strength and direction. Balanced forces sum to zero, causing no change in motion, while unbalanced forces result in changes in motion. - Text Box 2: Types of Interactions (PS2.B): Objects in contact	The layout of the DCI sub-ideas does not really reflect the DCI unpacking, which just lists the two major ideas listed in the PE that were asked to be unpacked before. In other words, the DCI map does not really include the unpacked sub ideas of the major DCIs in this PE.

Table 4-4 (cont'd)

	exert forces on each other. Arrows and Statements: Arrow from PS2.A to PS2.B: The direction and strength of forces (PS2.A) determine the interactions (contact/non-contact) between objects (PS2.B). Arrow from PS2.B to PS2.A: The interactions between objects (PS2.B) cause forces to be exerted, affecting motion (PS2.A). This setup focuses on the reciprocity between forces and interactions: how forces determine the nature of interactions, and inversely, how interactions modify the forces experienced by objects. Furthermore, understanding that balanced forces result in no motion change, while unbalanced forces result in motion change, forms a critical connection point between these sub-ideas. Generation Description: The created diagram lays out two main sub-ideas: "Forces and Motion" and "Types of Interactions" related to the performance expectation of investigating the effects of balanced and unbalanced forces. The relationship between them is illustrated via directional arrows indicating that the characteristics of forces (strength and direction) critically influence how objects interact with each other, and how these interactions, in turn, affect the forces that influence motion. This cyclic relationship underscores the interconnectedness of these core concepts in explaining how forces affect the motion of objects.'	
	Iteration 2	
For this iteration, I prompted the GPT-4 models to regenerate the IDM by: 1. explicitly pointing out the issues of the generated IDM. 2. provide specific sub-DCI ideas unpacked before in the prompt to	'For the Performance Expectation (PE) 3-PS2-1, we start by laying out the key Disciplinary Core Ideas (DCIs) that were identified in the unpacking of this PE, focusing on "Forces and Motion" (PS2.A) and "Types of Interactions" (PS2.B). Given this, the diagram for laying out and connecting the DCI sub-ideas could look as follows: 1. Text Boxes for DCI Elements and Subideas: - Force - Balanced Force - Unbalanced Force - Motion - Types of Interactions 2. Arrangement and Connections: - Force is connected to **Balanced Force** and **Unbalanced Force** to indicate that these are two categories or outcomes of forces acting on objects. - Balanced Force is connected to **Motion** with a statement indicating that balanced forces result in no change in motion.	The subideas of PS2.A are unpacked and presented based on the previous guidance. However, the ideas of PS2.B were not presented in the same way, which

Table 4-4 (cont'd)

model how to construct the DCI map; 3. providing examples to design DCI maps.	- **Unbalanced Force** is connected to **Motion** with a statement indicating that unbalanced forces cause a change in the object\'s speed or direction of motion. - **Types of Interactions** is connected to both **Balanced Force** and **Unbalanced Force** to show that the types of interactions between objects can result in either balanced or unbalanced forces affecting the motion. 3. **Succinct Statements for Each Arrow:** - From **Force** to **Balanced Force**: "When equal forces act on an object from opposite directions, they are balanced, causing no change in motion." - From **Force** to **Unbalanced Force**: "Forces acting on an object that do not equal out result in an unbalanced force, causing a change in motion." - From **Balanced Force** to **Motion**: "Balanced forces result in no change in an object's motion." - From **Unbalanced Force** to **Motion**: "Unbalanced forces cause a change in the speed or direction of an object's motion." - From **Types of Interactions** to **Balanced Force**: "Interacting objects can exert balanced forces on each other." - From **Types of Interactions** to **Unbalanced Force**: "Interactions between objects can lead to unbalanced forces, affecting motion." This diagram represents the relationships among the key concepts related to forces and motion as elaborated in the unpacking of DCI for PE 3-PS2-1. It visually illustrates how different types of forces and interactions between objects influence motion, serving as a foundation for integrating the SEPs and CCCs in step 2 of the mapping process. This structured approach ensures a clear understanding of the disciplinary content before integrating practices and crosscutting concepts.'	may reflects the prompts should be explicit and comprehensive enough to direct to detect the major requirement, analyze the information, take actions to complete the tasks and then learn from this process to transfer learning from this case to the following similar tasks.
Iteration 3		
1.Reiterated the importance of referring back to the IDM development procedure. 2. Address incorporate	'Given the detailed unpacking provided for PS2.A: Forces and Motion and PS2.B: Types of Interactions, let\'s synthesize these elements into a coherent DCI map explicitly designed for the grade 3-5 boundary, focusing specifically on the PE 3-PS2-1. **DCI Map Description for 3-PS2-1** **DCI Elements and Key Sub-ideas Association: **Force**: Central to both PS2.A and PS2.B, representing both acted upon and resulting actions between objects. Sub-ideas include strength, direction, balanced forces, unbalanced forces, and the types of interactions (contact vs. at-distance) that can generate force. **Motion**: Tied closely with force from PS2.A, emphasizing changes in an object's speed or direction caused by unbalanced forces. **Types of Interactions**: From PS2.B, focuses on the specific	The revised DCI map meets the requirements of the task.

Table 4-4 (cont'd)

the unpacked DCIs and their sub ideas. 3. Ask the GPT model to provide rationale of how and why it changed the IDM. 4. Point out the importance of addressing the grade appropriateness. 5. Emphasize the importance of concise and clearness for the IDM and visualization	nature of forces that objects exert upon each other, including gravitational and contact forces such as friction. **Explicit Connections:** 1. **From Force to Motion**: An arrow from "Force" to "Motion" can be labeled "Unbalanced forces cause changes," reflecting the understanding that any non-zero sum of forces (unbalanced forces) can directly alter an object\'s state of motion, either in speed or direction. This connection satisfies the core of PS2.A as it applies to the expectations for grades 3-5. 2. **From Types of Interactions to Force**: Multiple arrows come out from "Types of Interactions," each pointing towards "Force." These arrows can be labeled with the specific interaction types—gravitational, electric, magnetic, and contact. Each label serves to illustrate that these interactions are the source of forces acting upon objects. **Visual Representation:** Imagine the map structured with "Types of Interactions" at the top, branching out to "Force" in the center, and then leading down to "Motion" at the bottom. This visual flow guides the understanding from the source of forces through interactions, their nature and characteristics in the form of force, and finally to the effects of these forces as observed in motion. This structure directly ties back to the focal PE by illustrating the cause-and-effect relationship pivotal for understanding how different forces impact motion. **Revision Observations:** Given the PE\'s grade level (3-5), the map is tailored to emphasize qualitative understandings over quantitative. Thus, while the map outlines the types of forces and their effects on motion, it intentionally avoids delving into complex equations or concepts (e.g., Newton\'s laws in detail) that are beyond the recommended grade boundary. This revision ensures alignment with the grade-specific expectation that students can identify and describe forces and their effects but may not mathematically quantify those forces or effects. This DCI map provides a visual and conceptual foundation for designing learning experiences and assessments that align with the PE 3-PS2-1, ensuring grade-appropriate depth and encouraging students to explore the cause-and-effect relationships between forces and motion in their physical world.'	

The iterative process for generating the IDM for PE 3-PS2-1 was essential for refining and enhancing the quality of the output. Initially, the process began with providing the GPT-4 model with a rationale and purpose for the IDM, along with design principles and an explicit design process. The first iteration's output, however, lacked specificity and did not adequately reflect the detailed unpacking of the

DCIs, indicating the need for more precise guidance. Subsequent iterations addressed these shortcomings by providing detailed sub-ideas unpacked from the DCIs and examples of how to construct a comprehensive DCI map. This approach aimed to improve the alignment of the IDM with educational standards and specific task requirements. The second iteration showed progress but still did not fully integrate the detailed sub-ideas of PS2.B, revealing the need for even more explicit and comprehensive prompts. Further iterations emphasized referring back to the IDM development procedure, incorporating the unpacked DCIs and their sub-ideas comprehensively. This also involved asking the GPT model to provide a rationale for changes made to the IDM, ensuring the adjustments were grade-appropriate and clearly visualized. By reiterating these critical aspects, the final iteration successfully produced a detailed and coherent DCI map that met the task's requirements and educational standards.

Through the iterations, I directed the GPT models to generate the DCI map. Afterward, I prompted the GPT models to enrich the DCI map by following the guidelines of adding unpacked SEP and CCC of the PE. I provided prompts outlining the requirements for integrating SEPs and CCCs into the DCI map to generate the IDM. Figure 4-5 presents the prompts I provided for generating the IDM.

Figure 4-5. Prompts for enriching the DCI map by adding SEPs and CCCs to generate the IDM

```
{
        "role": "user",
        "content": "You just complete the first step of generating Integrated Dimension Map. Now let's move to the next step of constructing integrated
Dimension Map. Step 2: integrate the SEPs and CCCs with the DCI sub-ideas to complete the map: The relationships between sub-ideas comprise the backbone of the
integrated dimension map. Once the DCI terrain is mapped, we draw from both the first and second focus areas of SEP and CCC unpacking to identify SEPs and CCCs that
can elicit students' proficiency for each connection made between sub-ideas. These SEPs and CCCs are oftentimes the same ones that are represented in the PE bundle
and, accordingly, we refer to the SEP and CCC focus areas (SEP focus area 1 and CCC focus area 1) to earmark essential dimensional aspects that can work for each
connection. But sometimes other SEPs and CCCs are equally or more appropriate for engaging a specific disciplinary relationship expressed in the map. In these
instances, we draw from SEP focus area 2: Identifying productive intersections between the SEP and other SEPs and CCC focus area 2: Identifying productive
intersections between the CCCs and between CCCs and the SEPs. Note that the selection of the most appropriate SEPs and CCCs might also be informed by the scope and
sequence of curriculum materials or students' level of experience with particular dimensions. Please integrate the SEPs and CCCs with the DCI sub-ideas of 3-PS2-1
to complete the map based on the procedure of Step 2 and refer back to the previous unpackings of developing and using models, constructing explanations, and Caused
and effect. For example, an example of an integrated dimension map for the PE Bundle of MSPS1-2 and MS-PS1-5. Note that each connection between sub-ideas includes a
DCI relationship statement that conceptually links the sub-ideas and also includes the SEP and CCC dimensions to show knowledge-in-use. For example, the DCI
relationship statement, "chemical reactions produce new substances", links the two sub-ideas of chemical reactions and substances. From our SEP unpacking -focus
area 2 – we identified the practice of constructing explanations to be particularly aligned to this DCI relationship because a chemical reaction is a phenomenon
that can be explained based on evidence about the starting and ending substances. For the DCI relationship linking substances and characteristic properties, we
identified the practice of analyzing and interpreting data because students can analyze data on characteristic properties of substances to determine whether a
chemical reaction occurred. In this way, we identified two complementary SEPs that could be integrated with the disciplinary relationships between chemical
reactions, the nature of substances, and characteristic properties, one of which is an anchor SEP from the PE bundle and the other which is drawn from our short
list of related SEPs. When you generate the integrated dimension map, please ensure: 1. Add unpacked SEP and CCC to the designed DCI map of 3-PS2-1; 2. Utilize the
unpacked SEPs(developing and using models, and constructing explanantions) and unpacked CCC (Cause and Effect) to complete the integrated dimension map; 3. Provide
explicit description of your generated integrated dimension map."
    },
```

In the prompts, I first clarified the task goals, which were to generate the IDM by adding SEPs and CCCs to the DCI map. This integration aimed to clarify how the relationships among various sub

DCI ideas could be connected and developed by incorporating appropriate SEPs and CCCs, ensuring alignment with previous unpackings. The IDM was required to cover and meet the performance expectations for PE 3-PS2-1. After clarifying the task goals, I provided specific design principles to guide the GPT model in generating the IDM. Additionally, I offered an example of how to generate the IDM to facilitate learning. I directed the GPT model to refer back to the previous SEP and CCC unpacking to think about the integration. Finally, I further clarified three main goals for the task: to generate a clear IDM, to align with the grade boundaries, and to ensure explicit descriptions of the generated IDM. The iterative refinement resulted in a robust IDM for PE 3-PS2-1, presented in Table 4-5.

Table 4-5. IDM for PE: 3-PS 2-1.

Integrated Dimension Map for 3-PS2-1:
****DCI Elements Sub-ideas Association:**
Force and **Motion**: Central concepts from PS2.A, detailing that forces can cause changes in the motion of an object, emphasizing the strength and direction of forces, and differentiating between balanced and unbalanced forces.
Types of Interactions: From PS2.B, elaborating on the nature of forces that objects in contact exert on each other, such as gravitational, electric, magnetic, and contact forces, like friction.
****Integration with SEPs and CCC:**
1. **Force to Motion (PS2.A):
SEP: Developing and Using Models – Students can model how unbalanced forces affect an object's motion, providing a visual and conceptual understanding of the principles such as direction and strength of forces.
CCC: Cause and Effect – Highlighting that unbalanced forces (cause) lead to changes in motion (effect).
2. **Types of Interactions to Force (PS2.B):
SEP: Planning and Carrying Out Investigations – Students can investigate different types of forces (e.g., gravity, magnetism) to understand their effects on objects. They can explore, for example, how the distance between objects affects the strength of the force.
SEP: Constructing Explanations – Students explain how specific interaction types (gravitational, electric, magnetic, and contact) cause forces that can alter an object's motion.

Visual Representation: The map visually presents **Types of Interactions** at the top, branching out to **Force** in the center, which then leads down to **Motion**. The flow demonstrates how interactions lead to forces, and how these forces influence motion. Arrows from each type of interaction towards Force are labeled with specifics like "gravity causes a pull on objects" or "contact like friction resists motion," each tied back to investigations or models students can create or use. The Cause and Effect CCC explicitly connects all elements of the map, underscoring that each interaction and the subsequent application of force lead to observable changes in motion as effects.

Additional Notes: This revised map capitalizes on the natural curiosity of students in grades 3-5 to explore the physical world through hands-on investigations (Planning and Carrying Out Investigations), to visualize complex ideas through the development and usage of models (Developing

Table 4-5 (cont'd)

and Using Models), and to articulate their understanding through the construction of scientific explanations. The CCC of Cause and Effect serves as the thematic backbone of the curriculum content, intertwining it with real-world observations and promoting a deeper understanding of the fundamental principles of physics as they pertain to forces and motion.'

4.1.3 Generating Learning Performances Based on Unpacking and IDM

Upon concluding the unpacking of the three dimensions and creating the IDM, I proceeded to generate Learning Performances (LPs) that encapsulate the essential competencies students need to display to meet the PE. These LPs are detailed, three-dimensional statements of performance that act as practical indicators of student progression towards the PE, combining diverse disciplinary facets in an integrated manner. Each LP represents a crucial component of a PE that students need to master, ensuring their steady progression towards achieving the entire PE. The LP design process began with a briefing about the concept of LPs and their relationship to the PE. Following this introduction, I prompted GPT to conceive LPs for PE 3-PS2-1. The iterations and reflections on how I worked with the GPT model to design the LPs for 3-PS2-1 are presented in Table 4-6. This iterative approach ensured that the LPs developed were comprehensive, aligned with the unpacked dimensions, and reflected the integrated nature of the PE.

Table 4-6. Iterative process of designing LPs for PE 3-PS2-1

Input	**Output**	**Reflections**
Iteration 1		
1. clarify the task goal, which is designing LPs for 3-PS2-1 based on the unpacking and IDM; 2. provide sufficient information about LPs, including the definition of LPs and the significance of LPs. 3. provide specific guidelines for doing the task, which is how to articulate LPs based on the IDM. I provided specific clarifications about the designing process. 4. provide further explicit elaborations of the designing process by providing some examples of LPs and further clarify the relationship between LPs and PE. 5. provide guidelines on how to determine whether the LPs are needed to be included or excluded from the set of LPs. 6. provide criteria of judging the quality of LPs to allow the GPT models to reflect on the generations 7. provide guidelines	Given the unpacking and integrated dimension map for the PE 3-PS2-1, focusing on "Plan and conduct an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object," here are three learning performances articulated: Learning Performance 1:**Students planning and conducting investigations (SEP)** to observe and record the effects **of balanced and unbalanced forces (DCI - PS2.A)** on the motion of an object in the playground, such as a ball or a swing, and then **identify cause and effect relationships (CCC)** based on their observations. **Rationale:** This learning performance addresses the SEP of planning and carrying out investigations by directly engaging students in hands-on activities to explore the core idea of forces and motion. It incorporates the CCC of cause and effect, as students are asked to draw connections between the forces applied and the resulting motion of objects. This learning performance has a smaller grain size compared to the PE, focusing on a practical exploration of the DCI within a specific, easily observable context. Together with others, it covers part of the PE by providing foundational experiences with forces. Learning Performance 2:**Students developing and using models (SEP)** to represent **how objects in contact exert forces on each other (DCI - PS2.B)** and then **explaining how these interactions can cause changes in motion (CCC)** using everyday materials like magnets or friction blocks. **Rationale:** This learning performance allows students to apply the SEP of developing and using models to the DCI relating to types of interactions, specifically focusing on contact forces. It includes the CCC of cause and effect as students are expected to connect the interactions to changes in motion, aligning with the "how" and "why" questions at the heart of scientific inquiry. This performance contributes to covering the PE by deepening students' understanding of	1. The grain size of the generated LPs are not small enough to differentiate them from the original PE, especially for the LP1. 2.Each LP includes multiple DCI sub ideas that make the LPs too broad. For instance, LP3 discusses both the strength and directions of force in one LP, which does not align with the goals of LP. 3. The structure of the LPs are clear and the three dimensions integrated very well.

Table 4-6 (cont'd)

for the format of the LPs in terms of their structures. 8. further clarify the task goals and provide specific requirements for the outputs, including asking to provide rationale of the generated LPS based on the criteria I provided.	the variety of interactions that can influence motion. Learning Performance 3:**Students constructing explanations (SEP)** for **how the strength and direction of forces (DCI - PS2.A)** affect an object's motion, using evidence from classroom experiments involving pushing and pulling toys to recognize **patterns (CCC)** in motion. **Rationale:** Engaging students in constructing explanations enables them to connect empirical evidence to the scientific theory of forces and motion. This learning performance targets a specific aspect of PS2.A, focusing on the quantitative attributes of forces and their qualitative effects. By recognizing patterns in motion as a CCC, students are encouraged to abstract from specific instances to the underlying principles governing motion. This performance pushes towards the comprehensive understanding required by the PE, demonstrating an integrated approach to three-dimensional learning. Learning Performance 4:**Students using mathematics and computational thinking (SEP)** to predict **the effects of unbalanced forces on an object\'s speed and direction (DCI - PS2.A)**, and understanding that **these changes in motion can be predicted based on cause and effect (CCC)**. **Rationale:** This learning performance incorporates computational thinking, an essential skill for interpreting and predicting complex systems\' behavior in science. By focusing on unbalanced forces, which is a key concept in PS2.A, and linking it to mathematical predictions, students engage deeply with the disciplinary core idea. The cause and effect CCC reinforces the focus on the scientific reasoning behind motion changes. This performance culminates aspects of the PE concerning prediction and investigation, rounding out the set of performances for comprehensive coverage of the PE\'s scope.	
	Iteration 2	
1. Point out the issues of the designed LPs, including the scope and grain size of the LPs (e.g., LP1 is too similar to the original PE.) 2. Direct GPT model to	**Revised Learning Performances for 3-PS2-1 LP1**: Students plan and carry out investigations (SEP) to observe how different strengths of forces (PS2.A) affect the motion of an object, highlighting the cause and effect (CCC) relationship between the force applied and the change in motion LP2**: Students develop models (SEP) to represent	1. The CCCs in the LPs are not integrated well into the LPs. 2. need to

Table 4-6 (cont'd)

reflect its' generations and provide justifications on how the generations align with requirements set up before) 3. prompt the FPT model to consider revision 4. provide exemplar LPs to further enrich the fundamental information for GPT to learn from.	scenarios where balanced forces (PS2.A) act on an object, resulting in no change in motion, such as a book on a table or a balanced tug-of-war, to illustrate the cause and effect (CCC) in physical interactions. LP3**: Through constructing explanations (SEP), students use evidence from investigations to describe how objects in contact exert forces on each other (PS2.B), including friction, showing the interaction (CCC) between objects as a cause of motion changes. LP4**: Students describe (SEP) how magnetic or gravitational forces (PS2.B) act at a distance, using models to explain the cause and effect (CCC) of these non-contact interactions on an object's motion or change in motion.	correct the SEP expression (describe, represent, etc.) 3.I provided exemplary revisions to revise the LPs.
	Iteration 3	
1. provided the revised the LPs to the GPT model to consult its' opinions 2. Prompted the GPT model to review, judge, and revise the LPs if needed.	Given the focal PE (3-PS2-1) and the unpackings provided for PS2.A: Forces and Motion and PS2.B: Types of Interactions, along with the integrated dimension map and the nature of the science and engineering practices (SEPs) and crosscutting concepts (CCCs) involved, let's address your inquiries about the four revised Learning Performances (LPs). **Do the Four LPs Make Sense?** - Yes, the four LPs make sense within the context of 3-PS2-1. They are structured to cover different aspects related to forces and their effects on motion, addressing both contact and non-contact forces in a manner that is suitable for students in grades 3-5. 2. **How Do the Four LPs Align with the Unpacking and Integrated Dimension Map?** - The LPs directly relate to the key concepts identified in the unpacking of DCIs PS2.A and PS2.B. For example, LP1 and LP2 focus on the effects of balanced and unbalanced forces on motion, which is at the heart of PS2.A. LP3 explores the interactions between objects in contact, including friction, aligning with PS2.B. LP4 expands this understanding to non-contact forces such as magnetic and gravitational forces, which, while more subtly hinted at in PS2.B's unpacking for the elementary level, serves as an important bridge to understanding the various types of forces that affect motion. - Each LP incorporates SEPs and CCCs appropriately—LP1 uses planning and carrying out investigations to observe effects (a SEP), highlighting	LP1**: Students plan and carry out investigations to observe how different strengths of forces affect the motion of an object. LP2**: Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength

Table 4-6 (cont'd)

	cause and effect (a CCC); LP2 and LP4 incorporate model development (a SEP) for explaining phenomena (a CCC); LP3 emphasizes constructing explanations (a SEP) based on evidence from investigations, involving interaction (PS2.B) and cause and effect (CCC). 3. **Do These Four LPs Cover the Scope of PE 3-PS2-1?** - Collectively, these LPs adequately cover the scope of PE 3-PS2-1. They explore the core concepts of forces and motion, including the nuanced distinction between contact and non-contact forces and their effects on an object's motion—elements crucial for students' understanding of this PE. They provide a comprehensive approach to building students' understanding through investigations, modeling, and constructing explanations. 4. **Further Review or Revision?** - Regarding LP4 and the question on the criticality of 'non-contact forces' for this PE—Yes, non-contact forces are essential for a full comprehension of PS2.B: Types of Interactions and, by extension, PE 3-PS2-1. While the focus at the grade 3-5 band is often on direct, observable phenomena, introducing the concept of non-contact forces like gravity and magnetism helps broaden students' understanding of forces beyond what they can see and touch. It lays a foundational understanding that not all forces require direct contact to influence motion, preparing students for more complex physics concepts later. Thus, LP4 is indeed on the right track as it engages students with the concept of non-contact forces acting at a distance—essential for grasping the broader spectrum of forces that influence the motion of objects, in line with PS2.B's emphasis on types of interactions, including those that do not require direct contact. Considering the analysis, these LPs appear well-constructed and aligned with the goals of PE 3-PS2-1, emphasizing hands-on, exploratory learning that integrates key scientific concepts with practice and crosscutting concepts in a manner appropriate for the targeted grade level.	on both sides. LP3**: Students construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes. LP4**: Students use models to explain how non-contact forces (e.g., gravitational forces) on an object's motion or change in motion act at a distance.

The iterative process of designing LPs for PE 3-PS2-1 involved several rounds of refining inputs and outputs to develop comprehensive and educationally aligned LPs. The initial step was to clarify the task goals, which included designing LPs based on the unpacking and IDM of the PE. I began by providing the GPT-4 model with sufficient information about LPs, including their definition, significance, and their relationship with the PE. This involved giving specific guidelines on how to articulate LPs based on the IDM, providing examples of LPs, and clarifying the criteria for judging the quality of LPs. Table 4-7 presents the evolving version of the designed LPs for 3-PS2-1.

Table 4-7. The evolving version of the designed LPs for 3-PS2-1

1st iteration	2nd iteration	3rd iteration (final version)
LP1: Students planning and conducting investigations (SEP)** to observe and record the effects **of balanced and unbalanced forces (DCI - PS2.A)** on the motion of an object in the playground, such as a ball or a swing, and then **identify cause and effect relationships (CCC)** based on their observations.	LP1: Students plan and carry out investigations (SEP) to observe how different strengths of forces (PS2.A) affect the motion of an object, highlighting the cause and effect (CCC) relationship between the force applied and the change in motion.	LP1: Students plan and carry out investigations to observe how different strengths of forces affect the motion of an object.
LP2: Students developing and using models (SEP)** to represent **how objects in contact exert forces on each other (DCI - PS2.B)** and then **explaining how these interactions can cause changes in motion (CCC)** using everyday materials like magnets or friction blocks.	LP2: Students develop models (SEP) to represent scenarios where balanced forces (PS2.A) act on an object, resulting in no change in motion, such as a book on a table or a balanced tug-of-war, to illustrate the cause and effect (CCC) in physical interactions.	LP2: Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides.
LP3: Students constructing explanations (SEP)** for **how the strength and direction of forces (DCI - PS2.A)** affect an object's motion, using evidence from classroom experiments involving pushing and pulling toys to recognize **patterns (CCC)** in motion.	LP3: Through constructing explanations (SEP), students use evidence from investigations to describe how objects in contact exert forces on each other (PS2.B), including friction, showing the interaction (CCC) between objects as a cause of motion changes.	LP3: Students construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes.
LP4: Students using mathematics and computational thinking (SEP)** to predict **the effects of unbalanced forces on an object's speed and direction (DCI - PS2.A)**, and understanding that **these changes in motion can be predicted based on cause and effect (CCC)**.	LP4: Students describe (SEP) how magnetic or gravitational forces (PS2.B) act at a distance, using models to explain the cause and effect (CCC) of these non-contact interactions on an object's motion or change in motion.	LP4: Students use models to explain how non-contact forces (e.g., gravitational forces) on an object's motion or change in motion act at a distance.

In the first iteration, the generated LPs were too broad, with some being nearly indistinguishable from the original PE. For example, the first LP focused on planning and conducting investigations but did not sufficiently narrow the scope to differentiate it from the PE. Similarly, some LPs integrated multiple DCIs, making them too extensive. Despite these issues, the structure of the LPs was clear, and the integration of the three dimensions (DCIs, SEPs, and CCCs) was evident. Reflecting on these outputs, I identified the need for more explicit guidance.

In the second iteration, I provided specific prompts to address the scope and grain size of the LPs, guiding the model to reflect on its outputs and consider necessary revisions. By offering exemplar LPs and further enriching the information for the GPT-4 model to learn from, the quality of the generated LPs improved. The revised LPs were more focused and better aligned with the educational goals. They included: Planning and conducting investigations to observe how different strengths of forces affect motion; Developing models to explain how balanced forces result in no change in motion; Constructing explanations based on evidence to describe how objects in contact exert forces on each other; and Using models to explain how non-contact forces, such as gravitational forces, affect motion.

In the final iteration, I reviewed and judged the revised LPs, ensuring they were aligned with the unpacking and IDM. This involved confirming that the LPs covered the scope of PE 3-PS2-1 and appropriately integrated the three dimensions. The final set of LPs demonstrated a thorough understanding of forces and motion, exploring both contact and non-contact forces in a manner suitable for students in grades 3-5. The iterative process allowed for continuous refinement, with each cycle building on the previous one, integrating feedback, and enhancing the outputs' quality.

4.1.4 Design Blueprints ---- Evidence Statements

The generation of LPs laid the groundwork for the next phase of designing knowledge-in-use assessment: the development of Design Blueprints, which encompassed crafting Integrated Proficiencies (IPs), Evidence Statements, Essential/Characteristic Task Features, Variable Task Features, and considerations for Equity/Fairness (Harris et al., 2019). These blueprints would guide the development of knowledge-in-use assessment tasks targeting the LP. In this process, GPT was provided with comprehensive information about the components of the design blueprints, their interrelationships, and their alignment with the LP. In this section, I presented the iterative design process of generating evidence statements for each LP for the PE 3-PS2-1. Prior to generating evidence statements, designing integrated proficiencies can provide guiding information for designing evidence statements. But due to the similar contents, this section only focuses on the design of evidence statements. Table 4-8 presents the iterative design process of generating the evidence statements for the LPs for PE: 3-PS2-1.I take LP2 as an example to show the design process.

Table 4-8. Iterative process of designing evidence statements for PE 3-PS2-1

Input	Output	Reflections
Iteration1		
1. Recall the integrated proficiencies designed before for each LP; 2. Define the task goals; which is designing evidence statements for each LP based on the designed integrated proficiencies; 3. Discuss the critical process of designing evidence statements for each learning performance in knowledge-in-use assessments. Emphasize the importance of evidence statements in assessment and rubric design, detailing how they describe the necessary evidence or performance to evaluate if students have achieved the learning outcomes. 4. Explain how integrated proficiencies, encompassing the abilities students need to demonstrate a learning performance, guide the identification of required evidence. Describe the process of determining observable evidence to support proficiency claims in learning performances, using the abilities described in integrated proficiencies to specify this	**Integrated proficiencies** are: **IP1.** Ability to develop models illustrating that when equal forces act on an object from opposite directions, the motion of the object does not change. **IP2.** Ability to explain within models how balanced forces maintain an object's state of motion or rest. **IP3.** Ability to apply these models to everyday scenarios showing the effect of balanced forces. **Evidence statements** for LP2: **Evidence Statement 1 for IP1**: Students construct a model (e.g., diagram, physical representation) that visually demonstrates forces of equal strength acting on an object from opposite directions, effectively showing that	1. The evidence statements generated for LP2 reflect a certain level of proficiency performance that students should demonstrate. For instance, the performance indicator for evidence statement 1 specifies that the model constructed by students could be either a diagram or a physical representation. This model should demonstrate the forces acting on an object and the relationships among these forces, which should be equal in strength and opposite in direction. By integrating the relationships among the components and the components themselves, the model should explain how these forces cause the object's motion to remain unaffected. Generally speaking, evidence statement 1 includes the required performance evidence for LP2. However, to make it more accessible for measurement, it is crucial to further unpack evidence statement 1 into a smaller grain size.

Table 4-8 (cont'd)

evidence in evidence statements format. 5. Provide a detailed guide on creating evidence statements for knowledge-in-use assessments, emphasizing their significance in assessment and rubric design. Explain the role of evidence statements in determining whether students have achieved the learning outcomes by describing the necessary evidence or performance. Focus on how integrated proficiencies—combining SEPs, DCIs, and CCCs—inform the creation of these evidence statements. 6. Provide examples of how to design evidence statements; 7. Provide specific design guideline for designing evidence statements	the object's motion remains unaffected. **Evidence Statement 2 for IP2**: Within their models, students provide annotations or explanations that detail why an object at rest stays at rest or an object in motion maintains its motion when subjected to balanced forces. **Evidence Statement 3 for IP3:** Students apply their models to real-life situations (e.g., a book resting on a table, two people pushing a box from opposite sides with equal force) to demonstrate understanding of how balanced forces operate in everyday experiences.	2. Evidence statements 2 and 3 focus on using or applying models to explain phenomena. However, for third graders, the critical abilities in developing models include: clearly identifying and including relevant components of the model in the system that can be used to explain the phenomena; clearly discussing or presenting the relationships among the components; and collectively using the components and relationships among the components of the model to explain relevant phenomena.
Iteration 2		
1. Point out that the evidence statements 1 needs further explanations and needs to explain explicitly about what the components of the models are; their relationships and how they sue them collectively to explain the phenomena 2. Here is the prompts I provided for the GPT model: *revise these evidence statements below for LP2 : "Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides." for 3-PS2-1 to ensure they are specifically for 3rd grade*	**Evidence Statement 1 for LP2:** Students construct a model that visually and accurately represents an object under the influence of balanced forces, showing the object either in a static position or moving at a constant speed without any change in direction. Students create a simple drawing or diagram that shows an object (e.g., a book on a table, two children in a tug-of-war with a rope) where no movement is happening or where there's movement at a steady pace. The model should visually display forces acting equally from all directions, keeping the object still or moving without speeding up or slowing down. **The model should:** · Basic shapes or figures to represent the object(s) and the forces acting on them. · Arrows of equal length pointing in opposite directions to signify balanced forces. A brief label or symbol next to each arrow, indicating the type of force (e.g., push).	The revised evidence statements do not only reflect the major performance for developing models, but also include specific components, relationships among the components, and also provide the rationale for each evidence statement.

Table 4-8 (cont'd)

specifically for 3rd grade students' their developing model abilities, including identify and include the essential components of their model that could be used to explain the phenomena; the relationships among the components of the model by using labels or symbols; and finally use the components and the relationships among the components together of the developed model to explain the corresponding phenomena. Your evidence statement should include explicit evidence of what and how the students should provide/ perform to ensure they have the ability and skills. revise these evidence statements: "1. Students construct a model that visually and accurately represents an object under the influence of balanced forces, showing the object either in a static position or moving at a constant speed without any change in direction. 2. Students include annotations or keys in their model to delineate between balanced and unbalanced forces, explaining scenarios where the forces cancel out, resulting in zero net force on the object. 3. Students apply their model to real-life scenarios like a book on a table or a car cruising at a steady speed, explaining how these situations exemplify balanced forces resulting in no change in motion. "	**Evidence Statement 2 for LP2:**Students include annotations or keys in their model to delineate between balanced and unbalanced forces, explaining scenarios where the forces cancel out, resulting in zero net force on the object. In their models, students use annotations, keys, or simple sentences to highlight the presence of balanced forces and compare these with potential unbalanced force scenarios. For example, they might annotate a tug-of-war drawing with "equal pull from both sides keeps the rope still." **The model should:** · Clearly mark parts of their model to show where forces are acting. Use simple vocabulary to describe how these forces are balanced or what might happen if they weren't (e.g., "If one side pulls harder, the rope moves that way"). **Evidence Statement 3 for LP2:** Students apply their model to real-life scenarios like a book on a table or a car cruising at a steady speed, explaining how these situations exemplify balanced forces resulting in no change in motion. Students select a familiar scenario (such as a book remaining on a table without falling off or a toy car rolling at a steady pace) and explain how their model illustrates the balance of forces at play in that scenario. **The explanation should:** - Identify the forces acting on the object in the scenario (e.g., gravity pulling the book down, table pushing it up). - Describe how these forces balance out, using elements from their model as reference points - Conclude how the balanced forces result in no change in the object's motion (either staying still or moving steadily).	

Reflecting on the evidence statements design process, it is easy for the GPT model to generate general outputs. Although the outputs are still domain-specific and three-dimensional, without providing explicit and clear prompts, it is not easy for the GPT model to generate outputs that align with the requirements. However, with appropriate prompts, especially with specific framework guidelines (e.g., components of model, relationships among models, and explanation using the components and relationships among the components), the GPT models can usually generate outputs that meet the requirements. This reflects the potential of using GPT models to design evidence statements. But human experts still need to monitor the process to ensure the outputs meet the task goals.

4.1.5 Design Blueprints ---- Essential and Variable Task Features

Upon delving into the design process of essential task features for measuring the LP, our approach mirrored that of the LP design. Essential task features, also known as characteristic task features, are key attributes shared by all tasks aimed at assessing a particular claim. These features serve as the foundation for creating tasks that effectively measure the LP. Like the LP design process, I guided the GPT in generating essential task features by providing appropriate prompts and refining the responses based on limitations. Essential task features aim to answer questions such as "What are the task features that must be present to assess this claim?" and "What are the common features that all tasks need to include?" These features encapsulate the attributes shared by all tasks that assess a specific learning performance. The variable task features emphasize their importance in adjusting task difficulty and ensuring accessibility and fairness for all students. To facilitate this understanding, I posed thought-provoking questions such as "What are the features that can vary among tasks?" and provided context pertaining to the targeted LP. GPT's initial responses were insightful, suggesting modifications in interactive systems, types of evidence, scaffolding levels, response modes, collaboration levels, contextualization, language, and representation. These proposals demonstrated its understanding of tailoring task complexity and accessibility to individual learning styles and proficiency levels.

The iterative process of designing both essential and variable task features involved continuous feedback and refinement. This ensured that the final task designs were robust, equitable, and aligned with educational standards. For the essential task features and variable task features for LP2, refer to Table 4-9 below.

Table 4-9. Essential task features and variable task features for LP2

Essential Task Features
Task presents a scenario where an object is under the influence of balanced forces, resulting in no change in motion.

- Example: A book resting on a table or a tug-of-war game with equal strength on both sides.

Task provides data or observations from investigations highlighting the impact of balanced forces on an object's motion.

- Example: Data showing a stationary object with forces acting equally in opposite directions or an object moving at constant speed.

Task prompts students to use evidence from the provided data or observations to construct a model demonstrating balanced forces.

- Example: Students use arrows to represent forces acting on an object and explain the absence of motion change.

Task includes prompts for students to explain at a conceptual level how balanced forces result in no change in motion, encouraging them to connect evidence and reasoning.

- Example: Prompts asking students to describe why an object remains stationary or moves at a constant speed when forces are balanced.

Variable Task Features
Scenario Variety Task scenarios can vary by the type of objects and the nature of forces acting on them (e.g., different weights, types of forces like gravity and normal force).

- Example: Objects of varying mass, different surfaces, or forces such as gravity and tension.

Scenario Variety Task scenarios can vary in the complexity of the investigations (e.g., analyzing balanced forces in different situations such as a hanging picture, a floating balloon).

- Example: Different levels of difficulty in understanding balanced forces in static and dynamic contexts.

Modes of Representation: Tasks can vary in the mode of expression for students' models and explanations (e.g., written descriptions, oral presentations, multimedia presentations, or physical models).

- Example: Allowing students to choose how to present their understanding, such as through drawings, digital tools, or physical demonstrations.

Scaffolding Levels: Tasks can include different levels of scaffolding, such as guiding questions, partial models, or diagrams for students to complete.

- Example: Providing templates with partial models that students need to complete or questions that guide their thought process.

Scaffolding Levels: Tasks can adjust the demand for background knowledge related to physics concepts of force and motion.

- Example: Varying the complexity of the explanations required or providing additional resources and support for students with less background knowledge.

Equity and Inclusion Considerations
Offer scenarios that reflect a diversity of experiences to ensure all students find the task relatable. Ensure that the language and content are accessible and respectful to all students, promoting an inclusive learning environment.

The essential task features and variable tasks features were also designed by the collaboration between the human experts and the GPT models. Ensuring the clear task goals and providing explicit requirements to complete the tasks, and making sure human experts timely judge the output are critical for the design of task features.

Despite its comprehensive response, GPT's grasp of equity considerations, especially cultural relevance and linguistic accessibility, was not robust. While it advised leveraging students' background knowledge and experience, it didn't elaborate extensively on this. Likewise, it suggested language complexity adjustment and multilingual resource integration but didn't sufficiently address diverse learners' needs. To remedy these deficiencies, I supplied additional prompts centered on cultural relevance and linguistic accessibility. I solicited more in-depth responses regarding cultural and local integration in task design and how to customize language complexity for diverse learners. This process of iterative prompting aimed to enhance the inclusivity of generated task features and refine ChatGPT's ability to align with equitable educational practices.

4.1.6 Design blueprint for LP2 of PE 3-PS2-1

Synthesizing all the unpacking, LPs, evidence statements, essential task features and variable task features, Table 4-10 present the final version of the design blueprint of LP2 for PE 3-PS 2-1. This design blueprint for LP2 guides the task design and was sent out for the first round expert review.

Table 4-10. Design blueprint for LPP2 of PE 3-PS2-1

PE	**3-PS2-1: Plan and conduct an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object.** [Clarification Statement: Examples could include an unbalanced force on one side of a ball can make it start moving; and, balanced forces pushing on a box from both sides will not produce any motion at all.] [*Assessment Boundary: Assessment is limited to one variable at a time: number, size, or direction of forces. Assessment does not include quantitative force size, only qualitative and relative. Assessment is limited to gravity being addressed as a force that pulls objects down.*]
Focal LP: LP2	Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides.
Evidence Statements	1. Students construct a model that visually represents an object under the influence of balanced forces, showing the object either in a static position or moving at a constant speed without any change in direction. **The model should:** ● Basic shapes or figures to represent the object(s) and the forces acting on them. ● Arrows of equal length pointing in opposite directions to signify balanced forces. ● A brief label or symbol next to each arrow, indicating the type of force (e.g., "push," "pull," "gravity").
	2. Students include annotations or keys in their model to delineate between balanced and unbalanced forces, explaining scenarios where the forces cancel out, resulting in zero net force on the object. **The model should:** ● Clearly mark parts of their model to show where forces are acting. ● Use simple vocabulary to describe how these forces are balanced or what might happen if they weren't (e.g., "If one side pulls harder, the rope moves that way").
	3. Students apply their model to real-life scenarios like a book on a table or a car cruising at a steady speed, explaining how these situations exemplify balanced forces resulting in no change in motion. **The explanation should:** ● Identify the forces acting on the object in the scenario (e.g., gravity pulling the book down, table pushing it up). ● Describe how these forces balance out, using elements from their model as reference points. ● Conclude how the balanced forces result in no change in the object's motion (either staying still or moving steadily).

Table 4-10 (cont'd)

Essential task features	● Task presents a scenario where an object is under the influence of balanced forces, resulting in no change in motion. ● Task provides data or observations from investigations highlighting the impact of balanced forces on an object's motion. ● Task prompts students to use evidence from the provided data or observations to construct a model demonstrating balanced forces. ● Task includes prompts for students to explain at a conceptual level how balanced forces result in no change in motion, encouraging them to connect evidence and reasoning.
Variable task features	● **Scenario Variety** Task scenarios can vary by the type of objects and the nature of forces acting on them (e.g., different weights, types of forces like gravity and normal force). ● **Scenario Variety** Task scenarios can vary in the complexity of the investigations (e.g., analyzing balanced forces in different situations such as a hanging picture, a floating balloon). ● **Modes of Representation**: Tasks can vary in the mode of expression for students' models and explanations (e.g., written descriptions, oral presentations, multimedia presentations, or physical models). ● **Scaffolding Levels**: Tasks can include different levels of scaffolding, such as guiding questions, partial models, or diagrams for students to complete. ● **Scaffolding Levels**: Tasks can adjust the demand for background knowledge related to physics concepts of force and motion.
Equity and inclusion considerations	● Offer scenarios that reflect a diversity of experiences to ensure all students find the task relatable. ● Ensure that the language and content are accessible and respectful to all students, promoting an inclusive learning environment.

4.1.7 Task Design

The design process for constructing assessment tasks with GPT began by providing explicit instructions and guidelines based on the defined LP for which the tasks were intended. To ensure alignment with the LP, both essential and variable task features were communicated to GPT, enabling the development of multiple tasks within a 'family' that maintained fidelity to the LP while allowing for variations in variable task features.

The process was initiated by introducing the task goals and explaining how to utilize previously generated information to guide the task design. Detailed information regarding LP2 was shared with GPT, along with design principles and guidelines on how to design knowledge-in-use assessment tasks.

Specific requirements for the assessment tasks were also provided, encompassing various aspects such as the purpose of design blueprints, task characteristics, designing process steps, task design objectives, task scenarios and prompts, alignment with learning performances, variability in tasks, phenomena representations, equity and inclusion, creativity and motivation in task design, relevance of phenomena, connection with students, developmental appropriateness, three-dimensional integration features, engagement and interest, ethical practices, coherent narrative, language and accessibility, phenomena, and scenarios, and the creative process. The design steps included understanding each element of the blueprint and its intended collaboration, considering potential phenomena that match the blueprint elements and are universally relevant to students, integrating equity and inclusion considerations, and ensuring three-dimensional integration features. These task requirements were incorporated based on the critical aspects of the assessment design, including the three-dimensional nature of the assessment, language and accessibility, engagement and relevance, and more.

4.1.7.1 Assessment Task Design for LP3

After providing the task design requirements and guidelines, the assessment design process began. Using LP3 for PE 3-PS2-1 as an example, the iterative design process was demonstrated. To facilitate understanding, Table 4-11 presents the design blueprints for LP3.

Table 4-11. Design blueprints for LP3 of 3-PS2-1

LP3: Students construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes. **Integrated Proficiencies (IPs) for LP3** **IP1:** Ability to construct claims about the effects of contact forces, especially friction, on motion. This proficiency involves students identifying friction as a force that opposes motion and affects the speed and direction of moving objects. **IP2:** Ability to select and use evidence from investigations to substantiate claims about friction's role in motion alterations. Students should demonstrate competency in choosing relevant experimental or observational data that clearly show how friction alters motion. **IP3:** Ability to reason regarding how friction as a contact force instigates changes in the motion of objects. This entails connecting evidence to claims in a reasoned manner to explain the mechanism by which friction influences motion.

Table 4-11 (cont'd)

Evidence Statements for LP3 1.**Students formulate claims** that pinpoint friction as a significant force influencing motion during contact between objects. 2.**Students compile appropriate evidence** from experimental or observational data showcasing friction's impact on motion. 3.**Students logically link their selected evidence with their claims**, providing explanations on how friction modifies motion. **Essential Task Features** 1. **Scenarios**: Incorporate real-world contexts (e.g., a soccer ball slowing down on grass vs. concrete) that vividly depict friction's role in modifying motion. 2. **Data and Observations**: Use findings from simple, age-appropriate experiments (e.g., sliding various materials across different surfaces) that allow students to observe friction's effects. 3. **Construct Explanations**: Engage students with prompts that guide them to explain motion changes using evidence of friction from provided or gathered data. **Variable Task Features** 1. **Scenario Variability**: Include diverse situations demonstrating friction (e.g., braking a bicycle, pushing a book across a desk) to cover a broad understanding. 2. **Scaffolding Levels**: Offer scaffolded supports like guiding questions (What happens to the object's speed when it moves across different surfaces?) or data interpretation aids to facilitate evidence-based explanations. 3. **Modes of Response**: Allow students to present their explanations through various formats, such as short essays, oral presentations, posters, or digital slides. **Equity and Inclusion Considerations** 1. **Relatable Scenarios**: Choose scenarios that are universally relatable and do not presume specific prior experiences, ensuring all students can connect with the task's context. 2. **Accessible Language**: Utilize language that is clear, straightforward, and devoid of unnecessary jargon, with instructions presented both verbally and visually to cater to diverse learning preferences. 3. **Varied Expression**: Enable students to demonstrate their understanding in different ways, accommodating various learning styles and needs, and providing equal assessment opportunities.

After the above task design requirement and design principles, I provided the design blueprints of LP3 to the GPT 4 models. Table 4-12 presents the iterative design process for designing one assessment task for LP3 using the design blueprints presented in Table 4-10.

Table 4-12. First round iterative design process for LP3 relevant task design

Input	Output
1. Task design requirements and principles 2. Learning performance design blueprints	**Assessment Task 1: "Sliding and Stopping"** **Scenario**: The class has just learned about the force of friction and its effects on the motion of objects. The teacher sets up two different surfaces on the classroom floor: a smooth, polished wooden plank and a carpet. Students have toy cars and are challenged to investigate how the type of surface affects the distance a car travels after being pushed with the same force. **Prompts**: 1. **Engage**: Students predict which surface will allow the car to travel further and explain their reasoning. 2. **Investigate**: Students push the toy cars across both surfaces from the same starting line, using a consistent push force. They measure the distance traveled on each surface three times and record their findings. 3. **Explain**: Based on their observations, students construct claims about the role of friction in changing the car's motion on different surfaces.

The designed task aligns with the LP3, whose SEP is constructing explanations. However, while the designed tasks included the third prompts of explaining the role of friction in the investigation, the tasks do not only emphasize the SEP rather than emphasizing conducting investigations. One reason why the tasks were designed in this way could be the LP3 statement "*construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes.*" The GPT model may design the tasks aligning with the statements from conducting investigation to provide explanation. However, it could dilute the critical role of constructing explanation in this LP.

Another interesting observation is the two designed tasks are both about friction, which could because of the LP statement mentioned the friction. This still reflects the linear and algorithmic nature of the machine approach. The other reason why the two tasks and the LP3 include construct explanations following up with the investigation activity is because of the previous explanations about the progressive order of the LPs for this PE. Overall, the designed tasks are decent in addressing the LP3, but the scenarios are not engaging enough and the prompts do not uncover the framework of constructing explanations with appropriate scaffoldings.

Based on the observations and the goals of designing assessment tasks to measure LP3, I further prompted the GPT-4 models, which led to the second round of designed assessment tasks, see Table 4-13.

Table 4-13. Second round iterative design process for LP3 relevant task design

Input	Output
● Refine tasks to better emphasize the 3D nature of learning. ● Develop clear, engaging scenarios and prompts for each task, making them relatable to students' daily experiences and clearly introducing the concept of friction. ● Provide specific prompts that guide students to construct claims, gather evidence, and apply reasoning to explain how friction affects motion. ● Include a brief rationale for each task, explaining how it assesses students' understanding of LP3 and engages them in applying scientific knowledge to real-world problems. ● Ensure tasks adhere to the LP3 design blueprint, making them accessible, inclusive, and accommodating diverse learning styles and abilities. ● Promote equity and inclusion by designing tasks that are accessible and relatable for all students, providing necessary scaffolding to support diverse needs and successful engagement.	**Task 1: The Playground Slide Challenge** **Scenario**: Amy and Ela are at the playground, and Amy decides to slide down the big, shiny slide. She notices that when the slide is dry, she zips down quickly, making her giggle with joy. However, when the slide becomes slightly wet after a quick sprinkle, her speed slows down, and she doesn't reach the bottom as fast. Ela, watching from below, tilts her head, curious about what's happening. **Prompt**: Help Amy and Ela understand why the slide's surface condition affects Amy's sliding speed. Think about the slide and Amy's socks: 1. **Claim**: Make a guess (hypothesis) about why Amy slides slower on the wet slide compared to the dry slide. 2. **Evidence**: Use what you observed about Amy sliding on the dry and wet slide as evidence. 3. **Reasoning**: Explain how the water on the slide might change how Amy's socks rub against the slide's surface, causing the change in speed.

In the second iteration, the tasks were refined to better emphasize the 3D nature of learning. The scenarios became more engaging and relatable, connecting directly with students' daily experiences. This shift made the tasks more meaningful and increased student motivation. Clear, specific prompts were developed to guide students through the process of constructing claims, gathering evidence, and applying reasoning. This structured approach is crucial for helping students develop their explanatory skills and understand the role of friction in motion changes. Including a rationale for each task helped to articulate how the tasks assess students' understanding of LP3, ensuring they align with educational goals and standards. For instance, for the task above the rationale the GPT-4 model provided is "*This task immerses students in a familiar scenario that vividly illustrates friction's role in motion changes, aligning with the DCI about objects in contact exerting forces, including friction (LP3). Encouraging explanations based on observable phenomena reinforces the SEP of constructing explanations. The task highlights 'cause and effect' (CCC) by linking the wetness of the slide (cause) with the change in sliding speed (effect).*" The tasks were also refined to adhere more closely to the LP3 design blueprint, ensuring they were accessible, inclusive, and accommodating diverse learning styles and abilities. This focus on equity and inclusion was critical in making the tasks relatable and achievable for all students, providing necessary scaffolding to support diverse needs and promoting successful engagement.

Table 4-14. Two assessment tasks for LP3

<table>
<tr><td colspan="2">LP3: Students construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes.</td></tr>
<tr><td>Task 1: The Playground Slide Challenge

Scenario: Amy and Ela are at the playground, and Amy decides to slide down the big, shiny slide. She notices that when the slide is dry, she zips down quickly, making her giggle with joy. However, when the slide becomes slightly wet after a quick sprinkle, her speed slows down, and she doesn't reach the bottom as fast. Ela, watching from below, tilts her head, curious about what's happening.

Prompt: Help Amy and Ela understand why the slide's surface condition affects Amy's sliding speed. Think about the slide and Amy's socks:
● **Claim**: Make a guess (hypothesis) about why Amy slides slower on the wet slide compared to the dry slide.
● **Evidence**: Use what you observed about Amy sliding on the dry and wet slide as evidence.
● **Reasoning**: Explain how the water on the slide might change how Amy's socks rub against the slide's surface, causing the change in speed.</td><td>Task 2: Classroom Carpet Carper

Scenario: During a classroom activity, Amy rolls a small toy car across different surfaces - the smooth, shiny classroom floor and the fluffy classroom carpet. On the smooth floor, the car travels far and fast, but on the carpet, it quickly slows down and stops. Ela, always ready for fun, watches intently, her tail wagging with excitement.

Prompt: Explore with Amy and Ela why the toy car moves differently on the carpet than on the smooth floor. Consider the surfaces and the car's wheels:
● **Claim**: Predict why the car goes further on the smooth floor than on the carpet.
● **Evidence**: Discuss what happened when the car moved across the different surfaces
● **Reasoning**: Explain how the fluffiness or smoothness of each surface might affect the car's wheels and its motion.</td></tr>
</table>

Table 4-14 presents the final two tasks for LP3. Reflecting on the entire iterative process of using design blueprints to design assessment tasks for LP3, the process revealed significant insights and areas for enhancement. Initially, the tasks were broadly focused on both investigation and explanation, aligning with the LP3 requirements. However, this broad approach risked diluting the primary focus on constructing explanations. The tasks were subsequently refined to emphasize constructing explanations, ensuring alignment with LP3's core objectives. Engagement and relevance emerged as critical factors. The initial scenarios, while functional, were critiqued for lacking engagement. The second iteration introduced more relatable and vivid scenarios, making the tasks more engaging and enhancing student motivation. This shift underscored the importance of creating contextually meaningful tasks to foster deeper student engagement. Another key improvement was in the scaffolding provided to students. The initial tasks lacked sufficient guidance, which was addressed in the second iteration by including specific prompts that guided students through making claims, gathering evidence, and applying reasoning. This structured approach is crucial for developing students' explanatory skills in scientific contexts. Equity and inclusion were also explicitly considered in the refined tasks. The second iteration aimed to make the tasks accessible and relatable to a diverse student population, providing scaffolding and considering different learning styles and needs. This approach supports all students in successfully engaging with and understanding the content. The use of GPT-4 in designing these tasks demonstrated both strengths and limitations of AI in educational contexts. While the AI-generated tasks aligned with LP3, the approach sometimes exhibited a linear focus, particularly on the concept of friction due to its mention in LP3. This highlights the need for human oversight to ensure a balanced and nuanced approach to task design.

Overall, the iterative process led to more structured, engaging, and inclusive tasks that better aligned with LP3 goals. This process underscored the importance of continuous refinement, thoughtful scenario design, and adequate scaffolding in creating effective educational tasks. The insights from the AI-generated tasks also provide valuable lessons for integrating AI in educational design, emphasizing the need for human-AI collaboration to achieve optimal outcomes.

4.1.7.2 Assessment Task Design for LP2

Following the training of designing tasks for LP3, I prompted the GPT 4 models to design two tasks for LP2. Since this step is a follow up step with designing tasks for LP3, I provided the similar prompts with the prompts of LP3 but emphasize the critical reflections from the above exploration process. The iterative process of designing tasks for LP2 is shown in Table 4-15.

Table 4-15. The iterative process of designing tasks for LP2

Input	Output	Reflections
Iteration 1		
1. Recall the critical aspects of designing knowledge-in-use assessment tasks, including task requirements, goals, and critical aspects that need to be further taken care of. 2. Recall the information, LP statement, and the design blueprints for LP2.	**Task 1: "The Case of the Static Skateboard"** **Scenario**: Jamie and Alex are at the park with a skateboard. They each push on the skateboard from opposite ends with equal force. The skateboard doesn't move from its spot, no matter how hard they push. **Prompt**: ● **Model Construction**: Draw a model to represent the forces Jamie and Alex are applying to the skateboard. Show why the skateboard doesn't move using arrows to represent forces. ● **Explanation**: Explain in your drawing how the forces from Jamie and Alex are balanced and describe what that means for the motion of the skateboard. ● **Application**: Think of another example where balanced forces result in no change in motion and describe it below your model. **Task 2: "Balancing Act on the Sea-saw"** **Scenario**: In the playground, Mia and Kai are on a seesaw. They position themselves in such a way that the seesaw is perfectly balanced and horizontal. Neither side is moving up or down. **Prompt**: ● **Model Construction**: Sketch a seesaw in balance. Add Mia and Kai to your model and use arrows to represent the forces they exert on the seesaw. ● **Explanation**: Annotate your model to demonstrate why the seesaw remains in a balanced, horizontal position without tilting to any side. ● **Application**: Describe a real-life situation, apart from the seesaw, where objects remain in a balanced state due to equal forces acting on them.	The designed tasks have engaging phenomena, but the prompts are not clear for the modeling task.

Table 4-15 (cont'd)

Iteration 2		
Revise the two designed assessment tasks for LP2 of 3-PS2-1 to ensure the prompts focus on: Developing models that include: Components/Elements: Clearly identify the parts of the model. Relationships: Describe the interactions among the components. Explanation: Explain how the components and their relationships together can account for the phenomenon.	**Task 1: "The Book on the Table"** **Scenario**: Imagine a heavy science textbook resting on a table. Despite people walking around and occasional bumps to the table, the book remains unmoved.\n\n** Prompt**: Develop a model to explain why the science textbook doesn't move on the table despite slight disturbances. Your model should include: ● **Components/Elements**: Identify the forces acting on the book and table (e.g., gravity, normal force). ● **Relationships Among Components**: Show how these forces are balanced. ● **Explanation of Phenomena**: Explain using your model how the balanced forces result in the book maintaining its position without movement. **Task 2: "The Tug-of-War Standoff"** **Scenario**: Two teams of equal strength are in a tug-of-war competition. The rope between them doesn't move in either direction, and both teams stay in place.\n\n **Prompt**: Create a model to explain why neither team can move the other during the tug-of-war. In your model, consider: ● **Components/Elements**: Represent the forces each team applies to the rope. ● **Relationships Among Components**: Illustrate the interaction between these forces and how they balance out. ● **Explanation of Phenomena**: Use your model to clarify why the balanced forces prevent any movement from either side, keeping the rope stationary.	These revised tasks direct students explicitly to construct models with a clear focus on elements (forces), their relationships (balance or equilibrium), and how these aspects coalesce to explain real-world phenomena of static objects or equilibrium scenarios. The tasks are scaffolded to guide students through the thought process of identifying forces, understanding their interactions, and applying this knowledge to explain why objects do not change motion under balanced forces. This approach aligns with the three-dimensional learning objectives of NGSS, fostering a deep understanding of physical science concepts while practicing model development skills.

The iterative design process for LP2 assessment tasks, aimed at understanding and modeling balanced forces, showcased significant evolution in task clarity and instructional alignment. Initially, the tasks "The Case of the Static Skateboard" and "Balancing Act on the Seesaw" provided engaging scenarios but lacked clarity in their prompts, particularly regarding the construction of models. This issue was identified during the first iteration, highlighting the need for clearer guidance on how to develop and use models to explain phenomena. In the second iteration, the tasks were revised to focus more explicitly on the three key aspects of model development: identifying components, describing relationships among components, and explaining phenomena. The revised tasks, "The Book on the Table" and "The Tug-of-War Standoff," provided clear and structured prompts. These tasks required students to construct models that explicitly represent the forces at play, illustrate how these forces interact, and explain the resulting balance or equilibrium.

This refinement process demonstrated a shift from merely engaging students with interesting phenomena to guiding them through the detailed and precise construction of scientific models. For example, the original item prompts for Task 1 for LP2 is: "

- _**Model Construction**: Draw a model to represent the forces Jamie and Alex are applying to the skateboard. Show why the skateboard doesn't move using arrows to represent forces._
- _**Explanation**: Explain in your drawing how the forces from Jamie and Alex are balanced and describe what that means for the motion of the skateboard._
- _**Application**: Think of another example where balanced forces result in no change in motion and describe it below your model."_

The first prompt of "model construction" does not provide explicit scaffoldings for students to construct a model, and the third prompt is beyond the scope of constructing models to explain the corresponding phenomena. After the revision, the prompts were revised into: " _Develop a model to explain why the science textbook doesn't move on the table despite slight disturbances. Your model should include:_

- _**Components/Elements**: Identify the forces acting on the book and table (e.g., gravity, normal force)._
- _**Relationships Among Components**: Show how these forces are balanced._
- _**Explanation of Phenomena**: Explain using your model how the balanced forces result in the book maintaining its position without movement."_

By emphasizing the identification of components that should be included in the model and their interactions, the revised prompts are better aligned with the 3D learning objectives of the NGSS. They provided a scaffolded approach to help students understand and apply concepts of balanced forces in real-world contexts.

The iterative process underscored the importance of clear and focused prompts in assessment tasks. It highlighted how specific guidance can enhance students' ability to develop and use models effectively. This approach not only aids in grasping complex physical science concepts but also fosters essential skills in scientific modeling and reasoning. The reflections on this process reinforce the value of iterative refinement in educational task design, ensuring tasks are both engaging and instructional, thereby supporting deep and meaningful learning.

4.1.7.3 Tasks and Exemplar Responses for The Tasks

To ensure the tasks are engaging for 3rd grade students, I used DALL.E to generate the scenario images to provide visual support for students to understand and engage in the tasks. The images (Tables 4-16 and 4-17) were incorporated into the item stem to provide support for students to understand the tasks. I directed GPT to generate the exemplar responses by providing the task, evidence statements, and the grade level. After several round iterations, the exemplar responses were presented in the Tables 4-16 and 4-17 along with the tasks together.

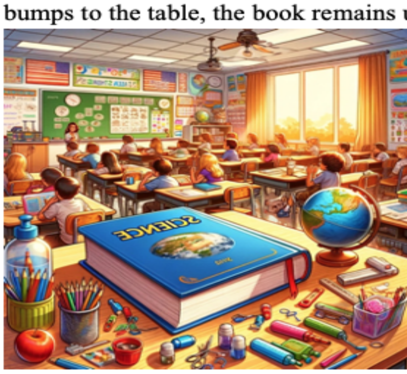
Table 4-16. Assessment tasks and their exemplar responses for 3-PS2-1 LP2

Tasks for LP2 of 3-PS2-1:
Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides.

***Task 1*: "The Mystery of the Still Book" (3-PS2-1-LP2-1)**

***Item Stem*:**

Have you ever noticed how some things don't move, no matter what? Here's a cool mystery: there is a heavy science textbook resting on a table. Despite you walking around and occasional bumps to the table, the book remains unmoved.


(Generated by DALL.E on March 11, 2024)

Item Prompts:

1. Draw to understand the Mystery: Develop a model to explain why the science textbook doesn't move on the table despite slight disturbances. Your model should include:
   - Identify the forces acting on the book and table (e.g., gravity).
   - Show the relationships of the forces that are acting on the book.

(Draw your model here)

2. Solve the Mystery with Your Drawing: After drawing, try to solve the mystery. Explain using your model how the forces result in the book maintaining its position without movement.

(Explain your model here)

Table 4-16 (cont'd)

Model description:

- Elements in the Model:
    - Book: Represent the book with a rectangle placed in the center of your model.
    - Table: Draw a larger rectangle underneath the book to represent the table.
    - Gravity Force: Draw a downward arrow above the book, labeled "Gravity", to represent the Earth's pull on the book.
    - Normal Force: Draw an upward arrow from the table surface, equal in length to the gravity arrow, labeled "Table Support Force", to represent the table supporting the book.
- Relationships Among Elements:
    - The gravity force and the normal force are a pair of interacting forces. They are equal in size but opposite in direction, creating a balance of forces. This balance of forces keeps the book stationary on the table, even if there are slight disturbances like nudging the table.
- Usage of Arrows and Labels:
    - Arrows are used to visually indicate the direction and magnitude of forces. In this model, two arrows of equal length pointing in opposite directions visually depict the state of force balance.
    - Labels "Gravity" and "Table Support Force" are used to clarify what each arrow represents. This helps students understand how these forces act on the book and how they counterbalance each other.

Explanation using my model:

So, here's what my drawing is telling me about our still book mystery. Even when we walk around or bump the table, the book doesn't move because the forces on it are balanced. Gravity pulls it down, but the table pushes it up just as much. It's like a tug-of-war game where both sides are equally strong, so the rope doesn't move. This balance of forces means the book stays right where it is, no matter the little shakes or bumps. That's why the book doesn't slide off or start floating into the air; all the pushes and pulls on it are in a perfect balance!

Table 4-16 (cont'd)

***Task 2*: "The Unmoving Tug-of-War Challenge"(3-PS2-1-LP2-2)**

***Item Stem*:**

Jinni's class and Cody's class are playing tug-of-war at the school playground. Both classes are equally strong, pulling the rope with all their might, but guess what? The rope doesn't move an inch! It's like an invisible force is keeping everything perfectly balanced. How can both classes pull so hard and yet nothing changes?


(Generated by DALL.E on March 11, 2024)

Item Prompts:

1. Develop a model to explain:

- What's Happening: Draw the push and pull both classes are giving to the rope.
- How They Work Together: Show how the push from one class and the pull from the other class are just right so that neither side move. You can use labels and/or symbols to show your ideas.

(Draw your model here)

2. Why No One Moves: Explain with your picture why, even though both classes are trying really hard, the rope doesn't move at all.

(Explain your model here)

Table 4-16 (cont'd)

Model description:

- Elements in the Model:
    - Rope: The central element of the tug-of-war, represented by a straight line across the model.
    - Jinni's Class Force: Drawn as an arrow pointing to the left from one end of the rope, labeled "Pull by Jinni's Class".
    - Cody's Class Force: An arrow of the same length as Jinni's class arrow, pointing to the right from the other end of the rope, labeled "Pull by Cody's Class".
    - Ground: The surface beneath the rope, ensuring context is provided for the tug-of-war scenario.
- Relationships Among Elements:
    - The forces exerted by Jinni's class and Cody's class are equal in magnitude but opposite in direction, demonstrating a balance of forces.
    - This equilibrium of forces results in no net movement of the rope, symbolizing how balanced forces maintain the status quo, irrespective of the strength applied from both ends.
- Usage of Arrows and Labels:
    - Arrows of equal length in opposite directions illustrate the concept of balanced forces exerted by both classes.
    - Labels "Pull by Jinni's Class" and "Pull by Cody's Class" clarify the source of each force, making it easier for students to comprehend how these forces interact to maintain balance.

Explanation using my model:

My drawing shows us the secret behind why the tug-of-war rope between Jinni's and Cody's classes stays put, even though everyone is pulling as hard as they can. Each class is pulling the rope towards themselves with the same strength, but because these forces are exactly the same and pulling in opposite directions, they cancel each other out. It's like when you and your friend push against each other's hands with the same strength; neither of you moves backward or forward because the forces are balanced. So, even with all that effort from both classes, the rope doesn't budge because the pushes and pulls on it are perfectly matched. This is a neat example of how things stay still when the forces on them are in perfect harmony!

Table 4-17. Assessment tasks and their exemplar responses for 3-LS4-3 LP2

Tasks for LP2 of 3-LS4-3:
Students engage in argument from evidence to support claims about which organisms can survive well, less well, or not at all in a specific habitat based on their characteristics and needs, using examples from various habitats to explore cause and effect relationships.

***Task 1*: " Squirrels and Their Search for the Perfect Home" (3-LS4-3-LP2-1)**

***Item Stem*:**

Amy's third-grade class is on a mission to discover what makes a perfect home for squirrels in their local park. To aid their investigation, they're equipped with a data table (Table a) capturing squirrel observations across different park areas over four weeks, alongside a brief on squirrel needs and park habitat characteristics (Table b).


(Generated by DALL.E on March 11, 2024)

Table a. Squirrel observation

Week	Trees Area (Squirrels Observed)	Open Area (Squirrels Observed)
Week 1	*15*	*3*
Week 2	*18*	*2*
Week 3	*20*	*4*
Week 4	*17*	*1*

Table b. Squirrel Needs & Habitat Characteristics

Area	Squirrel Needs & Habitat Characteristics
Trees Area	Abundant oak and pine trees (food sources: acorns and pine cones), dense foliage for shelter.
Open Area	Limited vegetation, few scattered trees, exposure to predators, and human activity.

Item Prompts:

1. Based on the data and information provided, choose whether you agree or disagree with the statement: "Squirrels prefer habitats with more trees."

2. Use the weekly observations from the data table to support your position. Include considerations about the squirrels' needs for food and shelter, and how the Trees Area and Open Area meet these needs.

3. Discuss how the presence of trees or lack thereof might affect squirrel behavior and habitat preference, connecting to their survival needs.

Table 4-17 (cont'd)

Exemplar Response for " Squirrels and Their Search for the Perfect Home"

1. Agree or Disagree: "I agree that squirrels prefer habitats with more trees."

2. In the data table, I saw that every week, more squirrels were found in the Trees Area compared to the Open Area. For example, in the first week, 15 squirrels were in the Trees Area and only 3 in the Open Area. This pattern kept going for all four weeks. This shows that squirrels are seen more where there are lots of trees."

3. Trees are super important for squirrels because that's where they find their food like acorns and make their homes up in the branches where it's safe. In the Trees Area, there are many trees for squirrels to eat from and hide in, which is why they like it better. The Open Area doesn't have many trees, so there's not much food or places for squirrels to build their nests, making it a not-so-good place for them to live. That's why the presence of trees makes the Trees Area a better home for squirrels. It has everything they need to be happy and safe!

Table 4-17 (cont'd)

***Task 2*: "The Mystery of the Growing Sunflowers"(3-LS4-3-LP2-2)**

***Item Stem*:**

In the school's garden, there are various sections where different plants flourish. One section is full of sunflowers that grow tall and healthy, while another section attempts to grow the same sunflowers, but they struggle and barely bloom. Mrs. Smith, the science teacher, presents this observation to her 3rd-grade class.

(Generated by DALL.E on March 11, 2024)

Mrs. Smith asks her students to explain why sunflowers in one section of the garden thrive while those in another section do not. She provides them with data (Table a) on sunlight exposure, soil moisture, and soil type for each garden section.

Table a. Garden information

	Phenomena	Sunlight	Soil Moisture	Soil Type
Sunny Spot	Sunflowers are tall and healthy.	Lots of sunshine all day.	Just right, not too wet or dry	Rich and full of nutrients
Shady Spot	Sunflowers are small and struggling	Very little, mostly in the shade	Quite damp, doesn't dry out much	A bit clayey and hard for roots to grow

Item Prompts:

1. Think Like a Scientist: Why do you think sunflowers in the sunny spot are doing so well compared to the ones in the shady spot? Remember, plants need certain things to grow big and strong.

2. Use the Clues: Look at the information about both spots in the garden.

- If you think sunlight is important, tell us why the sunny spot might be better for the sunflowers.
- If you're curious about the soil, share how different soil types might affect the sunflowers' growth.

3. Solve the Mystery: Using what you've learned, explain why you think the sunflowers are happier in one part of the garden than the other. Is it just about the sun, or is the soil also a big clue?

Table 4-17 (cont'd)

Exemplar Response for "The Mystery of the Growing Sunflowers" 1. I think the sunflowers in the sunny spot love getting lots of sun because plants need sunlight to make their food. It's like how we need to eat breakfast to have energy. The sunflowers getting more sunlight can make more food, so they grow big and strong! 2. The soil part is interesting! The sunflowers in the sunny spot have soil that's just right – not too wet and full of good stuff that helps them grow. It's like having a perfect bed to sleep in; you sleep well and wake up happy! But the shady spot's soil is too wet and hard, so I guess it's hard for sunflower roots to spread out and find food. 3. So, putting it all together, I think the sunflowers are way happier in the sunny spot because they get all the sun they need, and the soil is just perfect for them. In the shady spot, even though they try to grow, it's tough without enough sun, and the soggy, hard soil doesn't help either.

4.1.2 Critical Self-Reflections on Collaborating with GPT-4 for Assessment Design

Humans can collaborate with AI to design knowledge-in-use assessment tasks using the systematic NGSA approach. During the process, humans play a critical role by providing explicit guidance for prompting AI toward the task requirements and goals. Humans are also essential for monitoring the assessment design process by identifying critical areas where AI needs to focus. Meanwhile, humans guide AI in conducting critical reflections to help it detect and diagnose key principles and lessons that the AI can then learn from and apply in subsequent steps. Domain-specific information, including science content, 3D knowledge, and knowledge-in-use assessment design, is crucial for designing and monitoring the process. The iterative refinement and revision process is also vital for the design of assessment tasks.

This section reflects on the collaborative process used to train generative AI models, specifically GPT-4, for designing knowledge-in-use assessments, addressing RQ 1: How can generative AI models be effectively and iteratively trained to design these assessments? The discussion outlines a strategy for combining AI capabilities with human expertise to enhance assessment design, highlighting key components such as the need for clear instructions, leveraging domain-specific knowledge, engaging human expertise, refining processes iteratively, and fostering effective collaboration between AI and human experts.

4.1.2.1 Necessity of Explicit Guidance

Providing clear and detailed instructions emerged as a fundamental theme throughout the iterative design process. Explicit guidance was crucial for generating high-quality outputs from the GPT-4 model. Without well-defined guidelines and specific goals, the depth and quality of the AI-generated unpacking were limited. Explicit guidance involves providing the AI with precise instructions, clear task definitions, and well-articulated goals. This clarity ensures that the AI fully understands the task requirements and can produce outputs that meet the specified criteria. For instance, in the design of the Integrated Dimension Maps (IDM) for PE 3-PS2-1, initially, the AI produced a general map that lacked coherence and detail. The IDM needed to integrate various dimensions (DCIs, SEPs, and CCCs) into a coherent framework. By providing explicit guidance, such as detailing the elements to be included (observable components, relationships among these components, and limitations of models), the AI's output improved significantly. The detailed instructions ensured that the AI understood how to connect different components logically and meaningfully, resulting in a more comprehensive and coherent IDM. For instance, specifying that the map should clearly illustrate the cause-and-effect relationships between different scientific concepts ensured that the final output was both educationally valuable and practically useful.

Human experts play a critical role in this process. They clarify the objectives, provide detailed and explicit requirements, continuously monitor the AI's outputs, and adjust the guidance as necessary. The AI interprets the provided task goals, generates outputs that adhere to the defined objectives, and refines its understanding through iterative feedback. For instance, during the iterative process of designing LPs, human experts noticed that the AI's outputs were too broad. Then a human provided specific feedback emphasizing the need for detailed context and examples, which helped the AI refine its outputs to be more aligned with educational standards. Similarly, in the IDM design, human experts identified gaps in the AI's initial outputs and provided targeted feedback to ensure that the final map was detailed, coherent, and aligned with the educational objectives.

Explicit guidance is essential for training GPT-4 models effectively. Providing clear and detailed instructions ensures that the AI understands the task requirements, reducing the risk of producing irrelevant or low-quality outputs. Continuous monitoring and adjustment of the AI’s guidance is necessary to maintain alignment with educational standards and task requirements. Explicit guidance serves as the foundation for effective AI training, ensuring that the AI can produce high-quality, relevant outputs that meet educational standards. The iterative feedback loop between human experts and the AI plays a crucial role in refining these outputs, demonstrating the importance of clear and detailed instructions in the AI training process. This process not only enhances the AI's ability to perform specific tasks but also builds a collaborative framework where human oversight ensures that the AI's outputs remain aligned with educational objectives and standards. The importance of explicit guidance is underscored by the improvement in AI-generated outputs when detailed and precise instructions are provided, highlighting the need for continuous human involvement to guide and refine the AI’s contributions.

4.1.2.2 Importance of Domain-Specific Information

The importance of domain-specific information became apparent as a critical theme during the iterative design process. The AI's ability to generate detailed and accurate unpacking significantly improved when comprehensive and specific information was provided. Domain-specific information includes detailed knowledge about the subject matter, educational standards, and principles relevant to the task. This information helps the AI models understand the context and nuances of the task, enabling it to produce more accurate and relevant outputs.

When designing assessments for two PEs, 3-PS2-1 and 3-LS4-3, providing detailed information about the DCIs, SEPs, and CCCs was crucial. For example, in the case of PE 3-PS2-1, detailed explanations of forces, motion, and interactions between objects were provided, including Newton's laws and their applications to everyday phenomena. Despite GPT-4's ability to process general information, it lacked the depth needed for domain-specific tasks unless supplied with relevant details. By providing explicit descriptions of DCIs such as PS2.A: Forces and Motion and PS2.B: Types of Interactions, the AI could generate more nuanced and educationally relevant outputs.

An illustrative case was the unpacking of PS2.A: Forces and Motion. Initially, the AI-generated descriptions were general, such as "Forces can cause an object to move or stop." While this is correct, it lacked the depth necessary for a comprehensive educational framework. When detailed information about the grade levels of understanding and the specific disciplinary ideas and relevant ideas was provided, the AI's output transformed. The unpacking became richer, including descriptions of how balanced and unbalanced forces affect motion, and specific examples like "A book resting on a table demonstrates balanced forces, while a ball rolling down a hill shows unbalanced forces leading to acceleration." Another example involves PE 3-LS4-3, which focuses on environmental changes and their impact on organisms. Initially, the AI produced broad statements about environmental changes affecting living things. However, by providing detailed information on specific factors such as climate change, habitat destruction, and pollution, and their effects on various species, the AI could produce more targeted and relevant assessments. This included nuanced insights into how certain species adapt, migrate, or face extinction due to these environmental pressures.

In this process, human experts play several crucial roles. They select what information to provide to the system, ensuring it is both comprehensive and relevant. They monitor how well the AI perceives and processes this information, identifying areas where the AI's understanding might be superficial or incomplete. Human experts also pinpoint critical areas that need adjustment or further clarification and decide whether to iterate the current process or move to the next steps. This continuous oversight ensures that the AI remains aligned with educational standards and objectives. The AI, on its part, receives the provided background information and processes it to detect the prompts and understand the specific task requirements. It analyzes the information to summarize critical points and learns from the provided inputs and clarifications to retain important information and context for future tasks. This iterative learning process enables the AI to improve its performance over time, producing outputs that are increasingly aligned with educational standards. For instance, during the iterative unpacking of SEP related to "Developing and Using Models," the initial AI output was generic, merely stating that students should create models to represent phenomena. With detailed input about the various types of models (physical, conceptual, and mathematical), the specific criteria for evaluating these models, and examples of how these models can be used to explain phenomena, the AI's subsequent outputs became more sophisticated. It included specific strategies for students to develop and use models, criteria for assessing the models' effectiveness, and detailed examples illustrating the use of models in scientific inquiry.

Providing adequate background information is crucial for training GPT-4 models effectively. This step ensures that the AI has a solid foundation of domain-specific knowledge, enabling it to produce outputs that meet educational standards and task requirements. Continuous updating and refining of the AI's knowledge base are essential to maintain the quality and relevance of the outputs. For example, as educational standards evolve or new scientific discoveries are made, updating the AI with this new information ensures that it remains current and continues to produce relevant and accurate educational content. Moreover, the collaboration between human experts and AI in this context exemplifies a synergistic relationship where human intelligence provides the depth and context that AI needs to function effectively, while AI offers the capacity to process and integrate large volumes of information quickly and efficiently. This collaboration not only enhances the quality of the educational assessments but also accelerates the development process, making it more efficient and scalable.

4.1.2.3 Role of Human Experts

The role of human experts was underscored as a critical theme throughout the iterative design process. Human experts were indispensable in evaluating and refining the AI outputs, providing critical insights and feedback. Their application of domain knowledge and experience guided the AI models in generating high-quality outputs and ensuring alignment with educational standards and goals. For instance, when the initial LPs were too broad and lacked specificity, experts provided detailed feedback and examples to help the AI generate more focused and relevant LPs. This intervention transformed general statements into specific, actionable learning performances. An initial LP might state, "Students investigate the effects of forces on motion," which is broad and lacks detail. With expert guidance, this was refined to, "Students plan and conduct investigations to observe how different strengths of forces affect the motion of a ball rolling down a ramp, noting the differences in speed and direction." This level of specificity ensures that the learning performance is actionable and directly tied to observable student behaviors.

Moreover, human experts bring a nuanced understanding of educational contexts that AI currently lacks. They can interpret curriculum standards and translate them into specific, measurable learning outcomes. When unpacking a complex DCI such as PS2.A: Forces and Motion, experts not only provide the scientific content but also pedagogical strategies to effectively teach these concepts. This might involve suggesting inquiry-based learning activities, formative assessments, and differentiated instruction strategies to meet diverse student needs. For example, during the design of assessments for 3-LS4-3, which involves understanding how environmental changes affect organisms, the initial AI outputs were broad and lacked depth. Experts provided context about specific environmental changes such as deforestation, pollution, and climate change, and their impact on particular species. This allowed the AI to produce more detailed and contextually relevant outputs, such as, "Students analyze data on polar bear populations in the Arctic to understand the impact of melting ice caps on their habitat and survival rates."

Human experts evaluate the AI’s outputs, provide detailed feedback and examples to guide the AI in generating more accurate and relevant outputs, and ensure that the AI’s outputs align with educational standards and goals. The AI generates outputs based on the provided instructions and guidelines, incorporates feedback from human experts to refine its outputs, and learns from the provided inputs to improve the quality of its outputs. Human expertise is essential for training GPT-4 models effectively. Experts provide the necessary guidance and feedback to refine the AI’s outputs, ensuring that they meet educational standards and goals. This collaborative dynamic between AI and human experts enhances the overall quality and effectiveness of the generated outputs. This dynamic is not just about correction but also about enrichment. Experts provide insights that help AI models understand the broader educational landscape, including the integration of crosscutting concepts and practices that are essential for three-dimensional learning as advocated by the NGSS. Furthermore, human experts play a critical role in maintaining the ethical and equitable aspects of educational content. They ensure that the AI-generated outputs do not inadvertently reinforce biases or exclude certain groups of students. By reviewing and providing feedback, experts help ensure that the content is inclusive and accessible to all students, thereby promoting equity in education.

The role of human expertise in this collaborative process is multi-faceted. Experts provide detailed, domain-specific knowledge, offer pedagogical strategies, ensure alignment with educational standards, and uphold ethical and equitable principles in educational content. This partnership between human intelligence and AI results in high-quality, relevant, and effective educational assessments that are finely tuned to meet the needs of both educators and students.

4.1.2.4 Iterative Refinement

The iterative nature of the design process proved vital for continuous improvement. Each iteration builds on the previous one, incorporating feedback and additional information to enhance the quality of the AI-generated unpacking. Iterative refinement involves continuously evaluating and improving the AI's outputs through multiple rounds of feedback and adjustments. This process ensures that the outputs are progressively enhanced and aligned with the task requirements.

The iterative process was crucial in refining the AI's outputs for various tasks. For example, in designing the IDM for PE 3-PS2-1, each iteration involved re-evaluating the outputs, providing more detailed guidance, and refining the approach based on feedback. This process led to progressively better outputs, ensuring that the IDM became increasingly detailed and aligned with the task requirements. Reflective practice was also integral to this process. Human experts' reflections on the AI's outputs helped identify gaps and areas for improvement, guiding subsequent steps in the training process. For instance, when the initial unpacking was too general, experts provided detailed feedback and examples, leading to more accurate and relevant outputs in subsequent iterations.

Human experts continuously evaluate the AI's outputs, provide detailed feedback and examples to guide the AI in refining its outputs, and adjust their guidance based on the AI's performance. The AI generates outputs based on the provided instructions and guidelines, incorporates feedback from human experts to refine its outputs, and learns from the provided inputs to improve the quality of its outputs. Iterative refinement is essential for training GPT-4 models effectively. This process ensures that the AI's outputs are continuously evaluated and improved, leading to progressively better outputs. Reflective practice and feedback from human experts are integral to this process, helping to identify gaps and areas for improvement.

4.1.2.5 AI-Human Collaboration

The collaborative dynamic between AI and human experts facilitated the creation of high-quality educational tools. The AI's ability to learn from provided frameworks and examples, combined with human expertise, resulted in outputs that were not only accurate but also pedagogically sound. AI-human collaboration involves the combined efforts of AI and human experts in generating high-quality outputs. This collaboration leverages the strengths of both AI and human expertise to enhance the overall quality and effectiveness of the generated outputs.

Throughout the iterative design process, the collaborative dynamic between AI and human experts was evident. For instance, in designing the IDM for PE 3-PS2-1, human experts provided detailed feedback and examples, which the AI incorporated to generate progressively better outputs. This collaboration was also crucial in refining the AI's outputs for designing LPs. Human experts provided explicit guidance and detailed examples, which helped the AI generate more focused and relevant LPs. The iterative feedback loops and continuous refinement ensured that the final outputs were accurate and pedagogically sound. Human experts provide explicit guidance and detailed examples to guide the AI in generating high-quality outputs, evaluate the AI's outputs, identify areas that need refinement, and provide detailed feedback to help the AI refine its outputs. The AI generates outputs based on the provided instructions and guidelines, incorporates feedback from human experts to refine its outputs, and learns from the provided inputs to improve the quality of its outputs.

AI-human collaboration is essential for training GPT-4 models effectively. The combined efforts of AI and human experts enhance the overall quality and effectiveness of the generated outputs. This collaboration leverages the strengths of both AI and human expertise, resulting in outputs that are accurate, pedagogically sound, and aligned with educational standards.

4.2 RQ2. How Do Human Experts Across Different Disciplines Evaluate the AI-Generated Knowledge-In-Use Assessment and What Refinements Do They Suggest?

To respond to RQ2, "*How do human experts across different disciplines evaluate the AI-generated knowledge-in-use assessments, and what refinements do they suggest?*" a multidisciplinary expert panel review stage was conducted to collect feedback on the interim products. Their feedback focuses on the LPs, evidence statements, assessments, and rubrics, emphasizing 3D learning, engagement, language complexity, equity, and practical perspectives.

In this section, I report the analysis of the LPs and evidence statement design, as well as the assessment design feedback based on the different expert groups' input. First, I detail the composition and background of the expert panels for the two different PEs. Then, I explain how the data were analyzed, present the analytic results, and conclude with the major themes and suggestions derived from the reviewers' comments and feedback. Each report section includes a summary of both quantitative and qualitative analyses, emphasizing the critical takeaways and highlighting variations in feedback across different expert groups. This systematic method provides a thorough and detailed account of the expert evaluations and their recommendations, offering valuable insights into the assessment tasks' effectiveness and areas for improvement.

4.2.1 Expert Feedback on PE: 3-PS2-1

4.2.1.1 Themes of Feedback on LPs and Evidence Statements for 3-PS2-1

The quantitative analysis of expert feedback was conducted to assess the effectiveness of the LPs and Evidence Statements for 3-PS2-1 (see Table 4-18). The analysis involved collecting numerical ratings from experts on various dimensions and visually representing these ratings to identify trends and patterns. The feedback collection table includes six major dimensions that were designed to collect the feedback on the LPs and evidence statement using a Likert scale with scores from 1 (not at all) to 5 (completely).

These dimensions provide a comprehensive framework for evaluating the alignment and effectiveness of the LPs and evidence statements with the NGSS standards. These dimensions and their rationales are presented in Table 4-19.

Table 4-18. LPs and evidence statements for LP2 for 3-PS2-1 for review

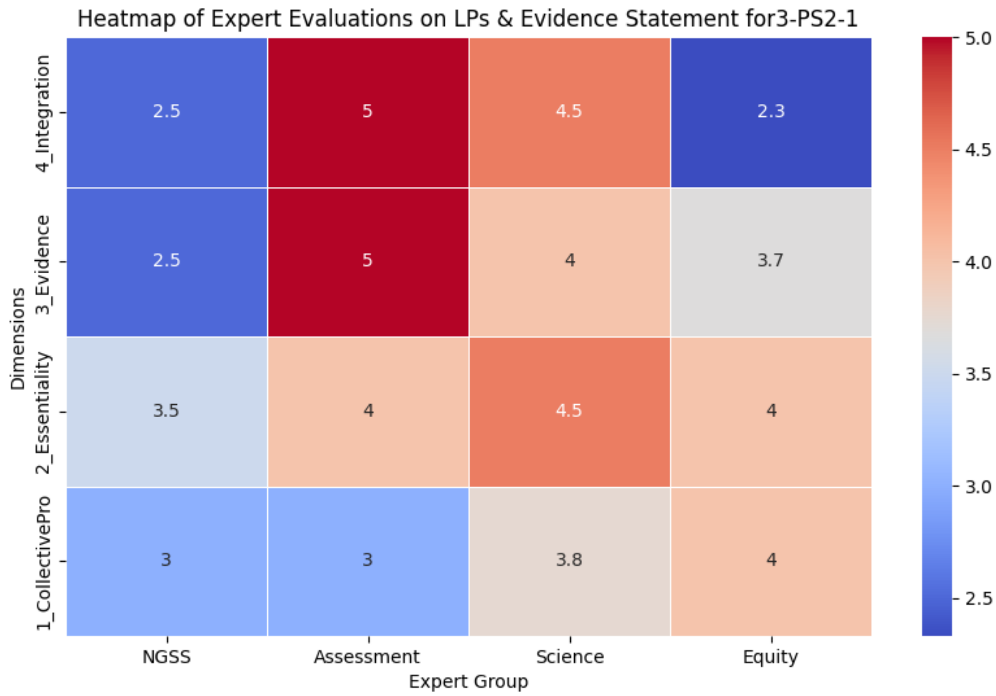
PE	3-PS2-1: Plan and conduct an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object.
LPs	LP1**: Students plan and carry out investigations to observe how different strengths of forces affect the motion of an object. LP2**: Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides. LP3**: Students construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes. LP4**: Students use models to explain how non-contact forces (e.g., magnetic or gravitational forces) on an object's motion or change in motion act at a distance.
Focal LP: LP2	Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides.
Evidence Statements	1. Students construct a model that visually represents an object under the influence of balanced forces, showing the object either in a static position or moving at a constant speed without any change in direction. The model should: ● Basic shapes or figures to represent the object(s) and the forces acting on them. ● Arrows of equal length pointing in opposite directions to signify balanced forces. ● A brief label or symbol next to each arrow, indicating the type of force (e.g., "push," "pull," "gravity").
	2. Students include annotations or keys in their model to delineate between balanced and unbalanced forces, explaining scenarios where the forces cancel out, resulting in zero net force on the object. The model should: ● Clearly mark parts of their model to show where forces are acting. ● Use simple vocabulary to describe how these forces are balanced or what might happen if they weren't (e.g., "If one side pulls harder, the rope moves that way").
	3. Students apply their model to real-life scenarios like a book on a table or a car cruising at a steady speed, explaining how these situations exemplify balanced forces resulting in no change in motion. The explanation should: ● Identify the forces acting on the object in the scenario (e.g., gravity pulling the book down, table pushing it up). ● Describe how these forces balance out, using elements from their model as reference points. ● Conclude how the balanced forces result in no change in the object's motion (either staying still or moving steadily).

Table 4-19. Feedback collection dimensions and rationale for LPs and evidence statements

Dimension	Statement	Rationale
Collective Representation of Proficiencies (1)	To what extent does the set of learning performances collectively represent the proficiencies that are necessary for attaining the performance expectation?	Understand how well the set of LPs collectively represent the necessary proficiencies for attaining the PE.
Essentiality of the Learning Performance (2)	To what extent does LP2 comprise an essential part of what is needed to achieve the performance expectation?	Understand whether the LP is essential for achieving the PE.
Sufficiency of Evidence Statements (3)	To what extent do the evidence statements of LP2 reflect obtainable pieces which, taken together, are sufficient for supporting a claim of student proficiency in this learning performance?	Determines if the evidence statements reflect obtainable pieces sufficient to support a claim of student proficiency.
Integration of Knowledge (4)	To what extent is LP2 an integrated 3-dimensional statement of knowledge-in-use?	Examines if the LP is an integrated three-dimensional statement of knowledge-in-use.
Gap Identification (5)	What gaps, if any, do you see in the set of learning performances (i.e., proficiencies required by the performance expectation that are not represented in the set of learning performances)?	Identifies any proficiencies required by the PE that are not represented in the set of LPs.
Overreach identification (6)	What overreach occurs, if any, in the set of learning performances (i.e., proficiencies that ARE NOT required by the performance expectation but that ARE required in the set of learning performances)?	Identifies any proficiencies that are not required by the PE but are included in the LPs.

The ratings were averaged, and the standard deviation (SD) was calculated to provide an overall assessment and measure of variability. Figure 4-6 presents the summary of the expert ratings on the LPs and evidence statements feedback.

Figure 4-6. Expert feedback on the LPs and evidence statements of PE 3-PS2-1



The heatmap visualizes the expert evaluations on the LPs and Evidence Statements for PE 3-PS2-1. Each column in this heatmap corresponds to a different expert group—NGSS, Assessment, Science, and Equity—reflecting their specific feedback. Similarly, each row represents a distinct dimension of the feedback, which are Collect/Interpret, Establishing Evidence, and Integration of the three dimensions.

The color scale on the heatmap ranges from blue to dark red, serving as a gradient to indicate the level of expert ratings. Dark red colors signify the highest ratings, approaching 5.0, which indicates strong agreement or high levels of satisfaction with the specific dimension evaluated. These areas suggest that the expert group views the LPs and evidence statements as being highly effective or excellently aligned with the designated criteria. In contrast, mid-range ratings are colored in shades ranging from light red to orange, spanning values between 3.0 and 4.5. These colors denote moderate satisfaction, suggesting that while some aspects are satisfactory, there are still opportunities for improvement in these areas. The presence of these colors on the heatmap points to dimensions where feedback suggests a need for adjustments or enhancements to better meet the PE standards or improve clarity and effectiveness.

Finally, blue represents the lowest ratings, near the value of 2.0, indicating significant concerns or dissatisfaction from the expert reviewers. These regions highlight specific dimensions where experts believe that the LPs and evidence statements fall short of expectations and require substantial revisions or reconsideration.

The layout and color coding of the heatmap enable a quick visual assessment of consensus and divergence among different expert groups. This helps in identifying areas of general agreement or satisfaction, where little modification might be needed, as well as pinpointing those aspects that require attention and likely intervention due to lower ratings. The visual format of the heatmap thus not only simplifies the comparison across multiple dimensions and expert groups but also aids in quickly locating areas of strength and those needing improvement, facilitating targeted adjustments to enhance the educational assessments.

The heatmap shows that the NGSS experts have more critiques on all of the dimensions, while the other groups of experts have more conservative scores on the first and second dimensions. The feedback reflects a mix of positive feedback and areas for improvement. The high ratings for sufficiency and integration suggest that the LPs are well-supported by adequate evidence and effectively integrate 3D learning. However, the lower ratings for collective representation and essentiality, particularly from NGSS experts, indicate a need to revisit these areas to ensure the LPs comprehensively cover the necessary proficiencies and are deemed essential for achieving the PE. Qualitative feedback provided deeper insights into these findings.

Coverage of DCIs and SEPs

There was a common theme regarding the omission or inadequate coverage of all the requisite proficiencies. The major concern is the set of generated LPs does not sufficiently cover the major ideas of “unbalanced forces” that are emphasized in the PE, and the LP4 addresses the idea of non-contact forces is reasonable due to some students often struggle with understanding “non-contact forces” but that may not meet the PE’s major expectation about the balanced and unbalanced forces. C also pointed out “LP4 has a DCI focus that appears outside the bounds of the PE – the foundation box emphasizes objects in contact (part of PS2.B) and the PE does not indicate that non-contact forces are a primary focus of the PE nor that distance apart of objects needs to be addressed. The DCI element in LP 4 seems better aligned to PS2-3.” The NGSS expert provided further insights, stating, "The unpacking of the DCI and SEP is not very good: 1. The DCI unpacking does not cover the major ideas of the DCI (the boundary of the DCI elements); 2. The SEP is beyond the grade level, such as the use of models to explain or describe. There is a need to readjust the SEP unpacking and the CCC unpacking." Interestingly, some experts believe that the set of LPs covers the DCIs in the PE very well, but more attention needs to be paid to whether the SEP in the PEs is well addressed, as only two LPs mention planning and carrying out investigations. This concern is echoed by assessment experts who commented, “LP1 most directly aims to address the SEP of the PE but does not sufficiently encompass the SEP (e.g., the LP requires students to observe in an investigation, but the PE requires that students produce/collect data that will provide evidence to make a claim).” Given the feedback, it is crucial to go back to reexamine the unpacking of the three dimensions, and redesign the LPs to ensure they cover all of the key ideas.

Integration and alignment of CCCs

Feedback indicated that while the LPs generally integrated CCCs, there were instances where the connections could be made more explicit. For example, assessment expert CH mentioned that "the evidence of a CCC is far less clear, but the phrase 'results in' could be evidence of cause and effect." Science content expert J remarked that "although stability and change are not as explicit as they could be in this assessment, there is still ample evidence of mastering this CCC in these evidence statements." Assessment expert P observed, "there is a bit more to be desired with the lens of cause and effect and how those are specifically developed. Could be spelled out more." NGSS experts also pointed out, "The CCC component is not clear and needs to be explicit." Additionally, they noted that "One of the LPs focuses on friction, which is not explicitly mentioned in the PE text, the clarification statement, the assessment boundary, or the DCI text. However, only gravity is explicitly mentioned, so it must be assumed that other forces would be discussed. There is an LP about students using models to explain the effect of balanced forces, but none of the LPs address the idea of the effect of unbalanced forces. This leaves a very significant part of the PE unaddressed." The DCI element for the PE even uses the phrases "they add to give zero net force on the object" and "forces that do not sum to zero."

Complexity and appropriateness of examples

Experts noted concerns about the complexity of some examples used in the LPs, suggesting that certain examples might be too advanced for the intended grade level. Science content expert S pointed out that "the scenarios of the assessment tasks may require different levels of knowledge: 'a book resting on a table or a tug-of-war game'—the difficulty of students interpreting these two scenarios may be quite different." Teacher expert Le criticized the example of "the motion of wheels of a cruising car" as not reflective of balanced and unbalanced forces due to external factors like motors and gasoline. In addition, experts suggested having more examples of moving objects. T mentioned, “Students often have a misconception that a force is necessary for an object to keep moving. Therefore, they often struggle with the idea that a moving object will keep moving if all of the forces are balanced. While the evidence statement does mention the idea of “...moving at a constant speed without any change in direction”, there are no examples where the object is in motion.” He suggested adding one example of moving objects.

Need for clarity in language and evidence statements

Several experts emphasized the importance of clear and precise language to avoid confusion, especially given the elementary education context. T noted that terms like "net force" and "different strengths of forces" might confuse educators and students, especially those with limited physics backgrounds. He suggested these terms could lead to inaccuracies in assessments and might not align well with the PEs. E similarly emphasized the need for clear and precise language, pointing out that some scientific terms might be too advanced for the target educational level. Language expert Su noted that "LP4 needs a verb. The 4 LPs indicate what students will do to show evidence from the results of the investigations they plan and conduct." Similarly, T suggested revising the LP4 as “Students use models to explain how non-contact forces act at a distance.” The integration of these findings in Table 4-19 suggests several key recommendations for refining the LPs and Evidence Statements.

Table 4-19. Integrated Analysis Results for PE 3-PS2-1 LPs and Evidence Statements

Theme	Key Points	Recommendations
Coverage of DCIs and SEPs	Effective coverage of force and motion concepts, gaps in addressing all necessary SEPs and DCIs	Redo the unpacking of the DCI to cover all major ideas and boundary elements (unbalanced forces, non-contact forces). Ensure the SEP levels are grade-appropriate, such as simplifying the use of models to explain or describe (explain or describe). Include LPs that address the effect of unbalanced forces to ensure comprehensive coverage of the PE.
Integration of CCCs	General integration of CCCs, need for more explicit connections	Make connections between CCCs and the LPs more explicit. Clearly define and illustrate concepts such as stability and change. Ensure the CCC component is clear and explicitly mentioned in the LPs. Include discussions of forces like friction and unbalanced forces as required by the PE.
Example Complexity	Concerns about the complexity of examples, some examples too advanced for grade level	Use simpler, more relatable examples that accurately reflect the concepts being taught. Avoid scenarios that could be misinterpreted or are too complex for the grade level. Provide examples that are within the students' understanding and experience, such as more straightforward comparisons or everyday situations.
Language and Terminology	Importance of clear and precise language, avoid confusing terminology	Simplify language to match the reading levels of elementary students and ensure consistent terminology across the LPs and evidence statements. Provide clear definitions and avoid complex phrases to enhance comprehension. Replace vague or advanced terms with simpler alternatives.
Gaps and Overreach	Need for alignment with NGSS standards, ensure developmental appropriateness	Review the LPs to identify and address any gaps or overreach. Ensure that all necessary SEPs are represented and that the LPs do not include proficiencies beyond what is required by the PE. Adjust the SEP and CCC unpacking to align with grade-level expectations.

4.2.3.2 Expert Feedback Analysis for PE: 3-PS2-1 Assessment Task 1

Expert panels received a protocol including three major sections (item stem, item prompt, and exemplar response) to collect their feedback on Task 1 designed for LP2 for the PE 3-PS2-1. See Figures 4-7 and 4-8 below about the task 1 and its' exemplar response. The protocol comprised 16 items across five dimensions for the item stem, 10 items across four dimensions for the item prompt, and 3 items for the exemplar response. Feedback was collected using a Likert scale from 1 (not at all) to 5 (completely) across these various dimensions. The collected data were then analyzed to provide a comprehensive evaluation of Task 1. Figures 4-9, 4-10, and 4-11 provide a visual representation of the expert ratings across various dimensions on the Task 1 item stem, item prompt and exemplar response.

Figure 4-7. Task 1 for LP2 of 3-PS2-1

***Task 1*: "The Mystery of the Still Book" (3-PS2-1-LP2-1)**

***Item Stem*:**

Have you ever noticed how some things don't move, no matter what? Here's a cool mystery: there is a heavy science textbook resting on a table. Despite you walking around and occasional bumps to the table, the book remains unmoved.

(Generated by DALL.E on March 11, 2024)

Item Prompts:

1. Draw to understand the Mystery: Develop a model to explain why the science textbook doesn't move on the table despite slight disturbances. Your model should include:

- Identify the forces acting on the book and table (e.g., gravity).
- Show the relationships of the forces that are acting on the book.

(Draw your model here)

2. Solve the Mystery with Your Drawing: After drawing, try to solve the mystery. Explain using your model how the forces result in the book maintaining its position without movement.

(Explain your model here)

Figure 4-8. Exemplar response for task 1 for LP2 of 3-PS2-1

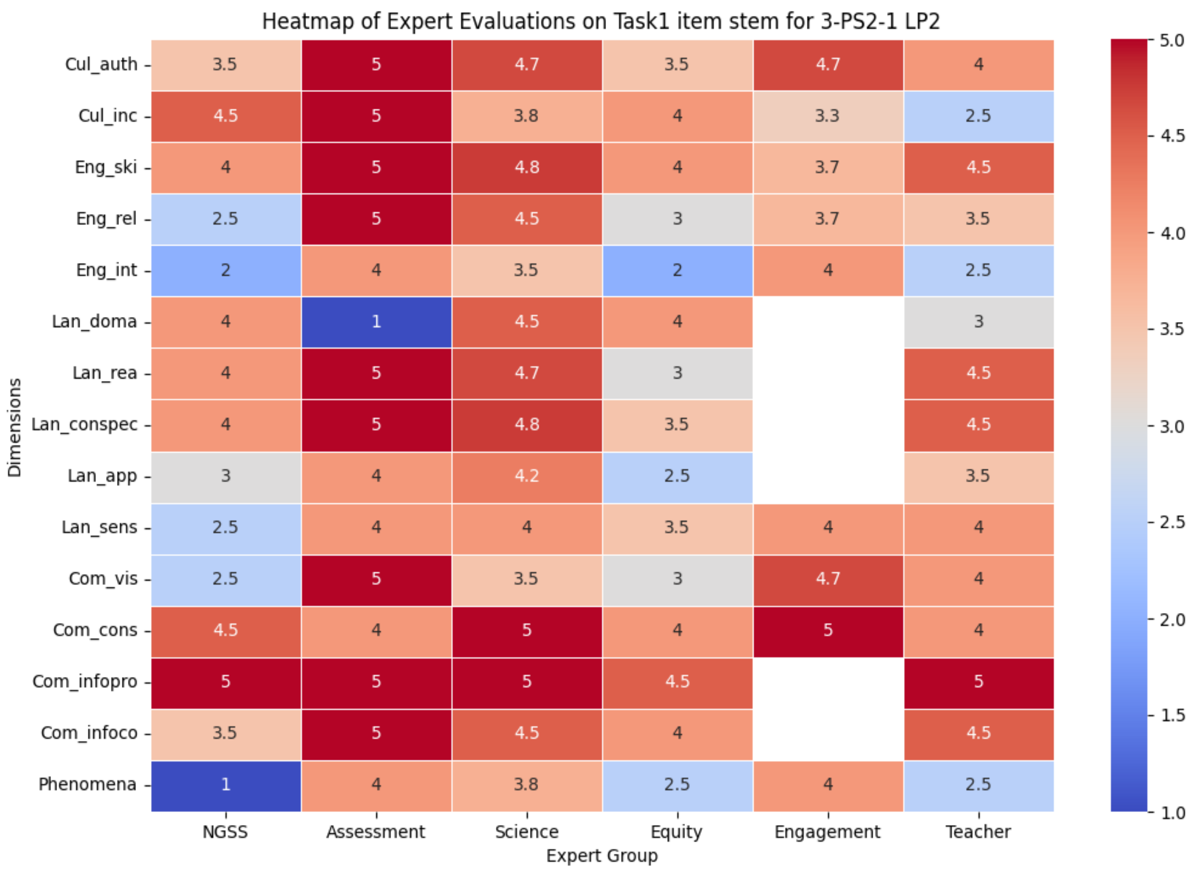
Model description:

- Elements in the Model:
    - Book: Represent the book with a rectangle placed in the center of your model.
    - Table: Draw a larger rectangle underneath the book to represent the table.
    - Gravity Force: Draw a downward arrow above the book, labeled "Gravity", to represent the Earth's pull on the book.
    - Normal Force: Draw an upward arrow from the table surface, equal in length to the gravity arrow, labeled "Table Support Force", to represent the table supporting the book.
- Relationships Among Elements:
    - The gravity force and the normal force are a pair of interacting forces. They are equal in size but opposite in direction, creating a balance of forces. This balance of forces keeps the book stationary on the table, even if there are slight disturbances like nudging the table.
- Usage of Arrows and Labels:
    - Arrows are used to visually indicate the direction and magnitude of forces. In this model, two arrows of equal length pointing in opposite directions visually depict the state of force balance.
    - Labels "Gravity" and "Table Support Force" are used to clarify what each arrow represents. This helps students understand how these forces act on the book and how they counterbalance each other.

Explanation using my model:

So, here's what my drawing is telling me about our still book mystery. Even when we walk around or bump the table, the book doesn't move because the forces on it are balanced. Gravity pulls it down, but the table pushes it up just as much. It's like a tug-of-war game where both sides are equally strong, so the rope doesn't move. This balance of forces means the book stays right where it is, no matter the little shakes or bumps. That's why the book doesn't slide off or start floating into the air; all the pushes and pulls on it are in a perfect balance!

Figure 4-9. Heatmap of the expert ratings on the Task 1 item stem for PE 3-PS2-1

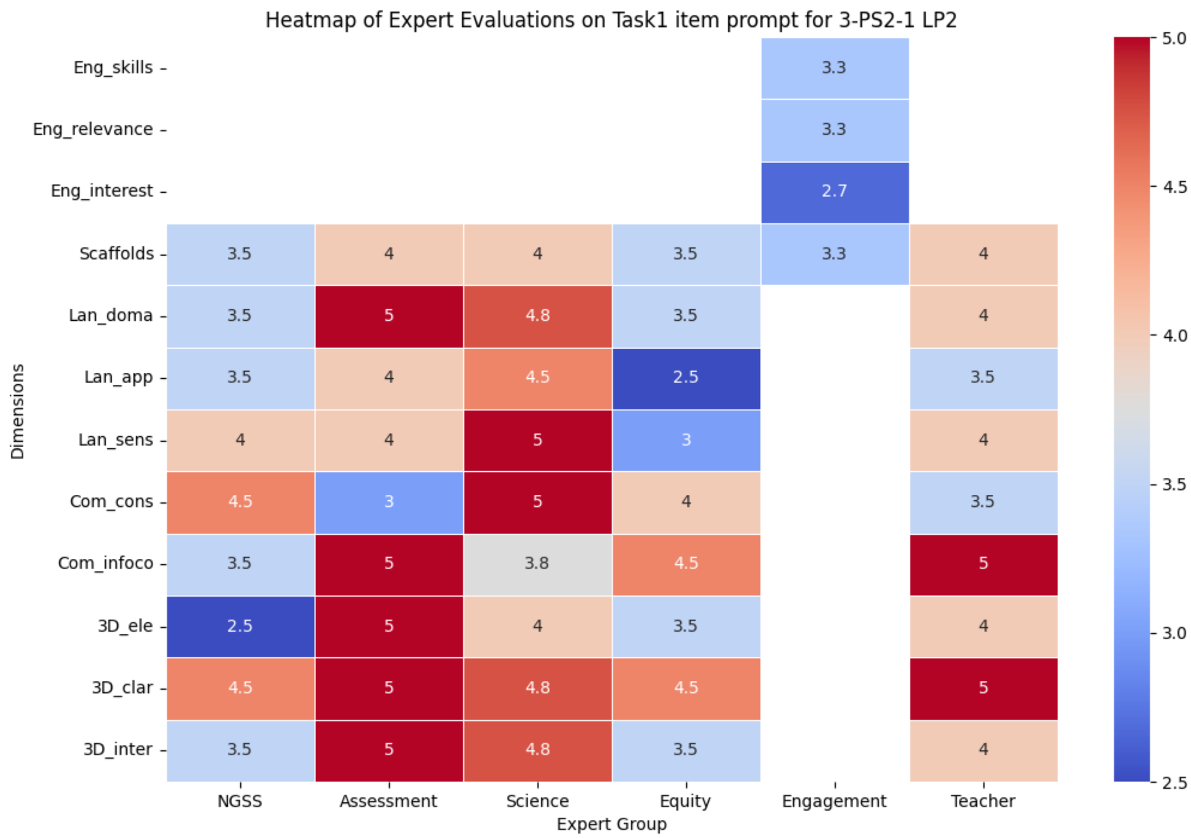


The most striking observation is the low score provided by NGSS experts, suggesting a significant concern with how the phenomenon is presented or utilized in the item stem. This contrasts with more moderate scores from other groups, indicating a discrepancy in how the phenomenon's relevance or clarity is perceived across expertise. Language-related dimensions received a wide range of scores. Notably, domain-specific language was rated poorly by NGSS experts, but much more favorably by other groups, highlighting a potential disconnect between NGSS content expectations and language used. The language sentence structure also varied, with lower scores indicating a need for better alignment with diverse student backgrounds. Engagement dimensions show moderate to high variability. Notably, engagement interest and relevance are scored lower by several groups, suggesting that the item may not effectively capture or maintain student interest or connect well with their real-life experiences. Comprehension scores are generally high, indicating that, structurally, the item supports student understanding and information processing. However, the visualization aspect received lower ratings from some groups, suggesting that visual aids or representations used in the item could be enhanced for clarity or effectiveness. Moreover, teachers have relatively lower scores on if the phenomenon in the task can be really interesting to students, although it is relevant to student life. The lower scores on phenomena engagement and language grade level appropriateness and sentence structure particularly stand out as areas needing immediate attention to ensure the item stem is both educationally effective and resonant with a diverse student population. The feedback underscores the necessity of enhancing the item's relevance, engagement potential, and language appropriateness ensuring it not only meets educational standards but also supports inclusive and equitable learning experiences.

The heatmap in Figure 4-9 displays expert ratings for the Task 1 item related to PE 3-PS2-1 across various dimensions. Each column represents a different expert group—NGSS, Assessment, Science, Equity, Engagement, and Teacher. Each row corresponds to specific dimensions of the assessment, such as 'Cultural Authenticity', 'Language Sensitivity', 'Comprehension', and 'Phenomena'.

The color scale of the heatmap ranges from blue to red. Blue indicates lower scores (1.0 to 2.0 range), signaling significant concerns or dissatisfaction. This suggests that aspects within this color range may require substantial revisions. Light red to orange represents moderate scores (2.5 to 4.0 range), indicating partial fulfillment of criteria and potential areas for improvement. Red signifies higher scores (4.5 to 5.0 range), denoting strong agreement or satisfaction with the dimensions evaluated, suggesting that these aspects are well-executed.

Figure 4-10. Heatmap of the expert ratings on the Task 1 item prompt for PE 3-PS2-1



The heatmap in Figure 4-10 provides a detailed view of expert evaluations across multiple dimensions of the Task 1 item prompt associated with 3-PS2-1, LP2. The evaluations were guided by specific criteria focused on 3D Prompt alignment, Comprehension, Language Complexity, the use of Scaffolds, and engagement by the engagement expert panel. The evaluations indicate varied perceptions of how well the prompts are 3D and align with integrated proficiencies. Particularly, the NGSS panel's lower scores (2.5) suggest concerns about the comprehensive integration or alignment of the prompts. The extent to which questions elicited by the prompts are motivated by the scenario described in the stem was also evaluated, with scores suggesting some alignment but also room for enhancement to clarify the connection between the scenario and the questions. Additionally, the accessibility of the prompts for novices was assessed, with varied scores indicating differing views on the prompt's suitability for students still developing relevant proficiencies. Experts assessed whether students have the necessary prior knowledge to understand and respond to the prompt, with generally high scores in related areas such as information coherence and consistency suggesting an adequate connection with expected prior learning. However, moderate evaluations in these areas also suggest that further clarification could be beneficial. Concerns were noted regarding the clarity and directness of sentence structure, as lower scores highlight a need for simpler language to aid comprehension. Mixed evaluations in vocabulary appropriateness and domain-specific vocabulary usage indicate that while some experts find the vocabulary suitable and well-integrated, others see a need for adjustments to ensure all vocabulary is accessible and clearly explained. The effectiveness of scaffolds in helping students navigate the complexity of the task received mixed perceptions. Some panels noted lower scores, suggesting that the scaffolds might not be effectively presented or sufficiently supportive for all students, particularly those with less background knowledge or proficiency. In terms of engagement, there is a clear need to make the prompts more engaging and interesting. This could involve integrating topics or scenarios that are more directly aligned with student interests or current societal issues, making the educational experience more engaging and motivating. By more directly linking the prompts to real-life applications and demonstrating how the skills and knowledge gained are applicable outside the classroom, the prompts could become more relevant and meaningful to students. Ensuring that the prompts not only introduce but also effectively integrate the three dimensions of learning will be crucial. This may involve revising the prompts to include clearer explanations or examples of how these dimensions are relevant and can be explored through the task.

Overall, there is a recognized need for improving language structures, ensuring vocabulary appropriateness, and enhancing term consistency to make the prompt more accessible and understandable for all students. The presentation and design of scaffolds also need revision to better assist students in understanding and engaging with the task, ensuring that scaffolds effectively break down task complexity and support learning.

Figure 4-11. Heatmap of the expert ratings on the Task 1 exemplar response for PE 3-PS2-1

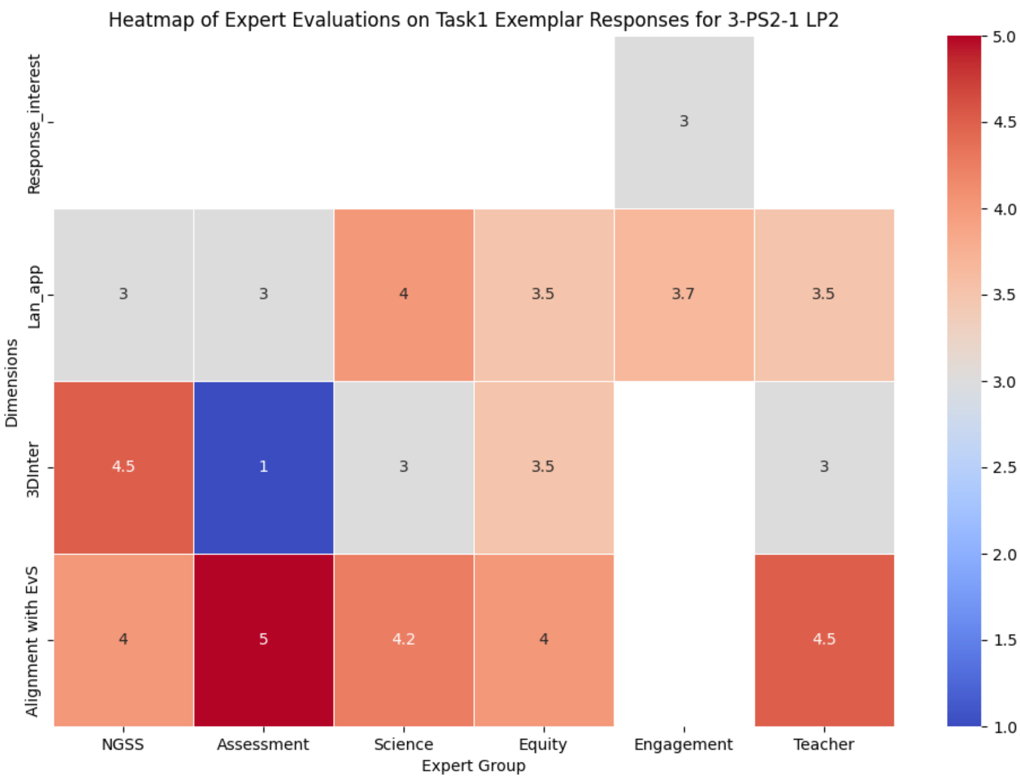


Figure 4-11 displays a heatmap of expert evaluations on the exemplar responses for Task 1. These evaluations are based on specific criteria focusing on the extent to which exemplar responses capture necessary evidence statements, address the integrated proficiencies entailed by the LP, and utilize grade-appropriate language. The scores of alignments with evidence statements reflect the extent to which exemplar responses capture all necessary evidence statements. The Assessment expert group rated this particularly high (5), indicating that the exemplar responses well represent the required evidence statements. In contrast, the NGSS group gave a lower score (4), suggesting some room for improvement in how comprehensively the responses cover all designated evidence statements. The 3D integration dimension assesses whether it is possible for students to provide accurate responses without necessarily attending to all integrated proficiencies required by the LP. The varied scores, with higher ratings from Assessment and Science panels (4.2 and 4 respectively) and lower from the NGSS (3), indicate differing views on how well the exemplar responses integrate or require engagement with the 3D learning components. The appropriateness of the language used in the responses concerning the grade level received mixed evaluations. The Teacher panel rated this dimension highly (4.5), suggesting that the language used is well-suited for the grade level. Conversely, scores from the NGSS and Engagement panels were lower, suggesting that some responses may not consistently meet the language level expectations for the target student group. The response interest evaluates how interesting the responses are likely to be for students, with varying perceptions among the expert groups. Lower scores from the NGSS panel (1) highlight a significant critique that the responses may lack elements that engage or captivate students' interests, which is crucial for maintaining engagement with the task.

Overall, the phenomena dimension received mixed ratings, indicating concerns about the compelling nature and comprehensibility of the phenomenon presented in the item stem. So do the assessment experts, equity/language experts, and the teacher experts. In contrast, engagement experts rated it significantly higher, suggesting that they found the phenomenon more engaging for students. This discrepancy highlights the differing perspectives among expert groups on what constitutes an engaging and comprehensible phenomenon for elementary students. It is important to collect further insights from teacher experts' insights on the engagement level of the item phenomena to gain better understandings. Additionally, the dimensions of language complexity and student engagement needs further exploring to understand the concerns or suggestions from the experts. Also, the level of exemplar responses were argued. There are concerns about students' ability to independently demonstrate proficiency and the appropriateness of the language used. These quantitative findings are further explored in the qualitative analysis. There are several themes that were identified based on the analysis.

Simplify language for clarity and grade-level alignment

Experts consistently noted that the language used in the item stem and prompts could be simplified to better match the reading level of third graders. Words like "despite" and phrases like "remains unmoved" were highlighted as potentially confusing. E suggested, "The second sentence is not appropriate for 3rd grade. Perhaps – The book does not move, even with slight bumps to the table." T noted, “The language in the stem creates a scenario that makes sense, but terms like ‘remains unmoved’ are too complex for third graders.” Assessment experts also suggest revising the item prompts from “Explain using your model how the forces result in the book maintaining its position without movement.” to “Use your model to explain how the forces result in the book staying in place without moving. OR Explain with your model how the forces keep the book from moving.” In addition, there were several comments about the need for clearer and more consistent terminology. For example, terms like "normal force" should be clearly defined. T noted, “Normal force is used in the example, but it is not defined in the question, nor used in any of the PEs, DCIs, or LPs.” C added, “Example of an inconsistency -- In the Stem, the phrase ‘occasional bumps’ is used; in the Prompts, the phrase ‘slight disturbances’ is used to mean the same thing.” Furthermore, most excerpts commented the exemplar response is “more of a 6th grade level explanation – the sentence structure and vocabulary is upper elementary 5-6 grade.”

Enhance engagement and inclusion

Most experts responded similarly to the failing “This (phenomenon) is fully comprehensible for students; perhaps not super compelling, but a very good relatable phenomenon for elementary students.” However, some NGSS and assessment experts have concerns if the phenomenon is compelling enough for students, especially for the NGSS expert who worried that the scenario will lead to misinterpretation of “side-to-side motion.” Engagement and teacher experts suggested that the task could be made more engaging by framing it as a story or using a hands-on demonstration. They also recommended using examples that are more directly relevant to students' everyday experiences. Sa proposed, “If we ran straight at a wall, we would bounce ‘off’ the wall instead of running ‘through’ it. What would it take for us to be able to run through a wall?” B mentioned, “Framing it as a story might be helpful for this age group.” Additionally, feedback also indicates the scenario is relatable to all students and that the language used does not inadvertently exclude any groups. For example, Sa commented, “The scenario is something students have experienced frequently.” But the experts also underline the importance of inclusion. Q noted, “scenario is culturally sensitive overall, but it’s important to ensure it’s inclusive for students with diverse backgrounds and abilities.”

Provide adequate scaffoldings

There was a strong recommendation to provide more scaffolding to help students understand the concept of balanced forces. This could include clearer instructions, visual aids, and step-by-step guidance. CH suggested, "Prompts scaffold for drawing the model and explaining the model. Small suggestion – you could scaffold students to use arrows." QL mentioned, "It would be better to tell them that they need to draw arrows and labels around the book. They may have no idea how to represent unseen forces." Several experts think the prompts do not provide adequate scaffolding for students to develop models. Table 4-20 below synthesizes the integrated findings of expert feedback on Task 1.

Table 4-20. Integrated Analysis Results for Task 1 of PE 3-PS2-1

Theme	Key Points	Recommendations
Simplify language for clarity and grade-level alignment.	Importance of clear, consistent, and precise language, avoid confusing terminology	Simplify language to match the reading levels of elementary students. Replace complex terms like "despite" and "remains unmoved" with simpler language, such as "even with slight bumps, the book does not move." Ensure consistent use of terms like "occasional bumps" and "slight disturbances." Define terms like "normal force" clearly in the context of the task Simplify language level of exemplar response.
Enhance engagement and inclusion	Suggestions to frame the task as a story or use hands-on demonstrations to increase interest. Ensure the scenario is relatable and inclusive for all students, including those with diverse backgrounds and abilities	Incorporate storytelling elements and opportunities for practical demonstrations. Use real-life examples and interactive elements to make the learning experience more dynamic. Frame the task as a relatable story or classroom event and include hands-on activities where students can physically manipulate objects to observe forces in action. Use examples and language that reflect diverse student experiences and backgrounds. Ensuring cultural sensitivity and inclusivity will help make the task more accessible to all students. For example, include culturally diverse names, contexts, and examples that reflect the backgrounds of the student population, making the task more engaging and relatable for all learners.
Provide adequate scaffolding	Need for clearer instructions and visual aids to help students understand balanced forces	Include step-by-step guidance, visual aids, and explicit instructions for modeling forces. Providing additional support materials, such as graphic organizers or visual aids, can help students organize their thoughts and responses effectively. For instance, provide clear diagrams with labeled arrows to show forces, and include detailed instructions that guide students through the process of modeling and explaining the forces at play.

4.2.3.3 Expert feedback analysis for PE: 3-PS2-1 Assessment Task 2

Feedback for Task 2 designed for LP2 of PE 3-PS2-1 was also collected using the same feedback protocols and same group of experts. Task 2 and its' exemplar response can be found from Figures 4-12 and 4-13. Figures 4-14, 4-15, and 4-16 visually present the expert feedback distribution.

Figure 4-12. Task 2 for LP2 of 3-PS2-1

***Task 2*: "The Unmoving Tug-of-War Challenge"(3-PS2-1-LP2-2)**

***Item Stem*:**

Jinni's class and Cody's class are playing tug-of-war at the school playground. Both classes are equally strong, pulling the rope with all their might, but guess what? The rope doesn't move an inch! It's like an invisible force is keeping everything perfectly balanced. How can both classes pull so hard and yet nothing changes?


(Generated by DALL.E on March 11, 2024)

Item Prompts:

1. Develop a model to explain:

- What's Happening: Draw the push and pull both classes are giving to the rope.
- How They Work Together: Show how the push from one class and the pull from the other class are just right so that neither side move. You can use labels and/or symbols to show your ideas.

(Draw your model here)

2. Why No One Moves: Explain with your picture why, even though both classes are trying really hard, the rope doesn't move at all.

(Explain your model here)

Figure 4-13. Exemplar response for task 2 for LP2 of 3-PS2-1

Model description:

- Elements in the Model:
  - Rope: The central element of the tug-of-war, represented by a straight line across the model.
  - Jinni's Class Force: Drawn as an arrow pointing to the left from one end of the rope, labeled "Pull by Jinni's Class".
  - Cody's Class Force: An arrow of the same length as Jinni's class arrow, pointing to the right from the other end of the rope, labeled "Pull by Cody's Class".
  - Ground: The surface beneath the rope, ensuring context is provided for the tug-of-war scenario.
- Relationships Among Elements:
  - The forces exerted by Jinni's class and Cody's class are equal in magnitude but opposite in direction, demonstrating a balance of forces.
  - This equilibrium of forces results in no net movement of the rope, symbolizing how balanced forces maintain the status quo, irrespective of the strength applied from both ends.
- Usage of Arrows and Labels:
  - Arrows of equal length in opposite directions illustrate the concept of balanced forces exerted by both classes.
  - Labels "Pull by Jinni's Class" and "Pull by Cody's Class" clarify the source of each force, making it easier for students to comprehend how these forces interact to maintain balance.

Explanation using my model:

My drawing shows us the secret behind why the tug-of-war rope between Jinni's and Cody's classes stays put, even though everyone is pulling as hard as they can. Each class is pulling the rope towards themselves with the same strength, but because these forces are exactly the same and pulling in opposite directions, they cancel each other out. It's like when you and your friend push against each other's hands with the same strength; neither of you moves backward or forward because the forces are balanced. So, even with all that effort from both classes, the rope doesn't budge because the pushes and pulls on it are perfectly matched. This is a neat example of how things stay still when the forces on them are in perfect harmony!

Figure 4-14. Heatmap of the expert ratings on the item stem of Task 2 for PE 3-PS2-1

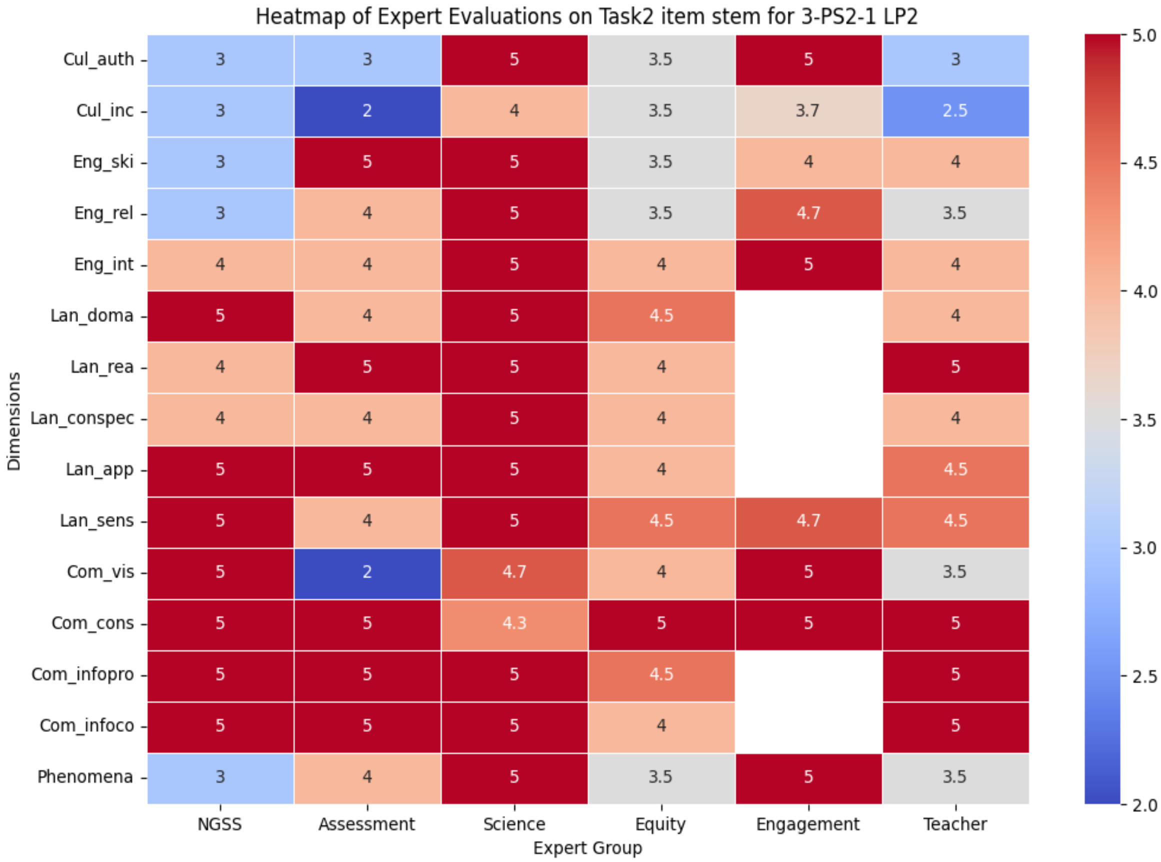


Figure 4-14 illustrates the heatmap of expert evaluations for the Task 2 item stem. The feedback revealed significant variations in how different expert groups perceived the phenomenon's clarity and engagement. Notably, NGSS experts provided lower ratings, suggesting that the phenomenon might not be as engaging or accessible to all students. This contrasted with higher scores from Science experts, who viewed the phenomenon as relatively engaging. This discrepancy indicates a need for adjustments to make the phenomenon more universally accessible and engaging. In terms of comprehension, all expert groups rated the consistency of terminology highly, which underscores the clarity in the use of terms, essential for student understanding. However, there was a notable variation in the perceived effectiveness of visual aids. Some experts recommended enhancements to improve comprehension, potentially including the addition of captions for better clarity. Engagement with the scenario, assessed through metrics of student interest and relevance to real-life and 3D learning, received moderate to high scores.

However, the variability in these scores suggests potential for further enhancing the scenario's captivation and relevance to the three dimensions of learning. Cultural sensitivity evaluations, particularly highlighted by lower scores from Equity experts, pointed to the task's limited resonance across diverse student groups. This suggests a pressing need for broader cultural considerations within the scenario to enhance inclusiveness and authenticity.

Figure 4-15. Heatmap of the expert ratings on the item prompt of Task 2 for PE 3-PS2-1

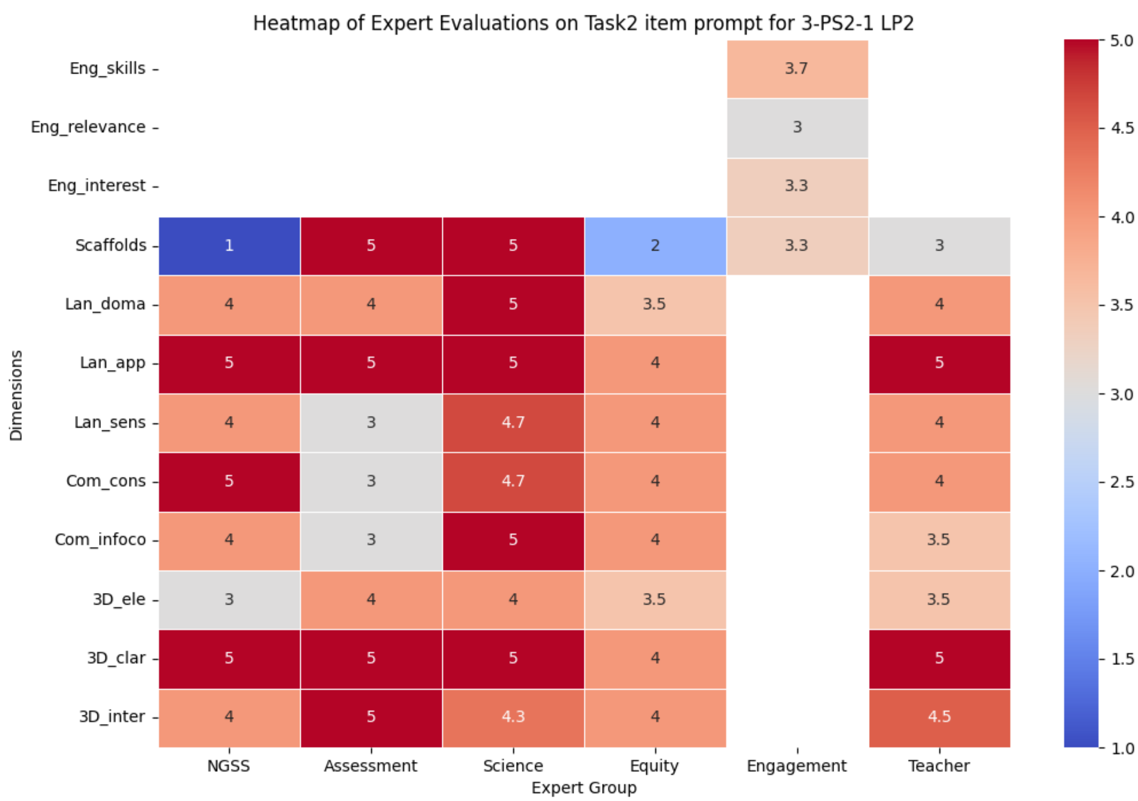


Figure 4-15 showcases a heatmap of expert evaluations for the Task 2 item prompt. The evaluation highlights two dimensions where the scores were notably low. The NGSS experts assigned a score of 1 to scaffolding and the equity and language experts assigned a score of 2, indicating significant concerns about the effectiveness of the scaffolding provided in supporting student understanding of the task. This suggests that the current scaffolding may not adequately help all students grasp complex concepts or engage deeply with the content. Enhancements in this area are crucial to ensure that the support structures are robust enough to facilitate comprehensive understanding across diverse student groups. Another low score was observed in the engagement relevance dimension, where the Engagement panel rated it 3, pointing to potential shortcomings in the scenario's ability to resonate with and captivate students' interests. This moderate score suggests that while the prompt has some engaging elements, it could be significantly improved to better capture and hold student interest, making the learning experience more compelling and relevant. In contrast, high ratings were consistently given by the Science panel across dimensions such as language appropriateness and 3D learning clarity, indicating that from a scientific and educational standpoint, the language and integration of learning dimensions are effectively executed. However, the mixed feedback across different panels, particularly the lower scores from the NGSS and Engagement experts, underscores the need for a more unified approach that aligns with NGSS standards and effectively engages students.

Figure 4-16. Heatmap of the expert ratings on the exemplar response of Task 2 for PE 3-PS2-1

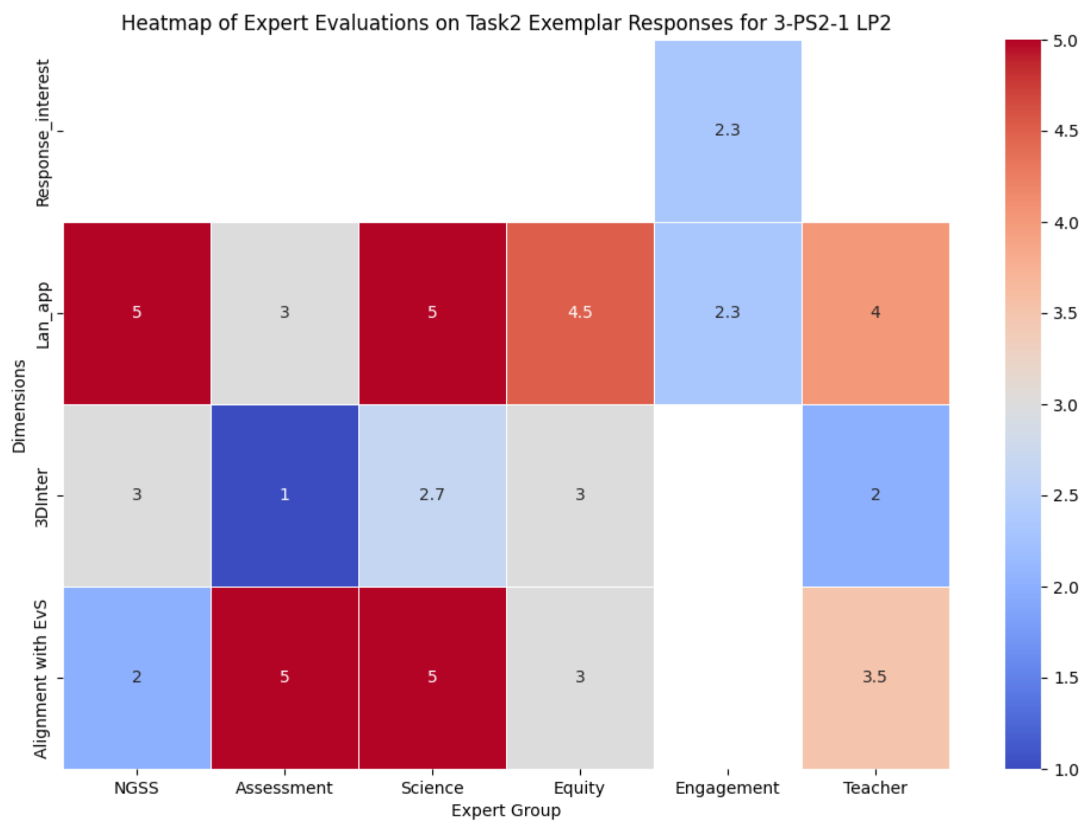


Figure 4-16 presents a heatmap of expert evaluations for the Task 2 exemplar responses. The analysis shows that the exemplar responses have significant disparities in how effectively they incorporate NGSS-required three-dimensional learning and engage students. The NGSS experts rated the Three-Dimensional Learning Integration at a notably low score of 1, indicating that the responses fall short in engaging students with the necessary competencies and learning dimensions mandated by the standards. Another area of concern highlighted by the Engagement experts is the low score of 2 for Response Interest. This suggests that the responses may lack elements necessary to capture and maintain student interest effectively, potentially impacting their educational effectiveness. Conversely, the responses were well-received in terms of their alignment with evidence statements, receiving high scores from Assessment and Science experts. This indicates that they accurately incorporate the required evidence statements. However, the Language Appropriateness received mixed evaluations, with a particularly low score from the Engagement group, suggesting the language may not be entirely suitable for all students.

With quantitative analysis, the next section provides qualitative analysis to further zoom in the specific suggestions. *Enhance engagement by considering individual experience.* Task 2 was generally found to be engaging and relevant, for instance, "very much comprehensible for students; a very good relatable phenomenon for elementary students." Experts also suggested enhancements to further consider individual experiences. Sa mentioned that the task stem is "clear and accessible" but emphasized the importance of considering individual interests. She stated, "I personally think this is quite interesting because children have likely played tug-of-war before, making it a good anchor point to expand their understanding of the phenomenon. However, interest is individual and exists on a scale. I'd be curious to ask: interesting in relation to what? As a standalone scenario, it works well as a classroom activity, but teachers might be better positioned to discuss if this is an interesting scenario for this age group as a whole or if there are more engaging activities." Sa also commented on the relevance, stating, "This will depend on the student. Generally, the stem has high 'interest value,' but the phenomenon's real-life application might depend on students' prior experiences with the game." E echoed this sentiment, cautioning that the task might not resonate with all students, especially those who have not played tug-of-war, suggesting, "Most students will not have done this before and would not relate to it."

Ensure coherent information across the task

Another theme is that incoherent information in tasks may lead to inefficient information processing. C mentioned that information incoherence exists in three areas: (a) the stem scenario presents a rope-pulling challenge, but the prompts require students to draw the push and pull that classes are giving to the rope; (b) the image in the scenario can mislead students to think each class is playing a separate tug-of-war game because it shows two games being played; (c) coherence is further disrupted by the prompt asking, "Show how the push from one class and the pull from the other class are just right so that neither side moves." This is echoed by science content expert P, who stated, " the stem uses one class 'pushes' and the other class 'pulls,' yet, in physics, both sides are pulling." Su similarly pointed out, "students would not suggest that one class is ‘pushing’ and one is ‘pulling’." These suggest the importance of emphasizing information coherence principles in the design process.

Simplify language for grade-level appropriateness

Experts emphasized the need to simplify the language used in the task to ensure it is suitable and understandable for third graders. Complex terms and phrases can hinder students' ability to engage with and comprehend the task. E noted that terms like "invisible force" could be confusing, suggesting instead, "'Invisible force' is pretty confusing for me and for kids." Similarly, P pointed out that phrases like "all their might" might not resonate with students, proposing it be changed to "really hard." SC also recommended simplifying vocabulary to better align with students' prior instruction and understanding. For the example response, C noted, "This is more of an upper elementary grade level explanation—the sentence structure and vocabulary are appropriate for grades 5-6." Language expert Su echoed this sentiment, stating, "... it is more sophisticated than expected for most third graders. It sounds more like text written by a middle schooler." However, the exemplar explanation remains student-friendly and appropriate for a student audience.

Ensure inclusive and clear visual representation

The visual depictions in the task received mixed feedback, highlighting concerns about their potential to mislead students and lack of inclusivity. E pointed out, "The image might be too sexist because the girls all have skirts on. And there are no students who are overweight." This underscores the need for visual materials to be inclusive and representative of diverse student populations. Su noted, "The photo with two groups in a tug-of-war is confusing, as some students might think it is showing two different tug-of-war events." Additionally, J raised concerns about inclusivity for students with disabilities: " it might be important to consider that students with physical disabilities might have more challenges with this question. I would keep this question but might change the image to include participation from students with physical disabilities, such as a kid in a wheelchair." This feedback emphasizes the importance of ensuring that visual aids are clear, unambiguous, and inclusive to prevent misinterpretation and promote understanding.

Ensure effective scaffoldings

The need for effective scaffolding to support students in understanding and completing the task was another key theme. Proper scaffolding can help students break down complex tasks into manageable parts and guide them towards successful completion. Q suggested providing example models, stating, "Including some sort of example model might be helpful for students." Similarly, C emphasized the importance of clear instructions and scaffolds, noting, "Scaffolds help with modeling – what to draw, what to show with the model, and encouragement to use symbols and/or labels." This feedback highlights the need for well-designed scaffolding to facilitate student understanding and engagement. Table 4-21 below synthesizes the expert feedback on Task 2.

Table 4-21. Integrated Analysis Results for Task 2

Theme	Key Points	Recommendations
Ensure coherent information across the task	Incoherent information in the task may lead to inefficient information processing	Ensure all information in the task is coherent and consistent. Align the stem, prompts, and images to present a unified scenario. Revise the language to consistently describe the actions of both classes as "pulling" to avoid confusion.
Simplify language for grade-level appropriateness	Complex terms and phrases could hinder students' understanding	Simplify the language to match the reading levels of third graders. Replace complex terms like "invisible force" with simpler alternatives. Use straightforward language and avoid jargon to ensure clarity.
Ensure inclusive and clear visual representations	Visuals could be misleading and lack diversity	Revise the images to ensure they clearly represent a single scenario and avoid depicting multiple games. Include diverse characters in the visual aids to reflect a variety of student backgrounds, including students with disabilities. Ensure that visuals are clear and unambiguous to prevent misinterpretation.
Enhance engagement by considering individual experience	The task is engaging but may not be relatable for all students	Consider individual interests and provide more relatable examples to increase student engagement. Incorporate storytelling elements to make the task more dynamic and captivating for students. Use practical demonstrations or hands-on activities to illustrate concepts. Ensure that the scenarios used are relevant to the everyday experiences of third graders, such as common playground activities or familiar classroom experiments.
Ensure effective scaffoldings	Need for more effective scaffolding to support student understanding	Provide example models and clear visual aids to help students understand the task. Include step-by-step guidance and explicit instructions for modeling forces, such as arrows and labels. Ensure that scaffolds are effectively integrated into the task, breaking down complex concepts into manageable parts. Provide additional support materials, like graphic organizers, to help students organize their thoughts and responses effectively.

4.2.4 Expert Feedback Analysis for PE: 3-LS4-3

4.2.4.1 Analysis of The Expert Feedback on LPs and Evidence Statements

Protocols distributed for feedback collection on PE 3-PS2-1 are the same, with the review content tailored specifically to the PE. Table 4-22 presents the LPs and evidence statements.

Table 4-22. LPs and evidence statements for LP2 for 3-LS4-3 for review

PE	3-LS4-3: Construct an argument with evidence that in a particular habitat some organisms can survive well, some survive less well, and some cannot survive at all.
LPs	LP1**: Students develop models to represent various organisms in a specific habitat and identify their basic needs for survival, illustrating the interdependence between organisms and their environment. LP2**: Students engage in argument from evidence to support claims about which organisms can survive well, less well, or not at all in a specific habitat based on their characteristics and needs, using examples from various habitats to explore cause and effect relationships. LP3**: Students analyze data to describe how certain adaptations help organisms survive in their habitats and explain the cause and effect relationship between specific adaptations and survival success. LP4**: Students predict the effects of minor environmental changes on the survival of organisms in a given habitat, identifying the cause and effect mechanisms that lead to these outcomes.
Focal LP: LP2	Students engage in argument from evidence to support claims about which organisms can survive well, less well, or not at all in a specific habitat based on their characteristics and needs, using examples from various habitats to explore cause and effect relationships.
Evidence Statements	1. Students collect and present specific evidence regarding the survival rates and adaptation mechanisms of organisms within varying habitats. - Students draw upon observable characteristics, inherent needs, and environmental factors influencing organismal survival.
	2. Students formulate clear claims regarding which organisms can thrive, survive less well, or perish in particular habitats, grounding their assertions in gathered evidence and understanding of habitat-organism interplay.
	3. Students succinctly explain, with examples, how specific habitat features afford or limit the survival capabilities of certain organisms, highlighting adaptation as a key determinant.
	4. Students predict survival outcomes for distinct species, explaining the role of physical and biological habitat components in determining these outcomes given comparative habitat scenarios.

Figure 4-17. Heatmap for expert feedback on LPs and evidence statements for 3-LS4-3

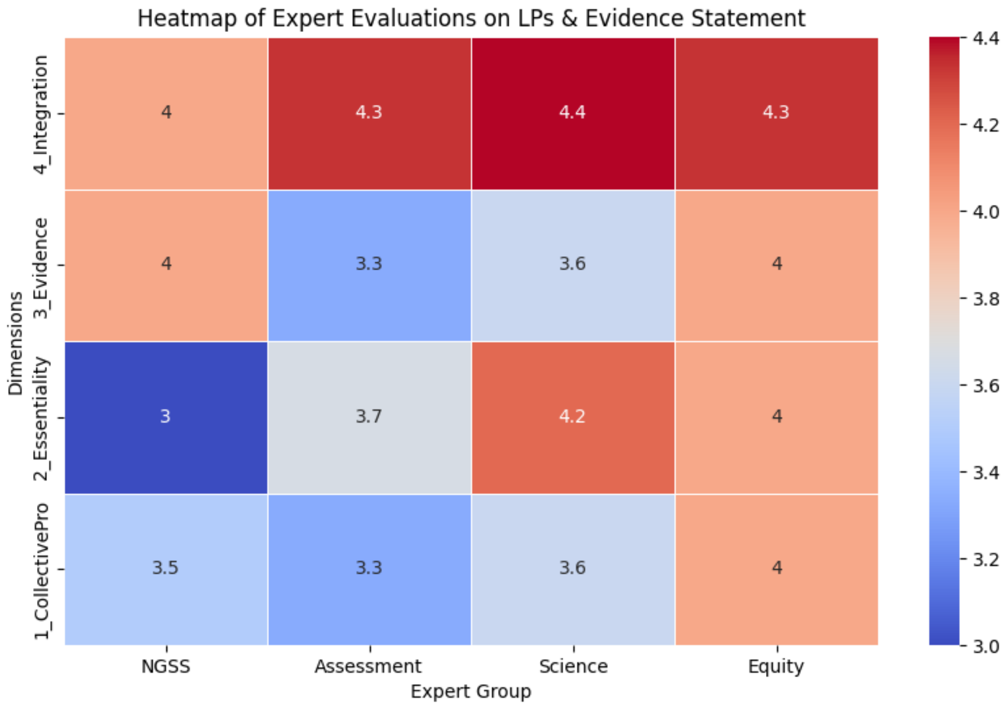


Figure 4-17 shows the quantitative analysis of expert feedback on the LPs and Evidence Statements for the PE. The expert evaluations reveal that the Collective Representation of Proficiencies scored a low of 3.5 by NGSS experts, highlighting a concern about the comprehensiveness of the LPs in covering all necessary proficiencies required by the performance expectation. This suggests significant gaps that may impede the achievement of the targeted educational outcomes. Furthermore, the Essentiality of the LPs received a modest score of 3 from NGSS experts, indicating some uncertainty about the critical nature of all components within LPs. This points to potential overreach in the current LPs, suggesting that some elements may not be essential for meeting the performance expectations and could be streamlined or eliminated. In terms of the Sufficiency of Evidence, the Assessment group's score of 3.3 raises concerns about whether the evidence statements adequately support claims of student proficiency. This feedback suggests that the evidence provided may not be sufficiently comprehensive or robust, necessitating enhancements to better support student assessments. Additionally, the Integration of Knowledge was rated slightly lower by the NGSS group at 4, implying that while the integration of knowledge generally meets educational standards, there is room for improvement in how effectively three-dimensional learning is incorporated into the LPs. These insights were taken to thematic analysis. Through the analysis, several themes emerged. The themes were also organized into Table 4-23 below.

Refine LPs to align with PE

Concerns regarding the appropriateness of the scope of LPs were often raised, highlighting that LP3 and LP4 included concepts beyond the intended scope of the PE. M noted, "Adaptation and environmental changes are not addressed in the PE," while C observed that "LPs 3 and 4 introduce DCI elements outside the PE," indicating a need for refinement to align these LPs more closely with the objectives of the PE. C further explained, "LP1 and LP2 accurately represent the proficiencies for the PE, with LP2 covering the entire scope. However, LPs 3 and 4 introduce unnecessary elements, such as adaptations resulting from selective pressures, which are beyond the third-grade curriculum." He added, "LP3 delves into adaptations that aid survival, more suitable for higher grades, and LP4, aligned with 3-LS4-4, focuses on environmental impacts, exceeding the intended PE focus." This trend of overreach was supported by J's comment that "LP4's predictions about environmental impacts extend beyond the PE's scope." Similarly, SS noted, "The detail in LP2 matches the original PE well, indicating its suitability, whereas LP3's adaptation content aligns with eighth-grade standards." NGSS expert E critiqued the excessive scope, saying, "LPs sometimes exceed necessary proficiencies. For instance, understanding 'adaptation mechanisms' isn't required; students only need to argue about organisms' varying survival likelihoods, such as comparing different aquatic species in Lake Michigan. The current focus on adaptation in LP3 is unwarranted." He also mentioned that the level of evidence required, as stated in the LPs, is often unrealistic for classroom settings. These critiques underscore the importance of tailoring content to be age-appropriate and directly aligned with PE goals. Adjustments should include scaling back advanced topics and simplifying explanations to ensure they are accessible to third graders, including English Language Learners. This approach will enhance clarity and relevance, ensuring LPs effectively meet the educational needs at the intended grade level.

Redefine evidence statements for enhanced clarity and filling gaps

The evidence statements associated with the LPs need substantial refinement to ensure clarity and age-appropriate alignment. Sm highlighted concerns with Evidence Statement 1, which suggests that "Students collect and present specific evidence regarding the survival rates and adaptation mechanisms of organisms within varying habitats." He concerned the clarity of the statement and questioned the realism of students "collecting" data, suggesting instead that they might be "identifying" data that qualifies as evidence. For Evidence Statement 2, which states, "Students formulate clear claims regarding which organisms can thrive, survive less well, or perish in particular habitats, grounding their assertions in gathered evidence and understanding of habitat-organism interplay," Sm criticized the vague language and called for more specific discussion about the organisms' needs relative to their environments. He appreciated the specificity in Statement 3 but pointed out that its focus on "adaptation" aligns with eighth-grade standards rather than third grade, indicating a misalignment with the intended curriculum. E also noted issues with the scope of the evidence described in these statements, particularly that the complexity of gathering such detailed evidence might not be feasible in many classrooms. Additionally, E highlighted that the term "adaptation" in Evidence Statement 3 does not align with third-grade expectations. C echoed these sentiments, noting that while the evidence statements address the three dimensions outlined in the LPs and are collectively obtainable, the breadth of evidence required sometimes exceeds the scope intended for the LPs. For instance, he pointed out that requiring students to compare different habitat scenarios goes beyond the narrow habitat focus expected at this grade level. C also identified a missing component in the SEP for grades 3-5, which includes critiquing explanations—a critical thinking skill not currently reflected in the evidence statements or the corresponding LPs. This comprehensive feedback highlights the need for more precise, age-appropriate adjustments to the evidence statements to ensure they effectively support the intended learning outcomes without overreaching the PE.

Refine the LPs and evidence statements for better accessibility and understanding

The integration of the three dimensions within the LPs has been well-received, exemplified by J's commendation of LP2 for effectively demonstrating this integration. However, the clarity and accessibility of the LPs and evidence statements remain critical areas for improvement. E from the NGSS expert group raised concerns about the use of complex language in the LPs, noting that terms like "succinct explanations" could be challenging, particularly for young learners and English Language Learners (ELLs). She expressed doubts about ELLs' ability to produce clear and succinct claims, highlighting the subjectivity of such requirements. Further, E critiqued the alignment of the DCI, suggesting a misunderstanding in the AI's interpretation of the PE, particularly with regards to adaptation. She pointed out that the essential idea is about animals meeting their needs in supportive environments, not adaptation per se. He also noted that the AI overlooked the simplicity intended in the PE, which is designed to be universally applicable across various educational settings. She also highlighted gaps in the practical implementation of the LPs, such as the unrealistic expectation for all activities to be conducted in a single well-known location like a schoolyard. She argued that the AI's design of the PE did not adequately consider the logistical and contextual realities of typical third-grade classrooms in the U.S. Furthermore, experts like Cn and Co emphasized the need for explicit instructions on evidence collection and analysis. Cn observed that the LPs lacked detailed guidelines for analyzing data, which is crucial for supporting students' arguments with evidence. Co added that the guidelines on how students should gather evidence were insufficiently clear, underlining the need for detailed and actionable instructions to aid students in their investigative processes. These insights call for a revision of the LPs and evidence statements to ensure they are not only aligned with the NGSS's 3D approach but also tailored to be clear, accessible, and practical for implementation in diverse educational environments. This includes simplifying language, clarifying expectations, and providing concrete, context-appropriate guidelines that accommodate the capabilities and realities of third-grade students, especially ELLs.

Table 4-23. Integrated Analysis Results for LPs and Evidence Statements for PE 3-LS4-3

Theme	Key Points	Recommendations
Align LPs with PE to Ensure Age-Appropriate Content	Concerns were raised about LP3 and LP4 including concepts beyond the intended scope of the PE, such as advanced adaptations not suitable for third grade. Experts noted that LP1 and LP2 align well with the PE, but LP3 and LP4 introduce unnecessary complexity. The grain size of LP2 is too similar to the original PE.	Refine LP3 and LP4 to eliminate advanced concepts not required at the third-grade level. Focus on simplifying content to ensure it is age-appropriate and directly aligned with the PE. Avoid overreach, especially for "adaptations." Further unpack LP2, especially focus on unpacking the meanings of survive well, not well, and not at all.
Refine Evidence Statements for Clarity and Educational Relevance	Evidence statements were criticized for their lack of clarity and realism in expectations. Concerns include the feasibility of students collecting data versus identifying data, and the vague language that doesn't specify organism needs in relation to environments. Also, there was a misalignment with grade-level standards, particularly with the use of the term "adaptation" which is more suited to eighth grade.	Revise evidence statements to be more specific and clear, ensuring they are age-appropriate. Replace "collecting" with "identifying" to better reflect realistic classroom activities. Clarify and specify the interplay between organisms and their habitats to enhance understanding and relevance. Exclude advanced terms like "adaptation" that align with higher educational standards.
Enhance Accessibility and Clarity in LPs and Evidence Statements	Integration of the three dimensions within LPs is well-received, yet the use of complex language and unrealistic logistical expectations highlights a need for simplification and practical adjustments. Concerns about the difficulty for ELLs to produce clear and succinct claims were noted, alongside issues with the practical implementation of evidence collection guidelines.	Revise LPs and evidence statements to simplify language and reduce complexity, making them more accessible, especially for ELLs. Ensure that the instructional materials and tasks are feasible within the common logistical and contextual boundaries of third-grade classrooms. Provide clear, actionable guidelines for evidence collection and analysis to support students in their learning processes effectively.

4.2.4.2 Expert Feedback Analysis for PE: 3-LS4-3 Assessment Task 1

Figures 4-18 and 4-19 represent Task 1 and its corresponding response for LP2 of 3-LS4-3. Expert panels were provided with the same review protocol reported above. Figures 4-20, 4-21 and 4-22 present the heatmap visualizations of expert ratings on Task 1.

Figure 4-18. Task 1 for LP2 of 3-LS4-3

***Task 1*: " Squirrels and Their Search for the Perfect Home" (3-LS4-3-LP2-1)**

***Item Stem*:**

Amy's third-grade class is on a mission to discover what makes a perfect home for squirrels in their local park. To aid their investigation, they're equipped with a data table (Table a) capturing squirrel observations across different park areas over four weeks, alongside a brief on squirrel needs and park habitat characteristics (Table b).


(Generated by DALL.E on March 11, 2024)

Table a. Squirrel observation

Week	Trees Area (Squirrels Observed)	Open Area (Squirrels Observed)
Week 1	*15*	*3*
Week 2	*18*	*2*
Week 3	*20*	*4*
Week 4	*17*	*1*

Table b. Squirrel Needs & Habitat Characteristics

Area	Squirrel Needs & Habitat Characteristics
Trees Area	Abundant oak and pine trees (food sources: acorns and pine cones), dense foliage for shelter.
Open Area	Limited vegetation, few scattered trees, exposure to predators, and human activity.

Item Prompts:

1. Based on the data and information provided, choose whether you agree or disagree with the statement: "Squirrels prefer habitats with more trees."

2. Use the weekly observations from the data table to support your position. Include considerations about the squirrels' needs for food and shelter, and how the Trees Area and Open Area meet these needs.

3. Discuss how the presence of trees or lack thereof might affect squirrel behavior and habitat preference, connecting to their survival needs.

Figure 4-19. Exemplar response for task 1 for LP2 of 3-LS4-3

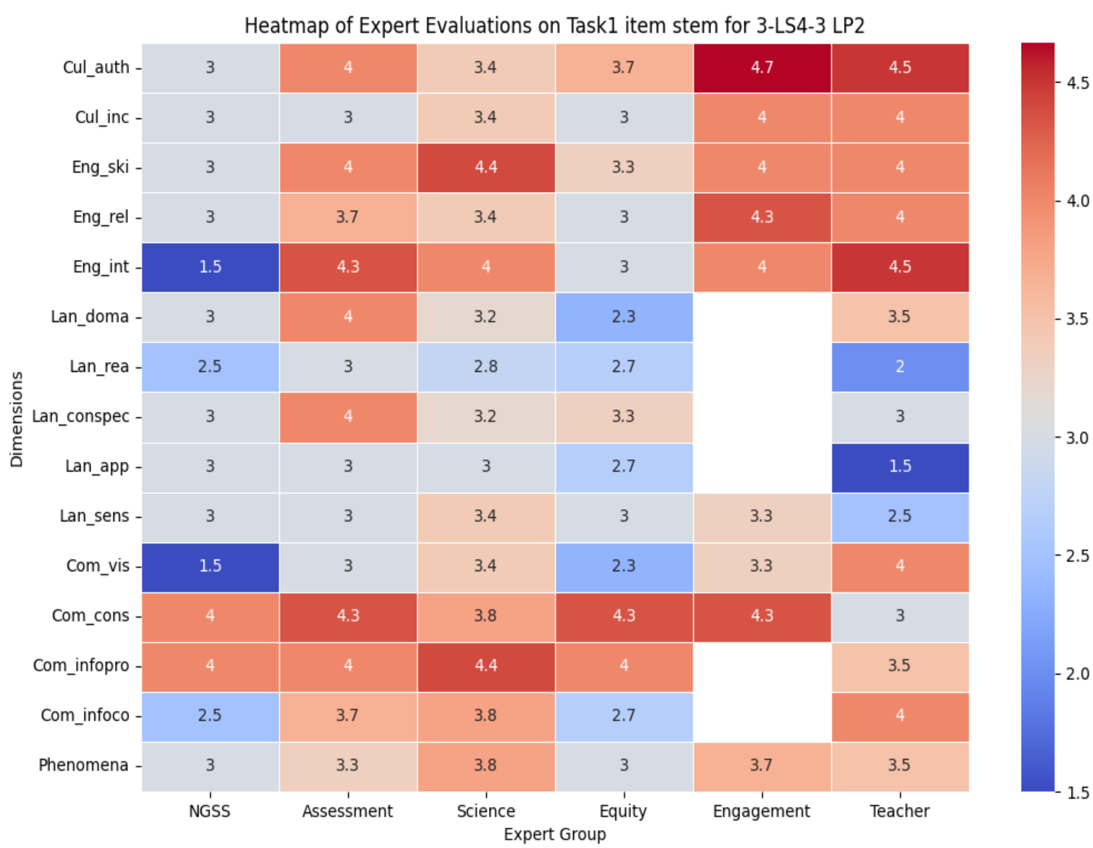
Exemplar Response for " Squirrels and Their Search for the Perfect Home"

1. Agree or Disagree: "I agree that squirrels prefer habitats with more trees."

2. In the data table, I saw that every week, more squirrels were found in the Trees Area compared to the Open Area. For example, in the first week, 15 squirrels were in the Trees Area and only 3 in the Open Area. This pattern kept going for all four weeks. This shows that squirrels are seen more where there are lots of trees."

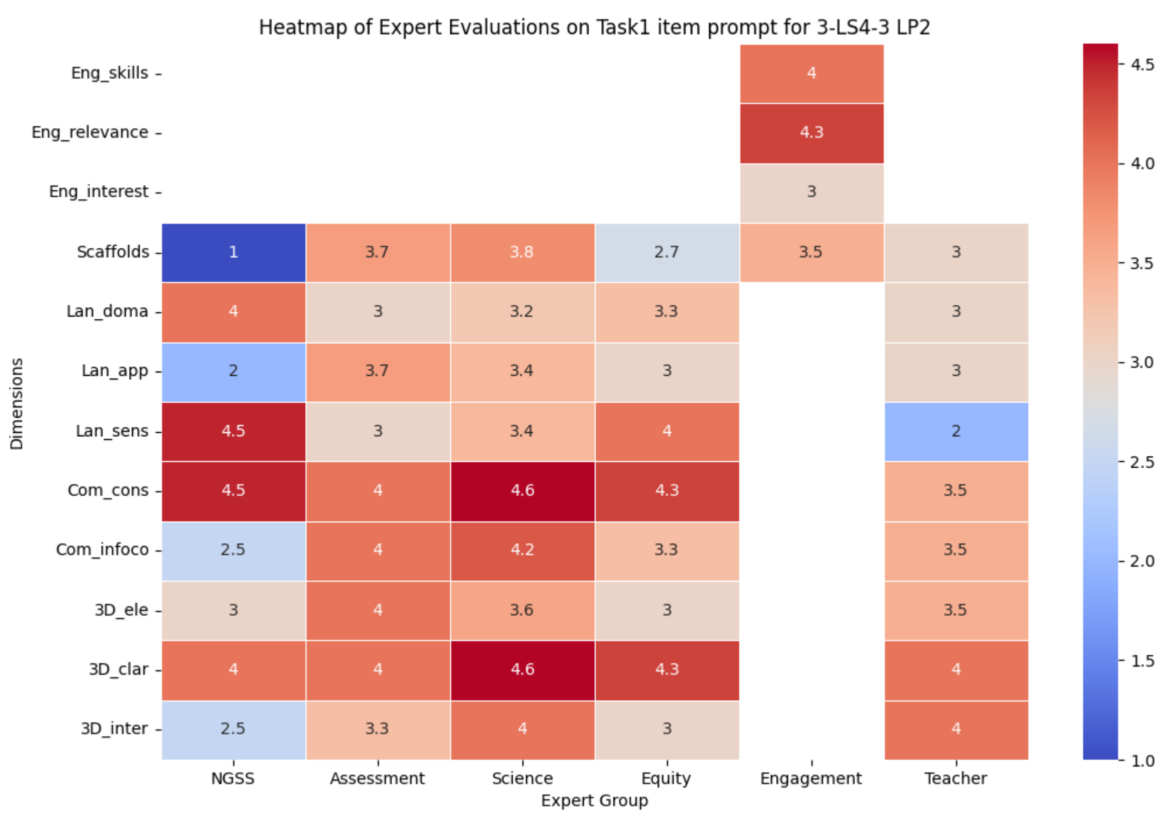
3. Trees are super important for squirrels because that's where they find their food like acorns and make their homes up in the branches where it's safe. In the Trees Area, there are many trees for squirrels to eat from and hide in, which is why they like it better. The Open Area doesn't have many trees, so there's not much food or places for squirrels to build their nests, making it a not-so-good place for them to live. That's why the presence of trees makes the Trees Area a better home for squirrels. It has everything they need to be happy and safe!

Figure 4-20. Heatmap of the expert ratings on the Task 1 item stem for PE 3-LS4-3



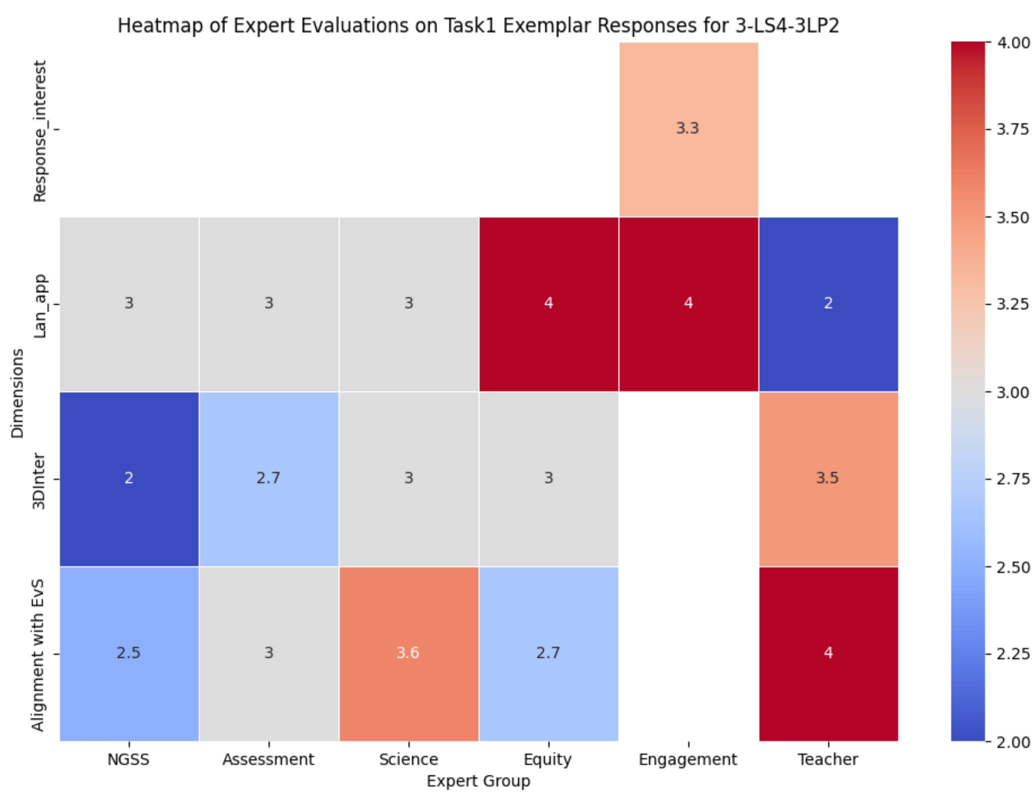
The expert evaluations for the Task 1 item stem reveal critical concerns and strengths. The NGSS group's feedback was particularly notable, emphasizing significant deficiencies in language complexity and information processing, which were underscored by low scores of 1.5 in both Engagement Interest and Visual Comprehension. These ratings highlight serious issues regarding the item stem's capacity to effectively engage students and the clarity of visual aids necessary for comprehension. While the terminology used was consistently rated highly across all expert groups, affirming its clarity, the effectiveness of visual aids was inconsistent, prompting suggestions for more descriptive captions to enhance understanding. Engagement assessments showed variability, with some groups noting the item stem's potential to captivate students and connect the material to real-life and three-dimensional learning, yet indicating that there is room to boost the stem's overall engagement appeal. Additionally, cultural sensitivity was rated lower, especially by Equity experts, pointing to the item stem's limitations in addressing the diverse backgrounds of all students.

Figure 4-21. Heatmap of the expert ratings on the Task 1 item prompt for PE 3-LS4-3



The evaluations highlight critical areas of concern particularly in the dimensions of Scaffolds and Language age-appropriateness, where notably low scores were observed. The NGSS group's feedback was especially critical, giving the lowest score in the Scaffolds category, suggesting that the scaffolds provided may not adequately help students break down the complexity of the task. This points to a potential disconnect between the scaffolding support and the students' ability to engage with and understand the content effectively. Language sentence structure received lower scores from teacher experts, indicating possible issues with the clarity and appropriateness of the vocabulary used. Moreover, low scores in 3D Integration from groups such as the NGSS suggest that the prompt may not fully align with the integrated proficiencies expected in the LP2.

Figure 4-22. Heatmap of the expert ratings on the Task 1 exemplar response for PE 3-LS4-3



The heatmap reveals a range of scores, with particular attention needed in the dimension of 3D integration. The NGSS group provided a low score of 2, indicating that the responses may not adequately integrate the required 3D aspects. This low score suggests a critical need for improving how the responses demonstrate interconnected scientific ideas and practices according to NGSS standards. In contrast, the evaluations for Alignment with Evidence Statements were generally favorable, with the Science expert group rating it at 3.6. This indicates that the exemplar responses reasonably reflect the necessary evidence statements, although there is room for further alignment to fully meet the expectations. Language appropriateness in the responses also varied, with the Teacher group giving a lower score of 2, highlighting concerns over whether the language used is direct and comprehensible enough for students. This feedback points to a need for simplifying the language or improving explanations to ensure that students can easily understand and engage with the content. Diving into the explicit feedback, several themes were identified for the expert feedback on task 1.

Enhancing engagement with real-world task scenarios

Experts from various fields stressed the need to make educational tasks both engaging and relevant. They suggested adding lively behaviors of squirrels to the tasks to ensure they relate to students' experiences and reflect accurate data, keeping students interested. For instance, NGSS experts noted that showing squirrels searching for homes is relevant, particularly for students familiar with squirrels living in trees. They suggested enriching the task by showing squirrels jumping and interacting with their environments. Assessment experts also emphasized the need for more captivating content. C noted that the task relates well to students' experiences with parks and animals, suggesting that focusing on different parks rather than areas within a single park could make the task more concrete and engaging. Sm recommended directly addressing how animals adapt to their environments to improve the task's relevance. Science content experts mentioned that squirrels are a common sight across the U.S., and most students find them interesting. They recommended adjusting the animals studied to better reflect the students' local wildlife, which could make the tasks more engaging. Sa proposed using characters like Pokémon to make activities more fun for young learners. This feedback highlights the importance of creating educational tasks that are engaging and closely connected to students' real-life experiences, ensuring a richer and more meaningful learning experience.

Simplifying language that aligns with grade-level language ability

Experts across various fields underscored the critical need for using clear and straightforward language to ensure educational tasks are accessible for third graders. NGSS experts advocated for consistent terminology, recommending more direct phrases like "to help their investigation" instead of "to aid their investigation," and clearer labeling such as "Squirrel observation data table." They also suggested replacing phrases like "support your position" with "support your choice" to simplify communication. Assessment experts emphasized the necessity of using simpler vocabulary to aid comprehension. They identified complex words such as "aid," and "equipped " as overly challenging for third graders, recommending they be replaced with simpler alternatives. P noted the importance of shortening long sentences to make them easier for young learners to understand, while C and Sm stressed providing clear definitions of scientific terms to help students, particularly those with lower reading proficiency or from non-English-speaking backgrounds. Science content experts concurred, pointing out that the task's language was too complex for third-grade students. They suggested substantial simplification, such as replacing "foliage" and "vegetation" with more straightforward terms. Consistency in terminology was highlighted by H, who advised that "habitat" should be the consistent term used throughout educational materials to avoid confusion. Equity and language experts also emphasized the need for clear and direct sentences, with E commenting on the need to simplify phrases like "to aid their investigation" to "to help their investigation." Co argued that the language in educational tasks should provide an accessible starting point for all students, particularly for those who are multilingual or have lower reading levels. Teacher experts reinforced these points, advocating for the use of simpler terms and shorter sentences to improve understanding. B and Le noted that certain phrases and terms (e.g., “equipped,” “alongside a brief,” “considerations,” “abundant,” and “dense foliage”) used in the tasks were not appropriate for third graders, suggesting more age-appropriate language and clearer instructions for data collection activities. This feedback emphasizes the essential role of using clear and straightforward language to ensure that educational tasks are accessible and comprehensible for third-grade students.

Enhancing visual clarity in assessment tasks

Experts across various disciplines emphasized the critical importance of using clear, accurate, and contextually appropriate visual aids to enhance student understanding and ensure scientific accuracy. The visual aids should closely align with the text and data, offering clear instructions to avoid confusion and make learning tasks more effective and accessible for all students.

NGSS experts stressed the necessity for visuals to accurately represent the described phenomena. They criticized some visuals for not adding value and potentially distracting students, such as images showing squirrels on the ground instead of in trees, contrary to the textual description. They underscored the importance of clear instructions and the definition of scientific terms to help students connect visuals with the accompanying text and data tables. Assessment experts noted inconsistencies between some images and the data presented, suggesting that visuals should be both clear and functional. For instance, C and Sm observed that one image showed more squirrels in an open area than indicated by the data table, potentially leading to student confusion. They advocated for visuals that are directly aligned with the data and include clear, directive captions to guide student interpretation. Science content experts and equity and language experts proposed using multiple images to accurately represent different scenarios described in the text, ensuring that visuals are not only scientifically precise but also culturally inclusive. For example, Cn suggested adding captions to enhance clarity, especially for students who may not be familiar with the subject matter. Teacher experts also highlighted the importance of visual accuracy. Le criticized images that misrepresented the data by showing squirrels in inappropriate settings, suggesting adjustments to better reflect the factual content or employing several images to depict varying habitats accurately. B recommended enhancing visual aids with captions like "Squirrel observation trip" to clarify the context and engage students effectively. This collective feedback from experts underlines the need for the assessment tasks to incorporate well-designed visual aids that are not only scientifically accurate but also tailored to support and enhance the learning experience.

Enhancing the accessibility of the assessment tasks with refining scaffoldings

NGSS experts have emphasized the necessity of straightforward terminology and the alignment of visual aids with textual and data information to reduce confusion and enhance learning. They noted instances where visuals did not accurately represent the described phenomena, suggesting that precise and informative captions could help link visuals to the underlying data effectively. Assessment experts, P pointed out the importance of labels and detailed captions to clarify visual aids, enhancing students' ability to connect these images with textual explanations. Science content experts remarked on potential discrepancies between images and data, which could lead to confusion. They recommended using multiple images to accurately represent different aspects of the data discussed, ensuring that all visual representations are scientifically accurate and aligned with educational goals. Additionally, equity and language experts advocated for the use of diverse and relatable visuals that cater to a broad range of student backgrounds, ensuring inclusivity in educational materials. Teacher experts underscored the need for visual aids that accurately match the data, suggesting adjustments to images to better align with the educational content and the use of captions to provide context and enhance understanding. The integration of effective scaffolding, including clear instructions, consistent terminology, and supportive visual aids, is essential. These elements help break down complex tasks into manageable parts, enabling students, especially those with lower reading proficiency for ELLs, to grasp and engage with the content more effectively. This tailored support is critical for fostering an accessible and inclusive learning environment.

Underline enhancing task relevance and inclusivity

Inclusivity and cultural sensitivity were central themes in the feedback from various expert groups. They emphasized the need to use examples and language that resonate with the diverse experiences and backgrounds of all students, ensuring tasks are accessible to multilingual learners and those with lower reading levels. Experts suggested incorporating more familiar animals and environments to enhance relatability and engagement. NGSS experts recommended making the tasks more inclusive by exploring various environments, providing context on the relevance of parks and animals like squirrels. They stressed the importance of connecting students' prior experiences with the content presented in the tasks. Assessment experts advocated for using different parks as focal points to make scenarios more tangible and engaging, while also aligning more closely with the PE through explicit discussions on adaptation mechanisms. Science content experts noted the potential irrelevance of tasks for students in regions without squirrels, suggesting the inclusion of universally familiar animals to ensure no student feels alienated. Cn highlighted that students in extreme urban or rural settings might find tasks centered around uncommon local flora and fauna less applicable. Equity and language experts focused on simplifying language and using clear visual aids to make tasks more engaging and comprehensible. They proposed modifications to the task instructions and content to make them clearer for younger students, suggesting the use of a checklist format to clarify expectations. Engagement experts advocated for adapting the content to reflect the animals and environments that are familiar to the students' own geographical backgrounds, arguing this would make the tasks more inclusive and engaging. They pointed out that assuming familiarity with squirrels and parks might exclude students who lack such experiences. Teacher experts echoed these concerns, emphasizing the need to adapt educational tasks to reflect the diverse environments and experiences of students. They suggested that students who have never visited parks or seen squirrels first hand would find such tasks less meaningful, advocating for the use of more relatable and accessible content. This collective feedback underscores the importance of designing educational tasks that are not only scientifically accurate but also culturally sensitive and inclusive, catering to the diverse educational needs and backgrounds of all students. Table 4-24 below synthesizes the analyses of expert feedback on Task 1.

Table 4-24. Integrated Analysis Results for Task 1 of PE 3-LS4-3

Theme	Key Points	Recommendations
Enhancing Engagement with Real-World Scenarios	Experts emphasized adding behaviors of squirrels to relate to students' real-life experiences and maintain engagement. They suggested using lively behaviors and familiar animals to make the tasks more engaging and relevant. NGSS experts specifically noted the relevance of squirrels in tree environments and proposed including dynamic interactions like jumping. Assessment experts suggested using different parks to make scenarios more tangible.	Include dynamic aspects of animal behavior to enrich tasks. Adapt the animal subjects to reflect the local wildlife familiar to students' geographical backgrounds, making scenarios more relatable and engaging. Use characters or elements like Pokémon to add fun and intrigue for younger learners. Focus on different parks rather than areas within a single park to provide concrete, engaging content.

Table 4-24 (cont'd)

Simplifying Language for Third Graders	Consistent and straightforward language is crucial. Experts across fields highlighted the need for clear language and terminology suitable for third graders. Complex terms and phrases like "aid," "equipped," and "foliage" were noted as problematic. The importance of breaking down long sentences and providing clear definitions was emphasized, particularly for students with lower reading proficiency or from non-English-speaking backgrounds.	Use simpler language and terminology that third graders can easily understand. Replace complex phrases with more direct alternatives, such as changing "to aid their investigation" to "to help their investigation." Ensure terminology consistency throughout the educational materials, using terms like "habitat" uniformly to avoid confusion. Provide clear, concise instructions and definitions to aid comprehension, especially for multilingual learners and those with lower reading levels.
Enhancing Visual Clarity in Assessment Tasks	Clear, accurate, and helpful visual aids are essential for supporting student understanding and ensuring scientific accuracy. Experts noted that visuals must align with the text and data and provide clear instructions. Inconsistencies between images and data, such as showing squirrels in incorrect settings, were highlighted as potentially confusing.	Ensure that visual aids accurately represent the described phenomena and align closely with the text and data. Use multiple images to represent different scenarios accurately, and include captions to enhance understanding and provide context. Adjust images to reflect factual content accurately and employ visuals that are both scientifically precise and culturally inclusive. Provide clear, directive captions to aid interpretation and ensure that visual aids are directly supportive of the educational content.
Refining Scaffoldings to Enhance Accessibility	Effective scaffolding is key to helping students navigate complex tasks. Experts stressed the importance of clear, consistent, and supportive scaffolding to aid comprehension and engagement. This includes ensuring that terminology and visual aids are straightforward and align with the textual and data information provided in the tasks.	Implement straightforward terminology and align visual aids with textual and data information to reduce confusion and enhance learning. Include precise and informative captions to link visuals to underlying data effectively. Ensure scaffolds are clear and directly supportive of the content, reflecting accurate data to avoid misconceptions. Provide detailed captions to clarify visual aids, enhancing students' ability to connect these images with textual explanations.

Table 4-24 (cont'd)

Enhancing Task Relevance and Inclusivity	Inclusivity and cultural sensitivity are crucial for making educational tasks accessible and engaging for all students. Experts suggested using examples and language that resonate with students' diverse experiences and adapting content to include familiar animals and environments. The importance of connecting students' prior experiences with the content was emphasized, as was the need to use clear visual aids and simple language.	Adapt educational tasks to reflect diverse environments and experiences, using familiar animals and settings to ensure no student feels alienated. Simplify language and use clear visual aids to make tasks more engaging and comprehensible. Provide an accessible starting point for all students, particularly for multilingual learners and those with lower reading levels. Adjust tasks to include animals and environments familiar to students' backgrounds to enhance engagement and inclusivity.

4.2.4.3 Analysis of Expert Feedback on Assessment Task 2

Figures 4-23 and 4-24 presents Task 2 and its corresponding exemplar response designed for LP2 of 3-LS4-3. Figures 4-25, 4-26, and 4-27 provide heatmaps of the expert feedback on item stem, item prompt and exemplar response for Task 2 for LP2 of 3-LS4-3.

Figure 4-23. Task 2 for LP2 of 3-LS4-3

Task 2**: "The Mystery of the Growing Sunflowers"(3-LS4-3-LP2-2)**

Item Stem**:**

In the school's garden, there are various sections where different plants flourish. One section is full of sunflowers that grow tall and healthy, while another section attempts to grow the same sunflowers, but they struggle and barely bloom. Mrs. Smith, the science teacher, presents this observation to her 3rd-grade class.

(Generated by DALL.E on March 11, 2024)

Mrs. Smith asks her students to explain why sunflowers in one section of the garden thrive while those in another section do not. She provides them with data (Table a) on sunlight exposure, soil moisture, and soil type for each garden section.

Table a. Garden information

	Phenomena	**Sunlight**	**Soil Moisture**	**Soil Type**
Sunny Spot	Sunflowers are tall and healthy.	Lots of sunshine all day.	Just right, not too wet or dry	Rich and full of nutrients
Shady Spot	Sunflowers are small and struggling	Very little, mostly in the shade	Quite damp, doesn't dry out much	A bit clayey and hard for roots to grow

Item Prompts:

1. Think Like a Scientist: Why do you think sunflowers in the sunny spot are doing so well compared to the ones in the shady spot? Remember, plants need certain things to grow big and strong.

2. Use the Clues: Look at the information about both spots in the garden.

- If you think sunlight is important, tell us why the sunny spot might be better for the sunflowers.
- If you're curious about the soil, share how different soil types might affect the sunflowers' growth.

3. Solve the Mystery: Using what you've learned, explain why you think the sunflowers are happier in one part of the garden than the other. Is it just about the sun, or is the soil also a big clue?

Figure 4-24. Exemplar response for task 2 for LP2 of 3-LS4-3

Exemplar Response for "The Mystery of the Growing Sunflowers"

1. I think the sunflowers in the sunny spot love getting lots of sun because plants need sunlight to make their food. It's like how we need to eat breakfast to have energy. The sunflowers getting more sunlight can make more food, so they grow big and strong!

2. The soil part is interesting! The sunflowers in the sunny spot have soil that's just right – not too wet and full of good stuff that helps them grow. It's like having a perfect bed to sleep in; you sleep well and wake up happy! But the shady spot's soil is too wet and hard, so I guess it's hard for sunflower roots to spread out and find food.

3. So, putting it all together, I think the sunflowers are way happier in the sunny spot because they get all the sun they need, and the soil is just perfect for them. In the shady spot, even though they try to grow, it's tough without enough sun, and the soggy, hard soil doesn't help either.

Figure 4-25. Heatmap of the expert ratings on Task 2 item stem for LP2 of PE 3-LS4-3

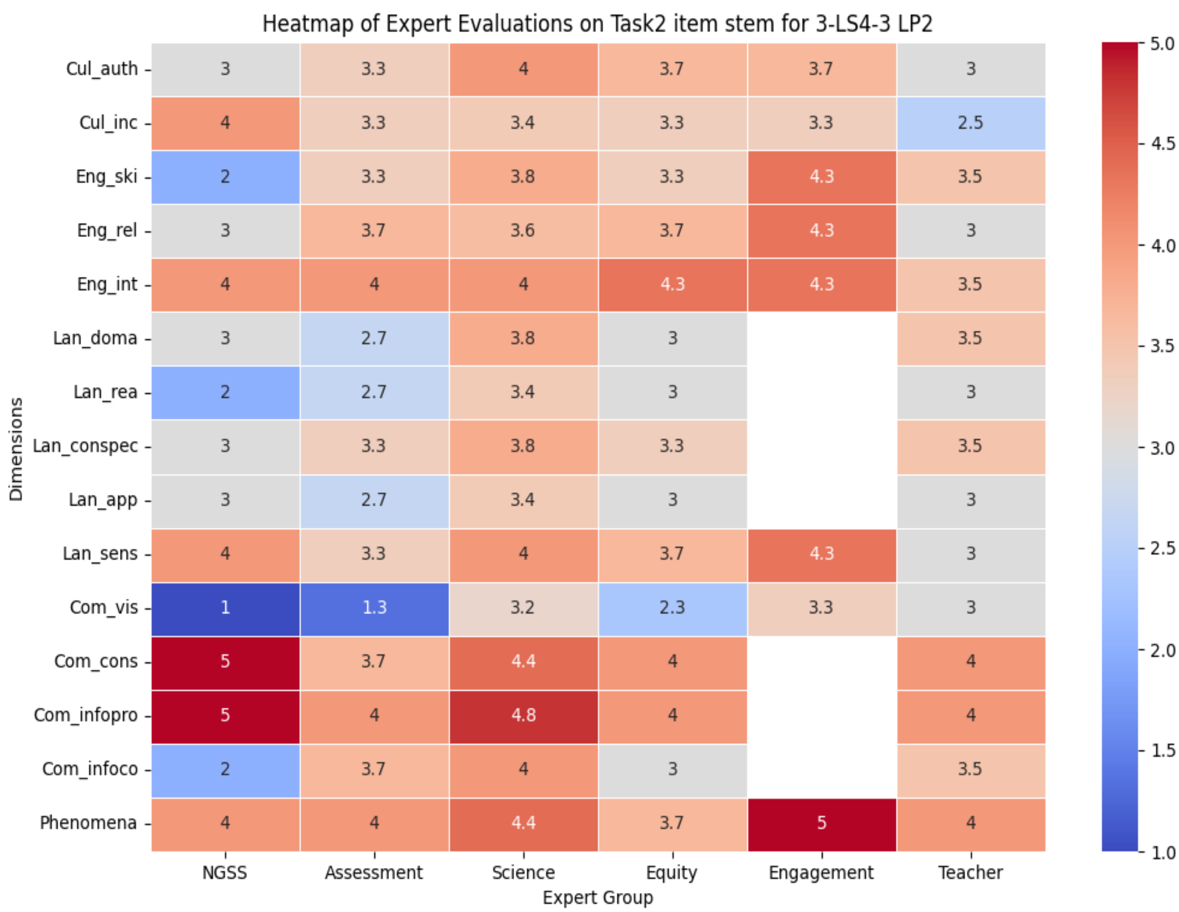
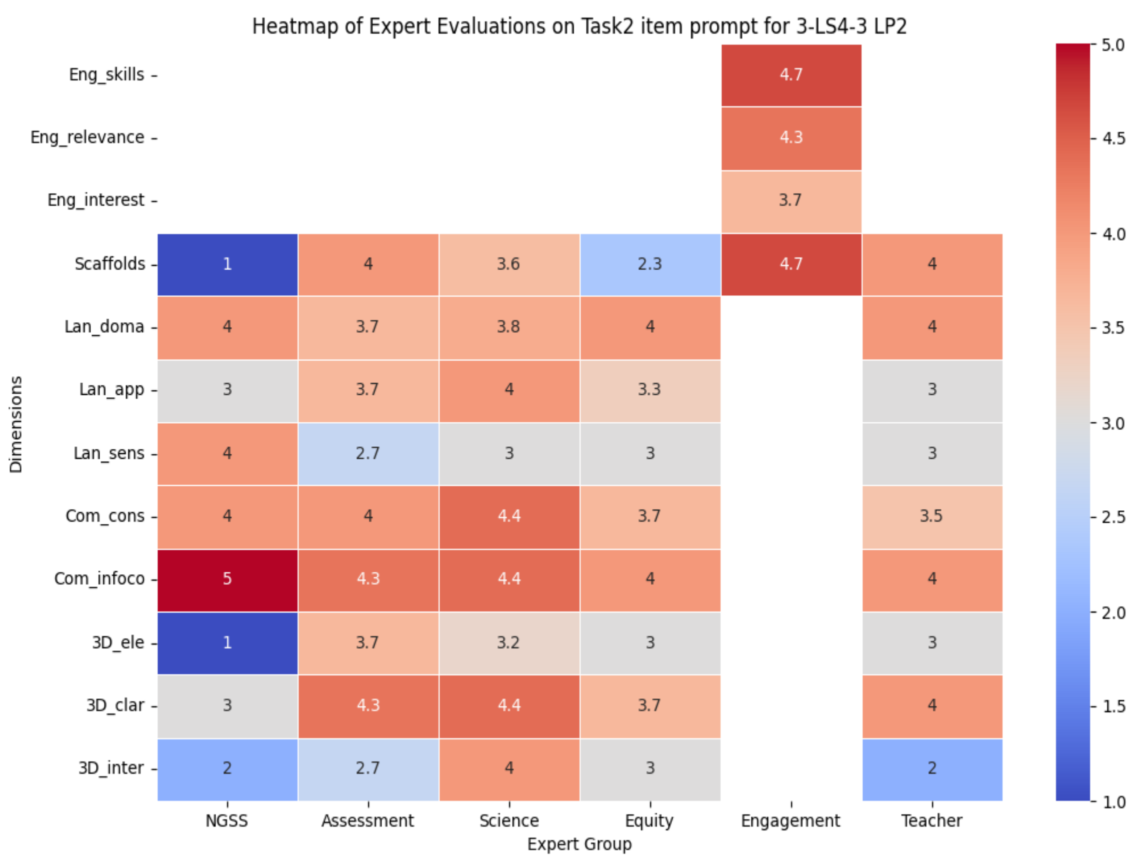


Figure 4-25 indicates several areas of concern that need addressing. Notably, Visual

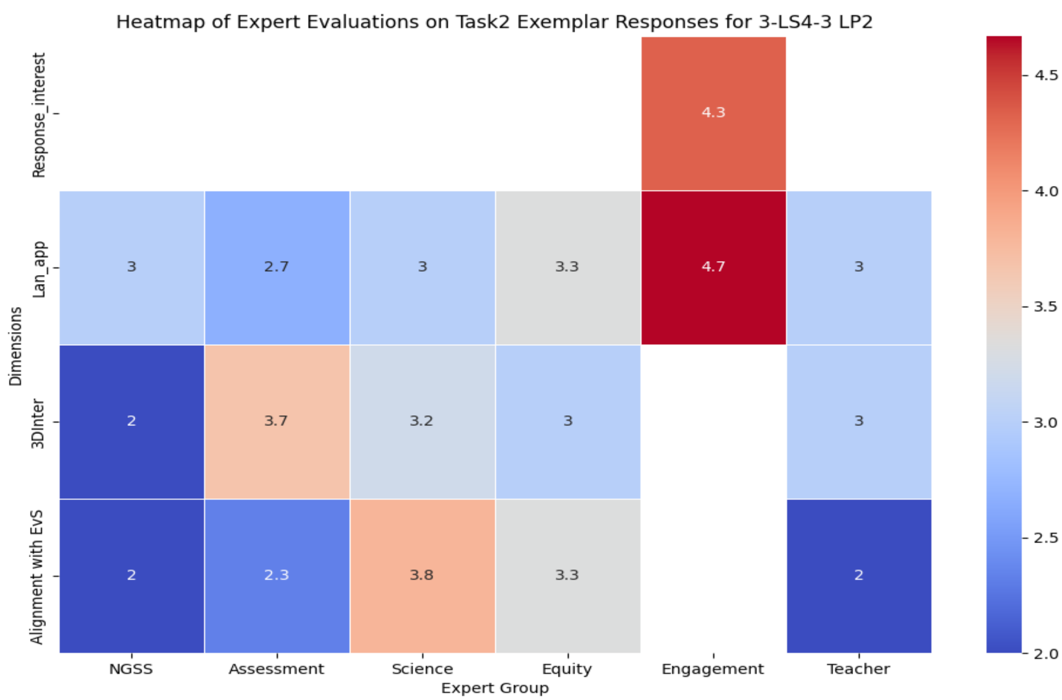
Comprehension received the lowest scores, with the NGSS group rating it at 1 and the Assessment group at 1.3. These scores indicate significant issues with the visual aids used in the item stem, suggesting that they do not effectively support student comprehension and may lack necessary detail or clarity. Language Complexity also showed variability, particularly in language reading level, where the NGSS group provided a low score of 2. This feedback suggests that the language used might not be accessible to all students, requiring simplification and clarification to ensure it is appropriate for the target grade level. Cultural Sensitivity received moderate scores, with the teacher group noting that the item stem might not fully resonate with or be inclusive of all student demographics. The Comprehension consistency dimension received higher ratings, particularly from the NGSS expert group, which rated it at 5. This suggests that the terminology used in the item stem is consistent and clear, aiding student understanding.

Figure 4-26. Heatmap of the expert ratings on Task 2 item prompt for LP2 of PE 3-LS4-3



The evaluations reveal several critical areas of concern and strength. The lowest scores were observed in the Scaffolds dimension, where the NGSS group rated it at 1, indicating significant issues with how scaffolds are used to support student understanding. This suggests that the scaffolds provided may not adequately help students break down the complexity of the task, pointing to a need for more effective scaffolding strategies. Language sentence structure also showed variability, with the assessment group providing a score of 2.7, suggesting that the language might not be fully accessible to all students. In contrast, the Comprehension Information dimension received high scores, particularly a 5 from the NGSS group, indicating strong alignment with what is being described in the scenario and clear focus on the essential information students need to respond to the prompt. Similarly, the dimension of 3D Clarity received favorable reviews, with scores of 4.3 and 4.4 from both the Assessment and Science expert groups. Engagement metrics were rated highly by the Engagement expert group.

Figure 4-27. Heatmap of the expert ratings on Task 2 exemplar response for LP2 of PE 3-LS4-3



The evaluations highlight significant concerns in the 3D Integration and Alignment with Evidence Statements dimensions, with the NGSS group rating both at 2. These low scores indicate that the responses do not adequately integrate NGSS's three dimensions of learning or align with the necessary evidence statements, suggesting a need for improvement. Language age appropriateness showed variability, with the Assessment group rating it at 2.7, indicating that the language may not be clear or appropriate for the target grade level, necessitating refinement for better clarity and accessibility. On a positive note, the Engagement dimension received high scores, particularly a 4.7 from the Engagement expert group, showing that the responses are engaging and relevant to students' lives. Qualitative analysis has similar themes.

Balancing engagement with realism in assessment tasks

A key observation from all expert groups was the critical balance between engagement and realism in Task 2. NGSS experts, E, found the task particularly engaging due to the distinctive "magic" sunflower element, which was highly rated for engagement. However, concerns were raised about its realism and alignment with NGSS standards, which emphasize real-world environments. E critiqued, "The sunflower picture is extraordinary. Not scientific, a different planet," highlighting the importance of using scientifically accurate visuals to avoid misleading students. Assessment experts, like C, also pointed out the potential confusion caused by the oversized sunflower image, noting that it detracted from the task's narrative coherence. Science content experts acknowledged the task's appeal, with J stating, "The phenomenon is compelling and relatable," yet emphasized the need for visuals that are both engaging and realistic to prevent student confusion. Teacher expert B appreciated the task's ability to engage students but cautioned that the large sunflower could cause confusion, suggesting the addition of a clarifying caption to better integrate the visual with the educational content. This feedback underscores the need for educational tasks to not only be attractive but also accurately reflect scientific standards and real-world scenarios.

Enhancing language clarity and consistency in assessment tasks

All expert groups highlighted significant concerns regarding the clarity and consistency of language in educational content, emphasizing the necessity for simplicity and adherence to scientific standards. NGSS experts, particularly E, criticized the use of terms such as "flourish" and "struggling" for third graders, citing their lack of scientific rigor and appropriateness. The consistency of terminology was rated as inadequate, receiving a score of 3 out of 5, prompting recommendations for simpler, scientifically accurate terms. Assessment experts like C pointed out the complex and inconsistent language usage, identifying terms such as "flourish," "exposure," and "moisture" as potentially confusing and overly advanced for third graders. Science content experts, including J and H, agreed, advocating for the simplification of vocabulary and sentence structures to align better with students' reading levels and improve comprehension. Teacher experts, such as BF and Le, emphasized the importance of straightforward, concise instructions to enhance comprehension and engagement. B noted that while most sentences were well-constructed, some words required replacement to better accommodate third-grade understanding. Equity and language experts like Sc further stressed the essential role of using grade-appropriate vocabulary to prevent confusion and ensure inclusivity, making educational tasks accessible and understandable for all students. This collective feedback underscores the need for careful language selection in assessments to support effective learning at the third-grade level.

Enhancing inclusivity and cultural relevance in assessment tasks

NGSS experts expressed concerns that the task's school garden setting might not resonate with students from urban environments who are unfamiliar with gardening. To enhance engagement, assessment experts, including P, suggested that the scenario could be more captivating with a clearer, more relatable question. Engagement experts, like Sa, recommended adapting the scenario to include plants that are indigenous to students' local environments, thus increasing its cultural relevance and inclusivity. Equity and language experts, such as Co, underscored the need for diverse visual representations and clearer distinctions in the task's data presentation. They pointed out inconsistencies in describing plant health and recommended labeling areas as "Spot A" and "Spot B" to better delineate different garden sections. Teacher experts highlighted the critical need for culturally relevant and diverse learning experiences, noting that students in urban areas might find the gardening scenario less applicable, due to their limited exposure to such environments. These insights from various expert groups underline the necessity of designing tasks that are not only scientifically accurate but also broadly accessible and culturally sensitive, ensuring all students can engage with and benefit from the learning experiences provided.

Enhancing assessment tasks through focused variables and effective scaffolding

Experts consistently emphasized the strategic use of key variables in enhancing assessment tasks. NGSS experts expressed concerns about the task's complexity, attributing it to the inclusion of multiple variables such as sunlight, soil moisture, and soil type. They argued that this complexity could overwhelm and confuse students. E specifically suggested reducing the number of variables, advocating for a focus on a single primary variable to enhance clarity. Similarly, assessment experts C and Sm observed that the multiple variables muddled the concept of "habitat," as the task had placed sunflower sections within the same garden environment. They recommended a design that contrasts distinct environments, thereby reducing ambiguity and enhancing student understanding. Science content experts, including J and H, stressed the importance of using realistic and instructive visual aids aligned with the data to prevent misconceptions. Teacher experts also emphasized consistent and accurate data presentation as crucial for helping students draw clear and meaningful conclusions. B noted the critical need for consistently using scientifically accurate terminology throughout the task. The discussion further highlighted the role of effective scaffolding in supporting student comprehension. Assessment experts C and Sm lauded the simplification of scientific terms but suggested further refinement to clarify the language. J commended the structured approach of the task, which methodically guides students through the scientific inquiry process, thereby enhancing their inquiry skills. Engagement experts noted the benefits of specific scaffolding elements, such as reminders about plants' needs, which effectively guide student learning. Teacher experts advocated for clear and direct scaffolding to improve learning outcomes and comprehension. They urged that instructions within the task be detailed and explicit, particularly for more complex sections, to ensure that students can fully engage with and understand the content. Collectively, these insights point to the necessity of carefully designing assessment tasks that focus on key variables and incorporate well-planned scaffolding strategies to foster a comprehensive understanding among students, especially in the context of science education. Table 4-25 presents the summary.

Table 4-25. Integration of Findings for Task 2 of PE 3-LS4-3

Theme	Key points	Recommendations
Balancing Engagement with Realism in Assessment Tasks	Experts noted the importance of balancing engagement with realism. NGSS experts found the "magic" sunflower element engaging but criticized its lack of realism. EM highlighted the sunflower as "extraordinary" and not scientifically accurate. Assessment and science content experts like C and J noted the oversized sunflower could cause confusion and detracted from realism. Teacher expert B suggested adding a caption to integrate the visual with the educational content better.	Ensure educational tasks are engaging yet accurately reflect scientific standards and real-world scenarios. Use scientifically accurate visuals and align narrative coherence to prevent confusion. Add clarifying captions to integrate educational content effectively.
Enhancing Language Clarity and Consistency in Assessment Tasks	All expert groups emphasized the need for clear and consistent language. NGSS experts like E criticized terms like "flourish" and "struggling" for lacking scientific rigor. Assessment experts identified complex terms as confusing for third graders. Science content and teacher experts advocated for simplifying vocabulary and sentence structures to match student reading levels and enhance comprehension. Equity and language experts emphasized the use of grade-appropriate vocabulary to ensure inclusivity and accessibility.	Simplify vocabulary and sentence structures to align with third-grade reading levels and scientific accuracy. Use straightforward, concise instructions to enhance comprehension and engagement. Replace complex words with simpler alternatives and ensure terminology consistency to support effective learning.
Enhancing Inclusivity and Cultural Relevance in Assessment Tasks	Experts highlighted the importance of inclusivity and cultural relevance. NGSS experts noted the school garden setting might not resonate with students from urban environments. Engagement experts suggested adapting scenarios to include local plants. Equity and language experts emphasized the need for diverse visual representation and clearer data presentation. Teacher experts pointed out the critical need for culturally relevant and diverse learning experiences, especially for students in urban areas with limited exposure to gardening.	Design tasks that are culturally sensitive and inclusive, using scenarios and visuals that resonate with diverse student backgrounds. Adapt scenarios to reflect local environments and include culturally relevant plants and animals. Label data clearly and use visuals that represent diversity effectively.
Enhancing Assessment Tasks through Focused Variables and Effective Scaffolding	Experts consistently emphasized the strategic use of key variables and effective scaffolding. NGSS experts expressed concerns about task complexity due to multiple variables. They suggested focusing on a single primary variable for clarity. Science content experts advocated for realistic visual aids aligned with data. Assessment and teacher experts noted the importance of precise language and clear scaffolding to support student comprehension.	Simplify tasks by focusing on one primary variable to enhance clarity and comprehension. Use visual aids that are realistic and aligned with data. Provide clear and detailed instructions, especially for complex tasks. Incorporate effective scaffolding strategies that guide students through scientific inquiries and enhance learning outcomes.

4.2.5 Cross-Case Synthesis: Summary of Expert Feedback on Refining Knowledge-In-Use Assessments

There are several major themes emerging after analyzing the experts' feedback. These emerging themes provide major guidelines and directions for further revising the products that are aiming for the knowledge-in-use assessment design. I reported the common and significant themes into two major sections, including themes related to LPs and evidence statement design and themes related to task design.

4.2.5.1 Themes Related to LPs and Evidence Statements Design

I first present the summary of the overall themes related to the LPs and evidence statement design in Table 4-26. Then, I specify each theme with detailed explanations.

Table 4-26. Summary of the themes related to the LPs and evidence statement design

Theme	Theme Description	Example Strategies
Ensuring Appropriate Grain Size	This theme focuses on designing LPs and evidence statements to accurately reflect the scope and complexity outlined in the PEs. It ensures that the content is neither too broad nor too narrow and adheres closely to NGSS standards.	1. Focus exclusively on the ideas specified in the PE, avoiding advanced topics beyond the grade level. 2. Identify elements in LPs that go beyond the scope of the PE. Simplify and align content with grade-level expectations.
Improving Integration of CCCs, DCIs, and SEPs	This theme emphasizes the synergistic integration of Crosscutting Concepts (CCCs), Disciplinary Core Ideas (DCIs), and Science and Engineering Practices (SEPs) to enhance students' understanding of scientific principles through a 3D learning model.	1. Ensure that LPs and evidence statements explicitly demonstrate the integration of CCCs, DCIs, and SEPs, showcasing their mutual reinforcement. 2. Create content that clearly defines and exemplifies the connections between CCCs, DCIs, and SEPs, making these links explicit in the learning material.
Ensuring Consistency in Terminology	This theme underlines the importance of using uniform terminology across all educational materials to prevent confusion and ensure a consistent learning experience. It focuses on the structured and logical presentation of information.	1. Establish and use a consistent set of terms across all LPs and evidence statements to avoid confusion and enhance clarity. 2. Align language with learning goals to ensure that terminology supports the understanding of key concepts.

Ensuring appropriate grain size of LPs and evidence statements that adhere to PE boundaries

This theme emphasizes the importance of designing LPs and evidence statements that accurately reflect the scope and complexity outlined in the PEs. "Grain size" refers to the level of detail and specificity within the LPs and evidence statements, which must be carefully calibrated to ensure they are neither too broad nor too narrow relative to the expectations set by the NGSS standards. The goal is to ensure that each LP and evidence statement fully captures the necessary concepts without introducing extraneous content or omitting crucial information. Adhering to PE boundaries means that the content must directly align with the defined standards, avoiding any extension beyond the intended scope or depth. This precise alignment is crucial for maintaining the integrity and focus of the assessment tasks, ensuring they truly measure what they are intended to measure. This theme also informs the first principle to refine the design. Table 4-27 summarizes the strategies and example prompts to refine the design. The strategies were generated based on the data and then I designed the prompts to instruct the GPT-4 models.

Table 4-27. Summary of the strategies and example prompts for coverage of the PE

Strategies	Description	Exemplar Prompt
Ensure Content Matches PE Requirements	Focus exclusively on the ideas specified in the PE, avoiding advanced topics beyond the grade level.	"Generate LPs for PE 3-PS2-1 that focuses solely on the concept of balanced and unbalanced forces. Ensure the content does not extend into advanced topics like gravitational fields, which are beyond the third-grade curriculum. Provide a clear explanation suitable for third graders."
Highlight and Correct Content that Exceeds PE Requirements	Identify elements in LPs that go beyond the scope of the PE. Simplify and align content with grade-level expectations.	"Review the following LP draft for PE 3-LS4-3: 'Students analyze how environmental changes can lead to plant and animal adaptation.' Identify and list elements in this draft that exceed the scope of third-grade expectations, focusing on the unnecessary inclusion of adaptation mechanisms, and suggest modifications to simplify the content."
Address Content that Falls Short of PE Requirements	Assess evidence statements to ensure they meet the required understanding as specified in the PE. Revise statements to directly tie to the core concepts and skills outlined in the PE.	"Assess this evidence statement for PE 3-PS2-1: 'Students describe how different objects move.' Indicate how this statement falls short of addressing the required understanding of forces and motion as specified in the PE. Propose a revised statement that directly ties object movement to the types of forces acting on them."

Table 4-27 (cont'd)

Correct Overreaching Content	Revise LPs to remove advanced topics that are not required by the PE. Focus on observable, grade-appropriate properties and processes.	"The draft LP includes the analysis of intermolecular forces in water samples. This topic is not required by the PE and is too advanced for the grade level. Please revise the LP to focus on observable properties of water like state changes and buoyancy, which align with the core ideas in the curriculum."
Revisiting the Unpacking Documents	Correct or refine the unpacking of the PE to ensure alignment with the intended learning goals and appropriate scope.	"Revisit the unpacking document for PE 3-PS2-1. Identify and correct any misalignments or overextensions beyond the grade-level expectations. Refine the unpacking to ensure it accurately reflects the core concepts and skills specified in the PE, providing clear and concise guidelines for the development of LPs and evidence statements."

Improving integration of CCCs, DCIs, and SEPs

The theme "Improving Integration of CCCs, DCIs, and SEPs" emphasizes enhancing the integration of CCCs, DCIs, and SEPs within LPs and evidence statements. This approach aims to deepen students' understanding of scientific principles through a 3D learning model advocated by the Framework for K-12 Science Education (NRC, 2012) and utilized by the NGSS. Such integration ensures that learning not only meets curricular standards but also connects more effectively with real-world applications, making scientific reasoning more intuitive and contextually relevant for students.

Experts underscore the necessity of developing LPs and evidence statements that not only cover individual components of CCCs, DCIs, and SEPs but also demonstrate their synergistic interaction. This integration is crucial for providing students with a cohesive understanding of scientific concepts. LPs and evidence statements should clearly articulate the relationships between CCCs, DCIs, and SEPs, clarifying how these dimensions interlink within the assessment materials. Further, it's important to address any gaps in the current integration of these dimensions within LPs to ensure that these elements are seamlessly woven into LPs. The complexity of content integration should be tailored to match the cognitive and developmental stages of the learners, ensuring that the material is both engaging and comprehensible at the intended grade level. Any integrations that are too complex or advanced for the target audience should be simplified, focusing on delivering clear, tangible, and relatable content. Regular review and refinement of unpacking documents are also recommended to ensure they accurately guide the development of integrated LPs and evidence statements (see Table 4-28).

Table 4-28. Summary of the strategies and example prompts for 3D integration

Strategies	Description	Exemplar Prompt
Comprehensive Integration of Dimensions	Ensure that LPs and evidence statements explicitly demonstrate the integration of CCCs, DCIs, and SEPs, showcasing their mutual reinforcement.	"Generate an LP for PE 3-PS2-1 that clearly integrates the concept of forces (DCI) with the practice of scientific investigation (SEP) and the concept of cause and effect (CCC). Provide an example that illustrates these connections in a scenario relevant to third graders."
Explicitly Define Connections	Create content that clearly defines and exemplifies the connections between CCCs, DCIs, and SEPs, making these links explicit in the learning material.	"Develop an evidence statement for PE 3-LS4-3 that exemplifies how changes in an environment (DCI) affect animal behaviors (CCC) and how students can investigate these changes through data collection (SEP)."
Address and Strengthen Integration Gaps	Identify areas where the integration of CCCs, DCIs, and SEPs is weak or unclear in existing LPs and revise them to strengthen these connections.	"Review the LP for PE 3-PS2-1 focusing on motion and forces. Identify where the integration of CCCs and SEPs could be enhanced to better illustrate the interplay of these dimensions. Propose revisions that enhance this integration."
Ensure Age-Appropriate Integration	Align the complexity of the integrated content with the cognitive and developmental level of the learners, ensuring it is appropriate for their grade level.	"Refine the LPs and Evidence statements to ensure the grade-appropriate level of DCIs, SEPs and CCCs integrations. For instance, the model should be a simple model , and evidence does not need to be sufficient."
Rectify Overly Complex Integrations	Simplify overly complex integrations that may confuse or overwhelm students, focusing on clear, tangible examples that reflect grade-appropriate learning.	"Revise the LP that currently integrates advanced genetic concepts into a third-grade curriculum on plant growth. Simplify it to focus on observable traits (DCI), pattern recognition (CCC), and basic data gathering (SEP)."
Revisit Unpacking Documents for Alignment	Review and refine the unpacking of standards documents to ensure that they accurately guide the development of integrations among CCCs, DCIs, and SEPs.	"Revisit the unpacking document for PE 3-LS4-3. Ensure that the descriptions accurately reflect how CCCs, DCIs, and SEPs should be integrated for third-grade students, providing a clear framework for developing LPs and evidence statements."

Ensuring consistency in terminology and coherence of information

The theme is critical for maintaining clear communication and logical progression within educational materials, particularly in LPs and evidence statements. It emphasizes the importance of using uniform terminology to avoid confusion and ensure that students receive a consistent educational experience across different topics. Moreover, coherence in content demands that information is presented in a structured and logical manner, which is essential for students to understand and build upon complex scientific concepts effectively.

Experts have suggested several approaches to enhance consistency and coherence. The adoption of a standardized glossary ensures that the same terms are used consistently across all materials, helping students to familiarize themselves with specific scientific language without the added difficulty of synonyms that might appear in different contexts. Structuring information logically allows students to follow the natural progression of ideas, which is crucial for grasping more complex theories and principles. Integrating concepts across various LPs can reinforce knowledge and show the interconnectedness of different scientific areas. Additionally, simplifying complex concepts makes the material more accessible, especially for younger students. Table 4-29 presents the summary of ensuring information coherently.

Table 4-29. Summary of the strategies and example prompts for information coherently.

Strategy	Description	Exemplar Prompt
Standardize Terminology	Establish and use a consistent set of terms across all LPs and evidence statements to avoid confusion and enhance clarity.	"Ensure that the term 'force' is uniformly used in all LPs related to PE 3-PS2-1, defining it clearly the first time it appears."
Align Language with Learning Goals	Adjust language to clearly reflect the learning goals and ensure that terminology supports the understanding of key concepts.	"Review the evidence statement for PE 3-LS4-3 to ensure that all terms align with the defined learning goals, adjusting language for clarity and educational alignment."
Enhance Coherence in Content	Ensure that content across LPs and evidence statements logically flows and supports a cohesive understanding of the curriculum.	"Create a sequence in LPs for PE 3-PS2-1 that progressively builds on the concept of forces, ensuring a coherent flow that facilitates deeper understanding."

Table 4-29 (cont'd)

Review and Refine Content Regularly	Periodically review LPs and evidence statements to maintain consistency and coherence, updating as necessary to align with evolving educational standards.	"Conduct a quarterly review of the LPs for PE 3-LS4-3 to check for terminological consistency and content coherence, making adjustments based on the latest educational research and feedback."
Simplify Complex Concepts	Break down complex ideas into simpler, understandable components while maintaining the integrity and accuracy of the scientific information.	"Simplify the explanation of ecological niches in the LP for PE 3-LS4-3, using straightforward examples and consistent terminology to enhance student comprehension."

4.2.5.2 Themes related to assessment task design

Table 4-30 is a summary table of the themes focusing on assessment task design. This table organizes each theme with a description and combines the strategies into a single column for clarity. Following the table, I provide explicit elaborations on each theme.

Table 4-30. Summary of themes, descriptions, and strategies for assessment task design

Theme	Theme Description	Strategies
Boosting Engagement	This theme involves connecting assessment tasks with students' real-life experiences to enhance understanding and retention. Tasks are designed to draw on familiar scenarios or intriguing contexts to increase motivation and engagement.	1. Integrate familiar contexts to make content relevant. 2. Connect concepts to real-world applications. 3. Incorporate interactive elements like hands-on activities or simulations.
Enhancing Clarity and Accessibility of Language	Focuses on simplifying complex scientific concepts through tailored vocabulary and sentence structures that are age-appropriate, ensuring the language used in assessment tasks is comprehensible for the target student audience.	1. Use age-appropriate vocabulary and structures. 2. Define technical terms clearly. 3. Test and refine for readability. 4. Align language with educational standards.
Enhancing Task Clarity and Guideline Precision	Centers on providing crystal-clear, straightforward instructions in assessment tasks to eliminate ambiguity, ensuring students understand exactly what is expected of them without confusion, aiding in effective demonstration of understanding.	1. Simplify instructional language. 2. Detail specific actions or steps. 3. Clarify task objectives. 4. Refine and test instructions regularly.

Table 4-30 (cont'd)

Incorporating Supportive Visuals and Scaffolds	Emphasizes the importance of integrating visual aids and scaffolding strategies into assessment tasks to make complex ideas more accessible and understandable, supporting textual information and promoting independent learning.	1. Use clear and relevant visuals like diagrams and graphs. 2. Provide structured step-by-step guidance. 3. Utilize interactive visuals for engagement. 4. Tailor scaffolds for varied needs.
Ensuring Cultural Sensitivity and Accessibility	Focuses on designing inclusive and reflective assessment tasks that resonate with students from diverse cultural backgrounds, promoting a more equitable learning environment and enhancing student engagement by incorporating culturally relevant content.	1. Include inclusive content selection. 2. Ensure language accessibility. 3. Represent diverse cultures in visuals. 4. Develop culturally relevant scenarios. 5. Implement feedback mechanisms for sensitivity.

Boosting engagement through relevant and contextual task design

Boosting engagement through relevant and contextual task design is essential for connecting assessment tasks with students' real-life experiences. This method transforms abstract scientific concepts into tangible and relatable challenges, enhancing understanding and retention. By designing tasks that draw on familiar scenarios or intriguing contexts, educators can significantly increase students' motivation to engage deeply with the content. Experts recommend several approaches to refine these tasks to ensure maximum engagement. First, integrating familiar contexts into the tasks helps make the content more relevant, as students can see direct links between their everyday lives and the scientific concepts being taught. Second, connecting these concepts to real-world applications clarifies their utility, boosting students' interest and the perceived value of their learning. Last, incorporating interactive elements into tasks, such as hands-on activities or simulations, makes the learning process more dynamic and engaging, fostering an active learning environment that is both educational and enjoyable (see Table 4-31).

Table 4-31. Summary of the strategies and example prompts for engagement

Strategy	Description	Exemplar Prompt
Incorporate Real-Life Scenarios	Use real-life contexts that students are likely to encounter to anchor the scientific concepts taught.	"Design an assessment where students analyze how playground equipment uses forces (PE 3-PS2-1) to function."
Connect Concepts to Daily Activities	Link scientific ideas to everyday activities to show their practical applications.	"Create a task asking students to describe how animals in their neighborhood adapt to seasonal changes (PE 3-LS4-3)."
Use Interactive Elements	Include components that require active engagement, such as simulations or hands-on experiments.	"Develop a simulation task that allows students to manipulate variables affecting the motion of an object on different surfaces."
Employ Storytelling	Craft scenarios as stories to draw students in and make the tasks more engaging.	"Write a story-based task where students help a character choose the best materials for building a kite, considering wind forces."
Highlight Relevance	Explicitly explain how the science topics students are learning about impact their lives.	"Ask students to investigate and present on how understanding of ecosystems can help improve local environmental practices."

Enhancing clarity and accessibility of language in the design of assessment tasks

It is pivotal for simplifying complex scientific concepts to ensure that the language used is suitable and comprehensible for the target student audience. This theme emphasizes tailoring vocabulary and sentence structures to be age-appropriate, minimizing the use of technical jargon unless it is necessary and clearly explained within the learning context. Simplifying the language in tasks to match the reading levels of elementary students while ensuring consistency across educational materials is crucial. This careful attention to language not only aids comprehension but also enhances the accessibility of scientific learning for all students.

To operationalize this principle in task design, GPT-4 can be instructed to prioritize simplicity in vocabulary and structure during content generation. The process involves creating prompts that explicitly require the avoidance of technical jargon or, if used, ensuring it is adequately defined in a context understandable to young students. For example, a prompt might state: "Generate a task description for PE 3-PS2-1 that explains how objects move, using simple language suitable for third graders without using technical terms such as 'net force.'" The outputs from GPT-4 should be rigorously tested for readability and clarity, with subsequent adjustments based on iterative feedback to ensure that they meet the developmental and cognitive needs of the target age group. Table 4-32 summarizes the strategies and example prompts to refine the design of tasks focusing on language clarity and accessibility

Table 4-32. Summary of the strategies and example prompts for language appropriateness

Strategy	Description	Exemplar Prompt
Use age-appropriate vocabulary	Focus on straightforward language that is easy for students to understand, avoiding complex phrasing.	"Generate a task for PE 3-PS2-1 using simple terms to explain the concept of forces acting on stationary objects."
Define Technical Terms Clearly	Provide clear definitions for any scientific terms used within the task to ensure they are age-appropriate for students.	"Create a task for PE 3-LS4-3 that involves animal adaptations, and include a sidebar that defines 'adaptation' in simple terms suitable for third graders."
Test and Refine for Readability	Continuously evaluate and refine the task descriptions to ensure they are understandable for the intended age group, based on feedback.	"Revise this task description for clarity: Simplify the sentence structure and ensure any scientific terms are clearly explained."
Align Language with Educational Standards	Make certain that the language used in the tasks aligns with the educational standards and learning objectives for the specified grade level.	"Review and adjust the language of this task for PE 3-PS2-1 to ensure it conforms to third-grade science standards and is understandable by students at this educational level."

Enhancing task clarity and guideline precision

It is important to focus on the necessity of providing assessment tasks with crystal-clear, straightforward instructions. This theme centers around crafting tasks in a way that eliminates any ambiguity, thus ensuring that students understand exactly what is expected of them without confusion. The precision of task instructions is critical in guiding students effectively through their responses, aiding them in focusing on demonstrating their understanding rather than deciphering the task requirements. By specifying exactly what steps to follow or what concepts to explore, students can easily access the assessment information, and enhance the function of assessment in students’ learning. Experts emphasized the importance of using clear and age-appropriate language in task instructions. They pointed out that complex terms and phrases can confuse students, especially at the elementary level. For instance, E noted that terms like "despite" and phrases like "remains unmoved" are too complex for third graders and suggested using simpler alternatives. Additionally, the experts highlighted the need for consistent terminology and explicit steps in task instructions to ensure clarity. P recommended that instructions should be direct and concise to match the reading levels of young students. Experts also suggested that tasks should include specific actions or steps that students need to follow. For example, detailing procedures and expected outcomes in a step-by-step manner. Regularly refining and testing task instructions based on student feedback was another key recommendation from the experts. They emphasized the need to adjust instructions to ensure they are clear and unambiguous. C and Sm pointed out that refining task instructions based on student responses can help identify and address any areas of confusion. See Table 4-33.

Table 4-33. Summary of the strategies and example prompts for clarity

Strategy	Description	Exemplar Prompt
Simplify Instructional Language	Use age-appropriate, clear and direct language in task instructions to avoid ambiguity and ensure students understand what is required.	"Describe how balanced and unbalanced forces affect an object's motion. Use simple language and diagrams to explain."
Detail Specific Actions	Provide explicit steps or actions students should take to complete the task, guiding them through the process.	"List the materials you will use to demonstrate balanced and unbalanced forces, describe the procedure step-by-step, and predict the outcome."
Clarify Task Objectives	Ensure that the goals and objectives of the task are explicitly stated so students understand the purpose and what they need to achieve.	"Explain the role of balanced and unbalanced forces in moving objects. State clearly what students need to demonstrate or explain."
Refine and Test Instructions	Regularly review and revise task instructions based on student feedback and performance to enhance clarity and precision.	"Based on student feedback, revise the instructions for the task to ensure they clearly convey the expected actions and outcomes."

Incorporating supportive visuals and scaffolds

It is crucial to emphasize the importance of integrating visual aids and scaffolding strategies into assessment tasks. This approach aims to support and enhance textual information, making complex ideas more accessible and understandable for students. Visual aids, such as diagrams, charts, and illustrations, provide concrete examples of abstract concepts, aiding in comprehension and retention. Scaffolding, which includes structured support like guided questions, step-by-step instructions, and checklists, helps students navigate through tasks that might otherwise be too challenging. These tools are crucial for building confidence and promoting independence as students progress in their learning and tackle more complex material.

Experts highlighted several key strategies to enhance the use of visuals and scaffolds in assessment tasks. They emphasized the need for clear, accurate, and contextually appropriate visual aids that directly support the textual information. NGSS experts stressed the necessity of visuals that accurately represent the described phenomena, noting that some visuals were either irrelevant or potentially confusing. For example, they criticized images showing squirrels on the ground instead of in trees, which did not align with the task's description. They also highlighted the importance of clear instructions and definitions of scientific terms to help students connect visuals with the accompanying text and data tables. Assessment experts pointed out inconsistencies between some images and the data presented, suggesting that visuals should be both clear and functional. C and Sm observed that one image showed more squirrels in an open area than indicated by the data table, potentially leading to student confusion. They advocated for visuals that are directly aligned with the data and include clear, directive captions to guide student interpretation. Additionally, they stressed the importance of providing step-by-step guidance and interactive elements in visuals to engage students actively with the material. Equity and language experts underscored the need for tailored scaffolds that cater to diverse learning needs, ensuring accessibility for all students. They recommended differentiated instruction sheets that include glossaries and cater to varying reading levels. Teacher experts highlighted the importance of visual accuracy and suggested adjustments to images to better reflect the educational content. see Table 4-34.

Table 4-34. Summary of the strategies and example prompts for scaffolding

Strategy	Description	Exemplar Prompt
Use Clear and Relevant Visuals	Incorporate diagrams, graphs, or images that directly relate to and help clarify the task's concepts.	"Show a diagram of balanced and unbalanced forces acting on an object to illustrate how they affect motion."
Structured Step-by-Step Guidance	Provide a breakdown of tasks into smaller, manageable steps that guide students through the learning process.	"Follow these steps to construct your model of a plant cell, starting with the cell wall and moving inward."
Interactive Visuals for Engagement	Utilize interactive elements in visuals that allow students to engage actively with the material.	"Use this interactive map to explore different ecosystems and their characteristics, and note how organisms adapt to their environments."
Tailored Scaffolds for Varied Needs	Adjust scaffolding techniques to meet the diverse learning needs of students, ensuring accessibility for all.	"Make sure each prompt has specific instructions using "building a claim", "identifying evidence" and "making reasoning."
Feedback Loops for Improvement	Integrate opportunities for feedback within tasks, allowing for adjustments and fostering deeper understanding.	"Submit a draft of your project for preliminary feedback on your use of scientific terms and concepts, and revise based on the feedback."

Ensuring cultural sensitivity and accessibility in task scenarios

It is important to focus on designing assessment tasks that are inclusive and reflective of the diverse cultural backgrounds and experiences of all students. This approach emphasizes the importance of using scenarios and contexts in assessments that resonate with students from different cultural perspectives, thereby fostering a more equitable learning environment. By incorporating culturally relevant content, educators can increase student engagement and promote a deeper connection with the material. Tasks that consider the varied experiences of students can help prevent cultural bias and ensure that all learners feel represented and valued in the learning process.

Experts suggested several strategies for improving cultural sensitivity and accessibility. They emphasized the need to incorporate familiar contexts and real-life applications in assessment tasks to make them more relatable and engaging. For example, experts noted that using culturally diverse and familiar scenarios, such as describing local natural resources, can significantly enhance student engagement. This is supported by experts who highlighted the importance of using scenarios that students can easily relate to, enhancing the relevance and engagement of the tasks. Experts also recommended using language that is age-appropriated, clear, and free from cultural bias, ensuring it is accessible to all students. This includes avoiding technical jargon and culturally specific terms that may not be universally understood. For instance, NGSS experts pointed out the need for tasks to use straightforward language and avoid complex terms that could confuse younger students. Furthermore, the importance of visual aids that accurately represent diverse cultures was highlighted. Experts suggested using images and examples in materials that depict a variety of cultural backgrounds to avoid cultural bias. As one expert noted, including visual representations of diverse characters and settings helps make the tasks more relatable and inclusive for all students. Additionally, scenarios should be relevant to students' daily lives and reflect their diverse experiences to make the tasks more engaging and meaningful. This involves creating tasks that reflect real-life situations familiar to students from different cultural backgrounds, making the learning experience more personal and engaging. Cultural sensitivity in task design not only enhances fairness but also enriches the educational experience by exposing students to different viewpoints and ways of understanding the world. It involves careful consideration of language, scenarios, and examples used in assessments to avoid stereotypes and biases. Moreover, ensuring that tasks are accessible to students with different abilities and learning needs is crucial for creating an inclusive classroom environment. This can be achieved through the implementation of feedback mechanisms that allow students to suggest improvements on cultural representation in tasks, ensuring that the educational materials remain relevant and sensitive to all students. See Table 4-35.

Table 4-35. Summary of the strategies and example prompts for cultural sensitivity

Strategy	Description	Exemplar Prompt
Inclusive Content Selection	Choose content that reflects a broad spectrum of cultures and experiences.	Describe how different communities use local natural resources to balance forces in engineering solutions.
Language Accessibility	Use clear language, ensuring it is accessible to all students.	Explain how balanced and unbalanced forces affect motion using simple terms.

Table 4-35 (cont'd)

Representation in Visuals	Include images and examples in materials that depict a variety of cultural backgrounds.	"Use illustrations showing children from different cultures engaging in activities that involve balanced and unbalanced forces."
Culturally Relevant Scenarios	Develop scenarios that relate to real-life situations experienced by students from diverse backgrounds.	Create a task where students investigate how animals adapt to their environments in different cultures, focusing on observable traits."
Feedback Mechanisms for Sensitivity	Implement systems for students to provide feedback on cultural relevance and sensitivity.	"Provide a feedback form for students to suggest improvements on cultural representation in tasks, such as balancing forces and adaptations."

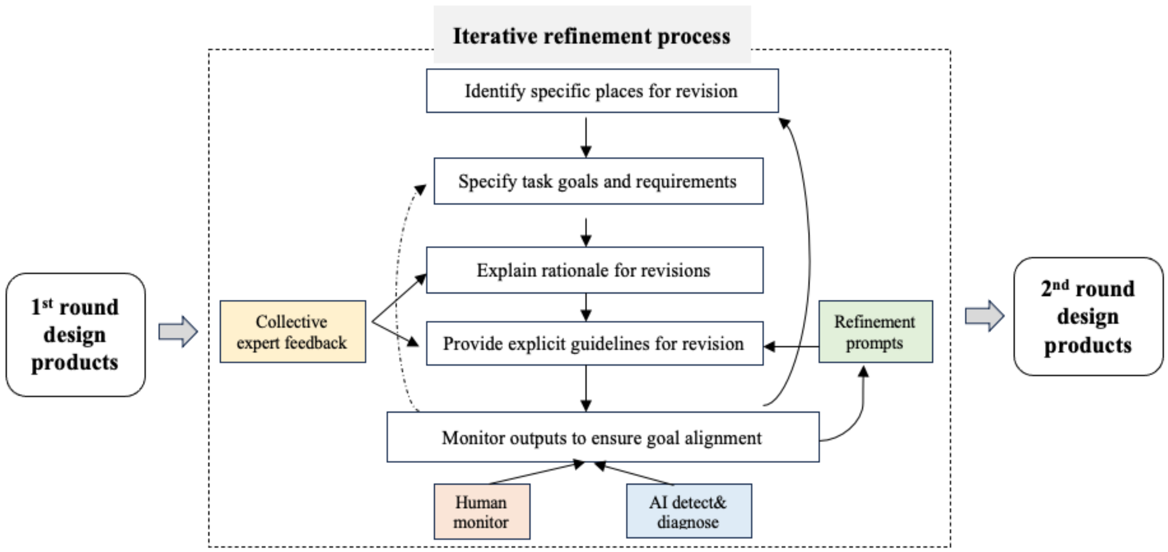
4.3 RQ3. What Is the Process of Refining AI-Designed Knowledge-In-Use Assessments Based on the Feedback Provided by Human Experts? Whether and How Are the Assessments Revised?

4.3.1 Refinement Model

After analyzing the expert feedback and synthesizing their critical suggestions, I formulated refinement principles and corresponding prompts to guide the revisions of LPs, evidence statements, and assessment tasks. These principles are detailed in Section 4.2.5, which presents the approach for both designing LPs and evidence statements and for the broader context of knowledge-in-use assessment design. A notable change in this phase was the shift from using a standard API to a customized GPT environment for the refinement process. The primary reason for this change was the rapid technological advancements and the limitations of the API in accessing external PDF documents. These limitations often restricted the availability of critical information. To ensure the AI had access to all necessary materials, including the NGSS framework and details of the assessment design process, I opted to utilize a customized GPT setup. This setup was enhanced by uploading the initial training script and essential reference materials to ensure the AI retained knowledge from the first round of training and was well-equipped to process the refinement tasks effectively.

Using the principles, I initiated an iterative refinement process with GPT-4. I collaborated with GPT-4 to refine the design products meticulously, one by one. Each refinement or revision adhered to the structured process depicted in Figure 4-28, ensuring that every modification aligned with the established goals and responded effectively to the expert feedback.

Figure 4-28. Design product refinement process.



In the refinement process, the primary goal is to integrate the expert panel's collective feedback to enhance and refine the knowledge-in-use assessment tasks. This refinement begins by identifying specific areas that require updates, which may include revisions to unpacking documents, LPs, evidence statements, or tasks. The next step involves defining specific task goals and requirements, which are derived from the collective insights gathered during the analysis of RQ2. Following this, the human operator provides a detailed rationale for each suggested revision, explaining why specific changes are necessary to improve the assessment tasks. This explanation is supported by the detailed guidelines for revision that were developed based on the collective feedback from experts and the refinement prompts generated from the RQ2 analysis.

The AI models then execute these revisions. Throughout this process, human operators continuously monitor the outputs to ensure they align with the established goals. The AI models are also prompted to reflect and provide explanations on how their generated outputs meet the task goals and requirements. This dual monitoring process allows for further refinement of the prompts in an iterative manner. Once the revisions sufficiently meet the task goals and adhere to the guidelines, the refinement process concludes with the production of the second-round design products. This iterative approach ensures that the final assessment tasks are well-aligned with expert feedback and educational standards.

The refined products were distributed to two distinct expert groups for evaluation and feedback. In addition to the original panel familiar with the AI's role in designing the tasks, I introduced a new group of experts who were not informed about the AI's involvement. This strategy was employed to mitigate any potential biases related to perceptions of AI-generated content. Below, I take two examples to illustrate the refinement process from LPs and evidence statement design stage and task design stage. I also discuss the enhancements observed in the refined products compared to those from the initial design round, demonstrating the effectiveness of this iterative approach in improving the quality and relevance of the assessment tasks.

4.3.1.1 Ensuring the Scope of the LPs and Evidence Statements by Revisiting Unpacking

Identify specific places for revisions

In revising the LPs for PE 3-PS2-1, the expert panel's feedback highlighted several key areas requiring attention. A primary concern was that the existing LPs did not adequately address the concept of 'unbalanced forces,' a central idea within the PE. For instance, while LP4 reasonably introduces 'non-contact forces'—acknowledging that students often struggle with this concept—it does not align with the PE's main focus on 'balanced' and 'unbalanced' forces. Moreover, there is a notable gap in addressing the effect of balanced forces on an object in motion, which should result in no change in motion. This is crucial since a common student misconception is that 'no force means no motion,' rather than the correct 'no force means no change in motion.' To illustrate this concept effectively, one proposed example involved a train car moving at a constant speed, which, despite experiencing friction, does not accelerate or decelerate—a practical demonstration of balanced forces at play. Additionally, an NGSS expert criticized the unpacking of the DCI for not thoroughly covering the essential ideas of the disciplinary core elements, pointing to a need for a comprehensive review and redesign of the LPs to ensure they fully encapsulate the key concepts as outlined in the PE. This feedback underscores the importance of revisiting and refining the unpacking of the three dimensions to ensure the LPs accurately reflect the standards and effectively address common misconceptions.

Specify task goals and requirements.

Based on the feedback, I set up the task goals and requirements, which are to revisit the DCI and SEP unpacking for PE 3-PS2-1 to ensure that the DCI unpacking covers all of the important DCIs that are emphasized in this PE; the unpacking includes the unbalanced forces and the effect of unbalanced forces; ensure the effect of balanced forces that cause the situation of objects in motion but without changes in motion; unpack the non-contact forces and the effect of the non-contact forces; specify the type of non-contact forces should be discussed in the PE, avoiding the magnitude forces; specify the grade boundary of the unpacking; ensuring the terms are used consistently and coherently through the unpacking. It is important to note that these task requirements and goals were introduced to the GPT model in a gradual manner. This phased approach was strategic, allowing for focused adjustments and ensuring that each aspect of the feedback was meticulously addressed to refine the educational materials effectively.

Explain the rationale for revisions.

I also explained the rationale of why these revisions need to be made by reiterating the importance and purpose of unpacking, which is to zoom in the larger grain size PE from looking into the three dimensions of scientific knowledge and skills to understand the critical sub-ideas and sub-skills of the DCIs, SEPs, and CCCs to break down the PE into smaller pieces and reorganize them into manageable and assessable grain size for effective science learning and teaching. Then it requires the unpacking to be thorough, comprehensive, and have clear performance expectations for each critical sub-idea and skill that align with the grade level requirements. The reason why the task goals and requirements were set up as above is based on the experts' feedback.

Providing explicit guidelines for revisions

The refinement process, as outlined in section 4.2.5, involves creating specific prompts based on the principles of refinement and the feedback gathered from expert reviews. For instance, one prompt I developed for refining the unpacking of the DCIs related to PE 3-PS2-1 was particularly focused on ensuring clarity and educational appropriateness. The prompt read: "*Please revisit the DCI unpacking of PE 3-PS2-1, ensuring that the unpacking aligns with the 3rd-grade boundary. For the unpacking of the DCIs, it appears not to be explicit enough. For example, under the major DCI of Force and Motion, you should list sub-major ideas including the following: a. force—defining it, discussing its characteristics like strength and direction, explaining balanced and unbalanced forces; b. motion—defining motion, emphasizing aspects like speed and direction, discussing factors affecting motion; and c. the impact of different forces on an object's motion. You don't need to use the terms I specified here but ensure that terms are consistent throughout your elaboration and meet the grade boundary.*" This directive aims to ensure that the unpacking comprehensively covers the range of sub-ideas within the DCI and is appropriate for the grade level. Additionally, I instructed the GPT models to simplify the language and ensure the terms are used consistently and coherently throughout the unpacking process. Table 4-36 presents the comparisons between the first and second rounds of DCI unpacking.

Table 4-36. Two rounds DCI unpacking for PS2.A: Force and Motion

1st round DCI unpacking	2nd round DCI unpacking
'Given the provided context and requirements for unpacking the first DCI (PS2.A: Forces and Motion) for the focal performance expectation 3-PS2-1, here's a detailed unpacking tailored for grades 3-5: Aspect 1: Elaborating the DCI Elements and Sub-ideas DCI PS2.A: Forces and Motion focuses on understanding how forces affect the motion of objects. This includes:** **Force Characteristics:** Forces are push or pull actions that can cause an object to start moving, stop moving, or change its motion's direction. Each force acts on a particular object with a specific strength (magnitude) and direction. **Interaction of Forces:** Multiple forces can act on an object simultaneously. When the forces are balanced (sum to zero), the object's motion doesn’t change; it remains at rest or continues moving at a	**DCI: PS2.A Forces and Motion** 1. **Force:** ○ Definition: A push or pull on an object. 2. **Characteristics of Force:** ○ Strength: How hard the force pushes or pulls. ○ Direction: The way the force pushes or pulls. 3. **Balanced Forces:** ○ Definition: Forces that are equal in size and opposite in direction, canceling each other out, resulting in no change in motion. 4. **Unbalanced Forces:** ○ Definition: Forces that are not equal and opposite, causing a change in the object's motion (speeding up, slowing down, changing direction). 5. **Motion:**

Table 4-36 (cont'd)

constant speed in a straight line. Unbalanced forces (those that do not sum to zero) cause changes in the object's motion, either accelerating it or altering its direction. **Newton's Third Law:** The mutual interactions between two objects can be described in terms of equal and opposite forces. When one object exerts a force on a second object, the second object simultaneously exerts a force equal in magnitude and opposite in direction on the first object. Aspect 2: Defining Boundaries At this grade level, the focus is on: Qualitative and conceptual explanations of how balanced and unbalanced forces affect motion, without delving into quantitative analysis or calculations of force magnitudes. Observations and descriptions of motion changes due to applied forces, without introducing technical terms like "magnitude," "velocity," or "vector." Aspect 3: Describing Prerequisite Knowledge **Students entering grade 3 should already have foundational knowledge, including:** Basic understanding of motion — knowing that objects move in various ways, such as straight, zigzag, round and round, back and forth, fast and slow. Familiarity with pushes and pulls as forces that can move objects. Aspect 4: Identifying Relevant Phenomena **Examples of relevant, real-world phenomena related to forces and motion include:** Playing on a seesaw demonstrates how applying force (pushing down) on one end causes the other end to move up, illustrating action and reaction forces. Observing how kicking a soccer ball (applying a force) changes its motion depending on the strength and direction of the kick. Aspect 5: Identifying Student Strengths and Challenges **Strengths:** Many students have practical experience with pushing and pulling objects, which can be leveraged to introduce the concepts of force and motion. **Challenges:** Students may believe that constant motion requires a constant force application, conflating force with	○ Definition: The change in position of an object over time. ○ Aspects: Speed (how fast an object moves) and direction (the way it moves). ○ Effects: Forces can change the speed and direction of an object's motion. **Defining Boundaries** 1. **Grade Level Scope:** ○ Focus on qualitative and conceptual understanding rather than quantitative calculations. ○ Students should understand the relative strengths and directions of forces. ○ Gravity is addressed qualitatively as a force that pulls objects down. ○ Technical terms such as magnitude, velocity, momentum, and vector quantity are not introduced at this level. **Describing Prerequisite Knowledge** 1. **Basic Understanding of Motion:** ○ Concepts of push and pull, start and stop. ○ Awareness that objects can move in different ways (e.g., straight, round, back and forth). 2. **Surface Interaction:** ○ Different surfaces affect motion (e.g., smooth vs. rough). **Identifying Relevant Phenomena** 1. **Examples of Phenomena:** ○ A ball starts to roll when pushed (unbalanced force). ○ A box remaining stationary when equal forces are applied from opposite sides (balanced forces). ○ Friction as a force that stops a moving object. **Identifying Strengths and Challenges**

Table 4-36 (cont'd)

motion. Some students might think that a larger object always moves slower than a smaller one, regardless of the forces applied, due to misconceptions about mass and force. By exploring and addressing these various aspects, educators can tailor their teaching strategies to effectively convey the foundational principles of forces and motion to elementary students, setting a robust groundwork for more advanced concepts in later grades	1. **Strengths:** ○ Direct experiences with pushing and pulling objects. 2. **Challenges:** ○ Misconceptions such as stationary objects not exerting force and the invisibility of forces like magnetism.
Major changes: ● The second round of unpacking significantly improved the grain size of the unpacked DCIs compared to the first round. In the first round of unpacking, it had a larger grain size, including Force Characteristics, Interaction of Forces, and Newton's Third Law. Although it included the major science ideas, it was still a larger grain size. The second round of unpacking has a smaller grain size. in addition, it emphasized the idea s of balanced and unbalanced forces. In terms of the grade boundary, the second-round unpacking further clarified the boundary of important ideas, including: 1. understand the relative strengths and directions of forces and 2. Gravity is addressed qualitatively as a force that pulls objects down.	

Table 4-36 shows that the second round of unpacking aligns with the requirements and goals I previously established. Using this updated unpacking, I continued refining the LPs to incorporate the experts' feedback regarding the initial set's coverage. I initiated the refinement by providing the GPT model with explicit instructions reflecting on the first round of LPs. I directed the GPT model with specific tasks:

> "Please refine the LPs to address the following parts: 1. Determine whether the forces acting on an object are balanced or unbalanced. 2. Determine that the effect of balanced forces on an object at rest is that it remains at rest. 3. Determine that the effect of balanced forces on an object in motion is that its motion does not change. 4. Determine the effect of unbalanced forces on an object at rest that its motion changes (it begins to move). 5. Determine that the effect of unbalanced forces on an object in motion is that its motion changes (it speeds up or slows down and/or it turns) Note: I have used the word 'Determine' above but this would be replaced with an appropriate practice such as: A. Carry out an investigation to show. B. Develop a model that shows… C. Construct an explanation to show… could you please design LPs based on the suggestions and our previous conversations."

After GPT produced its output, I further engaged it to identify and address any issues or to refine further. Example prompts include: *"Please ensure motion is appropriately unpacked in the LP, for instance, it could be the speed or the direction of the object change,"* and "*Please explain how the new set* *of LPs addresses the reviewers' comments and how it can correctly measure the PE? Can you reexamine this set of LPs to check if there is anything missed or overreached?"* GPT's response, labeled as 'Potential Refinements', was insightful: "LP-E01: No refinements needed; it is clear, well-scoped, and appropriately detailed. LP-E02 and LP-E03: Ensure that the models are simple and easily understood by elementary students. Provide guidance on the types of models that are suitable for the grade level." Following this analysis, I prompted GPT to continue the refinement process, focusing on simplicity and educational appropriateness: *"how to make sure: LP-E02 and LP-E03: Ensure that the models are simple and easily understood by elementary students. Provide guidance on the types of models that are suitable for the grade level. Please provide revised LPs."* Table 4-37 compares the refined set of LPs with those from the first round, illustrating the adjustments made.

Table 4-37. Two rounds of LPs for 3-PS2-1.

1st round LPs for 3-PS2-1	2nd round LPs for 3-PS2-1
LP1**: Students plan and carry out investigations to observe how different strengths of forces affect the motion of an object. LP2**: Students develop models to explain how balanced forces acting on an object result in no change in motion, using everyday scenarios such as a book resting on a table or a tug-of-war game with equal strength on both sides. LP3**: Students construct explanations based on evidence from investigations to explain how objects in contact exert forces on each other, including friction, showing the interaction between objects as a cause of motion changes. LP4**: Students use models to explain how non-contact forces (e.g., magnetic or gravitational forces) on an object's motion or change in motion act at a distance.	LP1: Students plan and carry out investigations to collect data on how balanced and unbalanced forces affect the motion of an object (speeding up, slowing down, changing direction). LP2: Students develop simple models to explain how balanced forces acting on an object at rest result in no change in motion. LP3: Students develop simple models to explain how balanced forces acting on an object in motion result in no change in motion. LP4: Students carry out investigations to observe how unbalanced forces cause objects at rest to start moving. LP5: Students construct explanations to show how unbalanced forces affect a moving object's motion (it speeds up or slows down and /or it turns).

Table 4-37 (cont'd)

Major changes: 1. Improve the coverage of the major ideas in the PE. 2. Further clarification of motion by including speeding up, slowing down, and changing direction. 3. Clarification that the model students are to develop is "simple" models. 4. The new LP2 emphasizes the relationship between balanced forces and an object at rest. 5. The newly added LP3 addresses the part of the PE about the relationship between balanced forces and an object in motion that the first round LPs do not include. 6. The new LP4 further emphasizes the original SEP in the PE and addresses the relationship between unbalanced forces causing objects at rest to start moving, which is not included in the first round LPs. The second round set of LPs removes the idea of contact forces and non-contact forces, which are not the critical ideas of the PE.

Throughout the process, I collaborated closely with GPT to reflect on and summarize the refinements made to the LPs and evidence statements, aiming to hone the underlying principles guiding their design. I engaged GPT in discussions to assess the evolution of the LPs, asking, "How do you see this new set of LPs different from the set of LPs?" These instructions were crucial for ensuring that the adjustments effectively addressed the feedback from the expert panel. To extract actionable insights from these revisions, I frequently prompted GPT to reflect on the principles that emerged from integrating the reviewers' feedback, asking, "*great! please recall the process of revising the LPs for 3-PS2-1, what principles would you extract about incorporating the reviewers' feedback from the process?*" This iterative dialogue helped identify core principles that could be applied to future revisions. In this role, I served as a questioner, pinpointing essential areas for reflection and guiding the GPT models to detect nuances and learn from the ongoing refinement process. Together, we worked to further enhance the design process, ensuring that each iteration of LPs and evidence statements was progressively aligned with educational objectives and expert insights.

4.3.1.2 Refining the Assessments by Enhancing the Realism and Relevance and Revising Language

Identify specific places for revisions

The second example of refinement is about task refinement for PE: 3-LS4-3. A major critique of the tasks designed for the LP4 for the 3-LS4-3 is the need for language simplification and enhanced realism in task design. Experts across various fields stressed the importance of using clear and straightforward language to make educational tasks accessible for third graders. They suggested replacing formal and complex phrases such as "to aid their investigation" with more direct and age-appropriate terms like "to help their investigation." The realism of visual aids used in the tasks was another significant concern. A notable example involved the use of an overly fantastical "magic" sunflower element in a task, which, while engaging, was critiqued for its lack of scientific accuracy. Assessment experts and science content experts further highlighted the potential for confusion caused by visuals that did not accurately depict scale or realistic biological processes. These insights led to targeted revisions that focused on simplifying language and aligning visual representations with scientific realities, aiming to enhance both the clarity and educational value of the tasks for young learners.

Specify task goals and requirements

To ensure the refinement of educational tasks meets specific requirements, a series of guidelines has been established. First, the tasks should use scientifically accurate visuals that align with the narrative to avoid confusion. Captions clarifying the content must be integrated effectively to enhance understanding. Furthermore, the scenarios used in the tasks should be familiar and relatable to the students. This could involve adapting the animal subjects to reflect the local wildlife known to students in their geographical areas and focusing on settings like various parks that offer diverse learning environments. Language and instructional clarity are critical. The tasks should use age-appropriate vocabulary and sentence structures that match third-grade comprehension levels without sacrificing scientific precision. Instructions should be straightforward and concise, eliminating complex terms in favor of simpler alternatives that maintain the accuracy of the scientific concepts being taught. Lastly, consistency in terminology is crucial. Terms should be used uniformly across all educational materials to prevent confusion.

Explain rationale for revisions

The necessity to refine assessment tasks stems from the overarching requirement to use visual aids that are both scientifically accurate and effectively tailored to enhance the learning process, ensuring that every aspect of the task is clear and relevant for all students. It is essential to employ clear and straightforward language, making the educational content accessible and comprehensible for third-grade students. During the review process, experts highlighted issues with the use of overly fantastical visuals, such as the sunflower depiction, which they felt was more appropriate for a fictional setting than a science-based task. This feedback highlighted a critical balance that needs to be maintained in educational materials: engaging students effectively while upholding the integrity of the scientific content presented. The revised tasks aim to captivate third graders with realistic and scientifically accurate scenarios, thus enhancing their overall learning experience.

Provide explicit guidelines for revisions

Start by recalling the design blueprint for LP2 of PE 3-LS4-3 and the overall assessment design process. I provided explicit guidelines focusing on several key areas: 1. Use of real images and visual aids: It's crucial to include high-quality, relevant images that clearly depict the scenario being studied. Visual aids should be seamlessly integrated into the tasks to enhance understanding of the phenomena; and 2. Provide adequate scaffoldings for younger students: To support younger learners, include step-by-step prompts or visual aids that guide students in representing forces and objects in their models effectively.

Addressing concerns about realism, I encouraged the GPT model to consider the authenticity of the presented phenomena, asking, "*Are these real-world phenomena, and is the data real? Are there real pictures available to illustrate the phenomena? The use of authentic images rather than generated graphics would greatly enhance the task's educational value.*" To ensure the tasks are age-appropriate and linguistically accessible, I guided the GPT model with specific restructuring directions: "*Revise the assessment task to start with a real and compelling phenomenon, accompanied by images and a data table. Follow this with a scenario where two students debate which organism will survive in a particular environment. The item prompts should then ask students to choose which argument they agree with, provide evidence, and reason their choice, ensuring all questions are tailored to be third-grade friendly.*"

Following this structured approach and refining the tasks based on feedback, the iterations led to significant improvements in clarity and relevance, as documented in Table 4-38, which compares task designs from the first and second rounds.

Table 4-38. Task 2 for LP4 for 3-LS4-3 across two rounds of refinement

1st round Task 2 for LP4 for 3-LS4-3

***Task 2*: "The Mystery of the Growing Sunflowers"(3-LS4-3-LP2-2)**

***Item Stem*:**

In the school's garden, there are various sections where different plants flourish. One section is full of sunflowers that grow tall and healthy, while another section attempts to grow the same sunflowers, but they struggle and barely bloom. Mrs. Smith, the science teacher, presents this observation to her 3rd-grade class.

(Generated by DALL.E on March 11, 2024)

Mrs. Smith asks her students to explain why sunflowers in one section of the garden thrive while those in another section do not. She provides them with data (Table a) on sunlight exposure, soil moisture, and soil type for each garden section.

Table a. Garden information

	Phenomena	Sunlight	Soil Moisture	Soil Type
Sunny Spot	Sunflowers are tall and healthy.	Lots of sunshine all day.	Just right, not too wet or dry	Rich and full of nutrients
Shady Spot	Sunflowers are small and struggling	Very little, mostly in the shade	Quite damp, doesn't dry out much	A bit clayey and hard for roots to grow

Item Prompts:

1. Think Like a Scientist: Why do you think sunflowers in the sunny spot are doing so well compared to the ones in the shady spot? Remember, plants need certain things to grow big and strong.

2. Use the Clues: Look at the information about both spots in the garden.

- If you think sunlight is important, tell us why the sunny spot might be better for the sunflowers.
- If you're curious about the soil, share how different soil types might affect the sunflowers' growth.

3. Solve the Mystery: Using what you've learned, explain why you think the sunflowers are happier in one part of the garden than the other. Is it just about the sun, or is the soil also a big clue?

2nd round Task 2 for LP4 for 3-LS4-3

***Task 2*: " Plants in Desert and Wetland Habitats "(3-LS4-3-LP4-2)**

***Item Stem*:**

Sam and Mia are on a field trip to study different plant habitats. They see that different plants grow well in different environments. Sam thinks plants with thick, waxy skin and deep roots survive better in deserts. Mia believes plants with broad leaves and floating roots do better in wetlands. They decided to collect some information to understand this better.

Table A. Plant Characteristics and Root characteristics.

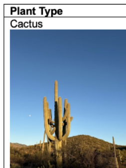
Plant Type	Root Characteristics	Description
Cactus  (Saguaro Cactus in the Arizona desert from the National Park Service)	Thick, waxy skin, deep roots	Cacti have thick, waxy skin to save water and deep roots to reach underground water.
Water Lily  (https://www.nps.gov/media/photo/view.htm?id=8F652E99-051B-4726-89A3-54EF140CBFBA)	Broad leaves, floating roots	Water Lilies have broad leaves that float on water and roots that anchor them in mud.
Grass  (https://grassbarrier.com/pages/grass-roots)	Shallow roots, flexible stems	Grass has shallow roots and flexible stems, helping it survive in both dry and wet places by quickly taking in water and nutrients from the soil surface.

Table B. Habitat Characteristics, Plant Needs, and Survival Rates

Habitat	Characteristics	Plant Needs	Survival Rates		
			Cactus	Water Lily	**Grass**
Desert	Hot, dry, sandy soil	Water conservation, deep roots	High	Low	Medium
Wetland	Waterlogged soil, high humidity	Floating ability, broad leaves	Low	High	Medium

Item Prompts:

1. Claim: Do you agree with Sam that "Plants with thick, waxy skin and deep roots survive better in deserts."?
2. Evidence: Use the observations from Table B to support your choice. Think about the root characteristics and how they help the plants in desert and wetland habitats.
3. Reasoning: Explain how the characteristics of the plants help them survive well in some habitats but less well or not at all in others. Connect this to what the plants need to live.

Major changes:

1. Underlining the authenticity of the phenomena and data included in the task. In the newly designed tasks, all data, phenomena, and images are sourced from actual research.
2. Providing more information to offer accessible opportunities for diverse students to engage in the tasks.
3. Revising the language of the assessment task prompts and stems to be age-appropriate and avoid unscientific terminology.
4. Enhancing prompt scaffoldings to better support students in engaging with and understanding the task.

Throughout the iterative refinement process, the human operator plays a crucial role in identifying, monitoring, and guiding the outputs. I prompt AI models to reflect, detect, and diagnose the outputs at certain critical points. This ensures that the AI learns from the iterative process and improves its capability for future design tasks.

4.3.2 Expert Panels' Feedback on the Refined Assessments

4.3.2.1 Unblinded Expert Panels

Table 4-39 presents the number of experts who provided the second-round feedback. In the following sections, I discuss the comparison between the first round and second round feedback on the LPs and designed tasks by PE.

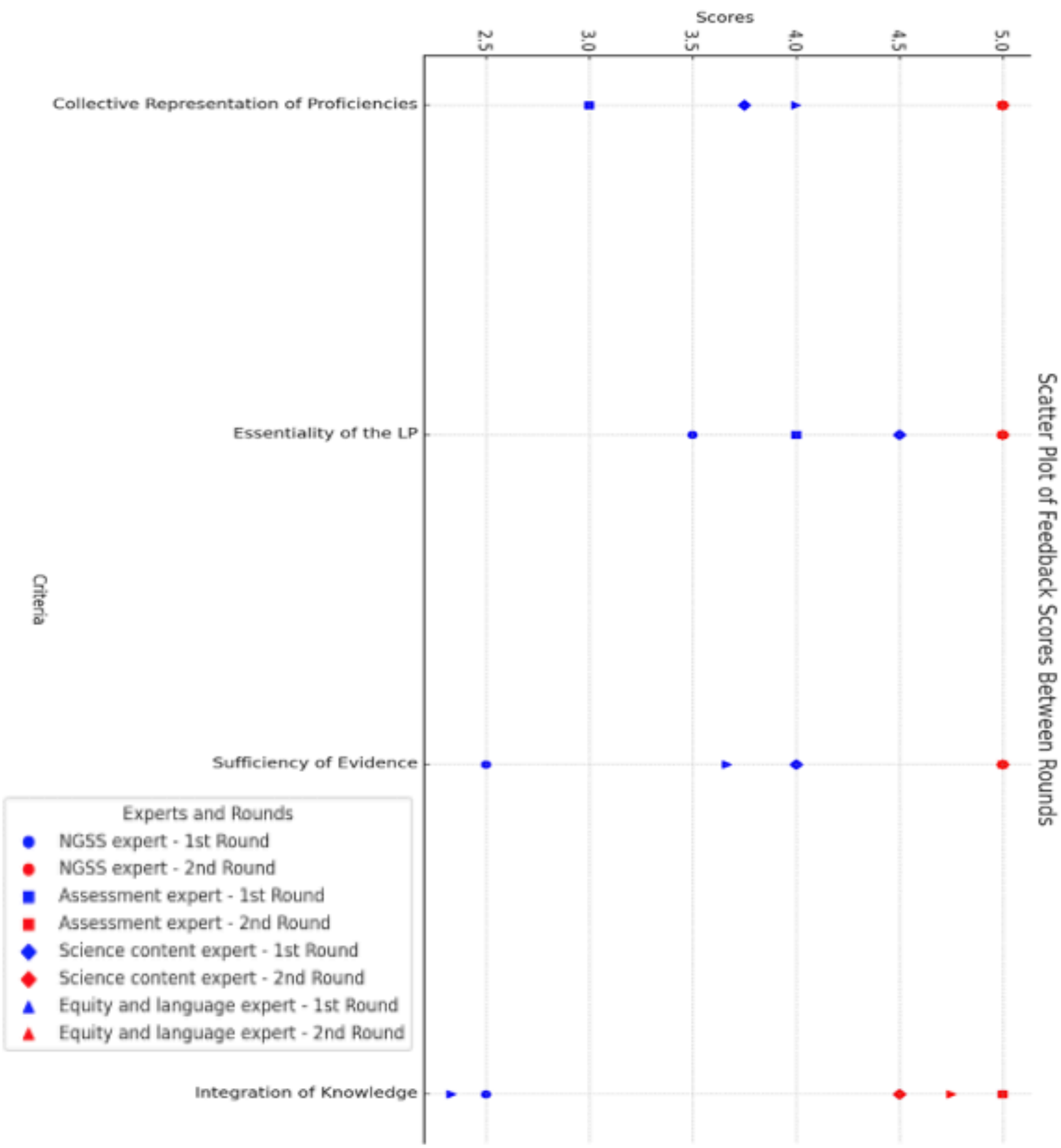
Table 4-39. Summary of experts on second round feedback

Expert group	NGSS Expert	Assessment Design Experts	Science Content Experts	Equity and language experts	Engagement experts	Elementary science teacher experts
3-PS2-1	1	2	4	2	3	2
3-LS4-3	1	2	5	3	3	2

3-PS2-1: Feedback on LPs and Evidence Statements

Figure 4-29 presents the feedback between the first and second round expert feedback on the designed LPs and evidence statements.

Figure 4-29. Scatter plot for the two rounds expert feedback on the LPs and evidence statements



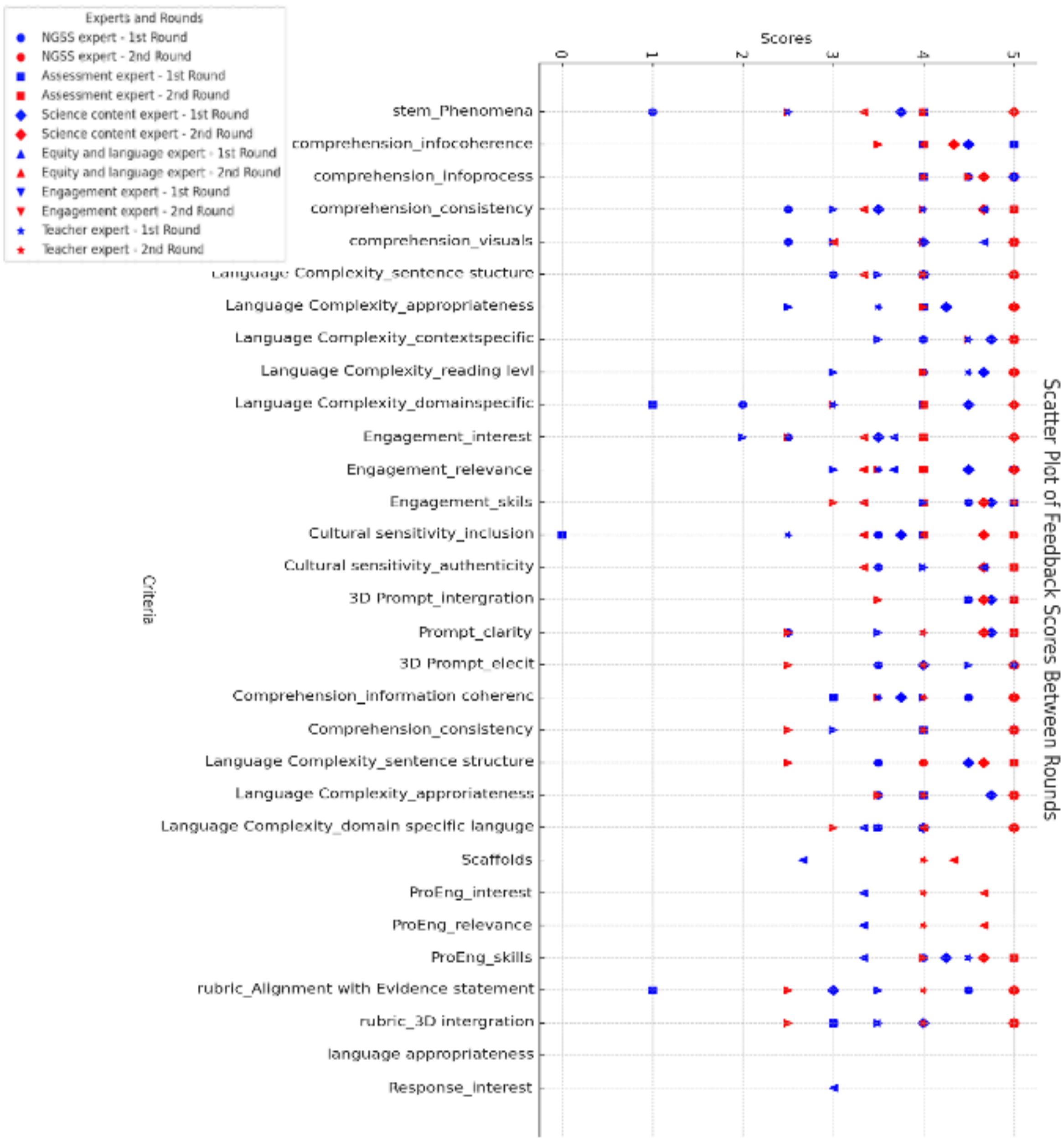
The scatter plot in Figure 4-29 displays feedback scores from a panel of experts on four dimensions on the LPs and evidence statements evaluated across two review rounds. The dimensions include Collective Representation of Proficiencies, Essentiality of the LP, Sufficiency of Evidence, and Integration of Knowledge. Four groups of experts—NGSS experts, assessment experts, science content experts, and equity and language experts—provided scores, visualized with distinct markers for each group. In the first round, shown in blue, feedback scores varied moderately across all criteria, with generally lower scores reflecting initial assessments of the LPs and evidence statements. By the second round, depicted in red, there was a noticeable improvement in scores for all criteria among all expert groups. This indicates a positive reception to the modifications made after the initial feedback. Specifically, the NGSS experts showed marked improvements in all areas, reaching the maximum score in the second round, indicating complete satisfaction with the changes made. Similarly, the assessment experts and science content experts gave higher scores in the second round, particularly noting improvements in the Sufficiency of Evidence and Integration of Knowledge. The equity and language experts, while also showing increased scores, provided slightly more conservative feedback in areas like Integration of Knowledge, suggesting areas where further refinements could be beneficial.

3-PS2-1: Feedback on Task 1

Figure 4-30 visualizes comprehensive feedback scores across a broad range of criteria from a diverse group of experts. These scores are compared between two rounds of review, with the first round represented by blue markers and the second by red markers. In general, the second round shows higher scores across most dimensions, indicating that the revisions made after the initial feedback were well received. For specific expert groups, NGSS experts demonstrated significant improvements in areas like Phenomena, Information Coherence, and Language Complexity-Sentence Structure, suggesting that revisions better aligned with NGSS standards in the second round. Assessment experts noted improvements in Engagement Relevance, Prompt Clarity, and Language Complexity-Domain Specific, reflecting a refinement in how assessments are designed to gauge knowledge accurately. Science content experts saw improvements in Criteria Sensitivity and Authenticity, indicating enhanced content accuracy and real-world relevance. Equity and language experts observed slight increases in Cultural Sensitivity/Inclusion and Language Appropriateness, making the content more inclusive and accessible. Engagement experts marked improvements in Engagement Relevance and Interest, highlighting better engagement strategies in assessment tasks. Teacher experts reported substantial improvements in

Response Appropriateness and Evidence Statement Alignment, suggesting that tasks became more effective for classroom use.

Figure 4-30. Scatter plot for the two rounds expert feedback on Task 1 for LP2 of 3-PS2-1

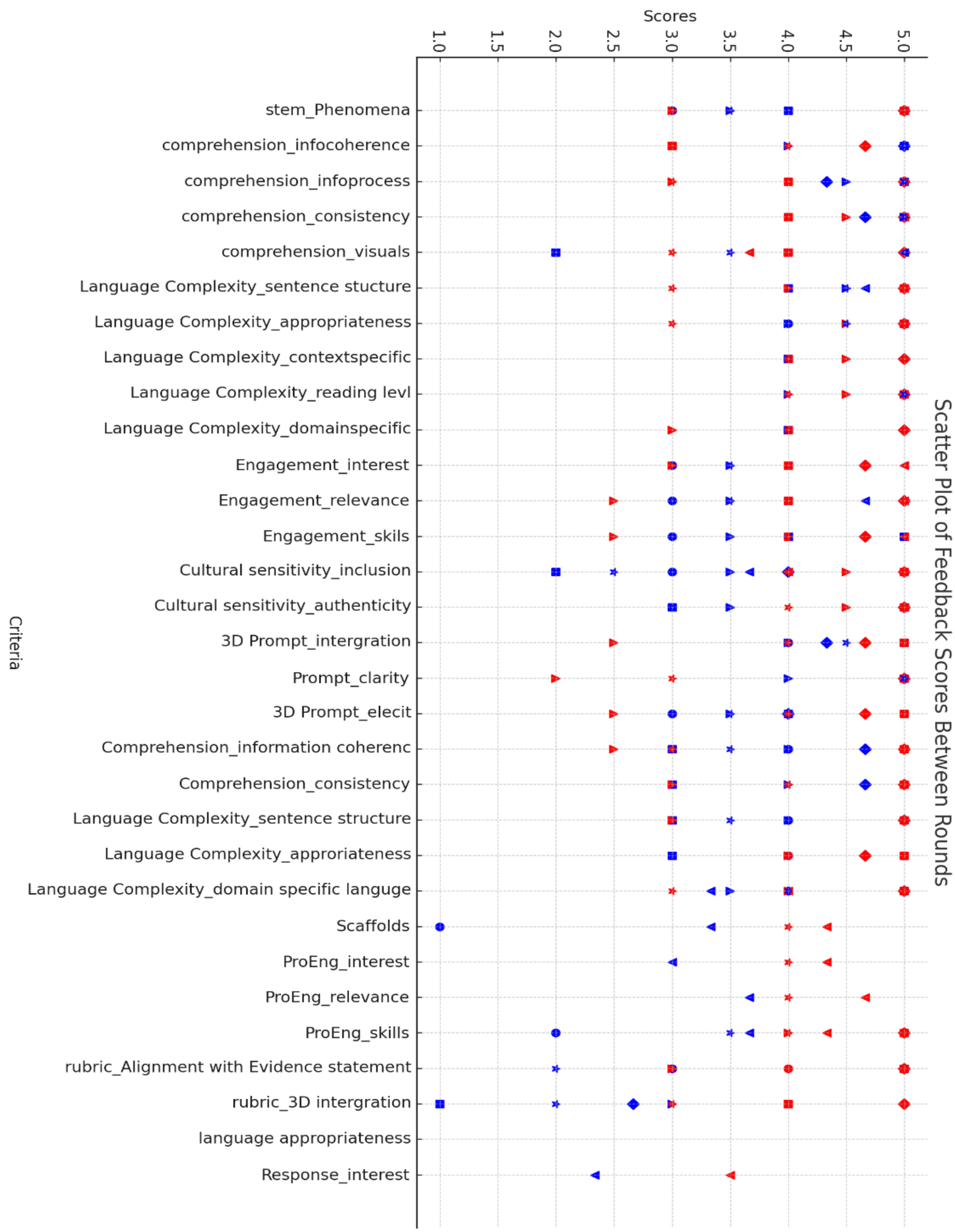


Despite these improvements, some criteria like "Language Complexity-Visuals" and "Procedural Skills" received relatively low and unchanged scores from specific expert groups like the teacher experts, signaling areas that may still need further attention. This detailed feedback from various expert perspectives allows for targeted improvements in future iterations of the assessment tasks, ensuring they are more effective and relevant.

3-PS2-1: Feedback on Task 2

Figure 4-31 presents the two rounds of expert feedback on Task 2 for LP2 of 3-PS2-1. In the first round, represented by blue markers, scores generally varied across different criteria, with several areas showing room for improvement. The second round, represented by red markers, shows a significant overall improvement in scores, suggesting that the feedback from the first round was effectively integrated into subsequent revisions. The NGSS experts showed marked improvements in areas related to the authenticity and relevance of STEM phenomena, demonstrating greater satisfaction with how these were presented in the second round. Assessment experts provided higher scores particularly in criteria involving engagement relevance and procedural skills, indicating that the assessments better captured student interest and effectively measured relevant skills in the second iteration. Science content experts, who focus more on the accuracy and depth of content, reflected increased scores particularly in language complexity and content coherence, suggesting enhanced clarity and alignment with scientific standards.

Figure 4-31. Scatter plot for the two rounds expert feedback on Task 2 for LP2 of 3-PS2-1

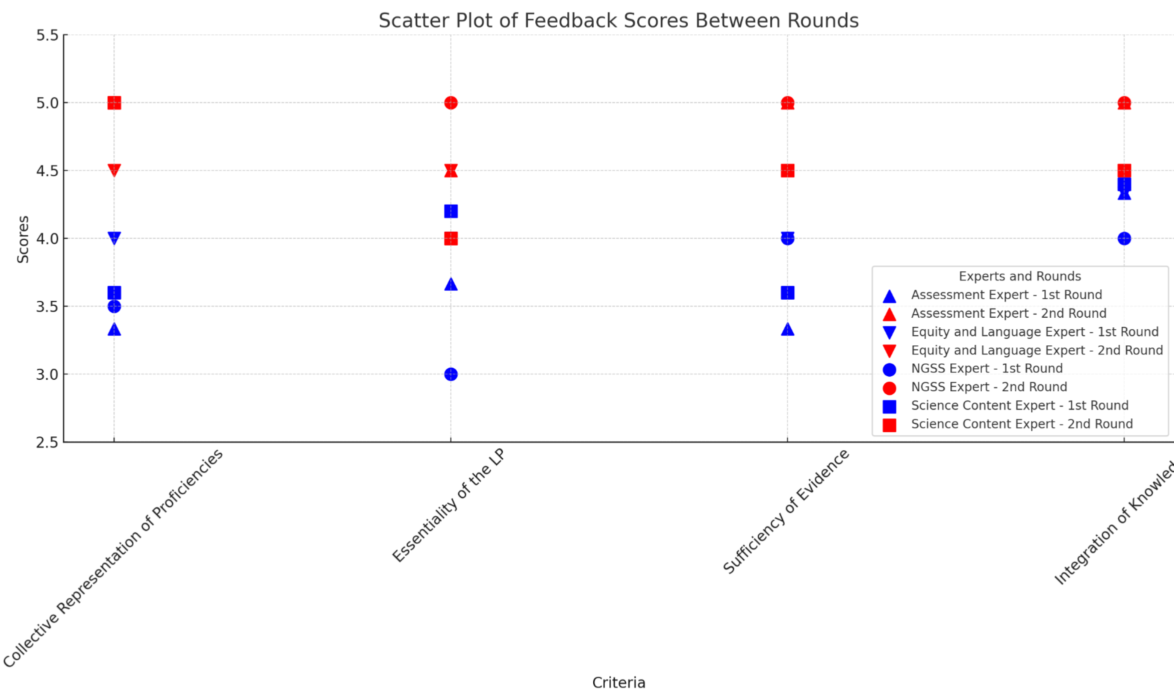


Equity and language experts, whose feedback is crucial for ensuring inclusivity and accessibility, noted better performance in cultural sensitivity and language appropriateness. Engagement experts, focusing on how engaging and relevant the content is for learners, recorded higher scores in engagement interest and skills relevance, highlighting more compelling and relevant content in the second round. Teacher experts, whose perspectives are vital for practical classroom application, also showed improvement, particularly in the clarity of prompts and the alignment of evidence statements with learning goals. This detailed comparison between the two rounds highlights the effective incorporation of expert feedback into enhancing the overall quality, relevance, and effectiveness of the assessment tasks.

3-LS4-3: Feedback on LPs and Evidence Statements

Figure 4-32 shows the two rounds feedback on the designed LPs and evidence statements of 3-LS4-3.

Figure 4-32. Scatter plot for the two rounds expert feedback on the LPs and evidence statements

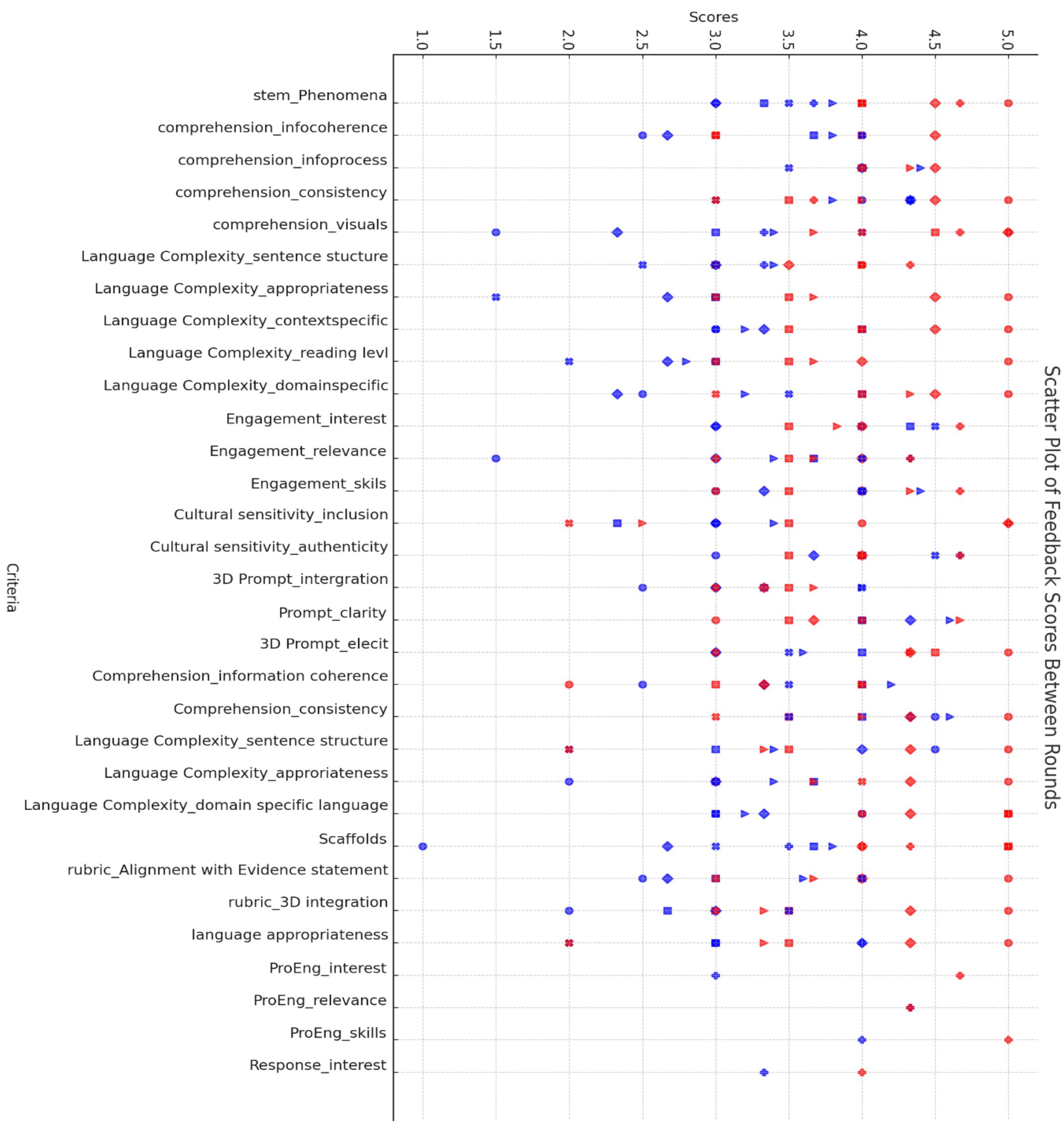


For the LPs and evidence statements for 3-LS4-3, Figure 4-32 shows a clear shift towards higher evaluations in the later round, reflecting the positive impact of the revisions. Initially, scores across the board were somewhat lower, indicating areas of concern or need for improvement. For example, the NGSS and science content experts provided stringent feedback particularly on the sufficiency of evidence and essentiality of the LPs, highlighting a need for more precise alignment with scientific standards.

Assessment experts mirrored these concerns, focusing on the overall effectiveness of the assessments. The second round shows an uplift in scores across all dimensions, signaling that the revisions were well-received. Enhancements in the sufficiency of evidence and the integration of knowledge were particularly well noted by NGSS and science content experts, showcasing an improved alignment with educational standards and better representation of proficiencies.

3-LS4-3: Feedback on Task 1.

Figure 4-33. Scatter plot for the two rounds expert feedback on Task 1 for LP2 of 3-LS4-3



In terms of Task 1 for LP2 of 3-LS4-3, Figure 4-33 provides a detailed comparison of feedback scores from a panel of experts across a wide array of assessment criteria over two review rounds. The plot reveals significant differences in scores between the rounds, indicating how expert feedback influenced the revisions of the assessment tasks. During the first round, feedback scores were generally lower across most criteria, suggesting initial concerns or deficiencies identified by the experts. For instance, lower initial scores in areas like engagement relevance, cultural sensitivity, and procedural skills highlighted the need for more focused adjustments to better cater to diverse learner needs and the practical application of knowledge.

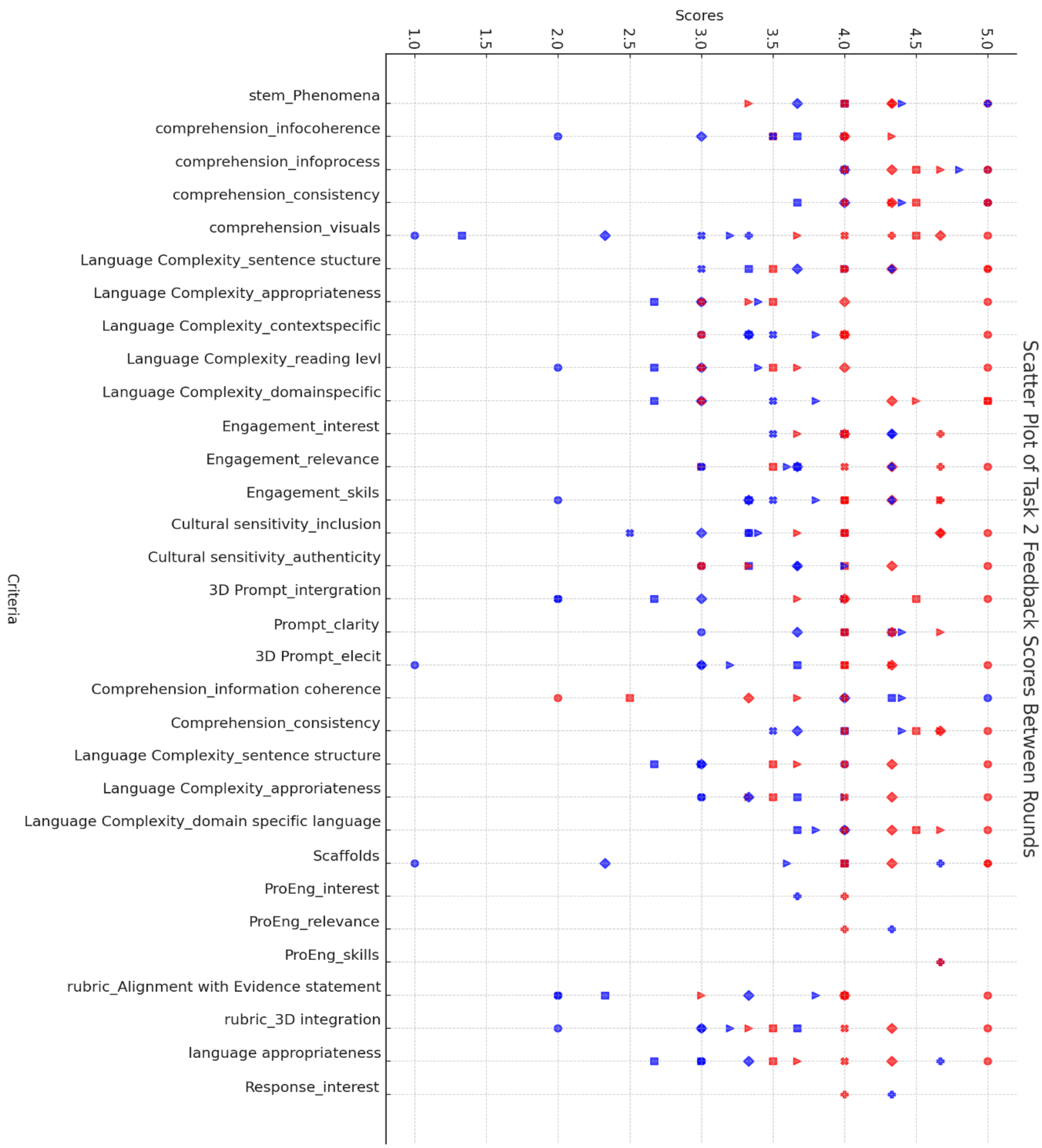
In the second round, improvements are evident across almost all criteria, with markedly higher scores indicating that the changes made were effective. NGSS experts, who focus on alignment with scientific standards, showed higher satisfaction in the second round, especially in criteria related to scientific accuracy and coherence. Similarly, assessment experts provided higher scores on criteria assessing the effectiveness of task design and alignment with learning goals, suggesting that the revisions better met assessment objectives.

The comparison of scores also illustrates the effectiveness of the iterative feedback process, where adjustments based on expert critiques led to enhancements in the assessment's design and content. This iterative process ensures that the assessments are not only comprehensive but also effective and appropriate for educational use, demonstrating a successful adaptation to the experts' insights.

3-LS4-3: Feedback on Task 2

The feedback on Task 2 is also similar to Task 1. Figure 4-34 provides a comparative analysis of feedback scores across a range of criteria related to educational task assessments, as reviewed by various expert groups over two rounds. From the initial to the revised assessments, there is a noticeable improvement in scores across most criteria, signaling that the modifications made were effective. This is particularly evident in the scores provided by NGSS experts and assessment experts, who showed increased satisfaction in areas like 'Information Coherence' and 'Language Complexity.' Their scores in the second round are consistently higher, reflecting their approval of the adjustments made in response to their initial critiques.

Figure 4-34. Scatter plot for the two rounds expert feedback on Task 2 for LP2 of 3-LS4-3



Science content experts, whose focus is the depth and accuracy of content, demonstrated increased scores in the 'Language Complexity' and 'Procedural Skills' criteria. This suggests that the revisions addressed their concerns about the clarity and application of scientific concepts in the tasks. Engagement experts, whose evaluations focus on how engaging the tasks are for students, showed higher scores in 'Engagement Interest' and 'Engagement Relevance' in the second round. These improvements indicate that the tasks were adjusted to be more engaging and relevant to students, aligning better with educational engagement goals.

Overall, the plot illustrates an effective feedback loop where expert critiques were taken into account, leading to substantial enhancements in task design and execution, which were acknowledged by higher scores in the subsequent review round. This iterative process underscores the value of expert feedback in refining educational assessments to better meet pedagogical objectives and improve learner outcomes.

4.3.2.2 Blinded Expert Panels

I report the expert panels who were not informed the assessments and the LPs were designed by AI. Below, I first present the experts' feedback on the PE 3-PS2-1 and then move to the feedback on the PE 3-LS4-3.

Feedback on LPs and evidence statement for PE3-PS2-1

The expert feedback on the LPs and evidence statements for 3-PS2-1 provided a rich source of insights. This feedback emphasized both the strengths and areas for improvement in the LPs, offering specific recommendations for future revisions. The expert reviews acknowledged the effective integration of 3D learning and commended the clarity and comprehensiveness of the LPs.

Positive feedback on LPs and evidence statements. Experts widely commended the alignment and effectiveness of the LPs in representing the essential proficiencies required to meet 3-PS2-1, emphasizing the depth and thoroughness of the educational framework. One reviewer praised the structured approach to three-dimensional learning, stating, "The LPs and components in the evidence statements—model elements, relationships, and explanations—are meticulously unpacked to ensure comprehensive coverage of the 3Ds, providing a robust educational framework that effectively fosters student understanding."

The clarity and precision in modeling and explanations within LP2 were highlighted as standout features. This clarity was particularly beneficial for addressing common misconceptions about forces on stationary objects. A reviewer elaborated on this strength, noting, "LP2 excels in demystifying the dynamics of forces, clearly illustrating how balanced forces interact on stationary objects, thereby countering the prevalent student misconception that stationary objects are not subject to forces."

Areas for improvement in LPs and evidence statements. Several experts have raised concerns about the learning performances, particularly highlighting the insufficient emphasis on "investigation" practices within the current framework. Most notably, only LP1 focuses on investigations, which might lead to confusion among students as they attempt to grasp the full scope of the curriculum requirements. One expert elaborated on this issue: "The set of LPs lack of emphasis on critical investigative skills, which are central to understanding forces. Students might struggle to differentiate the unique aspects of each performance, leading to a superficial understanding rather than a deep conceptual grasp."

In addition to the need for a broader focus on investigative practices, the lack of adequate scaffolding in LP2 was frequently noted. Experts are calling for more structured support systems to aid students in developing robust models and explanations. This concern was voiced by an expert who questioned the current educational supports: "I am curious as in LP2 would students be offered any scaffolding for the model and explanation."

The terminology used within the LPs also received feedback suggesting the need for refinement to better foster scientific inquiry. An expert specifically addressed this in the context of LP5, recommending a shift in the educational approach: "LP5 should ask for an evidence-based argument rather than an explanation, you do not explain empirical facts; you argue that the evidence indicates that the statement is true." This change aims to enhance the rigor and accuracy of the educational content, aligning it more closely with the practices of empirical science.

Further emphasizing the need for developmentally appropriate content, experts suggested that the evidence statements should accommodate the capabilities of third graders more effectively. This involves fostering evidence-based reasoning and argumentation that aligns with young students' understanding levels. For example, expert H proposed a model to illustrate balanced forces in a way that is accessible to younger learners: "Consider a ladder leaning on a wall, there are not equal arrows because the downward force of gravity (weight of ladder) is balanced by two upward forces (friction on the wall and upward force of ground on foot of ladder). Students create a simple drawing or diagram that shows an object at rest with balanced forces on it." This example not only aids in understanding the concept of balanced forces but also demonstrates the kind of practical, visual learning tools that can help third graders grasp complex scientific ideas.

These insights from the experts underscore the necessity for revising the LPs to incorporate a greater emphasis on investigative skills, provide more substantial scaffolding, refine the use of terminology to promote evidence-based reasoning, and tailor the evidence statements to better suit the cognitive abilities of third-grade students.

Feedback on task1 for PE3-PS2-1

Experts provided comprehensive feedback addressing the visual and conceptual aspects of the learning tasks. They focused on the representation of physical forces and the accessibility of the tasks for third-grade students, emphasizing the importance of aligning educational content with students' developmental levels and prior experiences.

*Positive feedback on task 1***.** The feedback from the experts highlighted several strengths in the design and execution of the task1, particularly in how the phenomena were presented to engage the students effectively. One expert emphasized the compelling nature of the phenomena used in the tasks, stating, "Students experience the effects of gravity all the time but seldom are confronted with objects at rest and balanced forces. My experience with elementary students is that they have found this an interesting phenomenon to make sense about." This observation underscores the relevance and engagement potential of the task, aligning well with the students' everyday experiences and curiosities. Another expert further supported this view by elaborating on the educational impact of the phenomena, mentioning, "The way the tasks introduce students to the invisible yet ubiquitous forces at play in everyday objects provides a foundational understanding that stimulates curiosity and critical thinking."

This additional insight highlights how the tasks not only align with what students observe daily but also challenge them to think deeply about the physical world. Moreover, the feedback also included appreciation for how these tasks are structured to promote scientific inquiry. As one reviewer pointed out, "By engaging students with scenarios that are both familiar and intriguing, the tasks encourage a deeper exploration of scientific concepts that they can see and feel but often do not notice." This quote reflects the thoughtfulness behind the task design, aiming to transform everyday observations into opportunities for scientific discovery and understanding.

Areas for improvement for Task 1. Representation of Forces: Concerns were raised about the level of abstraction in how forces are represented in the task. Experts have stressed the importance of **developing more intuitive and less abstract representations of forces** to better suit third graders' understanding levels. Specifically, an expert commented on the potential confusion this might cause for third graders, "The students may have been taught to represent the downward force of gravity and the upward force of the table by a single arrow but this is a very abstract representation for the third grader, maybe the student would represent the upward force by lots of little arrows because the table pushes up wherever the book touches it, not just at the middle of the book." This feedback underscores the need for more developmentally appropriate visual representations that align with third graders' conceptual understandings, suggesting a more detailed approach that mirrors how young students perceive interactions between objects. It also highlights the need for educational materials to align more closely with how young students visually and conceptually perceive physical interactions.

Visuals and Context: There is a strong push for **incorporating visuals that are more familiar and relevant to students' daily school experiences**. This comes in response to feedback about the current visuals not adequately engaging the intended audience. The visuals used in the task were noted to be less suitable for the intended audience. An expert pointed out the disconnect, saying, "This photo is for adults. Consider using a photo showing a book on a desk in a classroom setting." This recommendation emphasizes the need for visuals that are more relatable and understandable for young students, highlighting the importance of context and environment in educational materials.

Scaffolding and Terminology: **Enhancing scaffolding and support** in the tasks has been identified as crucial for aiding students' scientific reasoning and expression. The feedback also highlighted deficiencies in scaffolding and the need for clearer terminology. An expert raised a significant question regarding the representation of forces, asking, "Would you accept one big down arrow and multiple small up arrows as an 'accurate representation of the situation?'" This inquiry indicates a gap in the guidance provided to students on how to accurately model and discuss scientific ideas.

Feedback on task 2 for PE3-PS2-1

The feedback from experts on Task 2 of PE 3-PS2-1 centers around several key aspects related to the representation of forces and the design of the task. This feedback provides actionable insights, specifically focusing on the representation of forces and the contextualization of the task scenarios.

Positive feedback on Task 2. Engagement and contextual relevance: Experts appreciated the setup and context provided in Task 2, noting its potential to effectively engage students. One expert specifically commended the task's design and offered a suggestion to enhance its appeal to the students' interests: "I like this task. Picture is good, maybe the setup would be more interesting to students if instead of helping to clean Lisa and John want to move the table to make space to dance or put on a show… (it's easy enough to sweep under that table)." This feedback underscores the value of crafting scenarios that are directly relevant and exciting to students, ensuring they are more than just educational but also enjoyable. Another expert emphasized the importance of context in the learning process, stating, "Engaging students with scenarios that mirror their real-life experiences not only makes the tasks more interesting but also enhances their understanding of the scientific principles being taught." This viewpoint reinforces the strategy of using relatable and dynamic scenarios to promote a deeper connection between the students and the educational content, making learning a more integrative and enjoyable experience.

Integration of real-life contexts: The positive feedback was further enhanced by suggestions to make the tasks more practical and focused on core concepts. An expert advised on optimizing the instructional visuals, saying, "Why not ask the kids to represent the forces acting on the table on an outline version of the picture, rather than wasting their time redrawing the setup in a blank space." This recommendation underscores the importance of using visuals effectively to simplify the task execution and concentrate on sensemaking rather than on redundant activities. Another expert elaborated on this idea, emphasizing the educational benefits of such an approach: "Using pre-drawn outlines for students to annotate forces can significantly reduce cognitive load and allow them to focus more on the scientific principles involved." This insight suggests that reducing the complexity of tasks can help students better engage with and understand the scientific concepts being taught, making learning both efficient and effective.

Areas for improvement for Task 2. The feedback from experts highlighted concerns about the abstract nature of force representation, which might be confusing for younger students. One expert critically examined the typical methods of depicting forces, suggesting a more approachable method for third graders: "Again the question of what is an acceptably 'accurate' representation arises – would you accept a balance of the ¼ of table weight pulling it down and the floor pushing up at each place the table touches the floor, or would you only accept one downward force of gravity acting at the center of the table and four upward forces at the feet adding to the same total?" This expert's query points to the necessity for flexible and developmentally appropriate visual representations that resonate with young students' ways of understanding and visualizing forces.

The need for better scaffolding and clearer terminology was also emphasized by experts to enhance how students model and discuss scientific concepts. Highlighting a potential improvement in instructional guidance, one expert asked, "Is it OK to have the arrows not quite to scale with the forces because the model is just a sketch, if the student writes words to explain the balance of forces?" This query underscores the need for more precise guidelines and supports students in expressing their scientific observations accurately and meaningfully. This feedback suggests integrating clearer explanations alongside visual models, thus bridging the gap between abstract scientific models and students' understanding.

Table 4-39 below presents the summary of the expert feedback on LPs, tasks for 3-PS2-1.

Structured table summarizing the major themes from the expert feedback on Learning Performances (LPs), Task 1, and Task 2 of PE 3-PS2-1, categorizing the feedback into positive aspects, areas for improvement, and suggestions for each section.

Table 4-39. Summary of the expert feedback on LPs, tasks for 3-PS2-1

Aspect	Positive Aspects	Areas for Improvement	Suggestions
LPs and Evidence Statements	- Effective integration of 3D learning standards.	- Insufficient emphasis on "investigation" practices.	- Incorporate more investigative practices.
	- Clarity and comprehensiveness in modeling and explanations.	- Lack of adequate scaffolding, especially in LP2.	- Provide more structured support systems for developing models and explanations.
	- Alignment with essential proficiencies of 3-PS2-1.	- Need for terminology refinement to foster scientific inquiry.	- Refine terminology to enhance rigor and align with empirical science practices.
Task 1	- Engaging setup and relevance to students' experiences.	- Abstract representation of forces potentially confusing.	- Develop more intuitive and less abstract representations of forces.
	- Encourages exploration of scientific concepts.	- Visuals not adequately engaging for the intended audience.	- Use more familiar and relevant visuals to enhance relatability and comprehension.
	- Structured to promote scientific inquiry.	- Need for clearer guidelines and better scaffolding.	- Implement clearer guidelines and enhance scaffolding to support scientific reasoning and expression.
Task 2	- Effective engagement and contextual relevance.	- Abstract nature of force representation confusing for students.	- Use practical, contextually relevant scenarios and optimize visual aids.
	- Use of real-life contexts enhances understanding.	- Need for more intuitive methods for depicting forces.	- Allow flexibility in how forces are represented to accommodate different understanding levels.
	- Suggestions to improve task setup and execution noted.	- Need for clearer terminology and better scaffolding.	- Ensure terminology and scaffolding are appropriate for third graders' cognitive abilities.

Feedback on LPs and evidence statement for PE 3-LS4-3

The feedback from experts regarding the LPs and evidence statements for 3-LS4-3 provides critical insights. Below is a summary of the positive aspects and areas for improvement based on the feedback provided.

Positive feedback on LPs and evidence statements. Experts have recognized the thoroughness with which the LPs cover the essential proficiencies required for PE 3-LS4-3. The structured approach to integrating three-dimensional learning is particularly commended for its effectiveness in fostering a deep understanding of biological adaptations. An expert highlighted the breadth of coverage offered by the LPs, stating, "This one covers all main aspects," affirming that the LPs comprehensively address the core elements expected in the curriculum. Another expert elaborated on this point, noting, "The learning performances are well-designed to encompass a broad spectrum of critical concepts that are necessary for students to master the performance expectation, ensuring no key element is overlooked."

The design and implementation of multiple evidence statements linked to specific LPs have been positively received. Experts view this approach as a robust framework for educational design. One expert praised the clarity and utility of this structure, saying, "I find the concept of a learning performance with an explicit list of multiple evidence statements a very strong template for task design." This sentiment was echoed by another expert who emphasized the advantages of such a structured approach: "Using multiple evidence statements provides a clear pathway for students to demonstrate their comprehension and application of the learned material, which significantly aids in both teaching and assessing complex concepts."

Areas for improvement for LPs and evidence statements. Significant feedback has emerged regarding the "critique" section within the LPs, particularly concerning its appropriateness and execution. Experts argue that the current approach might not fully grasp the complexity of how traits contribute to an organism's survival. Consequently, it is recommended to alter the evidence statement to more thoroughly assess the validity of claims, focusing on evidence-based reasoning rather than comparative strength. This adjustment encourages students to evaluate various claims about survival traits within their environmental contexts, fostering deeper analytical skills. One expert suggested, "Students evaluate other claims about traits that contribute to survival of the organism in this environment and provide evidence-based reasoning as to whether or not they find the claim valid." This change aims to enhance the critical thinking aspect of the curriculum, ensuring students engage more profoundly with the material.

Feedback has also highlighted the need for improved integration of the three-dimensional learning framework, encompassing disciplinary core ideas, science and engineering practices, and crosscutting concepts. While two dimensions are generally well-integrated, the third, specifically the crosscutting concept of systems, often remains only implicitly addressed. To remedy this, it is recommended to ensure all three dimensions are explicitly incorporated and effectively elicited through student tasks. This enhancement will provide a more balanced and comprehensive educational experience, allowing students to better understand and apply complex scientific principles in varied contexts.

Concerns about the contextual relevance of certain examples, like the arctic fox, have been raised, particularly considering contemporary environmental issues such as climate change. An expert noted, "I wonder whether some students might say, ‘Because of climate change, the arctic fox does not need the thick fur.’" This feedback points to the necessity of updating and expanding the contexts and examples used within the LPs to ensure they remain relevant and reflective of current scientific and environmental understandings. Updating these examples will help prevent misconceptions and provide students with a more accurate and relatable learning experience.

Feedback on Task 1 for 3-LS4-3

Positive feedback on Task 1. Experts have acknowledged the task format's ability to engage students effectively by presenting phenomena in a manner that is both engaging and educational. The tasks are carefully designed to ensure that the phenomena are not only comprehensible but also intriguing, thereby fostering a deeper interest and engagement among students. However, despite the general appreciation for the task's format, some experts voiced concerns that might impact the perceived realism and engagement quality. One expert pointed out, "Maybe [it's] compelling but see my comment above," highlighting that while the scenario is designed to be engaging, there may be underlying issues that affect its effectiveness in conveying realistic scientific phenomena.

The clarity of language and the structural organization of the tasks have been highlighted as strengths in the feedback from educational professionals. The tasks are praised for their coherent structure, which aids in sequential information processing and enhances comprehensibility. An expert commented on the beneficial structure of the tasks, stating, "well structured info." This feedback underscores the success of the task design in aligning with educational standards, making the content not only accessible but also effectively sequenced to facilitate student understanding and learning.

Areas for improvement for Task 1. The realism and authenticity of the data used in educational tasks are crucial for maintaining credibility and fostering genuine scientific inquiry among students. Concerns regarding these aspects were notably raised by experts reviewing the task. One expert explicitly criticized the believability of the scenario based on professional experience, stating, "My problem with this task is that it appears to me to be unrealistic. Based on my experience as a docent at a biological preserve at Stanford this data is fake." This feedback points to a significant issue with how data and scenarios are presented to students, emphasizing the need for educational materials to either use authentic data or to clearly label hypothetical scenarios as such to prevent confusion and enhance educational integrity. The criticism underscores the importance of aligning educational tasks with realistic scientific standards to ensure they effectively prepare students for real-world scientific understanding and applications.

The ecological accuracy of educational tasks is fundamental in teaching students about biology and the environment effectively. Experts reviewing the tasks raised concerns about the appropriateness of the settings and biological descriptions provided in the tasks. Specifically, the portrayal of hummingbirds and their environmental interactions was highlighted as not fully aligning with known biological facts. One expert provided detailed feedback on the necessary improvements to enhance realism and accuracy, stating, "Hummingbirds need a place to perch and to nest, open meadow may be part of their habitat but they are more likely found in a mixed environment which you call woodland, perhaps open woodland would be a better term." This feedback emphasizes the need to adjust the environmental settings described in the tasks to reflect more accurate biological and ecological conditions. By doing so, the tasks will not only become more scientifically precise but also provide students with a more authentic understanding of how organisms interact with their environments.

The presentation of observations from two children as contradictory has been identified as a potential source of confusion for students, particularly those at the elementary level. This concern was addressed by experts who noted the importance of clear and accurate communication in educational settings. One expert specifically commented on the unnecessary complexity introduced by framing the children's observations as contradictory, emphasizing that such an approach could be misleading. The expert advised, "See above no need to frame two students' ideas as if they are contradicting one another, they are not." This feedback suggests a revision in the way student observations and arguments are presented within the tasks to ensure they are clear and supportive of the learning objectives, without inadvertently confusing younger students.

Cultural and contextual sensitivity in educational tasks is crucial to ensure that all students find the content relatable and engaging, regardless of their background. Experts have highlighted a gap in the current tasks concerning their relevance to students from diverse environments, particularly those from urban or non-forest areas. An expert deeply concerned about this issue provided pointed feedback, noting the disconnect many students might feel: "Children from urban and even some suburban environments may never have seen forest, woodland, or a natural meadow." This observation underscores the importance of designing educational materials that cater to a broad audience by incorporating a variety of environments that are familiar to different groups of students. This feedback calls for an expansion in the types of settings and scenarios used in tasks to ensure they resonate with students from various geographic and cultural backgrounds. By doing so, educational materials can better serve their purpose of educating a diverse student population effectively, ensuring that no student feels alienated due to a lack of familiarity with the content presented.

Feedback on Task 2 for 3-LS4-3

Positive feedback on Task 2. Experts recognize the task's structured approach and the comprehensive way it engages students, particularly noting its effectiveness. One expert comments, "It is compelling and comprehensive to students," emphasizing how well the task captures and maintains student interest through its educational design. The orderly presentation of information is specifically highlighted for facilitating efficient information processing. "The order of information in the item stem is very good," confirms another expert, underscoring the clarity and structured nature of the task. Additionally, the effective use of visual aids enhances comprehension, as another reviewer points out, "Yes, the images helped a lot, well presented plants with different organisms." These visuals play a crucial role in reinforcing the educational content, making complex concepts more accessible to students.

Areas of improvement for Task 2. While the task's design has been positively received, experts have raised concerns about its realism and approach to teaching complex ecological concepts. One expert expressed frustration with the oversimplification presented in the task, stating, "My problem with this task is it asks for generalization based on single examples...The grass is not helpful, actually grass does not grow in wetlands, its seeds are waterlogged and do not germinate." This critique highlights the necessity for tasks to feature more authentic or clearly hypothetical scenarios to accurately reflect ecological realities and prevent the formation of misconceptions. Additionally, there's an emphasized need for clearer objectives to deepen students' understanding of environmental science. An expert critically notes, "Success in the task comes because the answers are made obvious by the stem, not because you understand anything about the nature of the environments and what it takes to survive in each." This feedback suggests that to truly enhance educational outcomes, the task should be revised to better connect its purposes with the intended educational goals, potentially by refining the prompts to ensure they more effectively guide student inquiry and engagement in learning scientific concepts. Table 4-40 shows the summary of the expert feedback on LPs and tasks for 3-LS4-3.

Table 4-40. Summary of the expert feedback on LPs and tasks for 3-LS4-3

Aspect	Positive Aspects	Areas for Improvement	Suggestions
LPs and Evidence Statements	- Comprehensive coverage of essential proficiencies for PE 3-LS4-3. - Effective integration of three-dimensional learning. - Structured approach praised for fostering deep understanding of biological adaptations. - Multiple evidence statements provide a clear pathway for student assessment.	- "Critique" section needs better alignment with the complexity of trait contributions to survival. - Crosscutting concepts need clearer integration.	- Revise evidence statements to focus on evaluating the validity of claims rather than their strength. - Explicitly integrate all three dimensions of learning.
Task 1	- Engaging and educational presentation of phenomena. - Structure aids in sequential information processing. - Visual aids enhance comprehension and engagement. - Language and task organization align well with educational standards.	- Concerns about the realism and authenticity of data. - Ecological accuracy of settings and descriptions. - Presentation of contradictory observations.	- Use authentic data or clearly indicate hypothetical scenarios. - Adjust environmental settings to reflect accurate biological conditions. - Clarify contradictory statements.
Task 2	- Structured approach effectively engages students. - Visual aids enhance understanding. - "Order of information in the item stem is very good."	- Realism of ecological concepts questioned. - Oversimplification of complex concepts like plant survival in various environments.	- Feature more authentic or hypothetical scenarios that accurately reflect ecological realities. - Refine prompts to better guide scientific inquiry.

CHAPTER 5: CONCLUSIONS AND IMPLICATIONS

This dissertation employed design-based research to explore how humans can work with AI to design knowledge-in-use assessments for elementary students to support their science learning. Interdisciplinary expert panels with diverse expertise collaborated to provide feedback on the co-designed knowledge-in-use assessments and interim products that are critical for the assessment design. This dissertation explores three major research questions: discussing how to iteratively and effectively design knowledge-in-use assessments with AI, the role humans play in the designing process, and the role that AI plays in the process, including where and how the synergy occurs to design these assessments. The dissertation found that humans can collaborate with AI to design knowledge-in-use assessments. The designed assessments were distributed to interdisciplinary expert panel members for review, and their collective feedback provided comprehensive insights for the assessment refinement process. Incorporating the collective feedback, the human operator worked with AI again to refine the designed assessments by incorporating the expert panel feedback. Refinement principles and frameworks were generated during the process. Additionally, the refined assessments were distributed to two different expert panels for review, including a new group of experts who were unaware that AI was involved in the design process to mitigate potential bias. The revised assessments received higher evaluations from the original expert panel compared to the first round of assessments. Interestingly, the new expert panel, who were not informed that AI was involved in the assessment design process, provided even more positive feedback compared to the expert panels, who knew that the assessments were designed by human and AI collaboration. Below, I discuss each research question.

5.1 Enhancing Knowledge-In-Use Assessments Design through Collaborating with AI

This dissertation builds on the evidence-centered design (ECD) approach to develop knowledge-in-use assessments by adopting the Next Generation Science Assessment (NGSA) approaches through collaboration with collective human experts and AI models. This research contributes to the body of knowledge on designing formative assessment tasks for measuring complex cognitive constructs, as explored in prior work (Harris et al., 2019, 2024; He et al., 2023; Li et al., 2024) The findings underscore the efficacy of the NGSA approach to designing knowledge-in-use tasks that were previously developed by humans. Further, this dissertation extends the NGSA design approach from solely human collaboration to human-AI collaboration, which saves both time and labor from previous studies (Pellegrino & Hilton, 2012). The systematic evidence-centered design approach of NGSA can effectively guide AI models in designing knowledge-in-use assessment tasks with the guidance of human operators and collective expert intelligence, which adds on to the effectiveness of ECD in designing assessment tasks (Mislevy & Haertel, 2006; Wilson et al., 2005).

This dissertation builds on and extends current literature by demonstrating that while AI can generate valuable educational content, its effectiveness is significantly amplified when guided by explicit and detailed human instructions, which aligns with existing research (Luckin et al., 2016). By integrating human expertise, iterative feedback, and detailed guidance, AI-generated outputs can achieve a high level of detail and accuracy, meeting educational standards and supporting effective assessment design. This finding reinforces the importance of human oversight in the AI design process, underscoring the collaborative dynamic between human and AI that fosters a synergistic relationship (Bearman & Ajjawi, 2022), which enhances the overall quality and effectiveness of knowledge-in-use assessments. The dissertation also extends the literature by identifying emerging themes (Table 5-1) that highlight the importance of effectively and iteratively working with GPT-4 models in designing knowledge-in-use assessment tasks. Throughout this process, human experts play a critical role in guiding and refining AI outputs. This supports the notion that while AI has the capacity to learn from provided frameworks and examples, it is crucial for humans to provide comprehensive feedback to ensure the outputs are accurate and pedagogically sound (Fenwick, 2010). The dissertation illustrates how the collaborative dynamic between AI and human experts facilitated the creation of Instructional Design Models (IDMs) and LPs that align with educational standards and support effective learning. By addressing these themes, the iterative process showcased in this dissertation demonstrates how AI and human collaboration can produce high-quality educational assessments. This extends the current understanding of AI's role in educational design, showing that AI-human partnerships can create tools that are not only efficient but also pedagogically sound, enhancing the potential of AI in education.

This dissertation reinforces the importance of hybrid intelligence between humans and AI, further extending the role of hybrid intelligence to complex cognitive constructs and systematic design approaches in education (Dellermann et al., 2021; Holmes, 2020). The contributions of this dissertation are twofold. First, it emphasizes the collaborative effort required in designing knowledge-in-use tasks, necessitating experts from various domains to guide and monitor the assessment design, as noted by Harris et al. (2024) and the National Research Council (2006). Assessment is a systematic effort that must consider various levels of thinking about learning, particularly for formative assessments. This dissertation further solidifies the role of assessment as a crucial component of the educational system, requiring collaborative efforts to ensure that designed assessments accurately capture students' performance and effectively inform teaching and learning.

This dissertation extends the effectiveness of the NGSA approach in designing knowledge-in-use assessment tasks and highlights the significance of evidence-centered design in these tasks. The collaborative efforts gathered from diverse experts are pivotal for effective assessment design, adding to previous findings that emphasize the necessity of expert reviews in validating assessment tasks (Black & Wiliam, 1998; Shepard et al., 2018). Additionally, this study underlines several important aspects that should be addressed when designing knowledge-in-use assessment tasks, including the integration of 3D proficiencies (NGSS, Lead States, 2013), creating equitable assessments that ensure all students can access the tasks (Darling-Hammond & Snyder, 2000), designing engaging assessments relevant to students' lives, considering language appropriateness (Lee, Quinn, & Valdés, 2013), and ensuring assessments can elicit evidence to understand students' performance effectively (Furtak, 2017, 2023; Penuel & Smolek, 2019). Moreover, this dissertation has the potential to expand on how to engage students from diverse backgrounds in the assessment tasks/scenarios. Since engagement is influenced by students' personal experiences and cultural backgrounds, designing universally engaging tasks is challenging. The integration of AI in assessment design has the potential to more efficiently provide alternative and adaptive scenarios, enhancing the ability to engage a diverse student body (Baidoo-Anu &

Owusu Ansah, 2023). Future research should explore how to further develop adaptive assessments.

The second major contribution of this dissertation involves expanding the collaboration with AI in designing assessment tasks. By extending the collaborative partner from humans to AI models, this dissertation enriches distributed cognition theories and emphasizes the importance of hybrid intelligence in human and AI collaborations (Hutchins, 2000; Pea, 1993). This dissertation provides deeper insights into collaborating with AI models to design assessments that capture complex cognitive constructs. The iterative design process, detailed instructions, and collaboration with human experts of varied expertise broaden the scope of human-AI collaboration, making it a more holistic and comprehensive approach for integrating AI into education. The Hybrid Human-AI Collaborative Model (HHACI) exemplifies a collaborative approach for complex task design, elucidating how human and AI models can work together, particularly highlighting AI models' strengths in efficiency, flexibility, and vast information access (Johnson & Verdicchio, 2017). AI's capabilities to detect, diagnose, act, and learn from the human-AI collaboration process, and to reflect on these experiences to inform future tasks, underscore the potential of AI in education. More importantly, this dissertation reaffirms the irreplaceable value of human intelligence in the collaborative process, emphasizing that while AI can augment many aspects of educational tasks, the nuanced judgment and ethical considerations provided by humans remain indispensable (Li et al., 2023; 2024).

5.2 Leveraging Interdisciplinary Expertise to Enhance Knowledge-In-Use Assessment Design

The second research question investigates the type of feedback provided by interdisciplinary expert panels on the assessment design. This dissertation discovered that diverse expertise from various perspectives is crucial for designing effective knowledge-in-use assessments. Education is a complex domain, and 3D learning is an integrated approach toward complex higher-order skills and thinking, which is central to knowledge-in-use, as argued in this study. Designing effective assessments to measure and support this type of high-order proficiency is even more challenging due to the systemic perspectives required for assessment design. For instance, such design demands not only robust science content knowledge but also a deep understanding of science learning and teaching, the NGSS, 3D learning, student engagement, and language literacy, which may impact students' science learning (Pellegrino & Hilton, 2012; NGSS Lead States, 2013).

This dissertation investigates the type of feedback provided by interdisciplinary expert panels on assessment design, highlighting the importance of diverse expertise in creating effective knowledge-in-use assessments. Building on the existing literature, it reinforces the necessity of integrating multiple perspectives to address the complexities of 3D learning, which involves higher-order skills and thinking central to knowledge-in-use (Pellegrino & Hilton, 2012; NGSS Lead States, 2013). This dissertation extends current understanding by demonstrating that designing effective assessments requires not only robust science content knowledge but also a deep understanding of science learning and teaching, the NGSS, 3D learning, student engagement, and language literacy. This holistic approach ensures that assessments are accessible to students from diverse backgrounds and measure the three dimensions of scientific knowledge and skills comprehensively (Penuel et al., 2017). It is impractical for one or two developers to possess such extensive expertise, underscoring the value of interdisciplinary collaboration.

An interdisciplinary expert panel that included experts from various fields was assembled to address this challenge. This approach effectively gathers diverse expertise to collaboratively achieve the integrated goals of supporting knowledge-in-use proficiency development (NRC, 2012). This dissertation found that experts on the panel with different expertise can provide critical feedback in different areas. For example, NGSS experts can offer feedback on whether the LPs and evidence statements exceed the scope or grade levels of the PEs, and if the performance goals of SEPs align with the intended grade level. Such feedback on the coverage and overreach of the designed LPs and evidence statements from NGSS experts is invaluable. When it comes to the alignment and extent of integrating the three dimensions, assessment experts provided crucial feedback, including whether the designed evidence statements can be used to understand students' 3D learning, and if the designed assessments capture the integrated nature of the three dimensions rather than focusing on just one or two (Pellegrino et al., 2014; Wilson, 2005). Science content experts typically conducted critical examinations of the appropriateness of the science ideas and the accuracy of the science mechanisms presented in the tasks (Lee et al., 2021). However, equity and language experts in science education research can provide important examinations of the cultural sensitivity and inclusiveness of the assessment scenarios, while language experts are crucial for determining if the tasks' language level aligns well with grade-level appropriateness (Lee, Quinn, & Valdés, 2013). One significant group involved in this study is the engagement experts, who can provide slightly different perspectives on these designed assessments to examine if they can interest and engage students in the learning process when using the assessments. Interestingly, this study found that engagement experts often provide different feedback from other groups. They highlight the individual personal interest value and emphasize the importance of considering personal experiences in understanding the engaging level of the designed assessments (Hidi & Renninger, 2006). Another critical group of experts is teacher experts, who provide extensive practical feedback on the designed assessments, which are invaluable for understanding if the assessments can be used in the classroom (Heritage, 2010). Assembling interdisciplinary expert panels is thus critical for reviewing and refining assessment tasks.

The interdisciplinary expert panel provided valuable feedback on the designed assessments, including both positive feedback and suggestions for further improvement. For the positive feedback, most designed products were highly regarded for their integration of the three dimensions. However, they also received critical suggestions for further enhancement. For the LPs and evidence statements design part, the expert panel suggested: 1. ensuring appropriate grain size of LPs and evidence statements that adhere to the PE boundaries; 2. improving integration of CCCs, DCIs, and SEPs (NGSS Lead States, 2013); and 3. ensuring consistency in terminology and coherence of information. This feedback mainly came from the NGSS and assessment experts. In terms of task design, the expert panels suggested: 1. boosting engagement through relevant and contextual task design; 2. enhancing task clarity and guideline precision by providing assessment tasks with crystal-clear, straightforward instructions; 3. incorporating supportive visuals and scaffolds to emphasize the importance of integrating visual aids and scaffolding strategies into assessment tasks; and 4. ensuring cultural sensitivity and accessibility in task scenarios to create assessment tasks that are inclusive and reflective of the diverse cultural backgrounds and experiences of all students. This feedback is mainly from the assessment, equity/language, teacher, and engagement experts (Baidoo-Anu & Owusu Ansah, 2023; Furtak, 2017).

This dissertation adds to the literature by demonstrating the critical value of collaborative learning that can help with collective sense-making to solve complex problems beyond the capability of one or two team members, allowing them to learn from each other and expand the boundary of the zone of proximal development for each team member (Vygotsky, 1978). Similarly, this dissertation expands the notion of distributed cognition theory (Hutchins, 2000; Pea, 1993) from humans and tools to humans and humans with different focal expertise. When working with technology or AI, hybrid intelligence can be synthesized from different cognitive agents with different expertise to contribute to the intelligence system (Dellermann et al., 2021). This hybrid intelligence has the potential to design effective knowledge-in-use assessment within a short time frame and it also has the potential to design better assessments for diverse students. In the collaborative process, AI also detects, diagnoses, and reflects on the learning process which enhances its ability to design domain specific assessments. This is especially important in the age of AI, where it is challenging to expect everyone to have AI expertise, but it is one approach that can be used to leverage distributed cognition to augment human intelligence through a hybrid intelligence system (Luckin et al., 2016).

5.3 Integrating Expert Feedback through Human-AI Collaboration

The third research question explores the process of incorporating experts' collective feedback to refine the designed products. This dissertation proposed a refinement framework that highlights the collaboration between human operators and AI models. The human operator identifies critical places for revision based on the collective feedback and themes identified from RQ2, specifies the task goals and requirements, explains the rationales for revisions, and provides explicit guidelines for revision. More importantly, the human operator monitors the outputs to ensure goal alignment and also prompts the machine to detect and diagnose critical places that may be useful for the machine to learn from and act on for future tasks. The AI models learn iteratively and extract critical principles for their use in future refinement and action (Holzinger, 2016; Kamar, 2016). The extracted principles and lessons learned by the AI models can further inform the human's understanding of the task. This process demonstrates how AI and humans can collaborate to work on complex tasks by extending each other's ZPDs and even "cognition" (4E theory) (Malafouris, 2013; Vygotsky, 1978).

However, it is worth noting that, unlike typical collaboration processes that are often synchronous and interactive (Wenger, 1998), where team members can build on or add to each other's ideas to achieve productive engagement (Chi et al., 2018), in this dissertation, the interactive process occurs only between the human operator and AI models, which may limit the interactive and productive nature of the expert panels' feedback. Future research should explore bringing the experts into the interactive environment to achieve real collaboration (Lai et al., 2021).

Another interesting finding of this dissertation is that the group of expert panels who knew the assessments were co-designed by AI tended to provide relatively critical feedback compared to the group of experts who were not informed about the AI's involvement in the process. This could be explained by the potential bias of perceiving AI. Research indicates that preconceived notions about AI can influence expert judgments and biases towards AI-generated outputs (Jussupow et al., 2020). This phenomenon underscores the importance of transparency and managing perceptions in human-AI collaboration.

AI is good at learning extensive information quickly, but it lacks flexible thinking and empathy (Cope & Kalantzis, 2020). However, it can offer alternative phenomena or solutions, but all require human judgment before they can be used. It could perform some designs by working closely with human experts from multiple disciplines. More stakeholders need to get involved if we finally want to design a customized platform where the human plays a critical role in the entire process (Dellermann et al., 2021). For future research, this dissertation lacks the voice of students who can provide a deeper understanding of the accessibility, inclusivity, and engagement of the designed tasks. Future studies can explore how these assessments could be used in the classroom to seek further refinements. Engaging students can provide valuable insights into the practical application of assessments and ensure that they are tailored to meet diverse learning needs (Baidoo-Anu & Owusu Ansah, 2023; Furtak, 2017).

This dissertation also enhances the design of 3D assessment tasks by demonstrating the effectiveness of a collaborative framework that incorporates interdisciplinary expertise. The positive feedback from expert panels on the integration of the three dimensions in the designed products underscores the potential of human-AI collaboration to produce high-quality assessments. The iterative approach aligns with hybrid intelligence systems principles, emphasizing continuous improvement through mutual learning between human and machine agents (Luckin et al., 2016). This framework leverages the strengths of both human expertise and AI capabilities to create more effective and inclusive educational tools that support diverse learning outcomes (Stanford HAI, 2020).

For future research, this study suggests exploring the integration of feedback from a broader range of stakeholders, including educators, students, and policymakers, to enhance the development and implementation of AI-driven educational assessments. Additionally, involving students can provide valuable insights into the accessibility, inclusivity, and engagement of the designed tasks, ensuring they meet diverse learning needs (Baidoo-Anu & Owusu Ansah, 2023; Furtak, 2017). Future studies should also examine bringing experts into the interactive environment to achieve real collaboration, further enhancing the productive nature of expert feedback (Lai et al., 2021).

5.4 Major Themes

This dissertation builds on, reinforces, and extends the current literature by providing a comprehensive framework for human-AI collaboration in designing and refining educational assessments. It enhances the design of 3D assessment tasks by demonstrating the effectiveness of integrating interdisciplinary expertise and iterative feedback. The study also offers new insights into the capabilities and limitations of AI in educational contexts, highlighting the importance of human judgment and collaboration in creating high-quality educational tools. Table 5-1 summarizes the major themes found from this study and how they could contribute to the literature.

Table 5-1. Summary of themes and how they contribute literature

Domains	**Themes**	**Contribution to Literature**
Human-AI collaboration related themes	Explicit Guidance	Builds on existing literature emphasizing the need for detailed and clear instructions when working with AI (Kumar & Thakur, 2012; Spector & Muraida, 1993). Reinforces the importance of human-provided explicit guidance to improve AI outputs. Demonstrates how specific instructions enhance the AI's ability to generate detailed and coherent educational materials, such as Integrated Dimension Maps (IDMs) (Hwang et al., 2020; Pedró et al., 2019).
	Domain-specific information	Extends the literature on the critical role of domain-specific knowledge in AI training (Järvelä et al., 2022). Highlights how detailed content about DCIs, SEPs, and CCCs improves the AI's ability to generate relevant and accurate outputs. Shows the necessity of providing comprehensive and specific information for designing and monitoring the AI's outputs.
	Role of human experts	Reinforces the literature on the indispensable role of human expertise in evaluating and refining AI outputs (Dellermann et al., 2021). Emphasizes how human experts provide critical insights and feedback, ensuring that AI-generated content aligns with educational standards and goals. Highlights the nuanced understanding of educational contexts and pedagogical strategies human experts bring, which AI currently lacks (Fenwick, 2010).
	Iterative refinement	Extends the concept of iterative improvement in AI training (Kim et al., 2022). Demonstrates the importance of multiple rounds of feedback and adjustments to enhance the quality of AI-generated outputs. Highlights the role of reflective practice in identifying gaps and areas for improvement, ensuring continuous enhancement of educational content (Holzinger, 2016; Gregor, 2001).
	AI-Human collaboration	Builds on the literature that emphasizes the synergistic relationship between AI and human expertise (Bearman & Ajjawi, 2022; Dellermann et al., 2021;). Demonstrates how the collaboration between AI's processing capabilities and human expertise results in high-quality, pedagogically sound educational tools. Highlights the iterative feedback loops and continuous refinement that characterize effective AI-human collaboration (Knox, 2020; Seldon & Abidoye, 2018).

Table 5-1 (cont'd)

Assessment design related themes	Ensure language clarity and age-appropriateness	Reinforces the importance of age-appropriate language in educational materials (Lee, Quinn, & Valdés, 2013). Highlights specific examples of simplifying language to better align with students' reading levels, ensuring accessibility and comprehension.
	Enhance engagement and inclusion	Extends literature on designing engaging and inclusive assessments (Baidoo-Anu & Owusu Ansah, 2023; Furtak, 2017). Emphasizes the need for culturally relevant and relatable scenarios to boost student engagement, suggesting framing tasks as stories or hands-on demonstrations.
	Provide adequate scaffolding and supportive visuals	Reinforces the role of scaffolding in supporting student learning (Greene & Azevedo, 2007). Emphasizes the need for clear instructions, visual aids, and step-by-step guidance to help students understand complex concepts, particularly in modeling tasks.
	Coverage and integration of DCIs, SEPs, and CCCs	Builds on the need for comprehensive coverage of DCIs, SEPs, and CCCs (NGSS Lead States, 2013). Highlights the importance of ensuring that all requisite proficiencies are adequately covered, avoiding gaps and overreach.
	Consistency and coherence in terminology and information	Reinforces the need for consistency in educational terminology (Lee, Quinn, & Valdés, 2013). Emphasizes the importance of coherent information flow within tasks to facilitate efficient information processing and enhance student understanding.
	Enhance clarity and precision of task guidelines	Emphasizes the importance of clear and precise instructions in assessment tasks (Heritage, 2010). Highlights strategies to eliminate ambiguity, ensuring that students understand task requirements and can effectively demonstrate their understanding.
	Ensure cultural sensitivity and accessibility	Extends the literature on cultural inclusivity in education (Darling-Hammond & Snyder, 2000). Highlights the importance of designing tasks that reflect diverse cultural backgrounds and experiences, ensuring accessibility and engagement for all students.

5.5 Limitations and Future Research Directions

This section highlights the limitations encountered during the research process and outlines the directions for future research. Acknowledging these limitations is essential for interpreting the dissertation's findings and understanding the scope within which these conclusions are drawn.

5.5.1 Operator Bias

One of the limitations of this study is the human operator. I analyzed and synthesized the expert panels' feedback and crafted refinement prompts to work with machines to refine those assessments. As mentioned in the positionality section, although I have a chemistry degree, which gives me robust science content knowledge, especially in physical sciences areas, I also have extensive knowledge-in-use assessment design experience and teaching experience in an Asian country. However, I lack sufficient understanding of life science domains that may affect my judgment of the life science assessments and products. This could be supported by the reviewers' feedback on designed products for the two domains, life sciences and physical sciences. The feedback on physical sciences is generally higher compared to the life science products, especially concerning the science ideas and phenomena parts. Born and growing up in a different cultural background causes me to not have enough understanding of Western country culture, which may affect my judgment on the feedback regarding cultural sensitivity and inclusivity. Similarly, as an English language learner, it may affect my ability to judge the language level and appropriateness of the design. With that, I want to claim that this could be one of the limitations of this study. As the monitor or operator, I may bring biased judgment into the decision-making process.

To mitigate such biases in future work, it is crucial to involve a broader array of experts with diverse backgrounds and expertise to ensure a more balanced and comprehensive evaluation process, particularly when addressing complex or contentious feedback.

5.5.2 Lack of Student Input and Practical Implementation

The other limitation of this study is the absence of student input in the feedback process and the lack of real classroom implementation of the designed assessments. While AI excels at processing extensive information swiftly, it does not possess the flexible thinking and empathy that human judgment provides (Cope & Kalantzis, 2020). This study did not incorporate the perspectives of those most affected by the assessments—students. Their insights are crucial for understanding the accessibility, inclusivity, and engagement of the designed tasks. Future research should focus on how these assessments can be applied in classroom settings to obtain concrete evidence of their effectiveness and practicality. Engaging students directly can provide valuable feedback on the assessments' relevance and instructional validity, ensuring they are adequately tailored to diverse educational needs (Baidoo-Anu & Owusu Ansah, 2023; Furtak, 2017).

5.5.3 Future Research Directions

Moving forward, the integration of feedback from a broader range of stakeholders, including educators, students, and policymakers, will be vital for enhancing the development and implementation of AI-driven educational assessments. Including these voices can provide richer insights into the accessibility, inclusivity, and engagement of the designed tasks, making them more relevant and effective. Additionally, future studies should aim to bring these diverse experts into a collaborative environment to facilitate real-time adjustments and refinements, enhancing the productive nature of expert feedback. Testing these assessments in actual classroom settings will also be critical to evaluate their instructional validity and impact on student learning, bridging the gap between assessment design and practical educational application. This will ensure that all voices in education are heard, reflecting the systematic and dynamic nature of educational assessment design.

BIBLIOGRAPHY

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock, J. B., Lin, Q., Thomas, C. M., Penuel, W. R., & Welch, M. M. (2018). Designing educational systems to support enactment of the next generation science standards. Journal of Research in Science Teaching, 55(7), 1026–1052.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*(9), 1175.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62.

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences*, *4*(11), 417-423.

Bakker, A., & Gravemeijer, K. P. (2004). Learning to reason about distribution. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147-168). Dordrecht: Springer Netherlands.

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science education*, *94*(6), 1008-1026.

Barab, S., & Squire, K.(2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*(1),1–14.

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160-1173.

Bellman, R. E. (1978). An introduction to artificial intelligence: Can computers think?. Boyd & Fraser

Bertenthal, M. W., & Wilson, M. R. (Eds.). (2006). *Systems for state science assessment*. National Academies Press.

Berg, G. A. (2000). Human-computer interaction (HCI) in educational environments: Implications of understanding computers as media. *Journal of Educational Multimedia and Hypermedia*, *9*(4), 347-368.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, *24*(1), 61-100.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *1989*, *18*(1), 32-42.

Brown, J. S., & Duguid, P. (1993). Stolen knowledge. *Educational technology*, *33*(3), 10-15.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports*. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, policy & practice*, *5*(1), 7-74.

Bloom, B. S. (1968). *Toward a theory of testing which includes measurement-evaluation-assessment* (No. 9). University of California.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Ascd.

Card, S. K., Moran, T. P., & Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: Lawrence Erlbaum.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Coghlan, D., & Brannick, T. (2014). Doing Action Research in Your Own Organization. SAGE.

Collins, A., Joseph, D.,& Bielaczyc, K.(2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences,13*(1), 15–42.

Chi, M. T., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., ... & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive science*, *42*(6), 1777-1832.

Chiu, T. K. (2021). A holistic approach to the design of artificial intelligence (AI) education for K-12 schools. *TechTrends*, *65*(5), 796-807.

Chiu, T. K., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*, 1-17.

Coffey, J., Black, P., & Atkin, J. M. (Eds.). (2001). *Classroom assessment and the national science education standards*. National Academies Press.

Cope, B., & Kalantzis, M. (2020). *Making sense: Reference, agency, and structure in a grammar of multimodal meaning*. Cambridge University Press.

Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational philosophy and theory*, *53*(12), 1229-1245.

Dai, Y., Liu, A., Qin, J., Guo, Y., Jong, M. S. Y., Chai, C. S., & Lin, Z. (2023). Collaborative construction of artificial intelligence curriculum in primary schools. *Journal of engineering education*, *112*(1), 23-42.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied developmental science*, *24*(2), 97-140.

De Cremer, D., & Narayanan, D. (2023). How AI tools can—and cannot—help organizations become more ethical. *Frontiers in Artificial Intelligence, 6*, 109372.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2021). The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354*.

DiCerbo, K. (2020). Assessment for learning with diverse learners in a digital world. *Educational Measurement: Issues and Practice, 39*(3), 90-93.

Esposito, A. G., & Bauer, P. J. (2017). Going beyond the lesson: Self-generating new factual knowledge in the classroom. *Journal of experimental child psychology*, *153*, 110-125.

Fenwick, T. J. (2010). (un) Doing standards in education with actor-network theory. *Journal of Education Policy*, *25*(2), 117-133.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277-304.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education, 101*(5), 854-867.

Furtak, E. M., & Lee, O. (2023). Equity and Justice in Classroom Assessment of STEM Learning. *Classroom-Based STEM Assessment*, 69.

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of educational research*, *77*(3), 334-372.

Greengard, S. (2022). ChatGPT: understanding the ChatGPT AI . *eWeek. Archived from the original on January*, *19*, 2023.

Gregor, S. (2001). Explanations from knowledge-based systems and cooperative problem solving: an empirical study. *International Journal of Human-Computer Studies*, *54*(1), 81-105.

Glaser, R., Chudowsky, N., & Pellegrino, J. W. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Ha, M., & Nehm, R. H. (2016). The impact of misspelled words on automated computer scoring: A case study of scientific explanations. *Journal of Science Education and Technology*, *25*, 358-374.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, *38*(2), 53-67.

Harris, C. J., Krajcik, J. S., & Pellegrino, J. W. (Eds.). (2024). *Creating and using instructionally supportive assessments in NGSS classrooms*. NSTA Press, National Science Teaching Association.

Hatano, G., & Oura, Y. (2003). Commentary: Reconceptualizing school learning using insight from expertise research. *Educational researcher*, *32*(8), 26-29.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Life Sciences Education*, *11*(3), 283-293.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Heritage, M. (2010). Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity?. *Council of Chief State School Officers*.

He, P., Chen, I.-C., Touitou, I., Bartz, K., Schneider, B., & Krajcik, J. (2023). Predicting student science achievement using post-unit assessment performances in a coherent high school chemistry project-based learning system. Journal of Research in Science Teaching, 60(4), 724–760. https://doi.org/10.1002/tea.21815

Herrington, J., & Oliver, R. (1995). Critical characteristics of situated learning: Implications for the instructional design of multimedia. *Learning with technology*, *10*.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational technology research and development*, *48*(3), 23-48.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 1-23.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. *Brain informatics*, *3*(2), 119-131.

Hutchins, E. (2000). Distributed cognition. *International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science*, *138*, 1-10.

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, *1*, 100001.

Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J., & Shimada, A. (2024). Artificial intelligence in education: Implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning*, 1-18.

Järvelä, S., Nguyen, A., Vuorenmaa, E., Malmberg, J., & Järvenoja, H. (2023). Predicting regulatory activities for socially shared regulation to optimize collaborative learning. *Computers in Human Behavior*, *144*, 107737.

Johnson, B. A., Coggburn, J. D., & Llorens, J. J. (2022). Artificial intelligence and public human resource management: Questions for research and practice. *Public Personnel Management, 51*(4), 538-562.

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines*, *27*, 575-590.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

Kamar, E. (2016, July). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *IJCAI* (pp. 4070-4073).

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, *98*(4), 674-704.

Kemmis, S., McTaggart, R., & Nixon, R. (2014). The Action Research Planner: Doing Critical Participatory Action Research. Springer.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). *Explainable artificial intelligence in education*. Computers and Education: Artificial Intelligence, 3, 1000

Kim, T. W., Jiang, L., Duhachek, A., Lee, H., & Garvey, A. (2022). Do you mind if I ask you a personal question? How AI service agents alter consumer self-disclosure. *Journal of Service Research*, *25*(4), 649-666.

Kim, J. (2023). Leading teachers' perspective on teacher-AI collaboration in education. *Education and Information Technologies*, 1-32.

Krajcik, J., Schneider, B., Miller, E. A., Chen, I. C., Bradford, L., Baker, Q., ... & Peek-Brown, D. (2023). Assessing the effect of project-based learning on science learning in elementary schools. *American Educational Research Journal*, *60*(1), 70-102.

Knox, J. (2020). Artificial intelligence and education in China. *Learning, Media and Technology*, *45*(3), 298-311.

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, *4*, 622364.

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973-1018.

Kulgemeyer, C., & Schecker, H. (2014). Research on educational standards in German science education— Towards a model of student competences. EURASIA Journal of Mathematics, Science and Technology Education, 10(4), 257–269.

Kumar, K., & Thakur, G. S. M. (2012). Advanced applications of neural networks and artificial intelligence: A review. *International journal of information technology and computer science*, *4*(6), 57-68.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge: Harvard University Press.

Lawler, R. W., & Rushby, N. (2013). An interview with R obert L awler. *British Journal of Educational Technology*, *44*(1), 20-30.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational researcher*, *42*(4), 223-233.

Lee, H. S., Gweon, G. H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology*, *30*(2), 168-192.

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science. *Journal of Research in Science Teaching*, *60*(6), 1385-1389.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233.

Li, T., Chen, I., Miller, E., Miller, C., Schneider, B., & Krajcik, J. (2024). The relationships between elementary students' knowledge-in-use performance and their science achievement. *Journal of Research in Science Teaching.* 1-61 https://doi.org/10.1002/tea.21900

Li, T., Liu, F., & Krajcik, J. (2023) Automatically assess elementary students' hand-drawn scientific models using deep learning of Artificial Intelligence. Proceedings of *the Annual meeting of the International Society of the Learning Sciences (ISLS)*.

Li, T., He, P., & Peng, L. (2024). Measuring high school student engagement in science learning: an adaptation and validation study. *International Journal of Science Education*, *46*(6), 524-547.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed. *An argument for AI in Education*, *18*.

Luckin, R., Clark, W., Avramides, K., Hunter, J., & Oliver, M. (2017). Using teacher inquiry to support technology-enhanced formative assessment: a review of the literature to inform a new method. *Interactive Learning Environments*, *25*(1), 85-97.

Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. MIT press.

Mensah, F. M., & Chen, J. L. (2022). Elementary multicultural science teacher education. In *International handbook of research on multicultural science education* (pp. 1-39). Cham: Springer International Publishing.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice*, *25*(4), 6-20.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2002). Design and analysis in task-based language assessment. Language Assessment, 19, 477-496. Also available as CSE Technical Report 579. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TR579.pdf [ECD perspective on designing task-based language assessments. Includes examples of Bayes nets for tasks that tap multiple aspects or knowledge and skill.]

Morrison, K. R. B. (2002) School Leadership and Complexity Theory. London: Routledge Falmer.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review, 56*(4), 3005-3054.

National Research Council. (2011). Assessing 21st century skills: Summary of a workshop. National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2019). *Science and engineering for grades 6-12: Investigation and design at the center*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2021). *Science and engineering in preschool through elementary grades: The brilliance of children and the strengths of educators.* National Academies Press.

National Research Council. (2014). Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

Nguyen, T., Novak, R., Xiao, L., & Lee, J. (2021). Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, *34*, 5186-5198.

Organization for Economic Cooperation and Development. (2019). PISA 2018 assessment and analytical framework. OECD Publishing.

Osborne, J., & Wertheim, J. (2019). Supporting Coherence Across a System of Assessment for NGSS.

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(8), em2307.

Pea, R. D. (1993). Practices of distributed intelligence and designs for education. *Distributed cognitions: Psychological and educational considerations*, *11*, 47-87.

Pea, R. D., & Kurland, D. M. (1987). On the cognitive effects of learning computer programming. In R. D. Pea & K. Sheingold (Eds.), *Mirrors of minds: Patterns of experience in educational computing* (Report No. SE 043 964, pp. 147-342). Ablex Publishing Corporation.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*(6130), 320-323.

Pellegrino J. W., Hilton M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, *20*(2), 65-77.

Penuel, W. R., Allen, A. R., Coburn, C. E., & Farrell, C. (2015). Conceptualizing research–practice partnerships as joint work at boundaries. *Journal of Education for Students Placed at Risk (JESPAR)*, *20*(1-2), 182-19

Penuel, B., & Smolek, T. J. (2019). Reconceptualizing Alignment for NGSS Assessments.

Penuel, W. R. (2019). Infrastructuring as a practice of design-based research for supporting and studying equitable implementation and sustainability of innovations. *Journal of the Learning Sciences*, *28*(4-5), 659-677.

Penuel, W. R., & Shepard, L. A. (2016). Social models of learning and assessment. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 146-173.

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., ... & Allen, A. R. (2017). How school and district leaders access, perceive, and use research. *AERA Open*, *3*(2), 2332858417705370.

People's Republic of China Ministry of Education. (2014). Opinions on deepening curriculum reform and implementing the fundamental tasks of Lide-Shuren. http://www.moe.gov.cn/srcsite/A26/jcj_kcjcgh/201404/t20140408_167226.html.

Penuel, B., & Smolek, T. J. (2019). Reconceptualizing Alignment for NGSS Assessments.

Reason, P., & Bradbury, H. (Eds.). (2008). The SAGE Handbook of Action Research: Participative Inquiry and Practice. SAGE.

Roberts, S. T. (2021). Your AI is a human. In T. S. Mullaney, B. Peters, M. Hicks, & K. Philip (Eds.), *Your computer is on fire* (Chapter 2). The MIT Press. https://doi.org/10.7551/mitpress/10993.003.0006

Rowlands, M. J. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests*. John Wiley & Sons.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(5), 369-393.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119-144.

Salomon, G. (Ed.). (1997). *Distributed cognitions: Psychological and educational considerations*. Cambridge University Press.

Sanusi, I. T., Oyelere, S. S., Vartiainen, H., Suhonen, J., & Tukiainen, M. (2023). A systematic review of teaching and learning machine learning in K-12 education. *Education and Information Technologies, 28*(5), 5967-5997.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Sanusi, I. T., Ayanwale, M. A., & Tolorunleke, A. E. (2024). Investigating pre-service teachers' artificial intelligence perception from the perspective of planned behavior theory. *Computers and Education: Artificial Intelligence*, *6*, 100202.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world: Using next generation science and engineering practices.* NSTA Press.

Schreiber, L. M., & Valle, B. E. (2013). Social constructivist teaching strategies in the small group classroom. *Small Group Research*, *44*(4), 395-411.

Seldon, A., & Abidoye, O. (2018). *The fourth education revolution*. Legend Press Ltd.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Selwyn, N. (2016). Minding our language: why education and technology is full of bullshit… and what might be done about it. *Learning, Media and Technology*, *41*(3), 437-443.

Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, *19*(4), 639-659.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.

Siemens, G. (2005). Connectivism: Learning as network-creation. *ASTD Learning News*, *10*(1), 1-28.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). FOCUS ARTICLE: implications of research on children's learning for standards and assessment: a proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective*, *4*(1-2), 1-98.

Spector, J. M., Polson, M. C., & Muraida, D. J. (Eds.). (1993). Automating instructional design: Concepts and issues.

Spiro, R. J., Bruce, B. C., & Brewer, W. F. (Eds.). (2017). *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education* (Vol. 11). Routledge. Spiro, R. J., Bruce, B. C., & Brewer, W. F. (Eds.). (2017). *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education* (Vol. 11). Routledge.

Spiro, R. J., Feltovich, P. J., Gaunt, A., Hu, Y., Klautke, H., Cheng, C., ... & Ward, P. (2018). Cognitive Flexibility Theory and the accelerated development of adaptive readiness and adaptive response to novelty.

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual review of psychology*, *49*(1), 479-502.

Sternberg, R. J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of personality and social psychology*, *49*(3), 607.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257-285.

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 183-200.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21-34.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

Spiro, R., Feltovich, P., Jacobson, M., & Coulson, R. (1992). Knowledge representation, content specification, and the development of skill in situation-specific knowledge assembly: Some constructivist issues as they relate to cognitive flexibility theory and hypertext. In T. Duffy & D. Jonassen (Eds.), *Constructivism and the technology of instruction (pp. 121–128*). Hillsdale, NJ: Lawrence Erlbaum.

Stiggins, R. (2014). *Revolutionize assessment: Empower students, inspire learning*. Corwin Press.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication, 25*(1), 74-88.

Tan, Q., Soler, R., Pivot, F., Zhang, X., & Wang, H. (2020). Introspection of Personalized and Adaptive Learning. In *INTED2020 Proceedings* (pp. 8054-8061). IATED.

Tegmark, M. (2018). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433-460.

Vygotsky, L.S. (1978). Mind in society. Cambridge, MA: Harvard University Press.

Ward, P., Gore, J., Hutton, R., Conway, G. E., & Hoffman, R. R. (2018). Adaptive skill as the conditio sine qua non of expertise. *Journal of applied research in memory and cognition*, *7*(1), 35-50.

Wenger, E. (1998). Communities of practice: Learning as a social system. *Systems thinker*, *9*(5), 2-3.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Westbroek, H. B., van Rens, L., van den Berg, E., & Janssen, F. (2020). A practical approach to assessment for learning and differentiated instruction*. International Journal of Science Education, 42(6)*, 955-976.

Williams, P. (2023). AI, Analytics and a New Assessment Model for Universities. *Education Sciences*, *13*(10), 1040.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance*. Jossey-Bass Publishers, 350 Sansome Street, San Francisco, CA 94104.

Xia, Q., Chiu, T. K., Lee, M., Sanusi, I. T., Dai, Y., & Chai, C. S. (2022). A self-determination theory (SDT) design approach for inclusive and diverse artificial intelligence (AI) education. *Computers & Education*, *189*, 104582.

Yang, S., Hu, L., Yu, L., Ali, M. A., & Wang, D. (2024). Human-ai interactions in the communication era: Autophagy makes large models achieving local optima. *arXiv preprint arXiv:2402.11271*.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668.

Zhu, M., Lee, H. S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648-1668.

APPENDIX

I used ChatGPT-4 to edit the language and grammar of my dissertation writing. However, ChatGPT-4 did not write or alter the original meanings of the written texts.