Ding *et al. Discip Interdscip Sci Educ Res* (2023) 5:12
https://doi.org/10.1186/s43031-023-00080-7

Disciplinary and Interdisciplinary Science Education Research

RESEARCH Open Access

Refinement of an instrument measuring science teachers' knowledge of language through mixed method

Chenchen Ding[1,2*], Catherine Lammert[3], Gavin W. Fulmer[1,4], Brian Hand[1] and Jee K. Suh[5]

Abstract

Teachers must know how to use language to support students in knowledge generation environments that align to the Next Generation Science Standards. To measure this knowledge, this study refines a survey on teachers' knowledge of language as an epistemic tool. Rasch modelling was used to examine 15 items' fit statistics and the functioning of a previously-designed questionnaire's response categories. Cronbach's alpha reliability was also examined. Additionally, interviews were used to investigate teachers' interpretations of each item to identify ambiguous items. The results indicated that three ambiguous items were deleted based on qualitative data and three more items were deleted because of negative correlation and mismatched fit statistics. Finally, we present a revised language questionnaire with nine items and acceptable correlation and good fit statistics, with utility for science education researchers and teacher educators. This research contributes a revised questionnaire to measure teachers' knowledge of language that could inform professional development efforts. This research also describes instrument refinement processes that could be applied elsewhere.

Keywords Language, Instrument refinement, Rating scale model, Epistemic tool

Introduction

The Next Generation Science Standards (NGSS Lead States, 2013) emphasize classroom environments where students can generate new scientific ideas rather than merely replicate existing ones. In teaching informed by NGSS, students are positioned as learners who can make decisions and gain knowledge using scientific tools and methods (Campbell & Oh, 2015; Elgin, 2013; Stroupe et al., 2018). An environment where students can generate and validate knowledge is known as a 'Knowledge Generation Environment' (Fulmer, et al., 2021). Constructing classrooms as Knowledge Generation Environments benefits both students and teachers. Students have opportunities to generate and negotiate ideas, which deepens their understanding of scientific concepts (Bae et al., 2021). Additionally, making students' internal dialogs available to teachers provides a window into students' thinking, which in turn helps them continue to construct such environments (Gutierrez et al., 1995). Therefore, shifting toward Knowledge Generation Environments is essential for teachers to enact NGSS-aligned teaching.

It has long been established that language plays an essential role in the science classroom (Norris & Phillips, 2003), with the particular roles of language highlighted in Knowledge Generation Environments (Prain & Hand., 1996, 2016a). In science classrooms, students must interpret language and produce language—as text, graphics,

*Correspondence:
Chenchen Ding
chenchen-ding@uiowa.edu
[1] Teaching and Learning Department, University of Iowa, Iowa, USA
[2] Department of Curriculum and Instruction, Zhejiang Normal University, Jinhua, China
[3] Teacher Education Department, Texas Tech University, Lubbock, USA
[4] Northwest Evaluation Association (NWEA), Portland, USA
[5] Department of Curriculum and Instruction, University of Alabama, Tuscaloosa, USA

and spoken dialog—to engage with and express their ideas (National Research Council, 2012). Learners use language not only to communicate ideas (Duschl et al., 2007; Norris & Phillips, 2003), but also to create new ideas for themselves (Pinker, 2010; Wang et al., 2010). Language enables higher-order cognition (Pinker, 2010) and allows us to connect new knowledge with existing knowledge to improve our understanding (Wang et al., 2010). As Norris and Phillips (2003) have emphasized, there is no science without language.

Teachers' knowledge of language as an epistemic tool also underpins their ability to create Knowledge Generation Environments (Fulmer, 2021). In Knowledge Generation Environments, teachers use language to support students as they create their own knowledge aligned to disciplinary knowledge and validate these knowledge claims of science through both private and public negotiation. Teacher knowledge of language as an epistemic tool relates to both the pedagogical knowledge of how to use language as learning tool (Aguirre-Muñoz & Pando, 2021) and knowledge of how using language promotes a learner building understanding of the concepts (Grangeat & Hudson, 2015). Here, pedagogical knowledge encompasses teaching methods and instructional strategies for using language to drive learning (Aguirre-Muñoz & Pando, 2021). The epistemological perspective of language as a tool will shape how they will utilize language pedagogically, that is, if teachers believe language is about learning the correct language of science, then pedagogically the emphasis will be on vocab and not on using it epistemologically. Pedagogical knowledge for using language should be driven by the epistemological perspective that language is an epistemic tool that is necessary for students to build their own understanding of science.

Prain & Hand (2016a) have argued that language is an epistemic tool because through language students generate ideas and connect new knowledge with prior knowledge. Fulmer (2021) developed a questionnaire to measure teachers' knowledge of language as an epistemic tool, starting with a literature review to construct domains of understanding language as epistemic tool, creating item pools for expert review and revision, and finally, piloting the initial version of language questionnaire. However, subsequent applications of the language questionnaire to measure teachers' knowledge of language as an epistemic tool shows that some items do not fit the proposed measurement model (Fig. 1). This heightens the need to revisit the definitions of teachers' knowledge of language as an epistemic tool by studying the instrument functioning and examining other evidence from participating teachers. By further analyzing the existing instrument and comparing its findings with in-depth qualitative findings, the present study will provide a clearer picture of how the construct of language as an epistemic tool could be measured and interpreted.

Fulmer (2021) provided support for the internal aspects of validity for the language questionnaire, such as content, substance, and structure, through both theoretical and empirical evidence. For content validity, they conducted a domain analysis and sought outside experts' reviews. For substantive validity, they consulted experienced teachers on their thinking about the concepts addressed by the questionnaire. For structural validity, the responses were examined using the Rasch model to ensure the response patterns were predictable based on respondents' ability (i.e., that respondents with higher ability would endorse more difficult items and vice versa).

However, there are some weaknesses of the validation process for Fulmer (2021) that can be addressed through

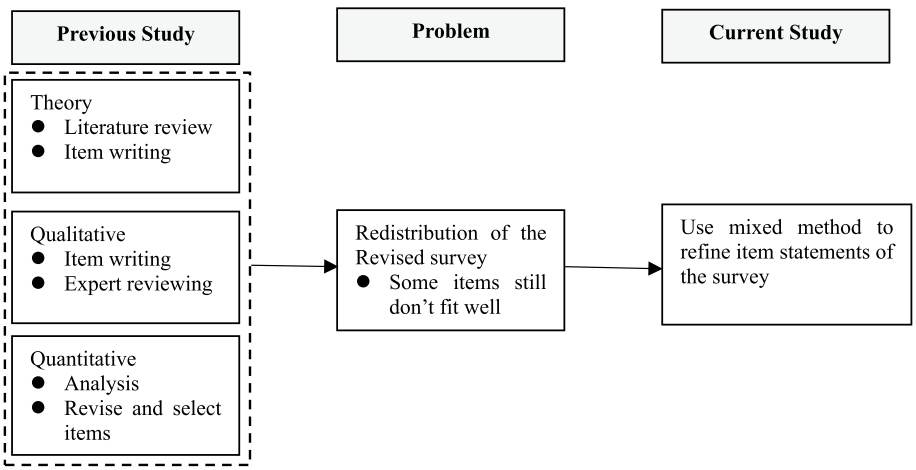

Fig. 1 Diagram of connection between current study and previous research

further study. First, even though content validity with expert judgment provided evidence about representative items to the content domain (Fulmer, 2021), we argue that interviews with teachers can provide additional information about their interpretations of the items (Singh & Rosengrant, 2003; Treagust, 1988). Furthermore, interviews with respondents could target specific topics of items' content and the construct that extend or contradict the feedback from expert review (Adams & Wieman, 2011). During interviews, respondents are free to speak openly about their interpretation of statements, which gives researchers insights to clarify the statements (Peterson et al., 2017). Moreover, the current work allows us to take more advanced steps in the applied Rasch measurement analysis to better understand how teachers respond to the items and provide insight for improving future applications of the instrument.

The paper builds on extant frameworks of mixed methods instrument development in order to refine the language questionnaire from two aspects: item statement and response categories. The key research questions are:

1. What themes emerged from interview data about teachers' interpretation of items in the language questionnaire? How do those themes inform the content and dimensionality of the language questionnaire?
2. What evidence of reliability and validity could be gained from applying Rasch measurement models to the quantitative data on the language questionnaire?
3. What refinement should be made for items on the language questionnaire based on the combined qualitative and quantitative analysis?

Literature

First, we review the development of the construct of language as an epistemic tool in learning science. Then we review the role of interview in developing questionnaires and one quantitative method to refine Likert scale.

The construct of language as an epistemic tool

In developing an instrument to measure teachers' knowledge of language as an epistemic tool, we have identified four language domains, based on the existing literature: language is essential, language is constitutive, language involves processes and products, and language includes multiple modes of representation (Fulmer et al., 2021).

Language is essential

The domain stating that language is essential stems from the view that, fundamentally, humans cannot think without representing ideas through some representational mode (Pinker, 2010; Vygotsky, 1978). One cannot imagine a teacher successfully teaching a science lesson without using any kind of language, including the everyday language students use outside of the classroom (e.g., casual phrasing and examples from daily life), as well as the domain-specific vocabulary, syntax and text structures unique to the sciences (e.g., formulas in chemistry). Everyday language may seem to be imprecise and unscientific compared to scientific terminology (Brock, 2015), but it is important for students' thinking and learning because it allows them to make connections between science concepts and their background knowledge (Warren et al., 2001). Prain & Hand, (1996) have demonstrated that prematurely forcing students' language into correct scientific forms negatively impacts learning.

Language is constitutive

In stating that language is constitutive, we suggest that, through the act of representing ideas through language, new knowledge can be built. Consider the case of a religious officiant saying, 'I now pronounce you husband and wife'. Through this language act, a new legally binding relationship has been created; thus, language does not just represent existing knowledge, but can be used for the act of novel creation. This domain emphasizes the learning process and the role of language as an epistemic tool (Prain & Hand., 2016a, Hand, Cavagnettto, et al., 2016, Hand, Norton-Meyer., 2016). Particularly in Knowledge Generation Environments, learning science is not just about memorizing concepts from teachers or books, but about negotiating meaning between new experiences and prior knowledge (Gee, 2000). There is no single best way to construct the ideas of science concepts, because each individual has unique prior knowledge (Anderson, 1992).

Language involves processes and products

Using language is about a process, not only the language product. Calkins (1994), an early leader in the process-writing movement, coined the phrase 'teach the writer, not the writing'. This adage has remained in regular use by teachers who are dedicated to the idea that, in the process of learning, students generate ideas and share those ideas with peers or teachers through written or spoken language (Norris & Phillips, 2003). This process may or may not result in improvement in students' final written products even as it helps them clarify their ideas. Like Calkins, we argue through this domain that the learning process is more important than what eventually ends up on the page or screen (Hand et al., 2001; Galbraith, 1999; Pelger & Nilsson, 2016). When the learning process is emphasized, students' understanding of science is enhanced (Prain & Hand., 2016b).

Language includes multiple modes of representation

Multiple modes of representation (MMR) not only include language in the form of written text, but also includes language in forms of speaking, pictures, diagrams, graphs, equations and tables to convey understanding or ideas of scientific concepts (Ainsworth & VanLabeke, 2004; Yore & Hand, 2010). MMR is an interplay of signs, interpretations and referents to convey meanings through an interpretation process, which helps students understand each other's ideas in communications with peers (Tang & Moje, 2010). Students will have a deeper understanding of science concepts if they can use MMR (Cikmaz et al., 2021). For example, when students include MMR in their argument writing in organic chemistry laboratory courses, they created more cohesive arguments in their reports than students who don't use MMR (Hand & Choi., 2010). Kohl and colleagues (2007) examined how multiple representations, such as force or motion diagrams of objects, affect students' learning. They found that college students who make extensive use of multiple representations to solve free-body problems have better performance than students who don't (Kohl et al., 2007).

The complexity of language, including its four domains, requires equally complex tools with which to measure it. Accordingly, we turn to discuss one method that could aid in refining a questionnaire to measure teachers' knowledge of this important construct.

Interviews for item interpretation

Interviews are widely used to refine instruments (Knafl et al., 2007), because they can provide evidence that the questions are able to measure what they intend to measure without misguiding test-takers (Chatzidamianos et al., 2021). In interviews, researchers can use structured or semi-structured interview protocols to probe participants' thinking processes, which provides additional information to survey data (Romine et al., 2017). Thus, interviews are one source of evidence of item validity (Padilla & Benítez, 2014). We apply this approach to further validate the existing questionnaire on teachers' knowledge of language as an epistemic tool.

Interviews are a common qualitative data source. When conducted for the purpose of instrument refinement, interview protocols are aligned with norms. In this context, rather than using the items to measure respondents' knowledge of the construct, respondents' interpretations of items are accessed (Knafl et al., 2007; Ryan et al., 2012). Brown et al. (2018) used interviews to refine a questionnaire by examining descriptions of terms, the difficulty of understanding, and ambiguous concepts and synonyms. This information is useful when revising items. In interviewing, it is not important that a large sample is interviewed; more important is that each interviewee is provided with each item and given extensive time and open-ended prompts (e.g., 'Say more about how you view that.') to elaborate on their thinking. For example, Ford et al. (2019) used interviews to improve content and face validity by interviewing just five participants. Based on those interviews, they found that most of their items had internal consistency and were easy to understand. Therefore, interviews can be a useful data source in mixed-methods approaches to the refinement of a questionnaire about teachers' knowledge of language.

Rating scale model

The Rating Scale Model (RSM; Andrich, 1988) is one member of the Rasch family of models. The RSM is a probabilistic model to estimate an unobserved construct by comparing observed response patterns in polytomous data to the expected response pattern according to the strict Rasch model (Lamprianou, 2019; Liu, 2020). The RSM assumes the discriminations are the same across items and calibrates item difficulty and person ability on the same scale (Bond & Fox, 2015). Difficulty estimates for polytomous responses are represented as thresholds, which are the location on the scale where the probability of a respondent endorsing two adjacent categories is equal. Person ability indicates the extent to which the person has a greater level of the measured trait, whether that is knowledge, skill, or an attribute. The higher the threshold on the latent trait location is, the greater level of person ability is needed to endorse it.

The main characteristics of RSM is that all items have the same threshold structure increasing in line with unique difficulty for each item (DiStefano & Morgan, 2010). The same threshold structure means that the latent trait intervals between two adjacent thresholds are the same across all items (Bond & Fox, 2015). This is suitable for items where there is an empirical or a theoretical rationale for assuming all the items have the same response structure (Lamprianou, 2019). The questionnaire in this study is a Likert-type scale, which assumes that the thresholds of each item are ordered, so RSM was applied for the data analysis. Even though not all response categories were chosen by participants by examining the response frequency table (Table 2), thresholds for each item should be ordered. The difference is that items have different numbers of thresholds. There will be less than four thresholds if not all five response options were chosen. In this study, RSM with TAM package was applied in the R statistical environment (Ihaka & Gentleman, 1996).

Methods

The current study focuses on instrument refinement rather than developing a new instrument, with both qualitative and quantitative methods being equally important. So, we adapt an exploratory sequential design (Creswell & Plano Clark, 2017) with three main steps: researchers begin with domain analysis with a qualitative phase, then follow that with a quantitative analysis for item statements and operation of response categories, and finally, integrate the results to inform the instrument refinement, as shown in Fig. 2. According to phases in Fig. 2, research question 1 was answered by phase one, research question 2 was answered by phases two, and research question 3 was answered by phase three. The qualitative phase provides evidence of dimensionality, based on which we chose the quantitative method to gain evidence for reliability and validity.

Phase one: qualitative analysis

The previous version of the language questionnaire (Fulmer, 2021) included 15 items and was claimed to have four sub-domains for the construct, language as an epistemic tool. Researchers tried to figure out how the four claimed sub-domains were represented in each item in the questionnaire based on the teacher's interpretation in the interview. In the interview, teachers interpreted each item and elaborated their understanding.

Method

To investigate teachers' interpretations of items, we developed a semi-structured interview protocol (Rubin & Rubin, 2011) with three questions that applied to each questionnaire item: 1) What does this mean to you? 2) Is anything unclear? 3) Did you have any questions in your mind as you read this? Interviews were concluded by asking participants if there were aspects of language use in science that were not represented by any items. The first author, who had no authority to evaluate the teachers or influence their performance review in their school, conducted all interviews via Zoom.

Data collection

Participants in interviews were selected with convenience sampling (Etikan & Alkassim, 2016), a sampling method that is appropriate for interviews when instrument refinement is the goal (Ford et al., 2019). We had email addresses of a list of science teachers from kindergarten to grade 7 who actively engaged in previous professional development workshops. Recruitment emails were sent to ask for volunteers to complete a half-hour interview about their interpretation of items in the language questionnaire. Four white female teachers volunteered for the interview. They were two Grade 2 teachers (Kelly and Hedy), one Grade 4 teacher (Ran), and one Grade 5 teacher (Bella). All names are pseudonyms.

Analysis

After being transcribed, interview data from one teacher was coded by a first round of structural coding process conducted separately by two researchers (Rubin & Rubin, 2011), then proceeded to code the remaining teachers sequentially. We identified that data saturation (Lowe et al., 2018; Saldaña, 2015) had been reached by coding the fourth interview as it did not contain any new codes not raised in interviews one through three; accordingly, we did not solicit additional interview participants. Structural coding frames interview data with conceptual phrases representing topics of related research (Saldaña,

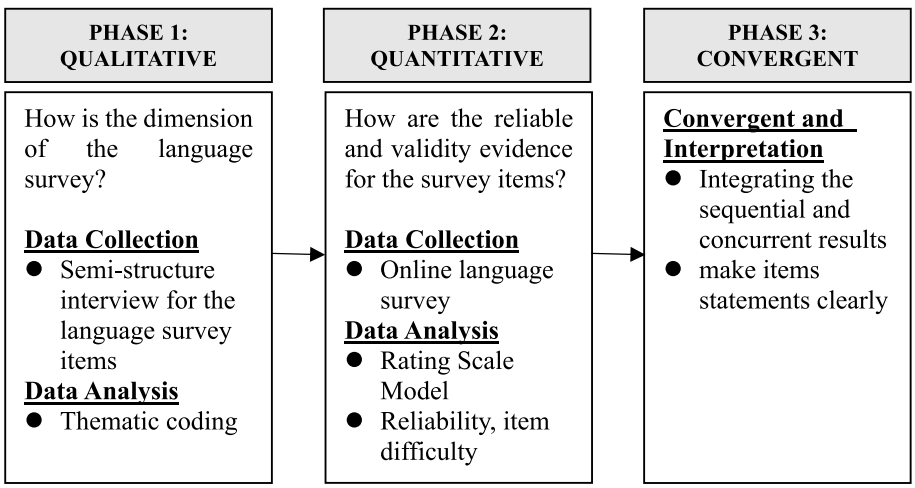

Fig. 2 Diagram of exploratory sequential design

2015). During this analysis, we did not evaluate teachers' level of knowledge of language as an epistemic tool; instead, we identified what domains of language as an epistemic tool they mentioned in their interpretation of each item. There were four domains as developed in the language questionnaire: language is essential, language is constitutive, language involves process and product, and language includes multiple modes of representation. In addition to the four domains of language, other topics related to science learning were also coded, which enables us to find new ideas and emergent themes related to these items. During first-round coding, the first and the second authors both independently identified a list of potential themes for each item. We negotiated differently coded items until a consensus was reached. Then, during the second-round coding, two researchers worked together to recode the first interview and to code all remaining interviews using the consensus codes based on which of the domains the teachers described.

Phase two: quantitative analysis

The language questionnaire was used to collect quantitative data, which were used for reliability and validity evidence. The questionnaire has 15 items with Likert-scale responses ranging from "strongly disagree" to "strongly agree" to numerical values from one to five, which is used to measure in- service teachers' knowledge of language as an epistemic tool in science teaching and learning (Fulmer et al., 2021).

Data collection

We distributed the questionnaire through online platform Qualtrics by email to in-service elementary science teachers who had attended the professional development workshops in the summer of 2020. The workshops occurred over six days each summer and four follow-up, half-day sessions during the academic school year, emphasized the role of epistemic tools, such as language, in creating Knowledge Generation Environments. There were 146 participating teachers from the Midwest and Southeast U.S., of which a total of 126 had no missing data and were retained for analysis. The participants in this study were overwhelmingly white and female. There were three male science teachers out of 126 teachers. The grade level of those participants ranged from K to 7. These teachers had experience ranging from 1 to 32 years in the classroom; taken together, they had 14 years of experience on average with SD as 9 years.

Data analysis

Based on the analysis of the dimensionality of language as a construct to measure teachers' understanding of language as an epistemic tool, the unidimensionality of the items in the questionnaire is corroborated. Then the RSM is used for quantitative analysis in order to provide evidence about the reliability of the items and fit statistics. First, item-total correlations were calculated for each item. Then, items fit statistics were estimated for item selection, such as infit t and outfit t. Fit statistics indicate how well the expected response pattern predicted by the model matches the observed responses. Infit t (the t-standardized value of the infit mean-square) and outfit t (the t-standardized value of the infit mean square) were used as item fit indices. Both infit and outfit t values can be either positive or negative, with positive values indicating that the observed response pattern has more variation and with negative values indicating that observed response pattern has less variation (Bond & Fox, 2015). Smith (2002) suggested that the mean-square value of infit and outfit (infit and outfit MNSQ) should not be outside the acceptable range for productive measurement (0.50 ~ 1.50). Meanwhile, the acceptable t values of outfit and infit ranged from − 2.0 to + 2.0 (Linacre, 2002). We used the TAM package with R language to run RSM for polytomous item responses, because this allows each item to have its own threshold pattern, to handle missing response categories (Robitzsch et al., 2020). Following the default from TAM, the Rasch estimates are constrained so that the average of the person ability estimates is zero. In the analysis, the input data matrix with item responses were coded as 0, 1, 2, 3, 4 for five-level Likert scale from 'Strongly disagree' to 'Strongly agree', and four backward-worded items (LQ20R, LQ22R, LQ26 and LQ100) were coded in the reverse.

Phase three: integration

Decisions and revisions of the questionnaire were made by combing analysis from qualitative and quantitative data.

Results

Qualitative results from the analysis of interview data and quantitative results from the analysis of questionnaire data are reported with the procedure of refining the questionnaire.

Results from qualitative analysis

The qualitative results are organized beginning with evidence of unidimensionality from systemic coding of interview data, followed by ambiguous interpretations of three items.

Unidimensionality of the language questionnaire

The analysis of interviews indicated that teachers often mentioned that the items made them think about the

domains on which they were based, but they also raised additional topics that related to science learning in general that went beyond the four domains of language as an epistemic tool. For example, even though teachers acknowledged that language is an epistemic tool that helps students learn, they described a multitude of other ways of learning science. Three teachers mentioned that students could learn science by doing, observing, or experiencing science. Here are some quotations:

> *Kelly: I would like to say, and experiencing activities about it, but that's probably not where you guys are going with this study.*
> *Hedy: They [Students] have a deeper knowledge of science by doing it and you know experiencing it.*
> *Ran: Something about either experiencing or observing … Because that's true they find out about hearing, reading, and writing about it, but they also can learn about it by experiencing it.*

Teachers also emphasized individual, private ways of learning, such as experiencing science and observing phenomena related to personal perceptions of nature. Aspects of language as an epistemic tool for each item were outlined in Table 1 and are explained in the following paragraphs.

Even though there were four domains in the theoretical framework of the original language questionnaire (i.e., language is essential, language is constitutive, language involves process and product, and language includes MMR; Fulmer et al., 2021), this analysis suggests that language is a unidimensional construct with four interwoven domains. As can be seen in our qualitative findings in Table 1, teachers' responses to nearly all items involved at least two theoretical domains. When this was discussed with respondents in interviews, it was clear that the four theoretical domains interweave together and cannot be separated exclusively, which reiterated the assertion by Fulmer (2021) that the subdomains are interrelated but also strongly suggests that they could be harder to disambiguate than conjectured. Take teachers' interpretation of LQ03, which intends to measure teachers' knowledge of the essential domain, as an example. In addition to the essential domain, four teachers interpreted LQ03 to relate to a general process of learning, such as sharing understanding with peers, writing in notebooks, and representing ideas in different ways.

> *Kelly: We want to communicate those [scientific] ideas … By sharing, I'm drawing, talking, writing, discussing, you know, having an argument with someone. So, all of those different modes, as a way to share.*
> *Bella: They [students] have to have some form of communicating, whether it's speaking with others, writing, demonstrating.*
> *Hedy: They're communicating in our class conversations and in their science notebooks, using those different parts of language. [The notebook] would be more drawings or writing…*
> *Ran: Students communicate to share ideas with another person, or share it in writing, or share it in drawings.*

Table 1 Qualitative coding outcomes: comparing theoretical constructs with interviews

Items Based on the Essential Construct					
LQ01R	MMR Constitutive Process and Product	**LQ03R**	MMR Constitutive Process and Product	**LQ20R**	Constitutive Process and Product
LQ22R	Constitutive Process and Product	**LQ24R**	Constitutive Process and Product	**LQ26**[a]	
Items Based on the Constitutive Construct					
LQ05	MMR Constitutive Process and Product	**LQ07R**	MMR Constitutive Process and Product	**LQ11**	MMR Constitutive Process and Product
LQ12R	MMR Constitutive Process and Product	**LQ16R**	MMR Constitutive Process and Product		
Items Based on the Product and Process Construct					
LQ06	MMR Constitutive Process and Product	**LQ18**	MMR Constitutive Process and Product		
Items Based on the Multimodal Representation Construct					
LQ100	MMR Process and Product	**LQ17R**	MMR Process and Product		

[a] Teachers' interpretations of LQ26 have no relation with aspects of language but their opinion about reading comprehension and learning science

The excerpt above not only demonstrated that teachers interpreted LQ03 from the perspective of the learning process, but they also attached multiple ways of representation to this item, such as drawing, talking, and writing. Another example of responses that engaged multiple domains is that teachers not only knew that they should engage students in the process of learning but also connected language use with the constitutive process of building an understanding of science:

> *Kelly: So, to me when you're communicating your ideas…that just seems a little bit more clear.*
> *Hedy: It's not perfect writing, they are second graders, but they are getting some, you know, the basic ideas.*

In addition to LQ03, the interpretation of other items also included more than one domain of language as an epistemic tool as shown in Table 1. Therefore, this provides empirical evidence from interviews to demonstrate that the four domains of language as an epistemic tool were interwoven in the teachers' interpretations of language. This supports the assertion that the construct of language as an epistemic tool is unidimensional.

Evidence of unidimensionality can also be checked by examining the relations between the four domains. First, we argue that the domain 'language is essential' is dominant over the other three. Since language in all its forms is necessary for learning, it is impossible to consider the constitutive nature of language, questions of process and product, or multimodality without first acknowledging the underlying necessity of language itself. Second, both processes *and* products are involved in representation, including MMR. In the process of choosing and using MMR to construct and represent ideas, students need to develop their ideas (i.e., processes) and write (i.e., products), which engages with the constitutive nature of language. Additional relationships exist between the domains. For example, students' everyday language often occurs through engagement with MMR (e.g., memes, emojis), so the constitutive nature of language and MMR are connected. Therefore, the four theoretical domains are interwoven, so the construct is unidimensional.

In conclusion, teachers' responses supported the proposed unidimensionality of the construct language as an epistemic tool. By way of further elaborating the interwoven nature of the four domains that comprise language, we present a model representing the role of language in science learning as described by the participants.

Three items with ambiguous interpretation

Four teachers had opposite interpretations for LQ06, LQ11 and LQ26, which may decrease the validity of items and make the items cannot measure the construct that they are intended to measure. We represent findings for each of these items.

Item LQ06 states, 'Students need to use specific scientific terms accurately'. This item was intended to measure teachers' knowledge of the domain 'language is constitutive'. We found that teachers at different grade levels held different ideas about academic language based on their lack of clarity on what is required at grade levels they did not teach. Hedy and Kelly, who were Grade two teachers, thought that using scientific terms accurately was not necessary for their lower-elementary students, but they speculated that it may be necessary for older children.

> *Hedy: Especially like again we have elementary students … I encourage them to use it [scientific terms] accurately [but] it's not something that we assess per se.*
> *Kelly: It would make it more clear when you have those students using the terms accurately … they just forget, or they are little, so they're mixed up.*

However, Bella and Ran, who taught upper-elementary students, thought that students in all grades should use scientific terms accurately to indicate full understanding and speculated this was important for younger children.

> *Bella: If they're using them [scientific vocabulary] in their language, they will express that they understand them accurately.*
> *Ran: In order to understand the concept so students need to use specific science terms accurately.*

This means that opposite interpretations of the same item exist for both lower- and higher-grade teachers, and the interpretation is not necessarily consistent with the domain from which the item was drawn. Therefore, LQ06 may not measure the construct because of disparities in teachers' interpretations of the item itself, and the item may not be measuring an underlying understanding of the relationship of everyday language to students' development of science knowledge.

Item LQ11 states, 'Students have to talk about and write their ideas to learn science'. This item emphasized the 'language is essential' domain and represented the idea that students have to talk and write to learn science for themselves, not only to listen and recall science ideas. The teachers' interpretations suggested that even though talking and writing are ways to learn science, this does not apply equally to all students. Kelly said that not all students like to share their ideas, but they still learn. Bella said that students can learn by observing, rather than talking or writing. Ran noted that students who are unable to hear (e.g., those with auditory disabilities) can still learn science.

Bella: I think if we're not talking and we're not writing we're still negotiating our surroundings … I'm thinking to myself, about how it relates to my prior knowledge or my understandings.
Kelly: Some introverted people maybe don't have to talk about … I think that the students who discuss and share their ideas and write about it are more confident in their ideas in science.
Ran: For some students that [talking] would not be true. I've had a student that can't talk. [But] they were still able to learn about science through videos and computer system, you know, that was sending us his feedback on it.

All teachers mentioned that there are many ways to learn science, so this item was difficult to agree with. The reason why this item fails to measure the construct it is designed to measure is that teachers interpreted the wording to emphasize the modality of language conflicting with learning processes being unique to individuals, rather than taking the broader view of language being essential for knowledge.

Item LQ26 states, 'Reading comprehension is not necessarily related to learning science'. This item is intended to emphasize that students with better reading comprehension would also understand science concepts better. There are two opposing interpretations of this item. One interpretation, from Bella and Hedy, is that reading comprehension relates to science learning. For instance, Bella argued that, since comprehension of science content can come through reading texts, being able to read well relates to succeeding in science class.

Bella: You need to be able to comprehend. If you're reading about science, especially you need to comprehend a text, and to know how to relate it to things that you already understand or prior knowledge.

A different interpretation, from Kelly and Ran, is that reading comprehension is not necessarily related to science learning. Kelly interpreted reading comprehension as general reading ability, which can be applied to many subjects.

Kelly: Reading comprehension is not necessarily related to learning science … Just because you have a high reading comprehension level doesn't necessarily mean you're going to understand all science concepts.

Kelly argued that higher levels of reading comprehension do not guarantee an understanding of science concepts, and she pointed out that students with low reading comprehension skills can still understand science. The two inconsistent interpretations indicate that item LQ26 has ambiguous meanings.

Results from quantitative analysis

The quantitative results are organized beginning with the frequency of response categories for each item of the questionnaire, followed by reliability and item fit statistics.

Frequency of response categories

The frequencies of response categories for 15 items were examined. Some items had missing response categories for the five-point Likert scale, such as LQ07R (Table 2).

Table 2 Frequency of five response categories and item means ($n = 126$)

Category	LQ01R	LQ03R	LQ05	LQ06	LQ07R	LQ100[a]	LQ11	LQ12R
	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)
0	2(1.6)	1(.8)	1(.8)	4(3.2)		1(.8)		1(.8)
1	25(19.8)	14(11.1)	10(7.9)	23(18.3)		6(4.8)	7(5.6)	3(2.4)
2	14(11.1)	15(11.9)	12(9.5)	21(16.7)	1(.8)	13(10.3)	14(11.1)	11(8.7)
3	59(46.8)	56(44.4)	70(55.6)	65(51.6)	52(41.3)	59(46.8)	64(50.8)	68(54.0)
4	26(20.6)	40(31.7)	33(26.2)	13(10.3)	73(57.9)	47(37.3)	41(32.5)	43(34.1)
Mean	2.65	2.95	2.98	2.48	3.57	3.15	3.10	3.18
Category	**LQ16R**	**LQ17R**	**LQ18**	**LQ20R[a]**	**LQ22R[a]**	**LQ24R**	**LQ26[a]**	
	n(%)	**n(%)**	**n(%)**	**n(%)**	**n(%)**	**n(%)**	**n(%)**	
0	5(4.0)		2(1.6)		4(3.2)	1(.8)	5(4.0)	
1	24(19.0)		1(.8)		12(9.5)		28(22.2)	
2	18(14.3)	1(.8)	13(10.3)	2(1.6)	19(15.1)	4(3.2)	13(10.3)	
3	60(47.6)	48(38.1)	62(49.2)	41(32.5)	60(47.6)	61(48.4)	58(46.0)	
4	19(15.1)	77(61.1)	48(38.1)	83(65.9)	31(24.6)	60(47.6)	22(17.5)	
Mean	2.51	3.60	3.21	3.64	2.81	3.42	2.51	

[a] Items are reverse items. They were reversed for descriptive analysis and following analysis, and data are coded as 0, 1, 2, 3, and 4

Reliability

Local independence and unidimensionality are two assumptions for conducting a Rasch analysis. The local independence was examined by the residual correlation of 15 items. We found no residual correlation higher than 0.3 for pairs of items. Therefore, local independence was satisfied by the data. For PCA analysis, the Cronbach's alpha reliability of the instrument was 0.62, which is above the accepted cut-off value for the group of teachers (Frisbie, 1988). In addition, the variance explained by PCA was 23% for 15 items with the eigenvalue as 3.506, which means that the instrument has internal consistency, indicating that the instrument measures one single construct, that is, knowledge of language as an epistemic tool.

Item correlations examine the extent to which scores on one item are related to scores on all the other items in a scale. The greater the correlation, the more consistent the item is with other items. As Table 3 shows, the range of item-total correlations was − 0.09 to 0.68. The only negative inter-item correlation came from LQ22R (alpha = − 0.09); this means that people who scored higher on LQ22R tended to have lower total scores. This is undesirable so this item is deleted in further analysis. Three more items (LQ06, LQ16R, and LQ26) had positive item-total correlations but less than 0.30.

Fit statistics

The item difficulty values in Table 3 indicate that item LQ20R was the easiest to endorse (δ = − 2.51), and item LQ06 was most difficult to endorse (δ = − 0.40). The average item difficulty was − 1.25, which indicates that item difficulty is generally low in the questionnaire—that is, teachers may find some of the items' statements easy to endorse. Using the accepted range of outfit and infit mean square from 0.5 to 1.5, LQ22R was out of range in Table 3; the high mean-square value indicates that the response is too unpredictable to contribute to good measures (Boone & Staver, 2020). Using the accepted range of t values of outfit and infit from − 2 to + 2, LQ16R, LQ17R, LQ22R and LQ26 were out of range in Table 3. In sum, two decisions were made: 1) LQ22 was deleted because of misfit and negative zero item-total correlation, 2) the three misfitting items (LQ16R, LQ17R, and LQ26) and three items with lower item-total correlation (LQ06, LQ16R, and LQ26) were examined in the qualitative analysis. The EAP reliability was 0.66 and the WLE reliability was 0.67. The item separation index is 1.42, indicating that one performance stratum can be identified (Wright, 1996).

To visualize the pattern of threshold for 15 items in the language questionnaire, the Wright map was generated as shown in Fig. 3. This map describes item thresholds on a latent trait and the distribution of item difficulty ranging from − 4.14 to 1.43 when considering the items and response thresholds. This indicates that the items and their thresholds cover a broad range of person abilities.

Table 3 Item difficulty, item fit statistics, and correlations ($n = 126$)

Item	Difficulty δ	Outfit MNSQ	Outfit.t	Infit MNSQ	Infit.t	Item-Total Alpha	Alpha if Item Deleted
LQ01R	-0.58	1.21	1.63	1.09	0.72	0.33	0.61
LQ03R	-0.96	1.08	0.58	1.05	0.37	0.52	0.57
LQ05	-1.00	1.08	0.56	0.99	-0.05	0.30	0.61
LQ06	-0.40	1.12	1.02	1.03	0.31	0.11	0.64
LQ07R	-2.26	0.76	-1.61	0.74	-1.69	0.51	0.60
LQ11	-1.19	0.85	-0.97	0.83	-1.14	0.63	0.57
LQ12R	-1.32	0.92	-0.49	0.87	-0.81	0.48	0.60
LQ16R	-0.43	1.26	2.13	1.18	1.51	0.18	0.64
LQ17R	-2.37	0.70	-2.06	0.73	-1.85	0.68	0.59
LQ18	-1.38	0.94	-0.32	0.94	-0.31	0.55	0.58
LQ20R[a]	-2.51	0.87	-0.80	0.89	-0.64	0.43	0.61
LQ22R[a]	-0.77	1.51	3.42	1.38	2.67	-0.09	0.67
LQ24R[b]	-1.85	0.77	-1.47	0.73	-1.71	0.68	0.58
LQ26[a]	-0.43	1.35	2.83	1.26	2.14	0.14	0.64
LQ100[a]	-1.27	1.30	1.75	1.16	1.02	0.30	0.62

(1) [a] means those items should be coded reversed in data analysis

(2) [b] Because LQ24R doesn't have category 1 response, recode 0 category as 1, so there is continuous with category 2, 3, and 4

(3) Item-total correlation was calculated by alpha () function in psych package in r language (r.cor = item-total correlation)

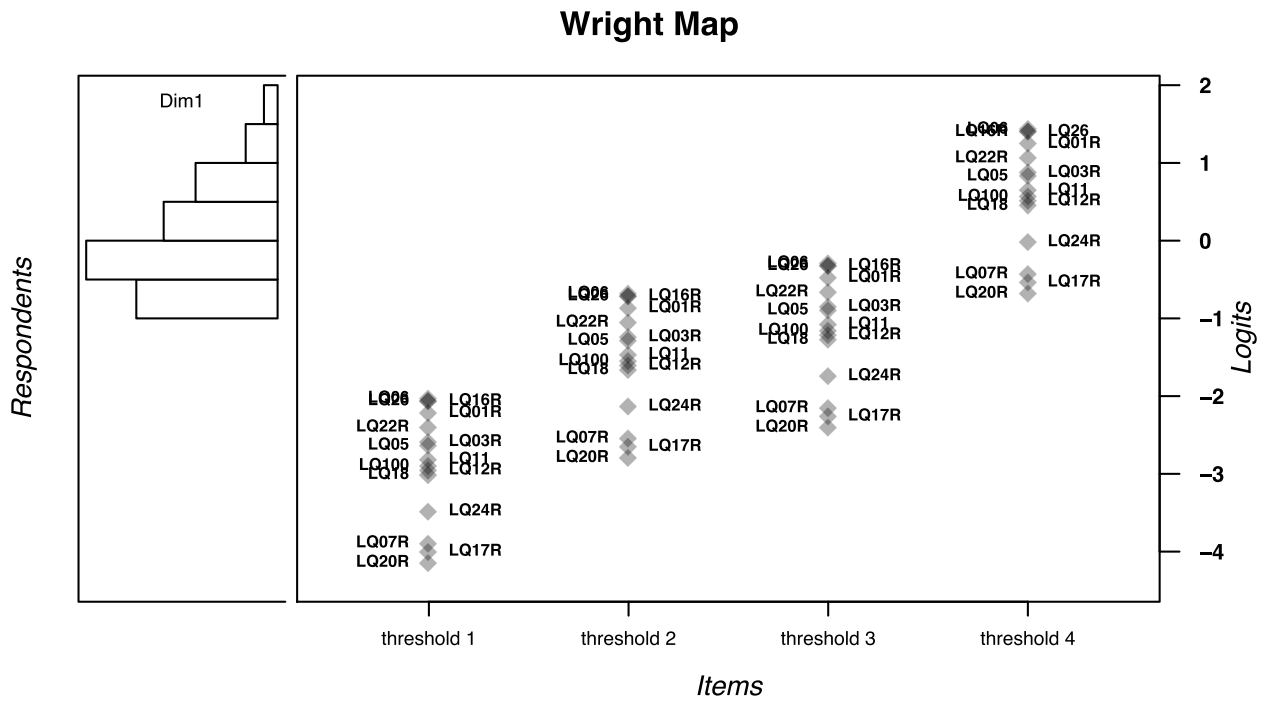

Fig. 3 Wright map of 15 items in Language Survey. Notes: (1) Diamonds demonstrate items' thresholds. For five-point Likert scale, there are four thresholds. Therefore, there are four sets of thresholds for 15 items distributing on the person's ability. (2) Because LQ24R doesn't have category 1 response, recode 0 category as 1, so there is continuous with category 2, 3, and 4

Integration and interpretation

Item removal was an iterative process. Items were removed one at a time, followed by an examination of fit statistics and item-total correlations. Based on interview and questionnaire data, LQ22R and LQ16R were deleted because of the low inter-item correlation and item misfit. LQ06, LQ11 and LQ26 were deleted because of their ambiguous meanings. The remaining items were used for a Rasch analysis. Finally, LQ17R was deleted because of misfit and overall item fit statistics was improved by deleting it (Table 4). Even though the outfit t value of LQ01R was beyond the acceptable range, the content of this item is important to understand language as an epistemic tool. Reliability was also within the acceptable range: both EAP reliability and WLE reliability were 0.71. The item separation index is 1.56, which is higher

Table 4 Fit statistics and thresholds of selected nine items ($n = 126$)

Item	Difficulty δ	Outfit MNSQ	Outfi.t	Infit MNSQ	Infit.t	τ 1	τ 2	τ 3	τ 4
LQ01R	-0.71	1.35	2.60	1.21	1.64	-2.55	-1.06	-0.60	1.39
LQ03R	-1.14	1.23	1.58	1.15	1.05	-2.98	-1.49	-1.04	0.95
LQ05	-1.20	1.26	1.74	1.13	0.90	-3.03	-1.54	-1.09	0.90
LQ07R[a]	-2.59	0.83	-1.12	0.78	-1.44	**-4.43**	**-2.94**	-2.48	-0.49
LQ12R	-1.56	0.94	-0.37	0.89	-0.68	-3.39	-1.91	-1.45	0.54
LQ18	-1.62	0.96	-0.25	0.97	-0.18	-3.46	-1.97	-1.51	0.48
LQ20R[a]	-2.86	0.88	-0.79	0.92	-0.46	**-4.70**	**-3.21**	-2.75	-0.76
LQ24R[b]	-2.14	0.73	-1.75	0.72	-1.84	**-3.98**	-2.49	-2.03	-0.04
LQ100	-1.49	1.29	1.77	1.19	1.24	-3.33	-1.84	-1.39	0.60

(1) [a]For two items (LQ07R, LQ20R), thresholds in bold font indicate that there is no data for such thresholds

(2) [b] Because LQ24R doesn't have category 1 response, recode 0 category as 1, so there is continuous with category 2, 3, and 4. There is no data for threshold in bold font

than the item separation index for the original instrument, indicating that two distinct strata can be identified (Fisher, 1992).

The purpose of this study was to refine and validate an instrument to measure teachers' knowledge of language as an epistemic tool. In sum, six items (LQ06, LQ11, LQ16R, LQ17R, LQ22 and LQ26) were deleted. The revised instrument consists of nine items as shown in Table 5.

Discussion

This paper presents an example of ongoing instrument refinement using a combination of further qualitative and quantitative work, in this specific case focusing on a questionnaire measuring science teachers' knowledge of language as an epistemic tool. Our findings indicated that the overarching construct of language as an epistemic tool was unidimensional as intended (Fulmer, 2021), in consideration of both the interview process and the iterative item analysis work. Because of the unidimensional nature of the language as an epistemic tool, the four subdomains are distinguishable yet interrelated. That underscores the importance of using an integrated view of language as an epistemic tool whether in instrument application or in teacher professional development work.

However, we found that some participating teachers' interpretations of the items varied from the measurement goals enough that it would likely affect the item, such as by focusing on language modalities rather than on fundamental aspects of language as an epistemic tool. This shows the value demonstrated by the widespread use of interviews to provide insight into participants' interpretations of items (Ryan et al., 2012). At a broader level, this also points to the importance of continued research on questionnaire use and interpretation to help improve understanding of the underlying construct and how it can be measured.

We also found that it was much harder to distinguish responses at the lower end of the 5-point response scale and for items with low overall difficulty. This may indicate that participants' ability might be higher than what the instrument initially aimed to measure. Creating more difficult items may give more differentiation for teachers' knowledge of language as an epistemic tool. Also, this may show that it is necessary to test alternative parameterizations and modelling approaches that make best use of the available data while also being consistent with a strict notion of good measurement such as the Rasch measurement model (Liu, 2020). Parameterizations reflect different types of response category structures, giving insights into the item and instrument function (von Davier & von Davier, 2013). Researchers could try different parameterizations representing different assumptions about what is measured. One-parameter models estimate thresholds, two-parameter models estimate thresholds and item discrimination for each item, and three-parameter models estimate guessing parameters in addition to thresholds and discrimination. Whereas the Rasch measurement approach emphasizes selecting items that show fit to a strict definition of measurement, other approaches allow researchers to find out which model fit data best by comparing different parameterization models (Brown et al., 2015). Model comparison not only gives different statistical outcomes but can also inform the interpretation of the construct itself.

Another reason why few participants chose the "strongly disagree" option in the survey might be social desirability. Social desirability response bias is another factor influencing participants' response patterns, which may affect their use of the full range on the response scale (Adams et al., 2005; Holbrook et al., 2003; Liang et al., 2006). Social desirability is the tendency of some participants to represent themselves on

Table 5 Refined language questionnaire items

Item	Statement
LQ01R	Students cannot think scientific ideas without language
LQ03R	Students cannot communicate scientific ideas without language
LQ05	Students are finding out about science by listening, reading, and writing about it
LQ07R	Students should be able to communicate their own ideas about what we have discussed in class
LQ12R	Producing language–writing, drawing, talking–is how students learn scientific knowledge
LQ18	Language is not only used to copy knowledge from the teacher or a textbook, but is also used to generate knowledge
LQ20R[a]	Filling in worksheets or templates from the curriculum is the most important use of language in science class
LQ24R	Writing to different audiences helps students to deepen conceptual understanding
LQ100[a]	Using multiple modes of representation would be confusing for students when we are learning science

[a] means reverse-coded item

self-reporting instruments or interviews in ways that are more favorable, socially desirable, or respectable during social interactions (Dodou & de Winter, 2014; Holbrook et al., 2003; Larson & Bradshaw, 2017). One advantage of the online questionnaire created a social distance between participants and instructors in professional development workshops. This may minimize social desirability (Holbrook et al., 2003) and reveal teachers' authentic knowledge of language as an epistemic tool. However, the effect of social distance may be cancelled out by the context of professional workshops and by the fact that they may attend the workshops in the future. As a result, some teachers may feel compelled to withhold their true opinions and instead choose options that align with the desires of PD workshop leaders (Larson & Bradshaw, 2017), such as "strongly agree" or "somewhat agree" in this study.

Taken together, this points to the importance of continued data collection and interpretation around proposed research instruments using a complementary variety of methodological approaches, particularly those addressing constructs and unmeasured effects that address domains such as language as an epistemic tool.

Implications

This questionnaire, and its design process, has many applications. The questionnaire could be used for professional development. Since it has been administered to teachers across a window in which they were receiving professional development related to the construct the questionnaire measures, we believe the instrument is sensitive enough to provide teacher educators with useful information about teachers' learning. It could also be used in preservice or in-service learning settings, or even as a tool for teachers' own reflection. This study does not dispute the previous instrument's development in the pilot study (Fulmer et al., 2021), even though the instrument refinement items differ. Differences in data sources have caused different conclusions in instrument refinement. In our pilot study (2021a), we developed an instrument to measure teachers' knowledge of language based on data collected from pre- and in-service teachers before the professional development was held. However, data in this study were collected from in-service teachers who have attended substantial professional development. There are two differences: the individual teachers in the population, and whether the respondents attended professional development. Therefore, the instrument in the pilot study may be applicable for teachers who are at an entrance level of knowledge of language as an epistemic tool, but the instrument in this study may be more useful for teachers who have been involved in professional development for learning about the role of language as an epistemic tool.

This study also reiterated the value of mixed-methods approaches to questionnaire refinement and the value of interviews given in tandem with RSM. In the item selection process, we deleted the least desirable items by considering both our quantitative and qualitative analyses. Even though some ambiguous items may have acceptable fit statistics from a quantitative point of view, they should also be examined from a qualitative point of view to ensure participants' interpretations match the intention. The combination of these methods allowed us to apply statistical tests to items and to dialog with respondents, the combination of which afforded us a complete picture of each item's utility and contribution to overall validity of the questionnaire.

We also found that unidimensionality of language as an epistemic tool was not only supported by statistical analyses using the Rasch model but was also exhibited in teachers' own words. The unidimensionality from multiple methodological perspectives also supports efforts in professional development to introduce domains and integrate them into an overarching view of language as an epistemic tool. Integrating the subdomains into the PD will help teachers understand the unidimensional nature of language as an epistemic tool, rather than overemphasizing discrete attention to narrow subdomains. For example, teacher educators could develop some activities to encourage teachers to see the connections among the different domains of language, which may help teachers understand the use of language as an epistemic tool. With this as a start, teachers could embed such strategies in their teaching.

Teachers' comments interacted with theoretical models of language as an epistemic tool in unexpected ways. For example, Hedy and Ran said that talking to learn is not applied to all students because some students were reluctant to talk. Hedy said that writing to learn is not applied to all students because lower-level elementary students are not good at writing. Kelly and Ran said that listening is another way to learn science. Even though this instrument measures language knowledge, all four teachers emphasized the popular idea that different intelligences or learning styles exist (e.g., Gardner, 2011), despite receiving professional development that highlighted universal principles that govern the use of epistemic tools for driving learning. Clearly, teachers brought their beliefs and experiences to professional development while they tried to learn a new approach. Teachers' own beliefs and experiences may result in negative effects on learning in professional workshops (Penuel et al., 2009). In this situation, they likely negotiated for themselves what was good for their teaching based on their own evaluation of what they had been taught. As instructors, we are excited to see teachers' growth in understanding our philosophy,

as captured by the instrument. However, it is equally important that teachers self-evaluate what they gain from professional development and find their own ways to incorporate new knowledge with their existing teaching philosophy and practices.

Limitations of the study

Our validated instrument cannot be directly generalized to a broader population beyond elementary science teachers. In our study, most participants were white, monolingual native English speakers from the Midwest and Southeast U.S., with demographic markers which are relevant to their knowledge of language in general. We also note that these elementary science teachers were typically content generalists certified to teach all curricular areas, including literacy. In the context of science, language has a distinct meaning, which may differ from that used to teach reading and writing. Researchers should always go through the validation process when a population changes; fortunately, a secondary contribution of this paper is a model of ongoing questionnaire refinement that could be used for this purpose.

For the current paper, we explored the unidimensionality, reliability, and validity of the questionnaire by eliminating some items, which showed that this results in a better fit to the Rasch model's strict notion of measurement (Liu, 2010). On the one hand, using the same data set to examine the model fit and revise it to improve the functioning of the instrument. Hergesell (2022) also used the same data set to select items to revise an existing instrument. On the other hand, validating the revised questionnaire with a new data set would strengthen by extending the work. One example of further study would be to administer the revised questionnaire to other groups of teachers with similar characteristics, which would offer a separate data set that could allow testing of the validity evidence for item fit statistics. This is an area we hope to pursue in the future to distribute the revised language questionnaire to additional samples of teachers for generalization.

In addition, this study focuses on teachers' knowledge of language as an epistemic tool, which was one purpose of the professional development, rather than their use of language as an epistemic tool in their teaching. Having more knowledge of language may be a prerequisite for using it in teaching (Yore & Treagust, 2006), but implementation is better measured through observational data sources and self-reporting. Further research might investigate this relationship between professional development and practice.

Conclusion

The purpose of this study was to refine and validate an instrument to measure teachers' knowledge of language as an epistemic tool in science classrooms. A revised list of items and revised recommendations for the number of response categories have been presented. In addition, we have outlined our instrument refinement process at length, to allow other researchers to follow it when designing instruments for measuring similar constructs. Our analysis of the questionnaire has revealed that language is a unidimensional construct. In describing how this is so, we have presented an emergent conceptual model.

While we view language as essential for doing and knowing about science (Pinker, 2010), we note that teachers stressed that language is not the only tool they use to drive learning. Different views of reading comprehension; differences in ideas about scientific vocabulary use between grade levels; and the teachers' challenge that all students can learn science, regardless of their language abilities, pushed us to reconsider elements of our framework. Future work that explores the ways the domains of language intersect is needed to advance science teaching toward the goals of the NGSS (2013).

Abbreviations

MMR	Represents multiple modes of representation, such as writing, drawing, using graphs or charts, etc.
RSM	Represents rating scale model
PCA	Represents principal component analysis
MNSQ	Represents mean-square value
EAP	Is an acronym of Expected A Posteriori
WLE	Is an acronym of Weighted Likelihood Estimates
TAM	Is an acronym of Test Analysis Modules, which is a package in R environments

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43031-023-00080-7.

Additional file 1. Language Questionnaire.

Acknowledgements

We would like to thank the teachers who attended the professional development workshops and took part in this study.
The Institutional Review Board (IRBs) at the University of Iowa approved the study. Approval number is 201804703. Our study meets the ethics/human subject requirements of IRBs at the time the data was collected.

Authors' contributions

CL, GWF, BH, and JKS collected quantitative data during workshops. CL and CD developed interview protocol, and CD collected interview data for this research. CD analyzed quantitative and qualitative data initially and drafted the manuscript. CL, GWF, and BH revised and refined the manuscript. All authors read and approved the final manuscript.

Funding

The project was funded by the National Science Foundation (Grant Agreement Number: DRL-1812576).

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors report no potential conflicts of interest.

Received: 9 December 2022 Accepted: 31 July 2023
Published online: 21 August 2023

References

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33*(9), 1289–1312.
Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., & Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology, 161*(4), 389–398.
Aguirre-Muñoz, Z., & Pando, M. (2021). Conceptualizing STEM teacher professional knowledge for teaching ELs: Initial impact of subject matter and disciplinary literacy PD on content knowledge and practice. *Bilingual Research Journal, 44*(3), 335–359.
Ainsworth, S., & VanLabeke, N. (2004). Multiple forms of dynamic representation. *Learning and Instruction, 14*(3), 241–255.
Anderson, O. R. (1992). Some interrelationships between constructivist models of learning and current neurobiological theory, with implications for science education. *Journal of Research in Science Teaching, 29*(10), 1037–1058.
Andrich, D. (1988). Rasch models for measurement. (Vol. 68). Sage
Bae, Y., Fulmer, G. W., & Hand, B. M. (2021). Developing latent constructs of dialogic interaction to examine the epistemic climate: Rasch modeling. *School Science and Mathematics*, 121(3), 164–174
Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd edition)*. Routledge.
Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences* (pp. 287–302). Springer.
Brock, R. (2015). Intuition and insight: Two concepts that illuminate the tacit in science education. *Studies in Science Education, 51*(2), 127–167.
Brown, C., Templin, J., & Cohen, A. (2015). Comparing the two-and three-parameter logistic models via likelihood ratio tests: A commonly misunderstood problem. *Applied Psychological Measurement, 39*(5), 335–348.
Brown, S. A., Tyrer, F., Clarke, A. L., Lloyd-Davies, L. H., Niyi-Odumosu, F. A., Nah, R. G. Q., Stein, A. G., Tarrant, C., & Smith, A. C. (2018). Kidney symptom questionnaire: Development, content validation and relationship with quality of life. *Journal of Renal Care, 44*(3), 162–173.
Calkins, L. M. (1994). *The art of teaching writing* (2nd ed.). Irwin.
Campbell, T., & Oh, P. S. (2015). Engaging students in modeling as an epistemic practice of science: An introduction to the special issue of the Journal of Science Education and Technology. *Journal of Science Education and Technology, 24*(2–3), 125–131.
Chatzidamianos, G., Burns, D., Andriopoulou, P., Archer, D., & du Feu, M. (2021). The challenges and facilitators to successful translation and adaptation of written self-report psychological measures into sign languages: A systematic review. *Psychological Assessment, 33*(11), 1100.
Cikmaz, A., Fulmer, G., Yaman, F., & Hand, B. (2021). Examining the interdependence in the growth of students' language and argument competencies in replicative and generative learning environments. *Journal of Research in Science Teaching*, 58(10), 1457–1488
Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed). Sage publications.
DiStefano, C., & Morgan, G. B. (2010). Evaluation of the BESS TRS-CA using the Rasch rating scale model. *School Psychology Quarterly, 25*, 202–212. https://doi.org/10.1037/a0021509
Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior, 36*, 487–495.
Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington: National Academies Press.
Elgin, C. Z. (2013). Epistemic agency. *Theory and Research in Education, 11*(2), 135–152.
Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1–4.
Fisher, W. (1992). Reliability, Separation, Strata Statistics. *Rasch Measurement. Transactions, 6*(3), 238.
Ford, E., Roomi, H., Hugh, H., & van Marwijk, H. (2019). Understanding barriers to women seeking and receiving help for perinatal mental health problems in UK general practice: Development of a questionnaire. *Primary Health Care Research & Development, 20*(e156), 1–8.
Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7*(1), 25–35.
Fulmer, G. W., Hwang, J., Ding, C., Hand, B., Suh, J. K., & Hansen, W. (2021). Development of a questionnaire on teachers' knowledge of language as an epistemic tool. *Journal of Research in Science Teaching*, 58(4), 459–490
Galbraith, D. (1999). Writing as a knowledge-constituting process. *Knowing What to Write: Conceptual Processes in Text Production, 4*, 139–164.
Gardner, H. E. (2011). *Frames of mind: The theory of multiple intelligences*. Basic books.
Grangeat, M., & Hudson, B. (2015). A new model for understanding the growth of science teacher professional knowledge. In Understanding Science Teachers' Professional Knowledge Growth (pp. 203–228). Brill.
Gee, J. (2000). Identity as an analytic lens for research in education. *Review of Research in Education, 25*, 99–125.
Gutierrez, K., Rymes, B., & Larson, J. (1995). Script, counterscript, and underlife in the classroom: James Brown versus Brown v. *Board of Education. Harvard Educational Review, 65*(3), 445–472.
Hand, B. M., Prain, V., & Yore, L. (2001). Sequential writing tasks' influence on science learning. In Writing as a learning tool (pp. 105–129). Dordrecht: Springer
Hand, B., & Choi, A. (2010). Examining the impact of student use of multiple modal representations in constructing arguments in organic chemistry laboratory classes. *Research in Science Education*, 40(1), 29–44
Hand, B., Norton-Meier, L., Gunel, M., & Akkus, R. (2016). Aligning teaching to learning: A 3-year study examining the embedding of language and argumentation into elementary science classrooms. *International Journal of Science and Mathematics Education*, 14(5), 847–863
Hand, B., Cavagnetto, A., Chen, Y.-C., & Park, S. (2016). Moving past curricula and strategies: Language and the development of adaptive pedagogy for immersive learning environments. *Research in Science Education*, 46 (1), 223–241
Hergesell, A. (2022). Using Rasch analysis for scale development and refinement in tourism: Theory and illustration. *Journal of Business Research, 142*, 551–561.
Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*(1), 79–125.
Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299–314.
Knafl, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, M., Dixon, J., & Grey, M. (2007). The analysis and interpretation of cognitive interviews for instrument development. *Research in Nursing & Health, 30*(2), 224–234.
Kohl, P. B., Rosengrant, D., & Finkelstein, N. D. (2007). Strongly and weakly directed approaches to teaching multiple representation use in physics. *Physical Review Special Topics-Physics Education Research, 3*(1), 010108.
Lamprianou, I. (2019). *Applying the Rasch Model in Social Sciences Using R and BlueSky Statistics*. Routledge.
Larson, K. E., & Bradshaw, C. P. (2017). Cultural competence and social desirability among practitioners: A systematic review of the literature. *Children and Youth Services Review, 76*, 100–111.
Liang, L. L., Chen, S., Chen, X., Kaya, O. N., Adams, A. D., Macklin, M., Ebenezer, J. (2006, April). Student understanding of science and scientific inquiry (SUSSI): Revision and further validation of an assessment instrument. In Annual Conference of the *National Association for Research in Science Teaching (NARST)*, San Francisco, CA (April) (Vol. 122).
Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.
Liu, X. (2010). Using and developing measurement instruments in science education: a rasch modeling approach. lap

Liu, X. (2020). *Using and developing measurement instruments in science education: a Rasch modeling approach (2nd edition).* IAP.
Lowe, A., Norris, A. C., Farris, J. A., & Babbage, D. R. (2018). Quantifying thematic saturation in qualitative data analysis. *Field Methods, 30*(3), 191–207.
National Research Council (NRC). (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington: The National Academies Press. https://doi.org/10.17226/13165
NGSS Lead States. (2013). *Next generation science standards: For States, by States*. The National Academies Press.
Norris, S., & Phillips, L. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224–240.
Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144.
Pelger, S., & Nilsson, P. (2016). Popular science writing to support students' learning of science and scientific literacy. *Research in Science Education, 46*(3), 439–456.
Penuel, W., Fishman, B. J., Gallagher, L. P., Korbak, C., & Lopez-Prado, B. (2009). Is alignment enough? Investigating the effects of state policies and professional development on science curriculum implementation. *Science Education, 93*(4), 656–677.
Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development, 50*(4), 217–223.
Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences, 107*(Supplement 2), 8993–8999.
Prain, V., & Hand, B. (1996). Writing for learning in secondary science: rethinking practices. *Teaching and Teacher Education*, 12(6), 609–626.
Prain, V., & Hand, B. (2016a). Learning science through learning to use its languages. In Using multimodal representations to support learning in the science classroom (pp. 1-10). Springer, Cham
Prain, V., & Hand, B. (2016b). Coming to know more through and from writing. *Educational Researcher*, 45(7), 430–434
Robitzsch, A., Kiefer, T., Wu, M. (2020). TAM: Test analysis modules. R package version 3.5–19. https://CRAN.R-project.org/package=TAM
Romine, W. L., Sadler, T. D., & Kinslow, A. T. (2017). Assessment of scientific literacy: Development and validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR). *Journal of Research in Science Teaching, 54*(2), 274–295.
Rubin, H., & Rubin, I. (2011). *Qualitative interviewing: The art of hearing data* (3rd ed.). SAGE.
Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small-and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414–430.
Saldaña, J. (2015). *The coding manual for qualitative researchers* (3rd ed). Sage.
Singh, C., & Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts. *American Journal of Physics, 71*(6), 607–617.
Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2), 205–231.
Stroupe, D., Caballero, M. D., & White, P. (2018). Fostering students' epistemic agency through the co-configuration of moth research. *Science Education, 102*(6), 1176–1200.
Tang, K. S., & Moje, E. B. (2010). Relating multimodal representations to the literacies of science. *Research in Science Education, 40*(1), 81–85.
Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169.
von Davier, M., & von Davier, A. A. (2013). Local equating using the Rasch model, the OPLM, and the 2PL IRT model—or—What is it anyway if the model captures everything there is to know about the test takers? *Journal of Educational Measurement, 50*(3), 295–303.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
Wang, J., Wang, Y., Tai, H., & Chen, W. (2010). Investigating the effectiveness of inquiry-based instruction on students with different prior knowledge and reading abilities. *International Journal of Science and Mathematics Education, 8*(5), 801–820.
Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 38*(5), 529–552.
Wright, B.D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9(4), 472. Retrieved from https://www.rasch.org/rmt/rmt94n.htm
Yore, L. D., & Hand, B. (2010). Epilogue: plotting a research agenda for multiple representations, multiple modality, and multimodal representational competency. *Research in Science Education*, 40(1), 93–101
Yore, L. D., & Treagust, D. F. (2006). Current realities and future possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education, 28*(2–3), 291–314.

Publisher's Note