Investigating Content Multidimensionality in a Large-scale Science Assessment: A Mixed Methods Approach

by

Cassandra N. Malcom

A Dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Quantitative Research Methods in Education

Dissertation Committee:

Dr. Kathleen Scalise, Chair and Advisor

Dr. Dianna Carrizales-Engelmann, Core Member

Dr. George Harrison, Core Member

Dr. Joanna Goode, Core Member

Dr. Beth Harn, Institutional Representative

University of Oregon

Spring 2024

© 2024 Cassandra N. Malcom
This work is licensed under a Creative Commons CC BY-NC 4.0 

DISSERTATION ABSTRACT

Cassandra N. Malcom

Doctor of Philosophy in Quantitative Research Methods in Education

Title: Investigating Content Multidimensionality in a Large-scale Science Assessment: A Mixed Methods Approach

Science, Technology, Engineering, and Math (STEM) skills are increasingly required of students to be successful in higher education and the workforce. Therefore, modeling assessment outcomes accurately, often using more types of student data to get a complete picture of student learning, is increasingly relevant. The Program for International Student Assessment (PISA) is promoted as a summative assessment opportunity that includes a science framework. As with many science assessments, the framework includes Life, Physical, and Earth science, which alone seems to imply multidimensionality, and also there are other sources of dimensionality that seem to be described conceptually in the framework. Using data from the 2015 PISA science assessment, a multidimensional item response theory (MIRT) model was fit to see how a multidimensional model operates with the data. Before developing the MIRT model, a qualitative review of the framework for multidimensionality took place and exploratory analyses were implemented for the quantitative data, including a data science technique to explore multidimensionality and some factor analysis techniques. After fitting the MIRT model, it was compared to several unidimensional IRT (UIRT) models to determine the model that explains the most variation. The qualitative analyses generated evidence of multidimensional science content domains in the 2015 PISA science framework, which should require a MIRT model, but quantitative analyses indicate a unidimensional model is more

practically significant. Once quantitative results were triangulated with the qualitative review of the framework for multidimensionality, the implications on equity and history of harm with regards to science assessments were discussed. Findings from the qualitative and quantitative aspects of the study were used to generate recommendations for different stakeholders.

Keywords: multidimensionality, item response theory, STEM education, summative assessment, large-scale assessment, qualitative framework review

CURRICULUM VITAE

NAME OF AUTHOR: Cassandra N. Malcom

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Southwest Texas State University, San Marcos

DEGREES AWARDED:

Doctor of Philosophy, Quantitative Research Methods in Education, (all but dissertation expected to be completed in 2024), University of Oregon
Master of Science, Biology, 2003, Southwest Texas State University
Bachelor of Science, Marine Biology, 2001, Southwest Texas State University

AREAS OF SPECIAL INTEREST:

Science Education
Measurement and Assessment
Collaboration and Inquiry in Science

PROFESSIONAL EXPERIENCE:

Graduate Instructional and Research Assistant, University of Oregon College of Education, 2020-Present

Science Content Writer/Reviewer Independent Contractor, Hurix Digital, 2022

Science Coordinator and Assessment Specialist III, Educational Testing Service (ETS), 2008-2021

Science Department Chair and Teacher, Robert G. Cole High School, 2004-2008

Science Teacher Intern, Nancy Ney Charter School, 2003-2004

Science Adventure Club Teacher, Witte Museum, 2002

Graduate Instructional and Research Assistant, Southwest Texas State (SWT) University

Biology Dept., 2001-2003

Undergraduate Instructional Assistant, SWT University Biology Dept., 2001

GRANTS, AWARDS, AND HONORS:

Spot Award, ETS, 2011-2013, 2017, and 2019

President's Award, ETS, 2012

Academic Excellence Award, SWT University Biology Dept., 2003

Ruth Strandman Field Biology Scholarship, 2000

Houston Livestock and Rodeo Scholarship, 1996-1998

Dean's List, SWT University, 1996-1998

Ford Scholarship, 1996

National Dean's List, 1996

Girl Scout Gold Award, 1995

PUBLICATIONS:

Scalise, K., **Malcom, C.,** & Kaylor, E. (2023). Chapter 8: Analysing and integrating new sources of data reliably in innovative assessments. In N. Foster & M. Piacentini (Eds.), *Innovating assessments to measure and support complex skills* (pp. 138-150). OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Scalise, K., **Malcom, C.,** & Kaylor, E. (2023). Chapter 13: A tale of two worlds: Machine learning approaches at the intersection with educational measurement. In N. Foster & M. Piacentini (Eds.), *Innovating assessments to measure and support complex skills* (pp. 216-224). OECD Publishing. https://doi.org/10.1787/d01eb8a4-en

Malcom, C. & ETS Data, Analysis, and Reporting (DAR) Group (2020, December). *National assessment of educational progress (NAEP) science 2019 operational assessment data* [Conference presentation]. National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB) NAEP IDQC, Princeton, NJ, United States.

Lavalli, K. L., **Malcom, C. N.**, & Goldstein, J. S. (2018). Description of pereiopod setae of scyllarid lobsters, *Scyllarides aequinoctialis*, *Scyllarides latus*, and *Scyllarides nodifer*, with observations on the feeding during consumption of bivalves and gastropods. *Bulletin of Marine Science*, *94*(3), 571-601. https://doi.org/10.5343/bms.2017.1125

California Department of Education (CDE) & **Malcom, C.** (2016, December). *California science tests (CAST) and the California alternate assessment (CAA) for science* [Conference presentation]. California Educational Research Association (CERA), Sacramento, CA, United States.

Malcom, C. N. (2007, September). *Description of the setae on the pereiopods of the Mediterranean slipper lobster, Scyllarides latus* [Poster presentation]. 8th International Conference and Workshop on Lobster Biology and Management, Charlottetown, Canada.

Malcom, C. N. (2003). *Setae on slipper lobster pereiopods* [Scanning electron microscope photographs and drawings]. In Lavalli, K.L., Spanier, E., & Grasso, F., *Behavior and Sensory Biology of Slipper Lobsters* (pp. 144, 165-167). CRC Press, 2007.

Malcom, C. N. (2003). *Description of the setae on the pereiopods of the Mediterranean slipper lobster Scyllarides latus, the ridged slipper lobster, S. nodifer, and the Spanish slipper lobster S. aequinoctialis* [master's thesis, Southwest Texas State University]. https://digital.library.txstate.edu/handle/10877/11901

Malcom, C. N. (2002). *Description of the setae on the pereiopods of the Mediterranean slipper lobster, Scyllarides latus, the ridged slipper lobster, S. nodifer, and the Spanish slipper lobster, S. aequinoctialis* [Poster presentation]. 8th Colloquium Crustacea Decapoda Mediterranea, Ionian University, Corfu Island, Greece.

Malcom, C. N. (2002). *Description of the setae on the pereiopods of the Mediterranean slipper lobster, Scyllarides latus, the ridged slipper lobster, S. nodifer, and the Spanish slipper lobster, S. aequinoctialis* [Poster presentation]. Marine Benthic Ecology Meeting, United States.

ACKNOWLEDGMENTS

I wish to express sincere gratitude to my chair and advisor, Dr. Kathleen Scalise, for her invitation to start this educational journey. Her mentorship has guided me throughout and helped me grow as a researcher. Best wishes to her as she starts her next chapter!

To my dissertation committee members: Dianna Carrizales-Engelmann, Joanna Goode, Beth Harn, and George Harrison, thank you all for being willing to serve. Especially since this process was done in a tight timeline and for a student based in another state. Your feedback has been invaluable and strengthened my writing. For the opportunity to continue my learning remotely I thank the University of Oregon's College of Education. To my gracious copy editors: Dr. Linda A. Malcom and Dr. Angelina Galvez-Kiser, also many thanks for pouring over the fine details.

Last, I'd like to acknowledge my positionality in writing this dissertation. I do so in counterpoint to the voices that say this type of statement "does no work" in a quantitative research study. For me personally, a positionality statement provides a lens into how a researcher views their world and all its data – it may even uncover biases of which the researcher is unaware. The positionality statement gives voice to underrepresented researchers and allows them to claim how they want to be recognized when too often labels are forced upon them. My positionality is such:

> This researcher identifies as a liberal, white, cisgender female whose formative education occurred primarily in Texas. Daughter of working-class parents, a high value was placed on STEAM education and reading in order to better oneself and

achieve dreams. Her love of science led her to studying and teaching science and biases her views on data in that she believes science can help answer any question.

DEDICATION

This dissertation is dedicated to my mother who walked a long, hard road to make sure I got here. Without her love, support, friendship, and her own dissertation journey, I might not have seen what all is possible. And to my cousin, who's faith in me convinced me that there was never any doubt about this journey's conclusion.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

LIST OF EQUATIONS

Equation Page

LIST OF ABBREVIATIONS

Term	Abbreviation	Definition[1]
Akaike Information Criterion	AIC	An estimate of model fit based on the number of model parameters and log-likelihood that favors more complex models and smaller sample size; the smaller AIC indicates better fitting model when two models are compared
Bayesian Information Criterion	BIC	An estimate of model fit that favors simpler models by adding a penalty for more parameters; the smaller BIC indicates better fitting model when two models are compared
Civil Rights Data Collection	CRDC	NA
Classical Test Theory	CTT	States an observed test score is the sum of a student's true score and random error
Computer Adaptive Testing	CAT	An online test designed to increase or decrease test difficulty based on a student's ability as shown during the test by providing an easier or harder item dependent on the score of the last item given
Computer-based Assessment	CBA	An assessment designed to be delivered and administered via a computer or tablet
Degrees of Freedom	df	Number of independent variables that are free to vary in a data sample
Diversity, Equity, and Inclusion	DEI	NA
Expected A Posteriori	EAP	An estimate of the predicted value for the latent trait posterior probability distribution
Exploratory Factor Analysis	EFA	Identifies the latent traits then builds a linear model of the variables
Gross Domestic Product	GDP	NA
Item Characteristic Curve	ICC	A graph of a probability of a correct response versus a student's ability
Item Information-weighted Fit	Infit	"Information" refers to the variance of the observations
Item Response Theory	IRT	Explains the relationship between a latent trait and an observable outcome
Local Item Dependence	LID	The items in an assessment may be related in that an answer on one item indicates a higher chance of answering other items in a similar manner, even when conditioned on proficiency estimate
Markov Chain Monte Carlo	MCMC	Generates a random sample of a target distribution with a large number of dimensions where each MCMC sample is dependent on the prior MCMC sample; can estimate the sum as either the mean or variance of drawn samples

[1] Statistical terms are provided with definitions that are summarized by the researcher. Terms that have no applicable statistical definition are marked NA.

Maximum Likelihood Estimation	MLE	Finding the parameter values that give a curve best fitting the data
Maximum marginal likelihood estimation	MMLE	Estimates the parameters that are most likely of the expected probability distribution based on observed data
Multidimensional IRT	MIRT	Can model an assessment measuring multiple traits
National Assessment of Educational Progress	NAEP	NA
National Center for Education Statistics	NCES	NA
Next Generation Science Standards	NGSS	NA
Office for Civil Rights	OCR	NA
One-parameter Model	1PL	Simplest model that describes the latent trait (i.e., ability) based only on the difficulty parameter
Organization for Economic Co-operation and Development	OECD	NA
Paper-based Assessment	CBA	An assessment designed to be delivered and administered via a paper form
Principal Component Analysis	PCA	Number of observed variables are reduced to a decreased number of principal components, which account for the most variance of the observed variables
Programme for International Student Assessment	PISA	NA
Root Mean Square Error	RMSE	Is the root of the mean of the squared errors between observed and predicted values; measures the error of a model when predicting quantitative data
Root Mean Square of the Residuals	RMSR	Is the square root of mean of squared residuals and measures badness-of-fit for a model (0 indicating perfect model fit)
Science, Technology, Engineering, and Math	STEM	Refers to education containing these subjects, or a required set of skills needed for the workforce. Sometimes the Arts are included, in which case the acronym becomes STEAM, which is out of scope in this dissertation.
Standard Deviation	SD	A measurement of the spread of data in relation to the mean of the population that indicates variability
Standard Error	SE	An inferential statistic that indicates the reliability of a sample population mean compared to the actual population mean
Three-parameter Model	3PL	Model that describes three parameters, difficulty, discrimination, and guessing, in relation to the latent trait
Two-parameter Model	2PL	Model that describes two parameters, difficulty and discrimination, in relation to the latent trait
Unidimensional IRT	UIRT	Models an assessment measuring a single latent trait
United States	U.S.	NA

CHAPTER 1. INTRODUCTION AND LITERATURE SYNTHESIS

Problem Statement

Three science subdomains (Life, Physical, and Earth and Space) that are commonly assessed in large-scale assessments are qualitatively describable as multiple dimensions based on teaching pedagogy and student ability. For a construct, in this instance science, to be considered multidimensional its dimensions[2] (e.g., the science subdomains) should be different from each other yet connected to the theorized construct – see section *Defining Dimensionality* (Ch. 2) for further information (Polites et al., 2012). However, data from national and global large-scale science assessments are typically quantitatively modeled with unidimensional item response theory (or UIRT) models rather than multidimensional IRT (or MIRT) models. IRT is a method of examining the relationship between something intangible, such as science ability, and how that latent trait manifests in, for example, scores on a set of science items – see section *Primer on Item Response Theory* (Ch. 2) for a more detailed description.

Some researchers advocate that we should more accurately model student data from large-scale science assessments by using MIRT models. The disconnect between how large-scale assessments report data, such as on science subscales[3], while the data is modeled unidimensionally, could impact policy decisions, instruction, and other aspects of the education experience. IRT models can allow teachers and students to consider and discuss how student ability is impacted by item parameters, such as difficulty (Uesaka et al., 2022), but only if the appropriate IRT models are used. This problem will be explored here using a mixed methods

[2] At least two are needed to be multidimensional.
[3] For 2015 PISA subscales were based on the science subdomains of life, physical, and Earth and Space systems – see Appendix C.

approach, where qualitative (document exploration) techniques are used to better understand the conceptual framework claims in a framework for which the United States (U.S.) sample of the quantitative data set is next explored quantitatively for some aspects of dimensionality.

STEM Education and U.S. Economy

Understanding the current status of science education with regards to the U.S. economy will help illustrate why there is a desire for Science, Technology, Engineering, and Math (STEM)[4] assessment research, such as the more accurate modeling noted above. Citizen science and the need in life for understanding and interpreting personal STEM contexts is very important for decision making in modern society. For instance being able to have enough STEM knowledge to understand and make decisions regarding implications for vaccination, masking, and other precautions in the recent COVID-19 pandemic would have helped citizens, especially in early days of the pandemic given the absence of a fuller understanding. This is one example, of which there are many others, of how decision making by individuals in society may interact with their base STEM understanding.

Students' cognitive skills and general knowledge can impact a nation's economy (Hanushek et al., 2008). A highly skilled workforce helps nations be competitive in the global marketplace (Hanushek et al., 2008). With regards to the economies of many nations, there is a long-standing link between more STEM education and greater productivity and creativity from the workforce in generating solutions, whether they be environmental, technological, or industrial (Hanushek et al., 2008). Over the years, this link has led to increased government interest in STEM education and in the status of the STEM workforce (Kelley & Knowles, 2016).

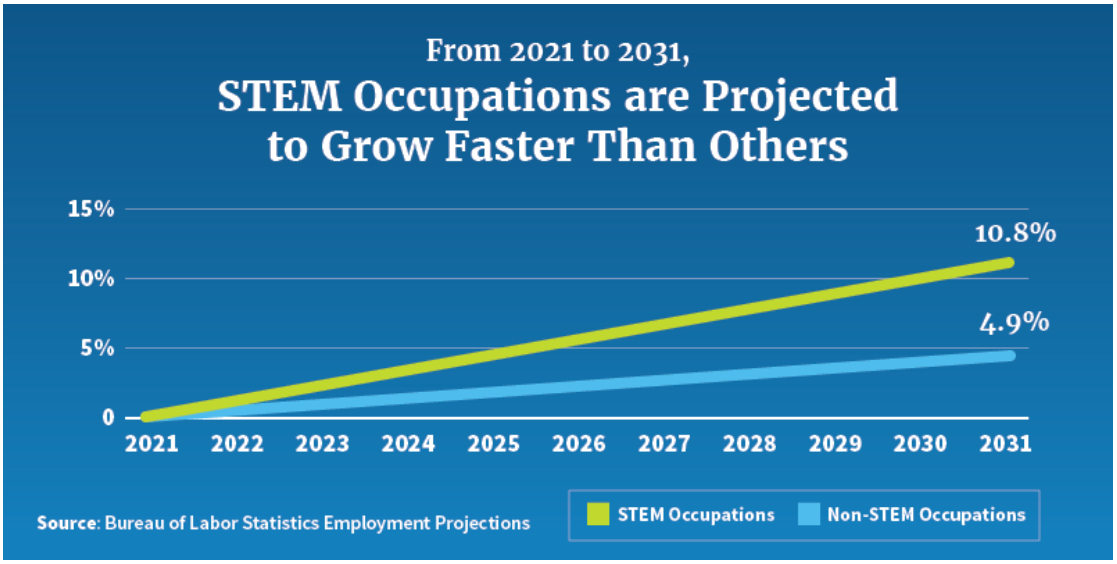
[4] Sometimes including Arts for STEAM

For example, in 2021, the U.S. government spent approximately 3.9 billion on STEM education (Lips & Moritz, 2023).

By 2031, the U.S. Department of Labor predicts the STEM workforce will grow by almost 11%, which is more than two times faster than the predicted growth rate of other occupations, see Figure 1 (Krutsch & Roderick, 2022).

Figure 1

STEM Job Predictions for 2031



Note. These predictions do not include careers related to just the Arts, as in STEAM where Arts are added to STEM. From "STEM day: Explore growing careers," by E. Krutsch and V. Roderick, 2022, *U.S. Department of Labor Blog* (https://blog.dol.gov/2022/11/04/stem-day-explore-growing-careers). Copyright 2022, U.S. Department of Labor.

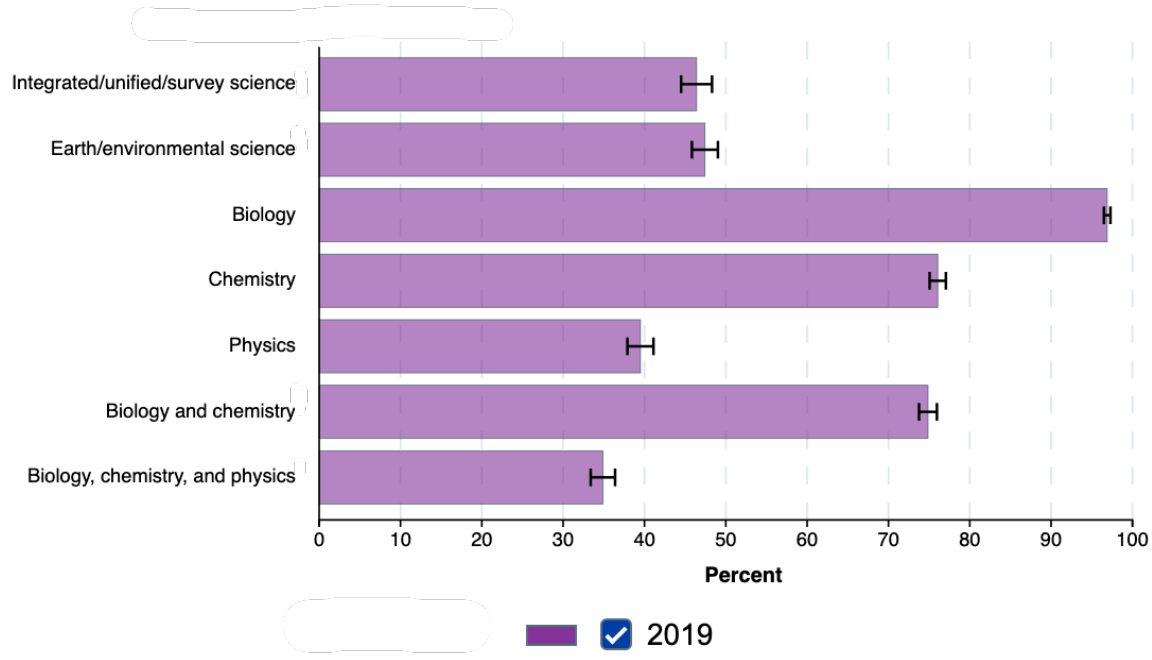
Generation of these new STEM jobs will lead to increased need for workers with STEM skills. This is especially true since the anticipated STEM jobs are expected to be enticing for students as they are also predicted to pay more on average (Krutsch & Roderick, 2022). In order to have enough STEM skilled workers, STEM education frameworks, associated teaching pedagogy, and

related assessments are being evaluated in the hopes that improvements in these areas will lead to more students taking STEM classes then entering a STEM career field.

Instead of more students entering STEM tracks in higher education or careers, students in the U.S. are currently falling behind other countries in the mastery of science skills (Hanushek et al., 2008). Under 40% of high school students in public and private schools took a biology, chemistry, and physics course in 2019 (National Center for Education Statistics [NCES], 2022). The majority of these students took at least one biology course, with chemistry as the second course students took most frequently, and physics plus Earth sciences lagged far behind in student enrollment – see Figure 2 (NCES, 2022).

Figure 2

2019 Science Course Enrollment



Note. Adapted from "High school mathematics and science course completion," 2022, *National Center for Education Statistics: Condition of Education* (https://nces.ed.gov/programs/coe/indicator/sod/high-school-courses). Copyright 2022 by the U.S. Department of Education, Institute of Education Sciences.

Evidence from the National Assessment of Educational Progress (NAEP) science scores shows students in grades 8 and 12 well below 50% proficiency (Stehle & Peters-Burton, 2019). The Programme for International Student Assessment (PISA) science scores rank the U.S. 25th with a mean score of 496, compared to the Organization for Economic Co-operation and Development (OECD) overall mean of 493, out of 72 participating economies and countries (Organisation for Economic Co-operation and Development [OECD], 2018). Due to this lag, both STEM educators and educational researchers are rethinking how STEM skills are taught and evaluated.

Integrating Science

Science educational pedagogy used to be focused on the retention of facts (Kaldaras et al., 2021; Pierson et al., 2019; Enger & Yager, 2009), such as memorizing all the parts of a cell. Newer pedagogy focuses on integrating science content and scientific inquiry (Kaldaras et al., 2021; Pellegrino & Hilton, 2013; Enger & Yager, 2009), along with soft skills like creative thinking (Csapó & Funke, 2017). This new pedagogy focus has led to the development of the Next Generation Science Standards (NGSS) in 2013[5], which advocate for crosscutting concepts, practices, and disciplinary core ideas to be taught, in a type of "sensemaking" effort. Crosscutting concepts in particular focus on applying knowledge across different science subdomains[6] (NGSS Lead States, 2013), and on drawing on these concepts to explore new STEM ideas as they arise in a student's life. This new focus could appear to call for science classes with integrated content, but the realization of that type of curriculum has been hard to achieve.

[5] The short timeframe between the NGSS being published, adopted by states, and then incorporated into classroom curriculum and the 2015 administration of the PISA left little room for any impacts on student learning to transfer to the 2015 science portion of the PISA.

[6] Scientific inquiry skills are found in the NGSS science and engineering practices and are also considered a separate dimension of learning that can apply across each science subdomain.

Policy and/or teacher preparation often leaves the coursework in separate “silos”. Assessments also need to change from measurement of constructs only requiring recall to this integrated content that requires students to apply knowledge and make sense of it (Kaldaras et al., 2021).

While there was a movement to combine the science subdomains into an integrated curriculum in the 1970s (Welch, 1977) and calls for this approach continue, most U.S. schools do not employ this method and the subdomains remain taught separately with little crossover in science content (Winarno et al., 2020). Winarno et al. (2020) provide several reasons why integration of the subdomains has remained difficult, including:

- Educators are often trained in only one of the science subdomains,
- Less professional development and college-level training is available for educators,
- High schools and state policy are designed around the idea that integrated curriculum in STEM does not yet strongly support higher education goals, since higher education courses often do not draw on integrated frameworks,
- Learning still tends to be lecture orientated when labs provide more connections for students, and
- Limited availability of integrated science textbooks.

In addition, the following factors, some from personal experience, also impact integrating science subdomain instruction.

The cost of implementation of integrated science courses is too high for most school districts to develop educator training on how to integrate, plus allow the district to purchase new lab equipment, supplies, and textbooks. A small suburban high school where I taught rejected a new integrated science course as the funds to restructure the science lab with the needed

equipment were quite high compared to the current cost of offering three subdomains and a few elective courses.

Developing a new curriculum that functionally integrates the 3 science domains can be time restrictive. Integrated science curriculum should not simply be, for example, a restructuring of a physics course to contain life science examples, but rather a deeper dive into learning how both physics and biology play a role together in a shared construct, like muscle movement. This would require a new way of approaching curriculum that provides opportunities for students to dive deep into how physics can help the understanding of the biology of muscles working with bones to create movement in a living organism. Considerable pre-planning, teacher coordination, and community feedback would need to occur for such a course to be successfully delivered.

Teacher resistance to the curriculum change, which can be due to the untested nature of how beneficial the new curriculum may be to student learning within the new learning environment, or a tendency to cling to what already works well. All educators have felt this way at some point or another – a dissatisfaction with being asked to change what works well for our students to the newest educational fad without enough teacher buy-in earlier in the process of curriculum redesign. While working on a NGSS-focused assessment redesign for a state, I was able to hear some of the frustrations teachers felt regarding how the three-dimensional learning required by NGSS was to be successfully implemented, especially for students with special needs.

Parent and student concern that integrated courses do not offer enough depth of subdomain content to prepare students for college. For example, a student may be more interested in

becoming a cell biologist and feel that less time will be devoted to cells in a course that has to cover major concepts from life, physics, chemistry, and Earth sciences. The more intricate details of cell biology may be missed unless the student takes additional courses, such as an advanced biology course. See Johnson's (2019) EdSource article[7] for insight into the parent and student concerns over integrated science courses at a California high school. In a similar vein, p*olicymaker and public resistance due to the belief that some science subdomains, when integrated, will lose teaching and learning minutes* (opportunity to learn issues) since available time will likely be divided up between several science subdomains. Instead of a student taking three different science courses spread out over three years, they may end up taking two integrated science courses spread out over two years. Students can often choose to take an elective science course, but counseling, the degree of informed choice, and access to such courses often varies by socioeconomics and can include substantial racial, ethnic, and gender disparities (Gao et al., 2019).

Assessments may not accurately show trends in data from the non-integrated year to the integrated year, so student growth can be harder to map. Many teachers develop their own assessments and unless a school requires a standardized formative assessment at the end of each science course the reasons for changes in student performance may be unknowable for many individuals. Finally, there is often *misuse of integrated science courses as low-level science classes*. In the past, I have heard school counselors refer to integrated science courses as a "dumping ground" for students who are not doing well in math and so cannot manage physics

[7] Located here: https://edsource.org/2019/how-one-high-schools-dispute-reflects-the-struggle-to-teach-californias-science-standards/618752

or chemistry, have less potential/desire for pursuing a science career, or who have special education needs. This filtering of students is obviously not the intended use of integrated science education and advocates for integrated science instruction will tell you that a better vision of an integrated science course is to prepare all students for dealing with science in their daily lives (Otarigho & Oruese, 2013).

All of the above restraints have led to integrated science curriculum not being fully applied in high schools throughout the U.S. Since high schools still offer distinct classes for each science subdomain, one might then expect large-scale science assessment data that is generated from an assessment of these subdomains to be modeled using MIRT, whereas UIRT is vastly employed as the standard.[8] By using an UIRT model the assessment designers may be impacting the assessment's consequential validity by assuming all students have the similar access to similar course content taught in a similar way.

As educators struggle to mirror the three-dimensional aspects of NGSS, so too do large-scale assessments like PISA wrestle with the challenges of modelling quantitative data. MIRT models will only be appropriate if what is anticipated by frameworks to be extensively multidimensional data actually shows such patterns. Alternatively, explaining why such data sets do not show such patterns if frameworks anticipate them is theoretically also a challenge. Educators and governments are often calling for the need to be able to assess these skills with the need to use more advanced modeling if necessary to evaluate the student data from those assessments, such as MIRT or hybrid IRT models. This study aims to first explore a framework

[8] While the goal of this study is not to determine which method of course delivery, e.g., integrated or nonintegrated, is most appropriate, the main method of course delivery in U.S high schools as nonintegrated subdomains of science may indicate multidimensionality.

for multidimensional claims, then use some data analytic and MIRT models to explore some 2015 PISA science scores from U.S. students to determine whether models can help showcase some of the implied multidimensionality. This might include how the three subdomains of the 2015 PISA science framework, i.e., physical, life, and Earth and Space systems[9], are distinct dimensions of student learning with supporting evidence from a qualitative analysis of the science subdomains in the 2015 PISA framework, or other aspects of multidimensionality. If a MIRT model substantially improves fit for some portions of the 2015 PISA science data better than unidimensional IRT models this could indicate that some of the latent trait skills are quite different from each other with regards to student ability. Not using a MIRT model when multidimensionality exists could potentially cause harm to students individually or in aggregated groups through invalid inferences being made about their ability in each subdomain (Spencer, 2004). In addition, harm could be caused to educational institutions trying to make policy decisions based on an incorrect understanding of student outcomes. These policies can be especially impactful if such decisions marginalize students of color.

History of Harm to Learning Equity by Science Assessments

Potential social impacts from how assessments are used is at the core of consequential validity (Iliescu & Greiff, 2021). Messick (1993, p. 5) defines consequential validity as an aspect of construct validity, which "appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to

[9] For those familiar with NGSS, these three subdomains fall into the NGSS dimension of core ideas. The NGSS dimensions of crosscutting concepts and science and engineering practices, such as inquiry, are outside the scope of this study. My use of the term "dimension" throughout this dissertation is meant to refer to aspects of multidimensionality in the PISA international framework such as science subdomains or other components in the framework, and not specifically to NGSS concepts, which tend to be U.S. embedded.

sources of test invalidity related to issues of bias, fairness, and distributive justice." When analyzing assessments for validity in general, Iliescu and Greiff (2021) advocate for validity to be applied to inferences, or claims made, rather than the instrument itself in order to focus on consequential validity. They further argue that more research is needed into the "social consequences of testing" as the effects of assessments on specific subpopulations and society in general directly draw a line to educational diversity, equity, and inclusion (DEI).

If the opportunity to learn hinges on the learning tools[10] available to a student, then any learning tool inequity can impact the validity of inferences made from an assessment (American Educational Research Association [AERA], 2014). Furthermore, the scores from that assessment may create additional ripples of inequity if student placement and future opportunities to learn hinge on those scores. We need to be able to redesign assessments to act as tools of equity rather than furthering existing inequities in education by assessing students on material they have not had the opportunity to learn. The Center for Professional Education of Teachers (CPET) (n.d.) suggests assessment practices can become more equitable if we:

- "Ensure our assessments align with what we actually teach
- Formatively assess students on a regular basis
- Differentiate assessment products whenever possible
- Offer a variety of ways to demonstrate mastery
- Be flexible (but not too flexible), and offer time to make up assessments
- Create relevant, engaging assessment methods

[10] Learning tools encompasses (but is not limited to) curriculum, technology, books, instructional aids, and lab equipment.

- Make assessments rigorous, not rote
- Develop and maintain a growth mindset
- Emphasize effort and progress, not grades
- Acknowledge and cultivate students' strengths and talents"

Not all of these suggestions apply to PISA and the intended use of its scores – "to evaluate education systems worldwide" in order to determine "how well students, at the end of compulsory education, can apply their knowledge to real-life situations and can therefore fully participate in society" (OECD, n.d.-a). For example, formative assessments align more with school responsibilities. However, the outcomes from some suggestions could be impacted by PISA. Due to the global nature of PISA, OECD cannot guarantee that it is aligned with what is taught in the schools of every country (OECD, n.d.-a), yet could OECD make its assessments more culturally responsive? With regards to differentiating assessments, PISA did not provide an adaptive[11] test that differentiates based on student ability for science in 2015 (OECD, n.d.-b), but is making progress on including more adaptivity (multi-stage) in later cycles in various content areas (OECD, n.d.-a). Taken together, substantial deficits in assessment equity could impact interpretation of assessments results even if the assessment's intended use is at a more aggregated level.

OECD (2016a) defines equity in education as calling for "opportunities to acquire these [science] skills[12] should be independent of students' backgrounds." PISA is used to analyze education equity through several lenses by: 1) examining "variation in the distribution of

[11] OECD does currently have an adaptive test for reading and math (OECD, n.d.-b).
[12] Students "have a basic understanding of science that will help them become informed citizens in a world shaped by science and technological progress. (OECD, 2016a)."

student outcomes, especially whether students acquire a baseline level of skills, as a way to assess the inclusiveness of school systems", 2) determining "impact of students' backgrounds on their outcomes at school", and 3) exploring if "access to educational resources and the incidence of sorting practices varies between students of different backgrounds as a way to identify some of the factors that mediate their association with performance. (OECD, 2016a)." For the U.S. 11.4% of the variation in science performance by students can be linked to their socio-economic status (OECD, 2016a). Language considerations mostly are outside the scope of this investigation, but there is substantial infrastructure in PISA within and across countries for multiple language support and translation, mostly addressed by national considerations regarding their country's students (OECD, 2017b).

In OECD's (2016a) report on *Country Note: Key Findings from PISA 2015 for the United States*, the performance and equity of the U.S. educational system is compared with those same aspects of other countries[13] that OECD has identified as showing high or improving levels. The following educational and equity conclusions (OECD, 2016a) were drawn:

- "The United States is a wealthy country."
- "The United States spends a large amount on education."
- "There is more variation in socio-economic status[14] in the United States than in the other four [comparison] countries."

[13] The four comparison countries identified by OECD are Canada, Estonia, Germany, and Hong Kong (China), which are shown in Table 2.

[14] OECD defined socio-economic status (SES) for the 2015 PISA by an index they developed from economic, social, and cultural status variables related to the family background of students (OECD, 2016a, p. 27). The variables included: parental education level, parental occupation, amount of wealthy possessions, and amount of books owned (OECD, 2016a, p. 27). The score is a composite determined from principal component analysis (PCA) and can be compared to other nations (each is weighted equally) since it is standardized to a mean of zero with a standard deviation of 1 (OECD, 2016a, p. 27).

- "The United States occupies an intermediate position in terms of the percentage of socio-economically disadvantaged students."
- "A large but not extreme[15] percentage of students in the United States have an immigrant background."
- "The United States is a large and complex country."

Another area where the U.S. experiences educational complexity is in the type of and access to science course offerings. Student enrollment in courses covering different science subdomains can vary by student ethnicity. Appendix A provides an overview of student enrollment percentages in U.S. high school science courses (biology, chemistry, and physics) by student ethnicity. Regardless of the science subdomain being taught the majority of students enrolled are White – see Figures 35-37. Therefore, it is important to consider if the inferences made based on PISA science scores can be accurately applied to other subpopulations of students. This effect could be further enlarged if the model used to determine how student ability relates to item parameters is inaccurately modeling the subdomains of science. If the U.S. bases educational policy changes or educational system reform on aspects of PISA ranking (rather than on needs of subpopulations) resulting from a model mismatch there could be unintended consequences for how PISA science data is used to inform economy, policy, and science education. An assessment can be evaluated both in terms of its use and its interpretive inferences, which can occur with probing the model for evidence of its accuracy (Messick, 1989).

[15] While OECD does not define what an extreme percentage would be, OECD does clarify that 23% of U.S. students are immigrants and that there are only 5 other OECD-member countries with a greater percent of student immigrants (OECD, 2016a).

For example, if policy makers determine that U.S. students are lagging behind other countries that are global competitors with regards to physical systems and Earth and space systems scores (see Appendix C) and earmark more funding for science education in these areas only, then the inference they are making on how to use PISA scores could be inherently flawed. Singer and Braun (2018, p. 39) describe this "mix of nationalism, fears about global competitiveness, and human nature" as leading to "unitary 'silver bullet' solutions based on highly aggregated data." This leads to an ecological fallacy, which is making inferences at an individual level, whether it be on a student's learning, a school/district's performance, or a state's educational requirements, based on population level data and can generate inaccurate conclusions. While OECD (2018) does clarify that PISA results should not be used to make inferences about individual students, it often makes inferences about school policy. As a country, the U.S. needs to clarify use of these inferences to educators, researchers, and most importantly to the media, which often discusses PISA scores incorrectly. OECD could also further clarify how survey data on school policy should be used per the AERA (2014, p. 23) standard for validity 1.3: "If validity for some common or likely interpretation for a given use has not been evaluated…that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations."

Overview of PISA

The OECD develops, administers, scores, and provides data for the PISA (OECD, n.d.-a). Results from PISA are used to rank countries by their students' mean domain score (OECD, n.d.-a). As mentioned earlier, these rankings can provide a measure for a nation's economy and are important to educational policy (Pokropek et al., 2022).

Historical Background

PISA was first administered by the OECD to students in 2000 (OECD, n.d.-a). The international program continues today, and countries can elect[16] to participate and receive information on 15-year-old students about their learning in several content areas (OECD, n.d.-a). The science assessment has been computer-based since 2015 (OECD, n.d.-b) for most countries (Jerrim et al., 2018). Students do not all receive the same items due to the assessment design (OECD, n.d.-b).

In the first cycle of PISA in 2000, 43 countries participated, including the U.S. (OECD, n.d.-a). Original participants included 29 countries that are members of OECD and 14 non-member countries (OECD, n.d.-a). PISA has since grown to include more than 90 countries and economies worldwide while serving around 3 million students (OECD, n.d.-b). The assessment is influential for policy development in some countries due to the country education ranking that OECD provides (OECD, n.d.-a). and has been linked to other national assessments in several countries, such as in the U.S. (OECD, n.d.-a).

Assessment Cycle

The "major" assessment domains found in PISA are traditionally math, reading, and science applied to "everyday activities," which usually rotate as the primary content in three-year cycles[17] (OECD, n.d.-a). For instance, science, as currently defined by OECD (see section *2015 Science Framework* below), last served as the major domain in 2015, followed by reading

[16] Student participation is also voluntary in the U.S. with an offer of a certificate of 4 hours service from the U.S. Department of Education (PISA USA, 2015). Students are chosen randomly by OECD from a list of all eligible students that is provided by each U.S. school that elects to participate with a goal of 42 students per participating school (PISA USA, 2015).

[17] Science therefore rotates as a major domain to be the focus of the assessment every nine years (OECD, 2017b).

in 2018 and math in 2021 (OECD, n.d.-a). COVID-19 has thrown off some of the administration dates such that science will next be administered as a major domain in PISA in 2025. Some content from the major domains usually reappears as a "minor" domain in PISA in each cycle to help extend trends in non-major years, along with a new innovative domain like collaborative problem solving chosen each cycle since 2012, and sometimes optional domains are delivered in some cycles, such as financial or digital literacy (OECD, n.d.-a). The degree of information available in non-major years, and the rotation of the innovative domains, have become somewhat problematic for some countries since they need information more often and for other reasons (OECD, n.d.-a).

2015 Science Framework[18]

A content framework like the 2015 PISA science framework drives the discussion around what educators worldwide might think should be taught for students to become proficient in science. There is a committee with science education experts from around the world that reviews the framework, as well as feedback from each participating country for each version of the framework. For the 2015 framework, the focus was on scientific literacy. Scientific literacy is defined as "the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen" (OECD, 2017a). OECD (2017a) claims that a comprehensive list of "all the ideas and theories that might be considered fundamental for a scientifically literate individual" has not yet been made. However, three competencies identified by OECD include: "explain phenomena scientifically", "evaluate and design scientific inquiry", and "interpret data and evidence scientifically" to provide evidence for scientific literacy (OECD, 2017a).

[18] See Appendix F for the full 2015 PISA science framework.

The framework also tries to describe what knowledge is assessable. The 2015 PISA science framework states that knowledge will be assessed if it meets the following criteria:

- "has relevance to real-life situations,
- represents an important scientific concept or major explanatory theory that has enduring utility, and
- is appropriate to the developmental level of 15-year-olds (OECD, 2017a)."

Content knowledge will comprise more than half of the assessment according to the developers (OECD, 2017a). The subdomains of science fall squarely into the framework's knowledge aspect that includes the following elements: "content", "procedural", and "epistemic" (OECD, 2017a). Content knowledge includes "facts, concepts, ideas, and theories about the natural world that science has established" (OECD, 2017a) that are historically learned in the classroom by teacher explanation and student exploration of a construct. Putting this altogether is sometimes called "sense making" in STEM. Procedural knowledge is what scientists follow to develop evidence supporting scientific knowledge via "practices and concepts on which empirical enquiry is based such as repeating measurements to minimize error and reduce uncertainty, the control of variables, and standard procedures for representing and communicating data" (OECD, 2017a). Sometimes this is considered to be "scientific practice" although defining specific practices of interest can widen and narrow this framing. Epistemic knowledge derives from "understanding science as a practice, which refers to an understanding of the role of specific constructs and defining features essential to the process of knowledge building in science" and "includes an understanding of the function that questions, observations, theories, hypotheses, models, and arguments play in science; a recognition of the variety of forms of scientific inquiry; and the

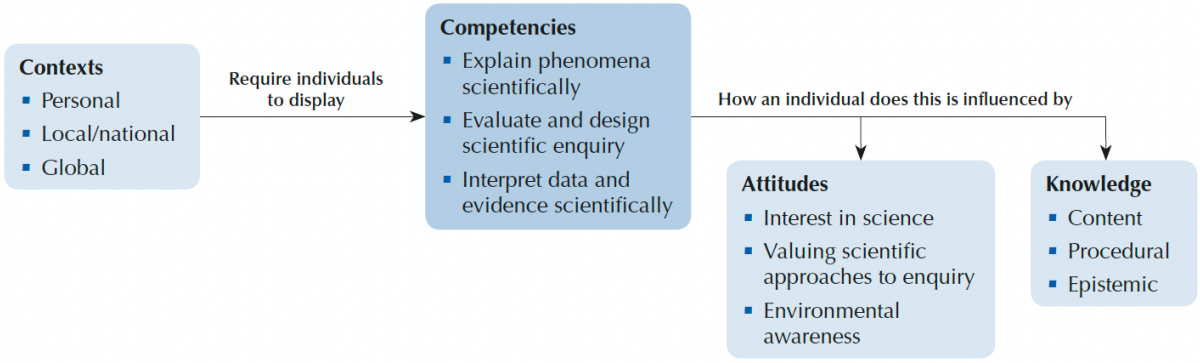
role peer review plays in establishing knowledge that can be trusted" (OECD, 2017a). Sometimes this is considered to include cross-cutting concepts, although defining specific concepts of interest can widen and narrow this framing.

In addition to knowledge, the PISA framework identifies three other aspects that it bases its view of scientific literacy on, which are: contexts, competencies, and attitudes. OECD (2017a) clarifies that the PISA science assessment "is *not* an assessment of contexts." Instead, the knowledge and competencies are assessed "*in* specific contexts" (OECD, 2017a) so that students must make sense of their STEM thinking in the applied context. Aspects and elements described above are provided in Figure 3.

Attitudes are thought of in PISA as possibly interacting with dispositions to learn or to use competencies and knowledge and may be addressed in the extensive PISA questionnaires depending on cycle and selection of questionnaire material, which are mentioned for clarity but are outside the scope of this research study.

Figure 3

Relationships among the Four Aspects



Note. Adapted from "PISA 2015 Assessment and analytical framework: science, reading, mathematic, financial

literacy and collaborative problem solving, revised edition," 2017, *OECD Publishing* (http://dx.doi.org/10.1787/9789264281820-en). Copyright 2022 by OECD.

As shown in Table 1**Error! Reference source not found.**, the science content knowledge is classified into three subdomains for this framework: life, physical[19], and Earth and space systems (OECD, 2017a).

Table 1

Three Science Subdomains in 2015 PISA

Subdomain	Code	Knowledge of the Content of Science
Physical Systems (PS)	PS1	Structure of matter (e.g. particle model, bonds)
	PS2	Properties of matter (e.g. changes of state, thermal and electrical conductivity)
	PS3	Chemical changes of matter (e.g. chemical reactions, energy transfer, acids/bases)
	PS4	Motion and forces (e.g. velocity, friction) and action at a distance (e.g. magnetic, gravitational and electrostatic forces)
	PS5	Energy and its transformation (e.g. conservation, dissipation, chemical reactions)
	PS6	Interactions between energy and matter (e.g. light and radio waves, sound and seismic waves)
Living Systems (LS)	LS1	Cells (e.g. structures and function, DNA, plant and animal)
	LS2	The concept of an organism (e.g. unicellular and multicellular)
	LS3	Humans (e.g. health, nutrition, subsystems such as digestion, respiration, circulation, excretion, reproduction and their relationship)
	LS4	Populations (e.g. species, evolution, biodiversity, genetic variation)
	LS5	Ecosystems (e.g. food chains, matter and energy flow)
	LS6	Biosphere (e.g. ecosystem services, sustainability)
Earth and Space Systems (ESS)	ESS1	Structures of the Earth systems (e.g. lithosphere, atmosphere, hydrosphere)
	ESS2	Energy in the Earth systems (e.g. sources, global climate)
	ESS3	Change in Earth systems (e.g. plate tectonics, geochemical cycles, constructive and destructive forces)
	ESS4	Earth's history (e.g. fossils, origin and evolution)
	ESS5	Earth in space (e.g. gravity, solar systems, galaxies)
	ESS6	The history and scale of the universe and its history (e.g. light year, Big Bang theory)

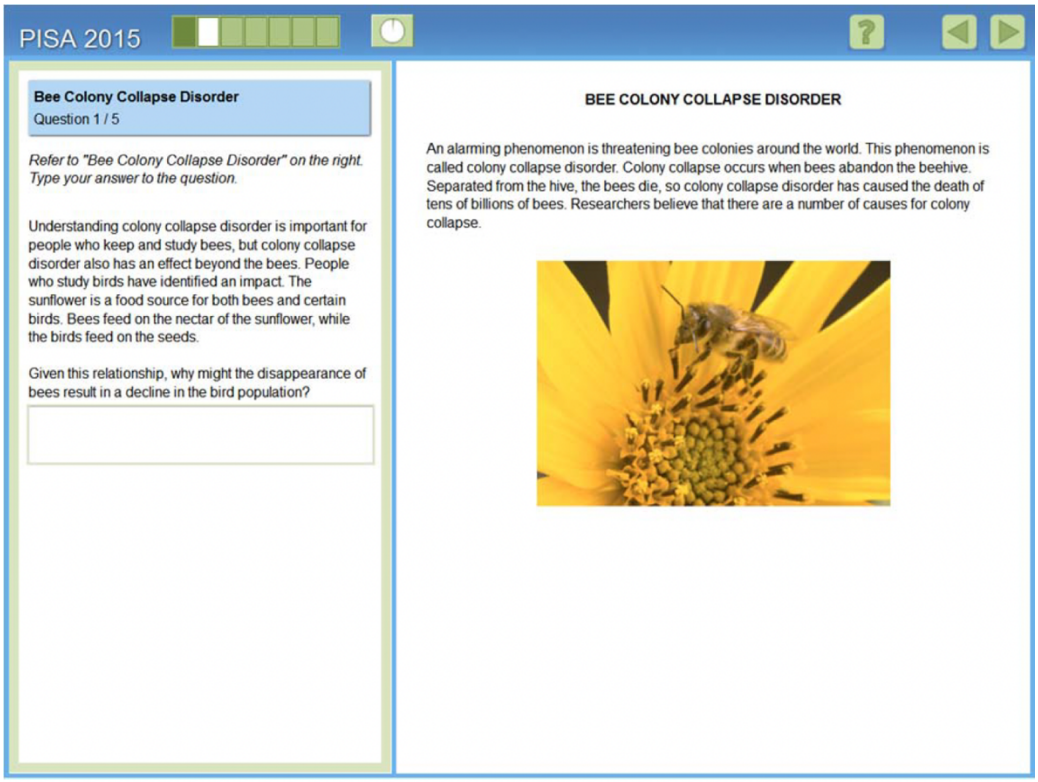
Note. Adapted from "PISA 2015 assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving, revised edition," 2017, *OECD Publishing* (http://dx.doi.org/10.1787/9789264281820-en). Copyright 2022 by OECD.

[19] The physical systems subdomain seems to encompass both physics and chemistry constructs.

The 2015 PISA science framework further clarifies that in the assessment approximately 36% of items will be physical, 36% of living, and 28% of Earth and space systems (OECD, 2017a). An example of a released science item with a focus on life science is provided in Figure 4.

Figure 4

Released 2015 PISA Science Item[20]



Item Number	CS600Q01
Competency	Explain Phenomena Scientifically
Knowledge – System	Content – Living
Context	Local/National – Environmental Quality
Cognitive Demand	Medium
Item Format	Open Response – Human Coded

[20] The correct response needed to reference "a flower cannot produce seed without pollination" (OECD, n.d.-c).

Note. From "PISA 2015 Assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving, revised edition," 2017, *OECD Publishing* (http://dx.doi.org/10.1787/9789264281820-en). Copyright 2022 by OECD.

Items were delivered in the following contexts: "health, natural resources, the environment, hazards, and the frontiers of science and technology (OECD, n.d.-c)." The contexts were set in "personal, local/national, and global settings (OECD, n.d.-c)." Items were developed to meet one of the following three cognitive demands[21]:

1. "Low - Carry out a one-step procedure, for example recall of a fact, term, principle or concept, or locate a single point of information from a graph or table.
2. Medium – Use and apply conceptual knowledge to describe or explain phenomena, select appropriate procedures involving two or more steps, organize/display data, interpret or use simple data sets or graphs.
3. High - Analyze complex information or data, synthesize or evaluate evidence, justify, reason given various sources, develop a plan or sequence of steps to approach a problem (OECD, n.d.-c)."

OECD (n.d.-c) defined theoretical difficulty for an item as "a combination both of the degree of complexity and range of knowledge it requires and the cognitive operations that are required to process the item."

2015 Assessment Design[22]

[21] These definitions were newly added to the 2015 PISA framework for the scientific literacy domain (OECD, n.d.-c)
[22] The complex design of test forms, item types, and sampling techniques used to calculate a country's rank are outside the scope of this study. Information provided in this section is to help orientate the reader to the higher-level details of the design of the science assessment. Complete details regarding assessment design are available in the PISA 2015 technical report (OECD, 2017b).

In terms of educational theory behind the PISA there is one unifying goal. PISA is intended to be:

> A collaborative effort among OECD member countries to measure how well 15-year-old students approaching the end of compulsory schooling are prepared to meet the challenges of today's knowledge societies. The assessment is forward-looking: rather than focusing on the extent to which these students have mastered a specific school curriculum, it looks at their ability to use their knowledge and skills to meet real-life challenges. This orientation reflects a change in curricular goals and objectives, focusing more on what students can do with what they learn at school. (OECD, 2017b, p. 22)

A "real-life challenge" faced by an individual in life might be the COVID-19 examples given previously. Such challenges would most likely include integrating content knowledge since rarely do we find successful scientific solutions in a vacuum. For example, for those in STEM careers, biologists rarely describe cellular processes without integrating organic chemistry knowledge. Physics is often used to describe the motion of animals and how planets align. However, as noted earlier, the integrated course model is not how science education is designed – rare is the course that incorporates both life and physical science in a K-12 setting. Neither are PISA or many other large-scale science assessments designed to be integrated with items assessing multiple science subdomains at once. The science assessment design includes items that are targeted to specific subdomains, Figure 4 (Bee Colony Collapse Disorder Question[23] 1), which is coded to *knowledge – system: content – living science* (OECD, 2017a). A

[23] Question # refers to a specific item in the order it appeared in a cluster.

lack of integrated content in items may negate some of the “forward-looking” aspects of PISA, although a fuller theoretical investigation of this topic is outside the scope of this dissertation.

The cognitive assessment (see Figure 5) included an additional domain of collaborative problem solving in 2015 (OECD, 2017b). Some of the countries took a paper-based assessment (PBA) version while others took a computer-based assessment (CBA) (OECD, 2017b). The CBA version of PISA was considered a field trial and included both trend and new items while the PBA version only had trend items (OECD, 2017b).

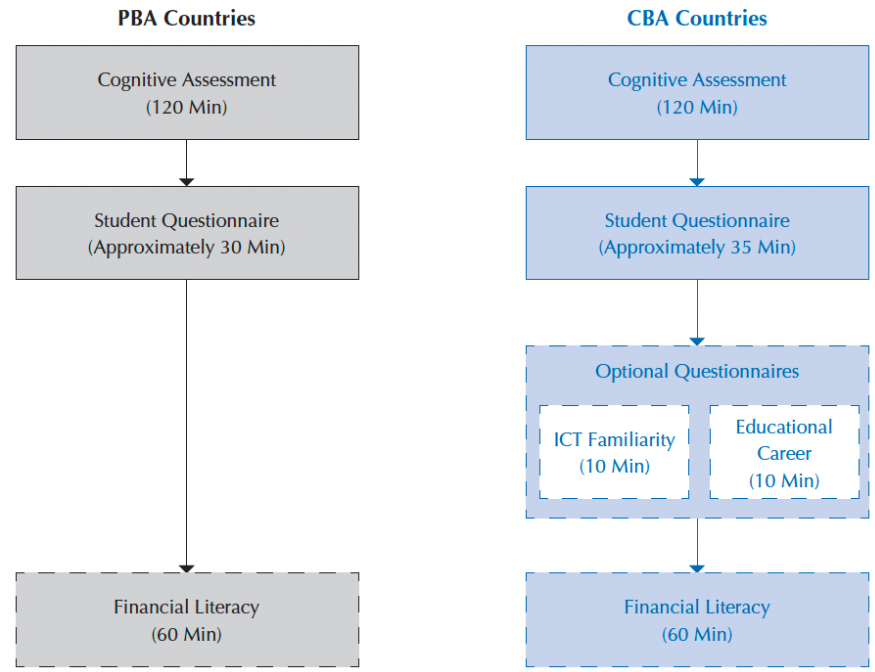
Items had several response formats: click on a choice, numeric entry, text entry, select from drop-down menu, and drag and drop and were dispersed among multiple test forms (OECD, 2017b). There was no common form with a common set of items that all students received (OECD, 2017b). Forms were randomly assigned (OECD, 2017b), with no linking information released. Within forms, the science items were organized into clusters (OECD, 2017a). There were 36 random science clusters combinations possible across the 66 CBA forms (OECD, 2017b). There was also a unique form referred to as the Une Heure (UH) form and is for students with special needs. This form included “easier items in each domain” with “a more limited reading load” and 50% of the items assessing science (OECD, 2017b, p. 42).

Approximately 61 physical items, 74 life items, and 49 Earth and space items for a total of 184 items were developed and chosen for the assessment, which is equal to 6 hours of test questions – only 85 were trend while 99 were new (Mostafa et al., 2018). However, the actual test was about 2 hours per student. As shown in Figure 5, students have 2 hours to take the cognitive portion of the assessment, which includes the major domain of science and the minor domains of reading, mathematics, and collaborative problem solving, and then they are offered

a questionnaire[24] with a shorter innovative assessment on financial literacy at the end (OECD, 2017b). The major domain, science for 2015, takes an hour of the time allotted to both the PBA and CBA cognitive assessment to complete (OECD, 2016b). Teachers had an optional short questionnaire that could be taken after the student questionnaire (OECD, 2017b). All elements of the assessment were offered in different languages according to PISA specifications to accommodate some language aspects of the participant's setting.

Figure 5

Comparison of PBA[25] and CBA[26] Assessment Designs



Note. From "PISA 2015 technical report," 2017, *OECD Publishing* (http://dx.doi.org/10.1787/9789264281820-en), p. 36. Copyright 2017 by OECD.

[24] OECD often refers to the cognitive assessment as a survey while referring to the student and teacher qualitative surveys as questionnaires in both the framework and technical report.
[25] While PBA countries were offered the optional financial literacy assessment none took it (OECD, 2017b).
[26] ICT refers to Information and Computer Technology Literacy Familiarity (ICT) questionnaires.

2015 Science Scoring

While items are described in the framework, these do not generate individual scores per student that are visible to the public (OECD, n.d.-b). Per OECD (2018), nearly 540,000 students globally finished the science assessment. There is not a theoretical minimum or maximum score for each of these students (OECD, n.d.-b). In this study's data, however, item level responses for students within the U.S. sample were used.

In the reported data[27] at the country level, scaling occurs with IRT and then is transformed for scores around normal distributions (OECD, n.d.-b). Means for OECD country participants are approximately 500 score points with a standard deviation (SD) of 100 score points (OECD, n.d.-b). Countries are then ranked according to the mean score (OECD, n.d.-b). Most students score between 400 and 600 points (OECD, n.d.-b).

Having mentioned earlier that PISA scores can be indicators of a country's education status and economy, it's important to consider that not all educators and researchers believe that PISA scores are good indicators of these factors (Strauss, 2019). Some educators feel that there is not a one-size fits all assessment and that assessment should be more inquiry-based where students actually show what they can do, which is only partially assessed on the PISA assessment from 2015. In addition, the scores reflect students' performance in different countries. OECD attempts to take into account cultures and policies in those countries but has not been able to fully resolve that these differences promote different knowledge, skills, and abilities (Strauss, 2019). This is partly what the OECD assessments are intended to consider, but in a large-scale assessment it is difficult to separate issues like cultural relevance from policy

[27] See Appendix B for science mean scores by country and Appendix C for science mean subscale scores by country.

variations that the countries feel might be important to detect. To better understand scores within the context of the science subdomains, a qualitative analysis of student level data, which is not possible given the nature of the PISA data and lack of access to all science items, might complement a quantitative analysis of 2015 PISA science scores. Instead, I am trying to identify conceptual multidimensionality via conceptual analysis of the framework with document analysis, followed by quantitative analysis of the student-level data set from the instrument representing that framework, to see if empirical evidence of patterns of quantitative multidimensionality exists.

Three MIRT Case Studies

The following studies exemplify the current state of MIRT model usage in assessment research and are similar to what was implemented in the methodology chapter of this study. Two of the studies provide original research addressing the use of MIRT in science assessments with only one being a large-scale case, while the third case study focuses on a large-scale English language proficiency assessment. All three are united in their use of a methodology to validate the use of MIRT models by comparing fit to other IRT models and by a theory that their educational constructs are expected to be multidimensional.

Yen and Leah (2007) - MIRT Model for Composite Scores

Similar to the PISA 2015 reading framework, the knowledge content of English language proficiency assessments is often multidimensional. The subdomains for this English language proficiency assessment were identified as speaking, listening, reading, and writing, but the researchers were interested in mainly the speaking and listening constructs. The speaking construct was further separated into four subsequent proficiencies while the listening construct

had three subsequent proficiencies. These proficiencies were considered as sub-subdomains in the analysis. Twenty items each were administered for the speaking and listening subdomains. Half of the speaking items were scored polytomously, but the rest were dichotomous, and all of the listening items were dichotomous. Students took the speaking portion of the test in about 10 minutes and the listening portion in about 15 minutes.

The assessment itself was considered large-scale because it was given to all eligible K-12 students in an unidentified location that was state, or country sized. However, the researchers pulled a sample of 12,008 student responses from only elementary school students. Only full sets of responses were analyzed as some students did not complete all of the items. The sample included slightly less female students than male students while one unidentified ethnic group dominated the sample.

The unidimensional IRT models that were analyzed included the 3PL and the two-parameter partial credit model. Calibrated together, both multiple choice and constructed response items were placed on the related subdomain scale. Marginal maximum likelihood was used to estimate the parameters at the same time for both item types. Importantly, the speaking and listening items underwent distinct calibrations, but an additional calibration was performed on these subtests together to build an “oral” composite scale.

Next an exploratory factor analysis was conducted to determine if the sub-subdomain proficiencies were actually multidimensional and to identify the number of dimensions that were present. Using the BMIRT computer program, four different MIRT models were then applied. One of the models [Model Type 1] assumed that speaking and listening subdomains are combined to measure one latent trait with several sub-subdomains while the other three

models [Model Type 2] assume that the speaking and listening subdomains measure distinct latent traits, each with their own set of sub-subdomains. The researchers pointed out that MIRT models can have parameter estimation issues due to the number of parameters, but the Markov Chain Monte Carlo (MCMC) method used by BMIRT helped alleviate this. Population parameters were fixed as a normal distribution to the mean of 0 and standard deviation (SD) of 1 for Model Type 1 and fixed as multinormal distribution with mean (0,0) for Model Type 2. Each model went through 10,000 iterations. Each model's fit was compared using the Akaike Information Criterion (AIC) and chi-square difference test. The Type 2 MIRT models had the better fit and the researchers concluded that multidimensionality existed in the assessment. The MIRT models proved successful even though assessments less than 100 items in length can have trouble with the discrimination parameter. Scalise and Clarke-Midura (2018) also had success in applying a MIRT model to assessment scores based on a multidimensional framework.

Scalise and Clarke-Midura (2018) - The Many Faces of Scientific Inquiry

The NGSS is a multidimensional framework used, with some adaptions, by over 30 states in the U.S. The researchers looked at whether an online/virtual performance task aligned with the Framework for K-12 Science Education (National Research Council [NRC], 2012), College Board Standards for College Success (College Board, 2009), and the NGSS inquiry practices and science content knowledge delivered evidence on more than one latent trait using a MIRT-Bayes model. The nature of scientific inquiry done by students is described as complex and often has to be less regulated to measure the full reasoning of students, which indicates that when performing inquiry students may use multiple abilities.

Items, both polytomous and dichotomous (multiple-choice), were developed by Harvard University learning scientists working in STEM innovations for two dimensions, inquiry and explaining. A sample of 1,986 student response were collected for 23 items. Less than 1% of students had missing data. Information on gender and ethnicity of the student sample was not provided. The entire assessment took students nearly 40 minutes to complete and process data on student actions was also collected during this time.

To model the data, researchers used an exploratory study design that included both unidimensional and multidimensional IRT models. MIRT models include a 2-dimensional partial credit and a hybrid-MIRT-Bayes model. For the 2-dimensional MIRT model the difficulty parameter was estimated freely once the means of the latent variables were set to zero. In the case of the hybrid MIRT model, a Bayes net was constructed first then the MIRT model was applied. The Bayes nets can help structure semi-amorphous data onto the constructs being investigated. Upon analysis, the unidimensional model was not as well fit to the data based on the significant deviance difference while item fit was acceptable for the 2-dimensional MIRT model. In addition, the two latent traits of inquiry and explaining were just moderately correlated, indicating student performance might be varied enough to justify the use of MIRT based on the concept framework, although an even larger scale sample could help show this more definitively. Both dimensions had expected *a posteriori* (EAP) reliability coefficients greater than 0.85. Researchers elaborated that theoretical support from this framework was validated by content experts, who designed the task to have items that would assess each ability/skill independently. In counterpoint to the Yen and Leah (2007) study, no items were found to load negatively, which would indicate negative discrimination. The hybrid MIRT model

appeared to fit even better with higher reliability estimates. Wright maps and standard error of measurement plots helping to visualize the data. Similarly, Li et al. (2012) explored how MIRT models can accurately depict data from items that cover several science domains.

Li et al. (2012) - Applying MIRT Models in Validating Test Dimensionality

Assessment dimensionality should be explored at the beginning of development to provide validity for the inferences made with regards to student ability/latent traits. Development of items is done with alignment to different anticipated dimensions so that each item targets one or more constructs. The researchers define dimensionality of the assessment "as the number of traits that must be considered to achieve weak local independence between the items" (Li et al., 2012, p. 3). The covariance between items should "approach zero as test length increases" to indicate weak local independence (Li et al., 2012, p. 3). The validation of dimensionality with regards to the learning domain or subdomain often occurs during the field test.

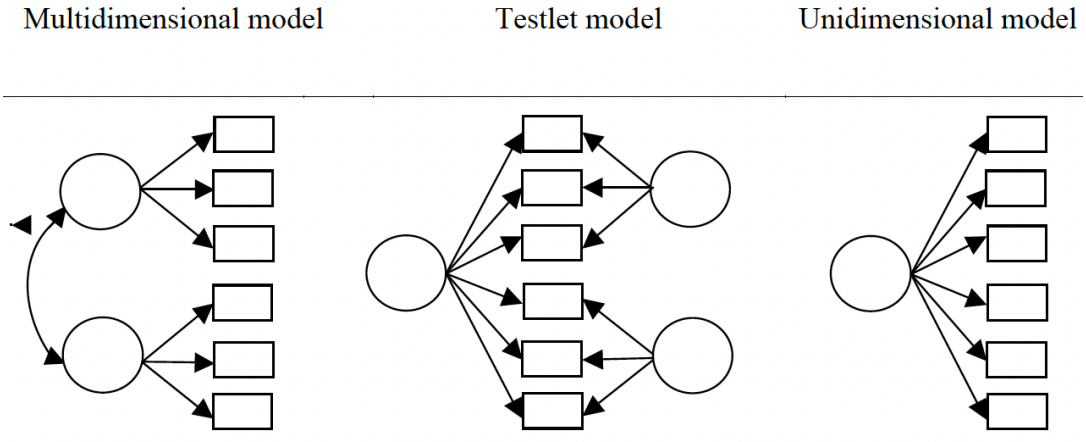
A random sample of 5,677 grade 5 students who had taken the 2008 Michigan state science assessment was selected in order to validate the dimensionality of four science subdomains: science processes, life, Earth, and physical sciences. A total of 45 science items were taken by students with no timeline given. Item type was not provided by the researchers, and it is also unclear if the items were field tested or operational. Gender and ethnicity of the students also was not declared.

To evaluate the dimensionality of the student data, the researchers selected three different IRT models: a unidimensional model, a "simple-structure" MIRT, and a testlet model. The testlet model treats different subdomains as different testlet residual dimensions to the

dominant dimension of general science ability. Figure 6 provides a general structure for each model that was evaluated.

Figure 6

Comparison of Models



Note. From "Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large-scale science assessment," by Y. Li, H. Jiao, and R. W. Lissitz, 2012, *Journal of Applied Testing Technology*, *13*(2), p. 9. Copyright 2012 by Journal of Applied Testing Technology.

Before applying the three models, a principal component analysis (PCA) and exploratory linear factor analysis (EFA) were conducted. Researchers selected eigenvalues greater than 1, which indicate those components account for more than mean total variance of items, in the PCA. A scree plot helped the researchers determine that at least two components did exist. Next the EFA confirmed that two factors existed and based on the reduction of root mean square error (RMSE) values either a four- or five-factor solution might be possible to adopt. A four-factor solution was used as the researchers felt that the statistical analysis should be compared against the theoretical dimensionality used in the assessment design.

The MIRT model was found to significantly fit the data better than the unidimensional model. AIC and Bayesian information criterion (BIC) were then used to compare fit between the MIRT model and the testlet model. The MIRT model was found to be the best fitting since it produced smaller information criteria indicating better model fit to the data. The conclusion reached was that if several abilities are measured then MIRT models should be used. Li et al. (2012) pointed out that at the time of their study no prior study on validating multidimensionality was located comparing the fit of the three models for a large-scale K-12 science assessment.

The study outlined in Chapter 2 is similar in methodology to the above 3 studies with a focus on multidimensionality of science assessments. There continues to be a lack of research on large-scale science assessments and their use of MIRT models, and few that also include other data analytic techniques from emerging data sciences such as shown here. This study will include analysis of data from the 2015 PISA large-scale science assessment.

By modeling data from PISA's 2015 science framework[28] and assessment, the following research questions are expected to be answered.

Research Questions

Research Question 1 (RQ1)

What evidence can be drawn from the 2015 PISA science framework to qualitatively support whether multiple dimensions are being described regarding student knowledge in science in the framework?

[28] See Appendix F.

Research Question 2 (RQ2)

Do quantitative indications of multidimensionality exist in the 2015 PISA science data?

- R2A: Does a data science cluster analysis[29] applied to the PISA 2015 data (U.S. sample at the item level) suggest multidimensionality in the student response data set? Regardless of dimensionality, how do the items cluster in the analysis?
- R2B: Does principal components analysis[30] (PCA) applied to the same data indicate multidimensionality? Does residual analysis employing standard tolerances used in the industry indicate multidimensionality that needs to be modeled in the U.S. sample? How do items cluster in the analysis, and how does this compare to R2A?
- R2C: By how much (practical significance) does a multidimensional IRT model applied to the same data indicate improved fit over less dimensionally complex models? Is fit statistically improved based on chi-square comparisons of nested models fitted for dimensionality? How does item clustering compare in R2A/R2B?

Research Question 3 (RQ3)

Depending on covariates available in the U. S. sample data set or linked data sets or overall population reports, what can be said about subgroup analysis in this data set and about aspects of classroom instruction relative to clustering patterns found in R2A-C? For instance, do multidimensional models yield results that showcase students have different levels of science

[29] A cluster analysis is a method for sorting items into groups based on their statistical relationship to one another. For example, if items are closely related because a student must have physic knowledge to answer them then all physics items may cluster together.

[30] PCA was chosen since it is a method of reducing the dimensionality of a large dataset into principal components (or dimensions) that retain as much variation as possible (Mailman School of Public Health, 2023). While EFA was considered, its goal is to show the correlation between variables is partially due to common latent variables (Mailman School of Public Health, 2023), which is not the dimensionality predicted for the science subdomains.

proficiency on different dimensions identified? Relative to demographic data, if available, what can be said about history of harm and employing or ignoring dimensionality in science data such as these? Whether contrasting or similar, what can be said about the findings viewed from both lenses?

CHAPTER 2. METHODS

Declaration of Interest: The author worked at one time but not now for Educational Testing Service (ETS), one of the vendors supporting some PISA efforts, and has also been the lead vendor of 5 companies developing, delivering, and analyzing NAEP in the U.S., for all domains including for NAEP Science.

Developing the Literature Synthesis

A literature review helped determine the state of research on multidimensionality in large-scale science assessments. The majority of searches were in Google Scholar and the University of Oregon library databases but were not limited to only peer-reviewed journals. The main search used the following key phrases and words: "multidimensional" + "IRT" + large-scale + "science" + "assessment"[31,32]. For the purposes of this literature review, the following definitions were used:

- Multidimensional IRT – a model estimating student ability containing more than one latent trait
- Large-scale Science Assessment – assessments measuring the science ability of a large proportion of the student population occurring at either the U.S. state or national level, at the country level for areas outside of the U.S., or at a global level

Nearly 7,560 results were generated in Google Scholar using the combined search phrase. In each iteration of searching, the Li et al. (2012) article was the top hit with very few

[31] Quotation marks were included to indicate for which terms the Google Scholar search platform should return articles with exact matches.
[32] Grade level and science subdomain were not used as eliminators.

other results matching the specific target. See Appendix D for Figure 39, which provides an overview of literature connected to the Li et al. (2012) study on multidimensional IRT model fit for student data from a large-scale state science assessment. Table 15 in Appendix E provides annotations for resources found during the literature review and used in the literature synthesis.

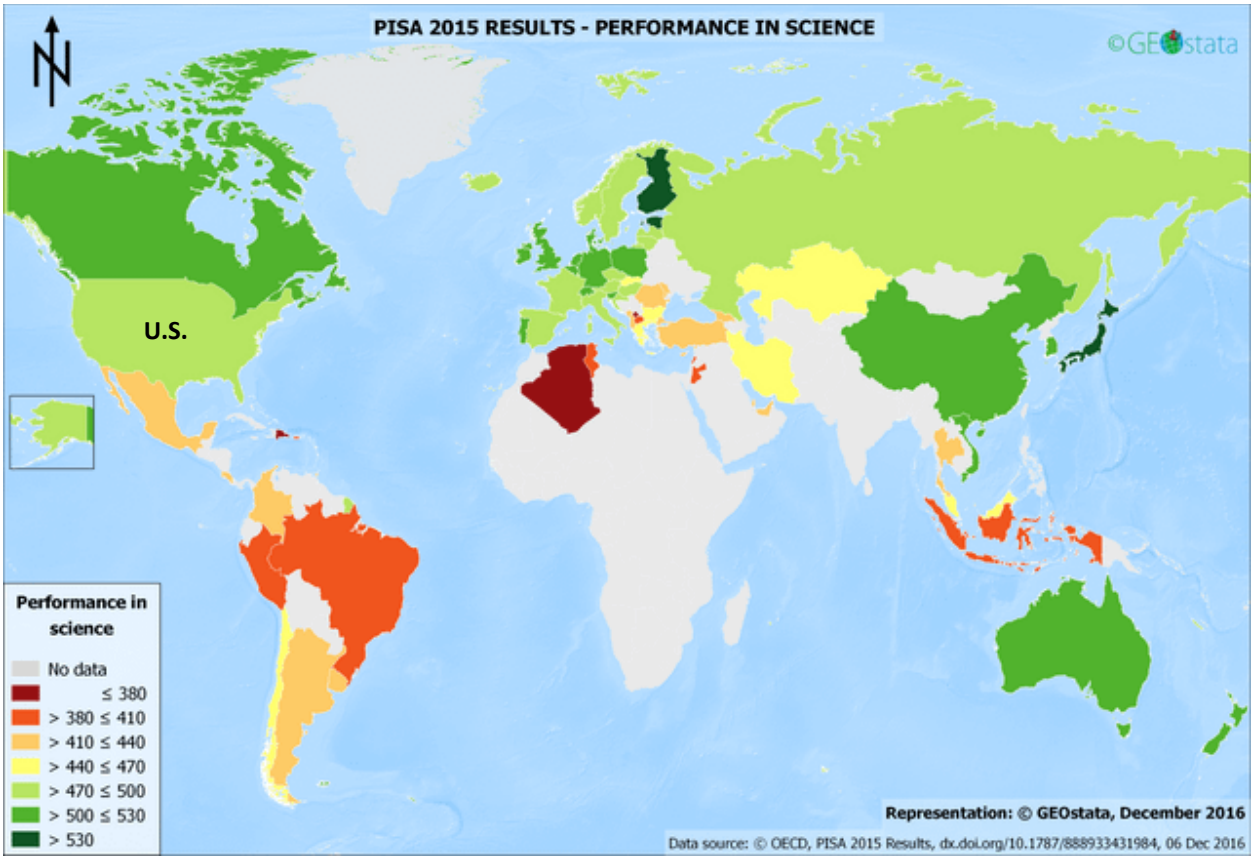
Additionally, the review focused not just on science education, but also on reading assessments that used MIRT to model dimensionality. This addition was made in part due to the lack of studies dealing specifically with multidimensionality of large-scale science assessments, but also because PISA declares reading to be multidimensional in the 2015 framework (OECD, 2017a). Therefore, reading offers somewhat of a mirror into multidimensionality that may be useful through which to view the science framework.

Setting

PISA is an international CBA and PBA that is available to all OECD countries or affiliates called partners (OECD, 2017b), and countries around the world are invited to partner if they are not already in the OECD. For 2015, the non-gray countries in Figure 7 chose to participate (GEOstata, 2016). The color scale illustrates how these countries ranked by their students' performance on the science assessment (GEOstata, 2016). The setting focus for this study will be the U.S. Note that in Figure 7 the U.S. is shown ranking in the 470 to 500 mean score range for science (GEOstata, 2016), which is near to but slightly above center.

Figure 7

PISA Science Performance by Country



Note. From "PISA 2015 Results – Performance in Science," 2016, GEOstata. Copyright 2015 by OECD.

Countries began testing in April 2015 in the following types of educational settings: educational institutions, vocational training or related educational programs, and foreign schools within a country (OECD, 2017b). Due to the range of countries participating in PISA the student population is very diverse.

Student Demographics

Around 540,000 global students completed the PISA in 2015 (OECD, 2018). Both full-time and part-time students were eligible to participate (OECD, 2017b). Students that were

home-schooled or taught in the workplace were not eligible to take the PISA (OECD, 2017b). Table 2 provides a breakdown of students in the U.S. compared to the highest performing and lowest performing countries. This table also includes demographics for the four countries to which the U.S. is compared by OECD (2016a). OECD member countries who are highlighted blue, and the remaining countries are OECD partners (OECD, 2017b).

Table 2

Country Demographic Comparisons

Country[33]	Student Sample Size (OECD, 2017b)	2015 Population Size[34] (in millions)	2015 Duration of Compulsory Education[35] (in years)	2013 Ethnic Fraction-alization[36]	Dominant Language/s[37,38]	PISA 2015 Mean Science Score[39]
Singapore	6,115	5.45	6	38.57%	*Mandarin (35%), *English (23%), *Malay (14.1%), Hokkien (11.4%) (2000 census)	556
Estonia	5,587	1.3	9	50.62%	*Estonian 67.3%, Russian 29.7% (2000 census)	534
Canada	20,058	35.7	10	71.24%	*English 58.8%, *French 21.6%, Other 19.6% (2006 Census)	528
Hong Kong (China)	5,359	7.3	9	6.2%	*Cantonese 90.8%, *English 2.8%, (2006 census)	523
Germany	6,522	81.7	13	16.82%	German[40]	509

[33] Countries, and in the U.S. the schools, elect to participate in PISA (NCES, n.d.-a).
[34] From https://datatopics.worldbank.org/world-development-indicators/
[35] From https://data.worldbank.org/indicator/SE.COM.DURS
[36] From https://worldpopulationreview.com/country-rankings/most-racially-diverse-countries
[37] From https://www.languagerc.net/languages-by-countries/
[38] Showing only those unofficial languages that are found at 10% or greater within the country's population.
[39] From https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf
[40] No census data provided.

United States	5,712[41,42]	320.7	12	49.01%	English (82.1%)[43], Spanish (10.7%) (2000 census)	496[44]
Dominican Republic	4,740	10.4	15[†]	42.94%	*Spanish[45]	332

Note. A Harvard study defined fractionalization as the probability that two people randomly selected from a country would be from different ethnic groups (Alesina et al., 2003). [†] This country's education requirement changed from 9 to 15 in 2010 – the other six countries have remained steady in their education requirements since the late 1990s through 2022. * Indicates official language/s of that country.

Even though slightly over half a million global students participated in the 2015 PISA, OECD did not select all of those students in their sample (OECD, 2017b). The U.S. alone had 4,220,325 15-year-olds in 2015 with 3,992,053 actually enrolled in school (OECD, 2017b). A few U.S. schools[46] (12,001) chose not to participate in 2015 at an exclusion rate of 0.30% (OECD, 2017b).

Data Collection

Data were collected during the 2015 PISA administration[47] then organized, analyzed, and reported by OECD in 2016-18. Student responses to each PISA form were collected online when students accessed the task through their computer or tablet and in a physical format when students took the assessment via a paper form. In addition to student responses on science content, formal and optional questionnaires collected information on student attitudes

[41] NCES (n.d.-a) reported 177 schools participated nationally in the U.S. with a student response rate of 90%.
[42] This is the national sample size and does not include the two states (Massachusetts/North Carolina) and one territory (Puerto Rico) sampled.
[43] While English is broadly used in the U.S., the U.S. has not designated an official language (USAGov, 2023).
[44] The U.S. mean score is not significantly different from the overall OECD mean score of 493 (OECD, 2018).
[45] No census data provided.
[46] As of February 2024, there were 20,318 high schools in the U.S. per https://mdreducation.com/how-many-schools-are-in-the-u-s/. Per NCES (n.d.-a), there were 27,144 public and private secondary and high schools in the 2015-2016 school year, but the national U.S. sample consisted of only 240 schools and not all of those participated.
[47] For the U.S., this occurred during October to November (NCES, n.d.-a)

toward science, along with student and teacher educational backgrounds. Student in-person interviews, cognitive lab data, focus group data, pilot and field trials were in many cases collected by OECD (2017b) but not released due to policy requirements with the countries.

Study Sample

For this study, the full PISA 2015 extant quantitative dataset was narrowed by selecting only the U.S. student population at the national, not state, level that responded to a CBA form[48]. This study used a science subset from that sample. The science subsample consisted of 5,712 students and 166 dichotomous items. Removal of students that lacked response to any item (so they had all NAs[49] for all 166 items) dropped the sample to 5,699 students. These 13 dropped students were considered as missing data at 0.2% for the U.S. CBA science subsample. OECD also performed casewise deletion for missing data (Mostafa et al., 2018).

There was no common form between all the students in 2015 PISA science. Items were spread across various forms as clusters of around 15 items each, with no equating sample or linking information available. Hence the item cluster S10[50] with the highest response rate was chosen for the first analysis, reducing the sample to 1,306 students and 15 items, all of which were new in 2015. As a confirmation of the model fit findings from the first analysis, a second analysis was conducted as a validity study, examining the second largest cluster, S11[51]. This was

[48] CBA was chosen as this is the mode of delivery most assessments are moving towards, including PISA. Also, eliminating the PBA forms allowed the subsample to be viewed through one mode lens rather than having possible effects on the data from different delivery modes.
[49] NA is the abbreviation for "not applicable"; however, when used in a data file it can mean many things. Note that the R program automatically codes blank/missing cells as NA. Per OECD (2017b), a cell coded NA by R in the PISA data file can mean missing data, an item not reached by a student, a data error, or skipped item.
[50] Cluster S10 as noted in the OECD (2017b) technical report's annex A had 17 items, but the 2 polytomous items were dropped from this study.
[51] Cluster S11 as noted in the OECD (2017b) technical report's annex A had 16 items, but the 1 polytomous item was dropped from this study.

identified as consisting of a subsample of 1,274 students and 15 items, all of which were new in 2015. Criteria used here for examining clusters was to select two of the 15 item clusters for an initial and then a validity study. Given the scope of this work as an unfunded dissertation, the study started with the largest sample size to help facilitate separately fitting the more complex models, and then selected the next largest.

Students in these cluster subsamples with some NAs indicating no response to an item during their attempt at answering the cluster had those NAs converted to 0 since they had the opportunity to attempt the item. There were no missing data in either of the science cluster subsamples using these criteria after the elimination of the 0.2% of U.S. responses indicated earlier. Hence, missing data rates were quite low, but they were not entirely eliminated. No imputation technique was undertaken for the 0.2% missing data due to insufficient additional information released. Also, numerous problems of imputation in assessment calibration made this a questionable statistical adjustment even if information had been available.

Data Analysis – A Mixed Methods Approach

Data analysis was divided into three steps. The first step directly supported RQ1 and involved qualitative data being analyzed, including the main document analysis of the 2015 PISA science framework. Step 2 supported RQ2 and consisted of the quantitative analysis[52] of 2015 PISA science student scores along with data triangulation to the Step 1 analysis. The last step supported RQ3 and involved analyzing with an equity lens how the best fitting IRT model might impact equity in regard to student outcomes for subgroups. All three steps are outlined in detail below within the context of a mixed methods approach to data analysis.

[52] Focus of this study is on the item level rather than the student level for each quantitative analysis.

When qualitative data exists to support, or contradict, quantitative data, a mixed methods design for research is essential. This type of design can meld findings from both methodologies into an effective solution (Johnson & Onwuegbuzie, 2004). For example, large-scale assessments like PISA often make available content frameworks, student and teacher survey responses, and other documentation like item specifications that can be analyzed qualitatively to determine narrative themes such as multidimensionality of science content. This is in addition to the quantitative scoring data, process data, and student demographic data from the assessment itself. Such narrative themes might add to a quantitative model of student science ability. For example, Claesgens et al. (2008, p. 66) had a mixed methods approach of "using IRT, the scores for a set of student responses and the questions are calibrated relative to one another on the same scale and their fit, validity, and reliability are estimated, and matched against the framework". In order to support such a mixture of data types and analyses, a researcher often needs a methodological design that incorporates different philosophical underpinnings.

Epistemology

An equivalent status design refers to a theory of research where both qualitative and quantitative epistemologies are valued for understanding constructs (Venkatesh et al., 2016; Johnson & Onwuegbuzie, 2004). This study used two epistemological foundations to pragmatically agree with Baskarada and Koronios (2018) that researchers should select philosophies that work best for the research questions to be addressed and the data to be analyzed. This choice was pragmatic in nature based on the requirements of this study. The chosen epistemological foundations were also aligned with Greene and Caracelli (1997), who

clarify that if methodological pragmatism requires different philosophies then each can be viewed as “logically independent and therefore can be mixed and matched” in order to achieve methodology that will work well in each stage of analysis. In the end, though, RQ1 and RQ2 were designed to consider triangulation, or in other words whether results tend to support each other or not across the techniques.

My epistemological approach to science education is mainly social constructivism. Constructivism is a viewpoint that students build their own learning with their reality built on experiences as a learner and the social aspect implies that learning is collaborative. Social constructivism describes students as dynamically constructing and reconfiguring knowledge while interacting socially (OECD, 2023). This educational theory is grounded in constructivism psychological learning theory, which views knowledge as something a learner must “actively construct” for themselves (McLeod, 2019). During social interactions, such as learning within a classroom, students often construct their knowledge in collaboration with others (Atkisson, 2010). The concept that students learn how to learn by interacting with others is a key foundation of social constructivism (Greenwood, 2020). Western Governors University (2020) describes several principles of constructivism:

- “Knowledge is constructed,
- people learn to learn as they learn,
- learning is an active process,
- learning is a social activity,
- learning is contextual,
- knowledge is personal, and

- motivation is key to learning."

Boon et al. (2022) also point out that a constructivist epistemology can suit "practice-orientated" research.

I have also chosen an approach that acknowledges a philosophy of scientific realism, or an organized reality (Moroi, 2020), which means that at some level I am acknowledging that there is a reality in STEM and it can be known (although in STEM, most if not all models can be disconfirmed to some degree at a different grain size, so the idea of scientific realism is that the philosophy acknowledges utility of the model at the grain size applied, such as in Atomic Theory and Quantum Physics). Similarly, I believe that we can try to measure concepts both independently of ourselves and in a rational manner, at a level that may provide some utility for decision making. This viewpoint pairs well with my social constructivist stance and as Maxwell and Mittapalli (2010) note these two philosophical paradigms can work well in a mixed methods approach. If multidimensionality is present in the 2015 PISA science framework and student data then it should be discoverable by a mixed methods analysis and can then be modeled based on the science content and student data. Of course, in a broader view, we know that all models fail at some level and do not incorporate all aspects of reality but such models can be useful if they provide gains in our understanding or utility in our context.

Purpose and Guidelines

Even though most researchers have an idea of the philosophy driving their research they still need to define its purpose. Venkatesh et al. (2013) advise identifying the purpose of mixed methods research early. Identifying the purpose/s will help establish the research goals and

later serve to inform reviewers of how to center any findings. They identify the seven purposes shown in Table 3.

Table 3

Defining Purpose of Mixed Method Approach

Purposes	Description	Illustration	Current Study
Complementary	Mixed methods are used in order to gain complementary views about the same phenomena or relationships.	A qualitative study was used to gain additional insights on the findings from a quantitative study.	Meets this purpose by triangulating the evidence from both methods and statistically the quantitative model complements the qualitative review of the science framework's multidimensional design. However, UIRT model was more practical with regards to improvement of model fit.
Completeness	Mixed methods designs are used to make sure a complete picture of a phenomenon is obtained.	The qualitative data and results provided rich explanations of the findings from the quantitative data and analysis.	NA
Developmental	Questions for one strand emerge from the inferences of a previous one (sequential mixed methods), or one strand provides hypotheses to be tested in the next one.	A qualitative study was used to develop constructs and hypotheses and a quantitative study was conducted to test the hypotheses.	NA
Expansion	Mixed methods are used in order to explain or expand upon the understanding obtained in a previous strand of a study.	The findings from one study (e.g., quantitative) were expanded or elaborated by examining the findings from a different study (e.g., qualitative).	Meets this purpose by expanding prior studies at the state and international levels that focused on quantitative model fit; typically, these studies did not have a qualitative review of the dimensionality of the science content within the framework.
Corroboration/ Confirmation	Mixed methods are used in order to assess the credibility of inferences obtained from one approach (strand).	A qualitative study was conducted to confirm the findings from a quantitative study.	Meets this purpose by corroborating other studies' findings of quantitative multidimensionality in science through qualitatively defining that science subdomains are multidimensional.
Compensation	Mixed methods enable compensating for the weaknesses of one	The qualitative analysis compensated for the small sample size in the quantitative study.	NA

	approach by using the other.		
Diversity	Mixed methods are used with the hope of obtaining divergent views of the same phenomenon.	Qualitative and quantitative studies were conducted to compare perceptions of a phenomenon of interest by two different types of participants.	Did not meet this purpose because sufficient student demographic data was unavailable for this data set. See research question reclarification in last chapter.

Note. Adapted from "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems," by V. Venkatesh, S. Brown, and H. Bala, 2013, *MIS Quarterly*, *37*(1), p. 26. Copyright 2013 by JSTOR.

For this research study 4 purposes were highlighted in green in Table 3 as scaffolding the work being done here: complementary since the PISA science assessment is built off the PISA science framework; expansion since the hope is that the qualitative data from the framework will illuminate the finding (or lack thereof) of multidimensionality in the student scores; corroboration since the qualitative data may or may not concur with any quantitative findings; and diversity because a lack of multidimensionality in the quantitative scores may be influenced by diversity issues. In Table 3's final column are the descriptions of how this study met, or did not meet, each purpose.

For example, while outside the scope of this study, a lack of latent diversity, or "diverse student mindsets" that can result from student diversity among other traits (Godwin, 2017, p. 13) could be masking the dimensionality of science subdomains if the student sample is not diverse and the students approach science problems in a similar manner. It is on diversity issues that I differ in opinion from Venkatesh et al. (2013, p. 22) as they state a mixed methods approach should be taken without regard to "cultural incommensurability" as long as it helps

the researcher answer their question. If the purpose of mixed methods research hinges in part on diversity, then cultural considerations should be taken into account when developing methodology for the research (Broesch et al., 2020).

Once purpose/s are identified the design of and data analysis for mixed methods research needs to be grounded in accepted qualitative and quantitative procedures (Venkatesh et al., 2013). Whether each type of research occurs concurrently or as in this case sequentially, Table 4 provides both general and validation guidelines that can be applied to any mixed methods research (Venkatesh et al., 2013).

Table 4

Guidelines for Mixed Methods Research

Guideline	Researcher Considerations
General	Carefully think about the research questions, objectives, and contexts to decide on the appropriateness of a mixed methods approach for the research. Explication of the broad and specific research objective/s is important to establish the utility of mixed methods research.
	Carefully select a mixed methods design strategy that is appropriate for the research questions, objectives, and contexts.
	Develop a strategy for rigorously analyzing mixed methods data. A cursory analysis of qualitative data followed by a rigorous analysis of quantitative data or vice versa is **not** desirable. Apply the same standard of rigor as typically used in analyzing quantitative and qualitative studies.
	Integrate inferences from the qualitative and quantitative studies in order to draw meta-inferences from mixed method results.
Validation	Discuss validation for both quantitative and qualitative studies.
	When discussing mixed methods validation, use mixed methods research nomenclature consistently.
	Mixed methods research validation should be assessed on the overall findings and/or meta-inferences from mixed methods research, not from the individual studies.
	Discuss validation from the point of view of the overall mixed methods design chosen for a study or research inquiry.
	Discuss potential threats to validity that may arise during data collection and analysis, along with any remedies.

Note. Adapted from "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems," by V. Venkatesh, S. Brown, and H. Bala, 2013, *MIS Quarterly*, *37*(1), p. 41.

Copyright 2013 by JSTOR.

Following these guidelines helped support validity in this mixed methods study and might help transferability to other contexts for inferences made based on the data (Venkatesh et al., 2013), although the limitations described earlier with regard to both linking and the documentation of the data set indicate additional work is needed for greater generalization.

Step 1: Qualitative Analysis

"Comparing and contrasting data is vital to qualitative analysis (Gale et al., 2013)." This is especially true when analyzing a framework, which could be considered a policy[53] at the state, national, or global level, and may incorporate multiple sources of information with various degrees of clarification. Since frameworks are documents, this can be done via a qualitative document analysis (Wach et al., 2013; Bowen, 2009).

Overview of Document Analysis. For digital and hard copies of documents, an organized procedure is needed for the analysis (Armstrong, 2021). Armstrong (2021) recommends beginning by identifying the objective of the document analysis and describes six common objectives: "defining concepts, mapping range and nature of phenomena, creating typologies[54], finding associations, providing explanations, and developing strategies." Once documents have been selected, it is important to not just "lift" text to be used in the report (Armstrong, 2021) or the analysis may be considered superficial. Rather, analysis should strive for deep understanding to develop meaning with regard to the construct.

Wach et al. (2013) outline the process of document analysis in several steps and notes:

1. Defining document inclusion criteria, which may be practical or strategic in nature,

[53] Wach, Ward, and Jacimovic (2013) defined policy as documents "that express official organizational aims and strategies."
[54] An analysis based on categories.

2. Gathering the document/s,
3. Outline analysis area/s,
4. Analyze the document/s using coding if applicable, and
5. Verify the analysis through an independent source (2nd reviewer) to increase reliability, impartiality, and dependability of the findings.

An important note about step 5 – an analysis is considered dependable if the second reviewer would have made the same conclusions while analyzing the document/s in the same manner. Finally, Wach et al. (2013) recommend that thought needs to be given if the organization owning the document actually delivered the proposed policy. In this case the policy would be the science framework claim of three subdomains in science.

There are several ways to present a document analysis. Mazzei and Jackson (2024) discuss "re-animating" documents in a visual format to uncover new "intensities", which can be interpreted as new ways of seeing content contained within the document. This could take the form of a visual aid, such as a logic model or flow chart, that details the structure of the data and points out claims.

Advantages and Disadvantages of Document Analysis. Many states and nations have conducted document analysis on content frameworks as an efficient means of uncovering content relatedness and connections to educational theory. Bowen (2009) describes advantages and disadvantages to this type of analysis in the list below. My counterarguments or agreements are in italics.

Document analysis provides:

- efficiency in data selection as collection from participants is not always required,

- *Agree, with caveat that sometimes documents can be difficult to get in entirety from institutions.*

- most documents as they are available in the public domain,
    - *Disagree, some documents may be public, but institutions can also keep many documents internal.*
- a decreased cost compared to other analyses,
    - *Agree, document analysis is not as high in cost as a recruitment of participants for a quantitative study.*
- less to no reactivity or obtrusiveness from documents since participants are not being observed,
    - *Agree, but Bowen (2009) also mentions that a researcher is less likely to influence the research due to lack of social interaction, which I disagree with as document analyzers can bring their own viewpoints into describing the meaning of the document.*
- documents are stable over time along with being exact in nature and researchers do not alter what is researched, and
    - *Disagree, documents often go through several versions that are not always reported to researchers; sometimes changes are left out of a document too.*
- broad coverage of material and historical events,
    - *Agree, documents allow researchers to peer into history.*

Document analysis is disadvantageous in that documents can:

- sometimes lack detail,

- be hard to retrieve, and
- become biased through the selection process.

Despite the disadvantages, document analysis can provide needed illustration. A document analysis of the framework's theoretical claims of multidimensionality will help to elucidate whether multidimensionality should be expected in the empirical results.

Conducting Document Analysis. The number of documents tied to the 2015 PISA is quite extensive and includes frameworks, reports on results, released items, technical reports, country level reports, webpage FAQs, brochures, and videos. To narrow this field, documents were selected following the process outlined by Voogt and Roblin (2012) who recommend screening for the goal of identifying the main theme, which in this study is science dimensionality. This is done by determining inclusion criteria *a priori* as per Wach et al. (2013). The inclusion criteria required the documents to mention: PISA 2015 science framework and any form of these words: domain, subdomain, or dimension, along with the document having been developed by OECD. This second screening requirement excluded any secondary sources that were not directly involved in science content development since content developers, including teachers and assessment designers, are closest to the intent of the framework. This led to two documents being identified for possible analysis – OECD's 2015 PISA Science Framework and PISA 2015 Technical Report. Next, a saturation evaluation on the selected documents was performed. Saturation evaluation, based in grounded theory[55], can be used in qualitative analysis to stop the data collection from a document if no additional data are found

[55] Grounded theory is an inductive qualitative methodology that allows new theory to be formed from the observed data (Ho & Limpaecher, 2021). That said, I acknowledge my prior experience teaching science may lead to deductive reasoning with regards to PISA science content standards and their fit into a dimensionality theory.

to code to the theme being analyzed that go beyond what has already been found, or in other words some degree of key saturation has been reached in the evaluation. (Saunders et al., 2018). The saturation evaluation eliminated the PISA 2015 Technical Report as it offered no substantial theoretical claims of multidimensionality in science.

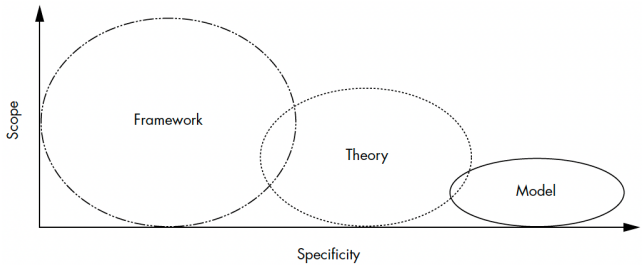
Hence, the 2015 PISA Science Framework document found during document collection was color-coded[56] by the main theme [dimensionality] and given relevant coder annotations in order to draw out any supporting evidence or subthemes [multi vs. unidimensionality] (Voogt & Roblin 2012). If no evidence was found supporting science multidimensionality that was also documented, along with any barriers to or reasons why multidimensionality is not present among the science subdomains. The subdomains were graphically connected to show any crossover between science content knowledge in the content standards (see Chapter 3, Figure 11, which was verified via a committee review as recommended by Wach et al. [2013]).

After qualitative analysis, the researcher may find among the coded themes evidence of a theory that supports a model.

Figure 8 illustrates how this theory building can come about (Carpiano & Daley, 2006).

Figure 8

From Science Framework Review to MIRT Model Development



[56] This was done by hand rather than a computer program.

Note. From "A guide and glossary on postpositivist theory building for population health," 2006, by R. M. Carpiano, and D. M. Daley, 2006, *Journal of Epidemiology and Community Health*, *60*, p. 566.

Carpiano and Daley's (2006) definitions of a framework, theory, and model associated with Figure 8 are adapted below to describe aspects of this study.

1. Conceptual Framework: A set of standards about content knowledge in science that students should be able to show mastery towards.
2. Theory: Grounded in educational pedagogy and learning philosophy, it indicates a relationship between the variables (science content knowledge) while diagnosing the science learning phenomena to predict an outcome (e.g., the science subdomains will present as multidimensional when scored since learning may be unique to each subdomain). This theory will hopefully be fully developed after qualitative framework analysis.
3. Model: Makes a specific assumption about the learning theory for science content knowledge that allows parameters (science content knowledge variables) to be tested quantitatively.

Carpiano and Daley (2006) further clarify that after articulating a theory the researcher could draw the model by detailing the constructs taken from the theory (such as PISA science scores), diagramming their flow from left to right with relationships shown using arrows[57], and indicating positive or negative relationships with a +/-.

[57] Double headed arrows indicate correlations, while single-headed arrows indicate causal relationships.

In order to predict the IRT model needed, evidence from Figure 11 comparing science content standards and from the framework document analysis that supported an educational theory on science dimensionality was then used to develop Figure 29 (similar to Figure 8). After the IRT model with the best fit was identified during quantitative data analysis it was compared to Figure 29's predicted model.

Step 2: Quantitative Analysis

"Models make assumptions around measurement explicit and testable (Lang & Tay, 2021)." IRT provides a group of statistical models that determine the probability of a student selecting a specific response (Immekus, 2019). Measurement using an IRT model is an attempt to explain this specific response as a continuous variable (Ayala, 2022) or set of variables (MIRT), such as student ability in the subdomain life science.

Primer on Item Response Theory. IRT is best described as a "formalized" statistical set of models for measuring skill/ability in an assessment (Lang & Tay, 2021; Wilson, 2013). This skill/ability is referred to as a latent variable since it is not directly observable and the scores derived from the assessment are manifest variables of this ability (Ayala, 2022; Lang & Tay, 2021). These manifest variables are empirically calibrated to be on an interval scale using the data since the difference between the values is meaningful (Ayala, 2022). Wilson (2013) further clarifies that IRT separates the scale from depending on the random number of items selected to be in the assessment. An item that has utility can differentiate well between student performance on different points of a trait continuum [the interval scale] (Ayala, 2022; Immekus, 2019; Mailman School of Public Health, 2023).

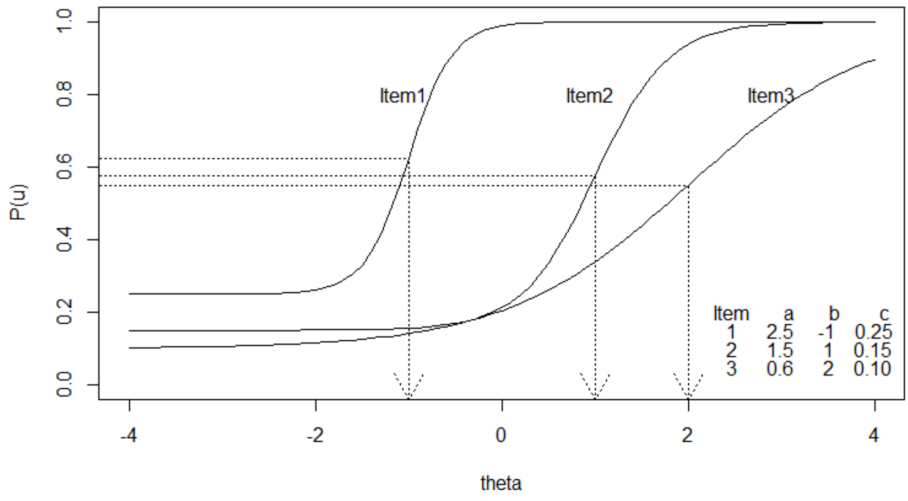
Item Parameters. Student performance hinges on several item parameters, which "define a blueprint for the model (Brooks-Bartlett, 2018)." These parameters, and their relationships to student ability, can be visualized with an item characteristic curve (ICC) graph.

Figure 9 shows ICCs for three simulated items based on an IRT model estimating three parameters (Park et al., 2020). Each item in

Figure 9 has a dichotomous score (0 for incorrect, 1 for correct).

Figure 9

ICCs Based on a Three-parameter Logistic (3PL) Model



Note. From "Technically speaking: Determining test effectiveness with item response theory," by S. Park, A. Reeger, and A. M. Aloe, 2020, *Iowa Reading Research Center*. Copyright 2023 by The University of Iowa.

In

Figure 9, P(u) represents the probability (P) of a student responding correctly to the item (u). Theta is a representation of student ability (this is the student's location and is also called θ). The item discrimination parameter or α is labeled (a) in the right bottom corner of

Figure 9 and provides the maximum slope steepness of each ICC. Steeper slopes indicated by higher α values point towards items with greater discrimination between student abilities (Harris, n.d.). In addition to item discrimination, item parameters can include item difficulty and guessing (Ayala, 2022; Mailman School of Public Health, 2023); the 4PL, not shown here, also adds an additional parameter to describe what tends to happen at the top end of the curve, and there are numerous other innovative IRT models; operationally however, usually no more than the three parameters per item shown here are used. In the same corner of the figure, (b) represents the item difficulty parameter as it relates to student ability and (c) represents the "lowest possible probability" of a student responding correctly to the item as an indicator of the student guessing parameter (Park et al., 2020).

Therefore, Item 1 discriminates well between students of different abilities, while Item 3 was the least discriminating. Item 2 was moderate in both discrimination, difficulty, and guessing. Finally, Item 3 was the most difficult and had the least probability of students guessing, while Item 1 was the least difficult and had the most probability of students guessing. Higher performing students tend to do better on more difficult items while lower performing students often answer easier items correctly, with discrimination and guessing empirically adding to what is sometimes known about how the item tends to perform.

While a difficulty parameter estimate is assigned to each item, an IRT model does not explain why an item is difficult for each student (Lang & Tay, 2021); only that empirically it was calibrated as difficult. An individual's cognitive process to reach an item response is also not described by an IRT model (Ayala, 2022), so this is done theoretically in large-scale assessment with expert panels in frameworks, and sometimes with practitioners in standard settings, or can

be done more empirically with qualitative analysis in cognitive laboratories or other small-scale settings using verbal protocol analysis (“think alouds”) or other techniques. For PISA, only the frameworks are released, but they are extensive documents.

***Assumptions*.** For an IRT model to estimate all three parameters successfully, the model needs to conform to several assumptions. Assumption 1 refers to monotonicity, which assumes that as the probability for a correct response increases a student’s ability also increases (Mailman School of Public Health, 2023). Assumption 2 is guided by conditional/local independence, which states item responses on an assessment are independent of each other based on a student’s location/ability (Ayala, 2022). Assumption 3 is based on unidimensionality, which assumes that only one continuous latent trait is measured (Ayala, 2022). Assumption 4 describes the functional form, which maintains that the data match the mathematical function described by a model (Ayala, 2022). Violating these assumptions can affect which model should be chosen as a best fit to the data. For example, if an assessment is measuring several proficiencies/latent traits, then a unidimensional latent variable may not be appropriate (Socha, n.d.).

Defining Dimensionality. Sometimes education attributes being assessed lack a pre-defined dimensional structure (Irribarra & Arneson, 2023; Reckase, 1990). The definitions for multidimensionality and unidimensionality need to be clearly defined, which Irribarra and Arneson (2023) argue has still not occurred in many domains even though dimensionality is often discussed in educational research. One way to begin a definition is to compare the

statistical versus psychological aspects of dimensionality. Psychological/theoretical[58] dimensionality is the hypothetical latent constructs, such as science ability, that in theory as identified by experts are required for performing well on an assessment (Irribarra & Arneson, 2023; Reckase, 1990). Statistical dimensionality is "the minimum[59] number of mathematical variables needed to summarize a matrix of item response data" (Reckase, 1990). Reckase (1990) further points out that statistical dimensionality is based on observable data, such as scores, and rests on the data matrix so is not a function of the assessment or the student population being assessed. Actionable dimensionality is described as a third aspect by Irribarra and Arneson (2023), which refers to the "number of values considered when making a decision based on an assessment." If the "art of assessing dimensionality" is finding the least number of latent abilities to preserve statistical prowess and construct meaning (Briggs & Wilson, 2003) then dimensionality depends on both the construct, the statistical model, and its intended usage. Multidimensionality over the content areas can be defined as three distinct science subdomains related to one theoretical construct – science ability; however other types of multidimensionality may also exist, such as in cognitive processing or item format as discussed earlier. IRT models have evolved as more research has shed light on dimensionality.

Development of IRT Models. Preceding multidimensional IRT models was classical test theory (CTT) and unidimensional IRT. MIRT models are often compared to unidimensional IRT

[58] Irribarra and Arneson (2023) prefer "theoretical" to "psychological". They redefine psychological dimensionality to theoretical dimensionality as "the number of relevant psychological attributes that can be reasonably conceived as quantities and are believed to be involved in generating responses to items to some extent (Irribarra & Arneson, 2023)."

[59] Irribarra and Arneson (2023) recommend removing the term "minimum" from the definition.

models in order to assess model fit. Therefore, an overview of the types of IRT models and their development is needed.

Classical Test Theory (CTT). CTT was developed first and used historically with achievement tests (Brandt, 2015). Similar to IRT, the latent variable is assumed to be continuous (Ayala, 2022). Opposite of IRT, CTT focuses on a student's whole score for an entire assessment (Ayala, 2022). In CTT, the item parameters depend on the sample from which they are taken (Immekus et al., 2019). In addition, CTT does not provide a reason for item difficulties, making linking between assessments with different item sets more challenging (Brandt, 2015). The CTT approach was updated to IRT in order to focus on the item and student relationship with a statistical model rather than the assessment in its entirety (Wilson, 2013). Now large-scale assessments, such as NAEP and PISA, use IRT models to avoid these disadvantages (Brandt, 2015).

Unidimensional IRT Models. Each UIRT model described in this section uses a logistic function to describe student ability with regards to item parameters to get the probability of answering an item correctly (Harris, n.d.). The simplest of the UIRT models is the Rasch model (Wilson, 2013; Ayala, 2022; Lang & Tay, 2021). The Rasch model assumes items are on a single continuum showing student ability via standardized z-scores. This allows student responses to an item to be compared based on proficiency level. The item discrimination parameter (α) in the Rasch model is set to a constant value of 1.0.

In contrast, the parameter α in the one-parameter logistic (1PL) UIRT model is allowed to vary from 1.0 and can be some other constant value across the items. Ayala (2022) describes this as a "philosophical perspective" where the Rasch model focuses on constructing the

variable and the 1PL UIRT model focuses on fitting the data. Both the Rasch model and 1PL UIRT model estimate the item difficulty parameter (item location or δ), which can vary. As the δ parameter increases the probability of a correct response decreases (Reckase, 2009; Lang & Tay, 2021; Ayala, 2022). The 1PL UIRT model[60] also has an advantage over CTT in that by fitting a logistic function the model no longer assumes measurement error is the same for each student as CTT does (Lang & Tay, 2021).

The two-parameter logistic (2PL) UIRT model allows the discrimination parameter to freely vary across items. Adding to the 2PL UIRT model, the three-parameter logistic (3PL) UIRT model estimates the guessing parameter (χ). The proportion of students in the lowest proficiency level choosing the right answer is the estimate used for the χ parameter (Reckase, 2009). After the 3PL UIRT model the next development in IRT was the 1PL MIRT model.

Functionality of MIRT. The complexity of educational constructs directly led to the development of MIRT models (Reckase, 2009). A MIRT model is able to relax the assumption of unidimensionality so that multiple correlated latent traits can be measured (Wang, 2021). An assessment, set of items, or even a single item may require students to use multiple abilities/latent traits, "especially in the compound areas such as the natural sciences" (Issayeva, 2022). A limitation of unidimensional models is they are not well fit for an instrument developed to be multidimensional (Immekus, 2019). Parameters are interpreted similarly to unidimensional IRT models, but they take the form of vectors and their direction in theta (θ) space will influence the interpretation (Socha, n.d.). This theta space is multidimensional and can be summarized as $\theta_j = [\theta_{j1} \ldots \theta_{jM}]'$ with *M* being the number of unobserved latent

[60] Other IRT models also have the same advantage over CTT.

dimensions needed to model a student's predicted response to an item (Immekus, 2019). The χ parameter is the exception because it is not a vector and retains its 3PL definition. As with unidimensional models, the maximum marginal likelihood estimation (MMLE) can calibrate item parameters (Ayala, 2022; Socha, n.d.).

MIRT models can be either compensatory or non-compensatory (Spencer, 2004; Issayeva, 2022; Socha, n.d.). Compensatory models are additive in nature and allows a student's high score in one dimension to make up for a low score in another (Socha, n.d.; Reckase, 1997). However, assessment designers may find it difficult to explain to test takers why scores on different dimensions are dependent on each other (Baghaei, 2012). A student's ability (θ) in one dimension does not counterbalance for their ability (θ) in a different dimension in a non-compensatory (or partially compensatory) model (Socha, n.d.; Reckase, 1997). Results from a non-compensatory model are the nonlinear sum of thetas (Duran, 2014). Non-compensatory models also tend to be simpler (Socha, n.d.) and may capture cognitive ability more accurately (DeMars, 2016). A drawback is that non-compensatory models have not been used as frequently due to issues with parameter estimation, especially the lack of efficient algorithms. although this is somewhat changing with more computing power becoming available (Ayala, 2022; DeMars, 2016; Spencer, 2004; Wang & Nydick, 2015).

Limitations and Benefits of MIRT Models. While an assessment's items may be multidimensional in the lens of a content framework, each dimension's strength may not be enough to change from a unidimensional model (Reckase, 1985; Socha, n.d.). Sometimes we say the dimensions may exist, but they are not sufficiently "separable" to matter. Another limitation to using MIRT models is that even though collecting data via online assessments is

less expensive than in-person, the increasing demand for data and its analysis are testing the limits of models like MIRT and other algorithms (Wang, 2021). There are also several sources of indeterminacy with a MIRT model. Metric indeterminacy results from the metric being relative (Ayala, 2022), that being that the item location and respondent location are relative to each other and not fixed until either the item mean, or respondent mean, for the calibration is set to zero, impacting getting calibration results on the fixed item or person. Rotational indeterminacy indicates the direction of each axis is not unique with regard to the item vectors; however, fixing the axes can help alleviate this problem (Ayala, 2022).

A benefit of a MIRT model is the ability to link calibrations in order to create a large pool of calibrated items, which then may be used to develop interesting adaptations, such as computer adaptive testing (CAT) with parallel multidimensional test forms[61] (Issayeva, 2022). Within-item multidimensionality can be used with a MIRT model to reduce test length because one item provides data on several dimensions, which is useful in CAT (Duran, 2014). Kose and Demirtasli (2012) found however, that longer tests (i.e., more items) and greater sample size are needed to reduce error and increase MIRT model sensitivity. Perhaps a greater benefit is more data about student ability on each dimension, which in turn can confirm or add to theories about educational constructs.

PISA IRT Models. The PISA *Results in Focus* 2015 report states that PISA's goal is not to determine the cause and effect of educational "practices and student outcomes" (OECD, 2018). However, the assessment does arguably make decisions about how student latent traits relate

[61] Test forms often refer to versions of the same assessment. For example, test form A may have items arranged differently from or contain only some of the same items in test form B.

to educational constructs, which is often used by policymakers to lay some of the evidentiary groundwork for educational practice and theory. PISA divided each domain, reading, math, and science, into several subdomains for the 2015 assessment indicating the constructs of each domain provided evidence of multidimensionality (Brandt, 2015).

Currently, and in 2015, PISA uses a unidimensional composite score for science developed via a 2PL Rasch model (Jerrim, 2016) while reporting scores in subdomains too. Student scores are developed by first choosing the IRT model to estimate item parameters and then using maximum likelihood estimate (MLE) to determine a latent trait ability level for each student (Jerrim, 2016). This is done in the main study after an extensive field trial to adapt instruments, usually by a keep and drop method, for which data does not get released. The difference in modeling for a multidimensional construct versus unidimensional scale scoring indicates that the assessment is being interpreted in both directions. There are both advantages and disadvantages to this approach per Brandt (2015) – see Table 5.

Table 5

Trade-offs Between Calibration Methods for a Unidimensional Score

Calibration Method for Multidimensional Data	Advantages	Disadvantages
Scale Score on Unidimensional Calibration	• Reliably allows calculation of individual scores using MLEs	• Overestimates reliability, plus biases difficulty and variance estimates, by neglecting local item dependence (LID) • Validity of the multidimensional constructs is reduced since assessment designed to be unidimensional • Framework may specify one set of weightings for each dimension, but the actual weight may change due to a need to drop items

Composite Score on Multidimensional Calibration	• Assuming items within a subdomain are separate dimensions the items within a dimension are more closely related than items between dimensions then this approach allows LID to be considered, which more accurately estimates reliability • Explicit and clear weighting of subdomains allows the unidimensional composite score to be developed	• Calculation of reliable individual scores via MLEs is not possible • Reliability of composite score is reduced since items not on a common scale • Not an appropriate calibration method if the Rasch model is used since dimensions cannot be set to equal variances (this leads to the need for standardization after model calibration increasing measurement error)

Note. Adapted from "Unidimensional interpretation of multidimensional tests," by S. Brandt, 2015, *Dissertation*, p. 28.

Construct validity is also a crucial issue here. Per Messick (1995) and Spencer (2004), ignoring any evidence, such as multidimensionality found in a framework, may negatively impact construct validity of the evaluation of meaning behind an assessment's results. Reading is the only domain that OECD (2017a) describes as multidimensional in the 2015 PISA Framework. Using a MIRT model could help validate the 2015 PISA science framework design if the subdomains are found to be indicators of multidimensionality. This finding would mean the current unidimensional scoring model could contain a construct misspecification, which leads to incorrect interpretation of PISA scores that could in turn impact education policy decisions for diverse student groups.

Conducting Quantitative Analyses. The analyses in this section were done for both the S10 and the S11 clusters of items and were all conducted using the R program in R Studio (RStudio Team, 2021). The PISA 2015 science data was first cleaned by removing student cases that contained only NAs. Then data was then run through descriptive analyses using the psych package (Revelle, 2024). Next histograms of average scores for the U.S. science sample and cluster S10 subsample were developed for the student population using the ggplot2 package

(Wickham, 2016). Item histograms showing number of each type of response (0 or 1) were plotted for the cluster S10 subsample. Due to the ordinal nature of the dichotomous data a polychoric matrix was developed using the psych package (Revelle, 2024) and saved for use in later analyses. The correlation coefficient rho was examined and developed into a table. Using the factoextra package (Kassambara & Mundt, 2020) a distance matrix heatmap was also developed from the polychoric matrix for the cluster S10 subsample. In order to evaluate the sensitivity of the cluster analysis the cluster S10 subsample was randomly split in half using the dyplr package (Wickham et al., 2023) and the random half subset was run through the same processes outlined in the Cluster Analyses and PCA subsections below, but not through IRT analysis as the subset's size was very small at 653 students.

Cluster Analyses. Each subsample was then run through a cluster analysis using the kmeans function of the stats package (R Core Team, 2023) to estimate 3 clusters. Since each subdomain should represent a unique dimension the logits for each dimension can be obtained via cluster analysis and then weighted to show each item's relationship to a dimension (Ayala, 2022). The results of the cluster analysis were graphed in a scree plot using the factoextra package (Kassambara & Mundt, 2020).

PCA. The two subsamples and random half subset were then also run through a PCA using the prcomp function in the stats package of R (R Core Team, 2023). The loadings, scores, and variances were analyzed to help develop several plots. Three-dimensional (3D) principal component plots were developed based on the PCA scores with the plot_ly function of the plotly package (Sievert, 2020) in R, while loadings were visualized with the barplot function of the graphics package (R Core Team, 2023) in R. Finally, principal component plots were loaded

publicly into Plotly Chart Studio (Plotly Technologies Inc., 2015) for later presentation due to their interactive nature.

IRT Analyses. A critical step in determining which model provides the most information about student learning is examining the fit of various models (Yamamoto, 1995). Therefore, student data was analyzed in an exploratory quantitative research design. The explorations consisted of a 1PL UIRT, 2PL UIRT, 1PL MIRT, and 2PL MIRT models – each model is described below.

Model 1 was a 1PL UIRT model and is described by Equation 1[62].

Equation 1 1PL UIRT

$$\boldsymbol{p\left(x_j = 1 \middle| \theta_s, \alpha, \delta_j\right) = \frac{1}{1 + e^{-\alpha(\theta_s - \delta_j)}}}$$

Where *p* is the probability of a value of 1 (a correct response) when the predictor is *x*, *e* is a constant of 2.7183 (i.e., the base of the natural logarithm), θ is a person's location (i.e., ability), δ is the item's location (i.e., estimated difficulty) of item *j* for student *s*, and alpha (α) is allowed to vary from 1 while kept constant across items (Ayala, 2022).

Model 2 was a 2PL UIRT model and is described by Equation 2.

Equation 2 2PL UIRT

$$\boldsymbol{p\left(x_j = 1 \middle| \theta_s, \alpha_j, \delta_j\right) = \frac{e^{\alpha_j \theta_s + \gamma_j)}}{1 + e^{\alpha_j \theta_s + \gamma_j)}}}$$

The discrimination parameter, α, is now allowed to vary across items (Ayala, 2022). Note that the intercept is represented by gamma (γ), which is constant for this model, and is a representation of the item's location and discrimination parameters' interaction (Ayala, 2022).

[62] Equations 1, 2, and 4 are from Ayala (2022). Equation 3 is derived using DeMars (2016).

"In a proficiency assessment situation γ_j would be interpreted as related to an item's difficulty/easiness (Ayala, 2022, p. 393)."

Model 3 was a 1PL MIRT model and is described by Equation 3.

Equation 3 1PL MIRT

$$\boldsymbol{p(x_{ij}=1|\underline{\theta}_s,\underline{\alpha}_j,\gamma_j)=\frac{e^{\underline{\alpha}_j'\underline{\theta}_s+\delta_j)}}{1+e^{\underline{\alpha}_j'\underline{\theta}_s+\delta_j)}}}$$

In order to determine an item to dimension relationship that can change along 2 or more dimensions the logits are weighted (Ayala, 2022). The item slopes (α) are fixed to a constant number across dimensions (DeMars, 2016). Note that an underscore indicates a vector and a prime symbol (') indicates a row vector.

Model 4 was a 2PL compensatory MIRT model and is described by Equation 4.

Equation 4 2PL MIRT

$$\boldsymbol{p(x_{ij}=1|\underline{\theta}_s,\underline{\alpha}_j,\gamma_j)=\frac{e^{\underline{\alpha}_j'\underline{\theta}_s+\delta_j)}}{1+e^{\underline{\alpha}_j'\underline{\theta}_s+\delta_j)}}}$$

Model 4 differs in that α can now vary.

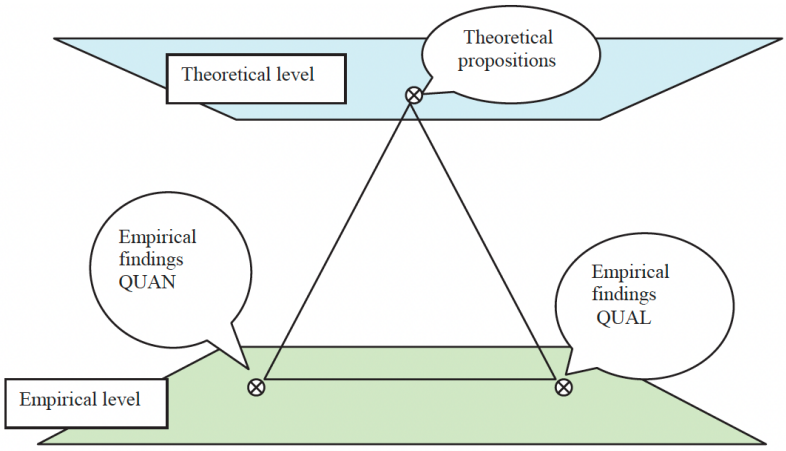
After model selection the dichotomous data (coded as 0 or 1) was analyzed using the TAM package (Robitzsch et al., 2022). This package was chosen primarily because of its use by OECD (Kiefer et al., 2015) and other researchers comparing IRT models, along with its ease of use. The TAM package allows for UIRT and MIRT models, but only compensatory MIRT. A table was developed that shows the statistics for each model and a second table comparing the models' fit was also generated. Based on the best fitting model, a wright map and ICCs were also developed using the wrightMap function of the WrightMap package (Irribarra & Freund, 2014) and the plot function of the graphics package (R Core Team, 2023) in R respectively.

Data Triangulation

After data analysis, the researcher should begin to build inferences from both types of data. Meta-inferences are defined by Venkatesh et al. (2013) as "theoretical statements…from an integration of quantitative and qualitative strands of mixed methods research." The pathway for this study flowed as follows: comparing (merging) qualitative + quantitative findings → meta-inference/s. The other pathways consist of either qualitative or quantitative findings leading individually to the next set of findings, which would be the opposite of the first step, then to a meta-inference. After an analysis path is chosen Venkatesh et al. (2013) state that researchers should take either a bridging or bracketing research path to develop the meta-inference/s. Bridging is described as a consensus between the two types of findings (qualitative + quantitative) while bracketing uses alternate views of the phenomenon to report differences between the two types of findings (qualitative vs. quantitative). Bridging was used for this study. An analysis path can be taken further to develop these qualitative and quantitative data into a triangulation, or mapped, to each other so that the data from one method supports or contrasts conclusions drawn in the other. Östlund et al. (2011) adapted the diagram shown in Figure 10 from Erzberger and Kelle (2003) to illustrate triangulation.

Figure 10

Triangulation for Mixed Methods Research



Note. From "Combining qualitative and quantitative research within mixed method research designs: A methodological review," by U. Östlund, L. Kidd, Y. Wengström, and N. Rowa-Dewar, 2011, *International Journal of Nursing Studies*, *48*(2011), p. 371. Copyright 2010 by Elsevier.

Figure 10 showcases that the empirical findings for both quantitative (QUAN) and qualitative (QUAL) can be used to support theory, which may be newly developed. The sides of the triangles represent connections between the theory and each set of findings, along with the finding to one another. Depending on the outcome of the research, the triangle sides can differ in appearance. First, the triangle sides may remain convergent (as shown in Figure 10), which is when findings from both methods support theory and lead to the same conclusion. Second, the triangle sides may become parallel to each other when the findings from both methods compliment or support one another. Third, the triangle sides may be divergent indicating the findings are different for each method or may even contradict one another, in which case the contrasts should be explored to understand why different lenses indicate different directions.

As noted in their research, Östlund et al. (2011) stated, that while a mixed method approach is gaining ground with other researchers, the triangulation process described above was only beginning to be used at that time. It has since expanded. Triangulation can help clarify results in mixed methods research by clearly identifying the interactions between different types of data. Bowen (2009) and Armstrong (2021) agree that document analysis is a way to triangulate with other methods of research. They also describe how triangulation can provide a solid foundation for the design of and theory behind the mixed methods approach (Östlund et al., 2011). A data triangulation figure based on Östlund et al.'s (2011) procedure shown in Figure 10 was built for this study.

Step 3: Equity Investigation

OECD did not make publicly available information on ethnicity/race in 2015 and only collected minimal information on school location and student economic status in the survey questions. Therefore, survey questions were not used to investigate equity issues. Instead, student ability levels between models were compared to determine which model type had a greater range of ability levels. Using item cluster S10 with Models 1b and 3b thetas were mapped in bar plots. Ability level (theta) range was then analyzed to determine where historically marginalized student groups might be located to determine if a less complex model sacrifices information about these groups for a more pragmatic design.

CHAPTER 3: RESULTS

The following results are reported in three sections by research question (RQ). Triangulation results from the qualitative and quantitative analyses is then reported after the RQ2 results. See section Research Questions for the details of each RQ.

Results Relating to RQ1

The saturation evaluation of the two documents identified for analysis, OECD's 2015 PISA Science Framework and PISA 2015 Technical Report, yielded mixed results. Neither a broader view of science educational theory nor multidimensionality for science knowledge is described in the 2015 PISA science framework. This theory was teased out during the qualitative analysis and is proposed at the end of this section.

Coding on dimensionality was accomplished for the first document (i.e., the science framework) and theoretical saturation was achieved at the end of that document analysis. Saturation occurred because the second document, while mentioning science and multidimensionality upon initial analysis, did not actually refer to science content dimensionality, but rather the possible dimensionality between new and trend science items. Any evidence resulting from color codes highlighting the two subthemes are reported in Table 6 below for the only document analyzed. Note that the evidence column for subtheme unidimensionality is deliberately left blank as no portions of the science framework were found to code to unidimensionality. Missing evidence occurs in the subtheme multidimensionality (see yellow-filled cells).

Table 6

Evidence Supporting Dimensionality Themes

Document Section/Feature (Pg. #)	Subtheme Unidimensionality	Subtheme Multidimensionality	Annotation
Box 2.1 Scientific knowledge: PISA 2015 terminology (Pg. 21)		"…three distinguishable but related elements."	While OECD does appear to be referring to multidimensionality with this phrase it is not about science content knowledge, but rather the types of knowledge: content[63], procedural, and epistemic.
Section: Organizing the domain of science and Figures 2.1-2.2 (Pg. 25)			Science literacy is referred to as a domain of interrelated aspects, but its science content dimensionality is not clarified.
Section: Scientific knowledge and Figure 2.5 (Pg. 28)		"Given that only a sample of the content domain of science can be assessed in the PISA 2015 scientific literacy assessment, clear criteria are used to guide the selection of the knowledge that is assessed. The criteria are applied to knowledge from the major fields of physics, chemistry, biology, earth and space science…"	OECD refers to "fields" of science indicating that these content areas are not one single subdomain. Figure 2.5[64] clearly separates the content knowledge required for each "field" of science visually by them out with a blue banner.
Table 2.2 (Pg. 29)		"Desired distribution of items, by content"	OECD defined the required percentage of items by science content (physical, living, Earth and space). This is similar to other state and national assessments that intend to report on subdomains of science to showcase student learning in each content area.

[63] This is the only type of knowledge relating to the science subdomains of life, physical, and Earth and space systems.

[64] Knowledge of physical systems seems to combine both chemistry and physics science content.

Document Section/Feature (Pg. #)	Subtheme Unidimensionality	Subtheme Multidimensionality	Annotation
Table 2.3 (Pg. 30)		"Desired distribution of items, by type of knowledge"	OECD lists content knowledge as a dimension separate from procedural and epistemic knowledge. This does not clarify if content knowledge by itself is built of separate subdomains.
Table 2.4 (Pg. 31)		"Desired distribution of items for knowledge"	OECD provides required item counts needed for the three types of knowledge by science content subdomain indicating each subdomain is its own dimension.
Figure 2.17 (Pg. 36)		"Framework categories"	Knowledge type is referred to as content indicating that science content is its own domain, but subdomains indicating multiple dimensions is not reported. However, as shown in the released item in this study's **Figure 4**, the item developers do document to which subdomain an item should be coded.

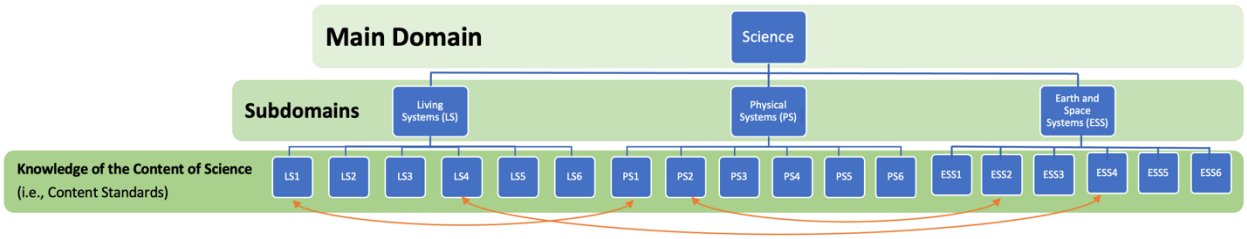
Note. Page number refers to the page numbers shown at the bottom of each page of the PISA 2015 Science Framework provided in Appendix F.

The 2015 PISA Science Framework document's Figure 2.5 seems to provide a strong piece of evidence for multidimensionality of the science subdomains in the form of the differentiated and unconnected pieces of content knowledge that PISA assesses for science. Based on this knowledge, Figure 11 was developed to visualize any crossover of knowledge that was shown in **Error! Reference source not found.**, which replicates and codes OECD's Figure 2.5. Any crossover noted below is not a concrete occurrence on the PISA science items. Item developers often carefully consider aspects like content crossover and develop the required

items in a manner that prevents cluing of other items. My subject matter expertise was used to provide examples in Figure 11 where content knowledge in one subdomain may benefit from content knowledge in another subdomain, thus potentially affecting multidimensionality by making the subdomains less distinct. The orange curved arrows indicate this possible content knowledge crossover in the content standards. For example, PS1 is the knowledge of the structure of matter and having this knowledge might increase a student's understanding of LS1, which is the knowledge of cells.

Figure 11

Possible Connections between 2015 PISA Science Content Knowledge



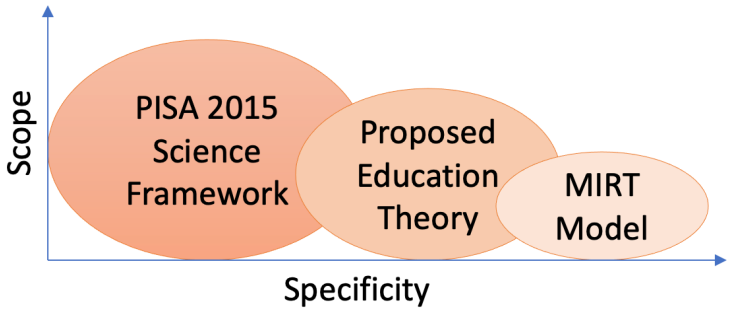
Since the qualitative results above seem to indicate that science is multidimensional with the three subdomains being assessed in 2015 PISA having little crossover, it seems a multidimensional IRT model might be warranted. However, there is insufficient information in the framework on the expected separability (or "difference") between the theoretical elements seen – is there expected to be enough difference to perceive in a large-scale assessment? Also, there is little or no discussion of confounds. For instance, are high performing respondents in one area expected to be high performing in the others? While this is not necessarily considered theoretically true for U.S. NGSS disciplinary core ideas specifically in the three subject matter

areas, it is expected theoretically to be more applicable by their nature for scientific and engineering practices (SEPs)[65] and cross-cutting concepts (CCCs)[66].

The continuum between the framework, the proposed educational theory, and its indicated model is provided in Figure 12. The proposed educational theory is that distinct science subdomains require students to have differentiated knowledge to demonstrate mastery of each subdomain, which may be more accurate for disciplinary core ideas than for practices and cross-cutting concepts.

Figure 12

Proposed Continuum



Note. Adapted from "A guide and glossary on postpositivist theory building for population health," 2006, by R. M. Carpiano, and D. M. Daley, 2006, *Journal of Epidemiology and Community Health*, *60*, p. 566.

Results Relating to RQ2

The descriptive statistics, histograms, and correlations are provided first then in subsection RQ2A: Cluster Analyses Results the cluster analysis and its sensitivity test are reported. In subsection RQ2B: PCA Results PCA results are provided. Subsection RQ2C: IRT

[65] OECD refers to content similar to this as procedural knowledge in the 2015 framework (OECD, 2017a).
[66] OECD refers to content similar to this as epistemic knowledge in the 2015 framework (OECD, 2017a).

Results contains IRT model fit for two different item clusters. Lastly, triangulation results are reported. Unless otherwise noted all results are for the whole student subsamples taking each item cluster (S10 and S11) analyzed rather than for the full U.S. sample. If results pertain to the half subset of the cluster S10 subsample that is also noted.

Descriptive Statistics

Table 7 below provides several statistics for each item and the three added variables of number of items attempted, student raw score, and average score. Curran et al. (1996) recommend for multivariate normality that skew not range outside of +/-2 and kurtosis outside of +/-7, which the data in Table 7 do not violate. The mean for the item variables remains stable, in other words the actual mean does not stray greatly from 0.5, and indicates all the items are on the same scale (a mean of say 15 might indicate an item on a different scale when all other item means hover between 0.1 and 0.8).

Table 7

Descriptive Statistics for Item Cluster S10 Full Subsample

Variable	n	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
DS625Q01C	1306	0.52	0.50	1.00	0	1	1	-0.09	-1.99	0.01
CS625Q02S	1306	0.64	0.48	1.00	0	1	1	-0.58	-1.67	0.01
CS625Q03S	1306	0.58	0.49	1.00	0	1	1	-0.31	-1.90	0.01
CS615Q07S	1306	0.29	0.45	0.00	0	1	1	0.94	-1.11	0.01
CS615Q01S	1306	0.82	0.38	1.00	0	1	1	-1.71	0.91	0.01
CS615Q02S	1306	0.48	0.50	0.00	0	1	1	0.09	-1.99	0.01
CS615Q05S	1306	0.18	0.39	0.00	0	1	1	1.65	0.73	0.01
CS604Q02S	1306	0.49	0.50	0.00	0	1	1	0.06	-2.00	0.01
DS604Q04C	1306	0.29	0.45	0.00	0	1	1	0.93	-1.14	0.01
CS645Q03S	1306	0.51	0.50	1.00	0	1	1	-0.02	-2.00	0.01
DS645Q04C	1306	0.57	0.50	1.00	0	1	1	-0.27	-1.93	0.01
DS645Q05C	1306	0.14	0.35	0.00	0	1	1	2.04	2.18	0.01
CS657Q01S	1306	0.71	0.46	1.00	0	1	1	-0.91	-1.18	0.01
CS657Q02S	1306	0.42	0.49	0.00	0	1	1	0.32	-1.90	0.01
CS657Q03S	1306	0.47	0.50	0.00	0	1	1	0.14	-1.98	0.01
Number Attempted	1306	15.00	0.00	15.00	15	15	0	NaN	NaN	0.00
Raw Score	1306	7.09	3.37	7.00	0	15	15	0.12	-0.91	0.09
Average	1306	0.47	0.22	0.47	0	1	1	0.12	-0.91	0.01

Figure 13 is a histogram of average scores for the full U.S. science sample while Figure 14 provides the same information for the cluster S10 subsample. Both populations seem fairly normal in distribution, but the subsample population does show less variability in average scores. Figure 15 is a set of histograms showing frequency of 0 and 1 scores for each item. A distance heatmap using the correlations from the polychoric matrix is shown in Figure 16 with the blue color indicating high similarity (close distance together) between items while the orange color indicates low similarity (far apart from each other) between items. Table 8 provides the polychoric correlations with the majority of items being weakly and negatively correlated – a negative correlation indicates that as a student does well on one item they score less on the other item.

Figure 13

Histogram of Student Average Scores for Full U.S. Science Sample

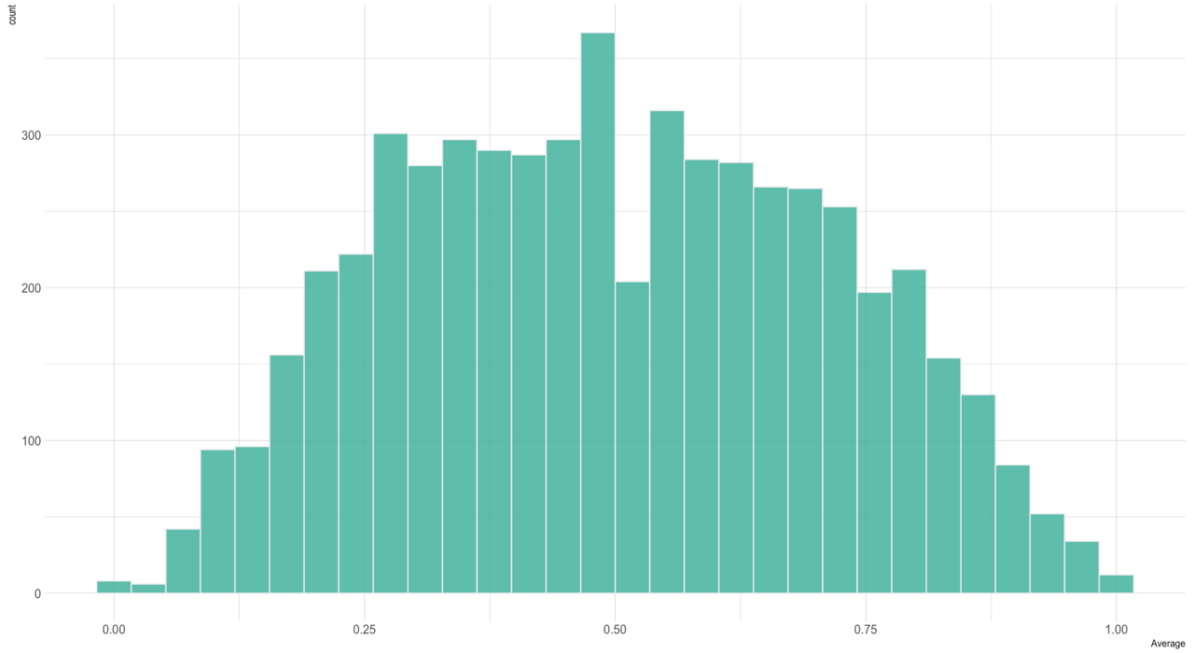


Figure 14

Histogram of Student Average Scores for Item Cluster S10 Full Subsample

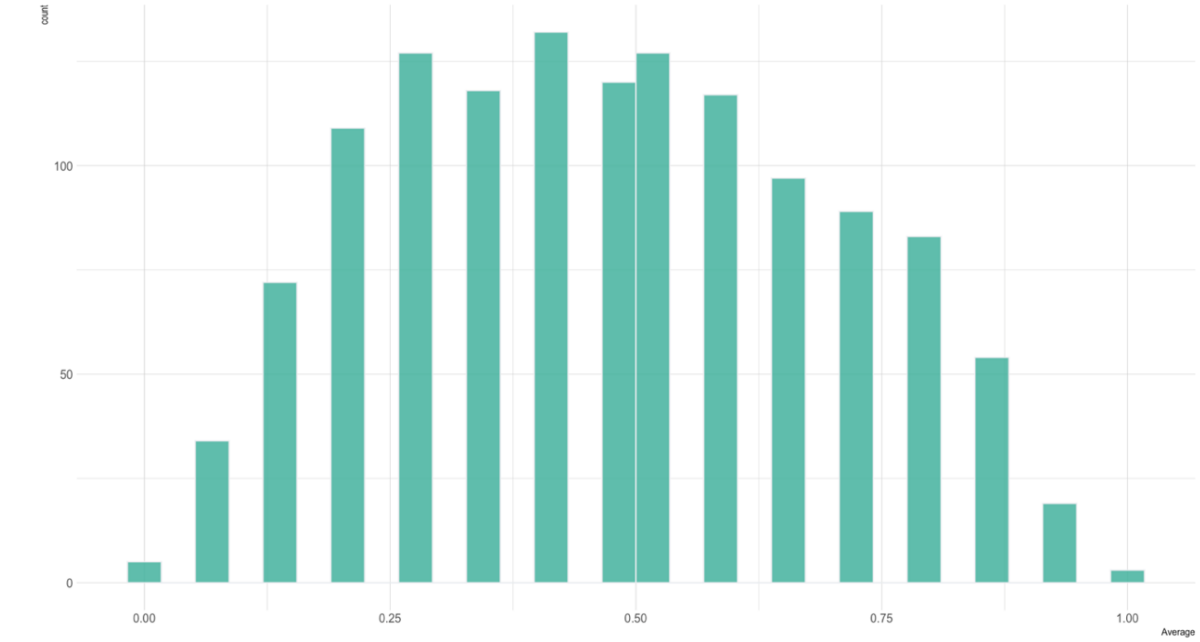


Figure 15

Histograms of Student Score Point Frequency for Item Cluster S10 Full Subsample

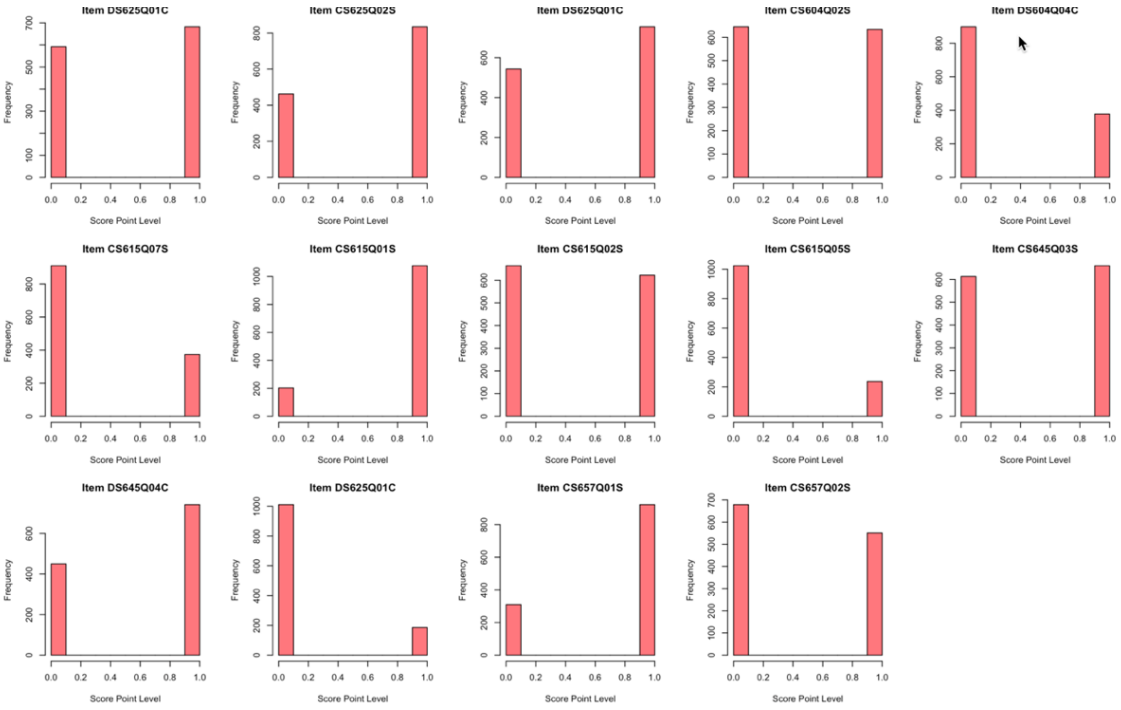


Figure 16

Distance Heatmap for Item Cluster S10 Full Subsample

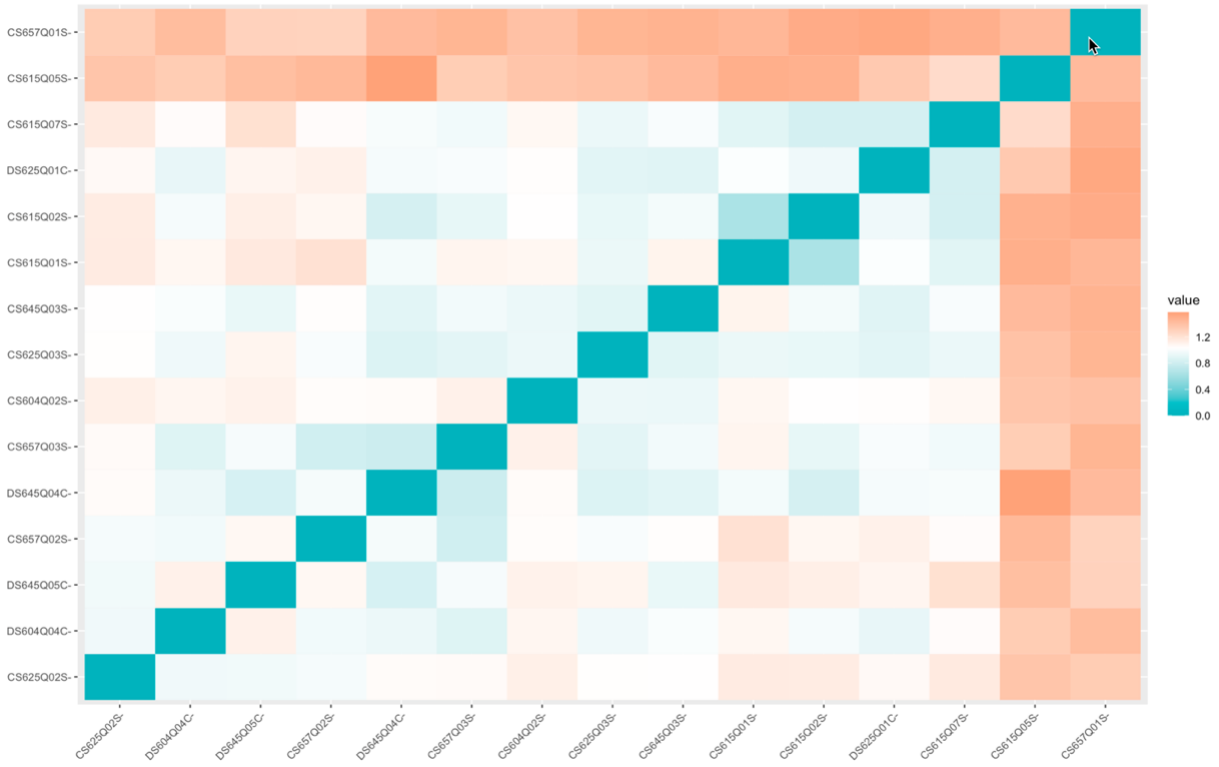


Table 8

Means (M), Standard Deviations (SD), and Correlations with Confidence Intervals (CI) for Item Cluster S10's Full Subsample

Item	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. DS625Q01C	0.35	0.21														
2. CS625Q02S	0.32	0.20	-.02													
			[-.53, .50]													
3. CS625Q03S	0.37	0.19	.22	-.00												
			[-.33, .66]	[-.51, .51]												
4. CS615Q07S	0.37	0.20	.40	-.08	.13											
			[-.15, .75]	[-.57, .45]	[-.41, .60]											
5. CS615Q01S	0.33	0.23	.13	-.07	.32	.38										
			[-.41, .60]	[-.56, .46]	[-.23, .71]	[-.17, .75]										
6. CS615Q02S	0.37	0.21	.21	-.07	.17	.40	.65**									
			[-.34, .65]	[-.56, .46]	[-.38, .63]	[-.14, .76]	[.21, .87]									
7. CS615Q05S	0.20	0.24	-.08	-.24	-.15	.09	-.42	-.26								
			[-.57, .45]	[-.67, .31]	[-.61, .40]	[-.44, .58]	[-.77, .12]	[-.68, .29]								
8. CS604Q02S	0.32	0.20	.09	-.07	.19	-.02	.11	.12	-.20							
			[-.44, .58]	[-.56, .46]	[-.36, .64]	[-.53, .50]	[-.43, .59]	[-.42, .59]	[-.65, .35]							
9. DS604Q04C	0.34	0.20	.26	.18	.09	.01	.11	.20	-.11	-.12						
			[-.29, .68]	[-.37, .63]	[-.44, .58]	[-.51, .52]	[-.42, .59]	[-.34, .65]	[-.59, .43]	[-.59, .42]						
10. CS645Q03S	0.34	0.21	.41	.11	.24	.15	-.01	.15	-.19	.17	.14					
			[-.12, .76]	[-.42, .59]	[-.31, .67]	[-.39, .62]	[-.52, .50]	[-.39, .62]	[-.64, .36]	[-.37, .63]	[-.40, .61]					
11. DS645Q04C	0.41	0.19	.29	.12	.33	.24	.34	.41	-.40	.16	.22	.31				
			[-.26, .70]	[-.42, .60]	[-.22, .72]	[-.31, .67]	[-.21, .73]	[-.13, .76]	[-.76, .14]	[-.39, .62]	[-.33, .66]	[-.25, .71]				
12. DS645Q05C	0.29	0.22	-.06	.21	-.02	-.26	-.15	-.10	-.24	.01	-.08	.27	.39			
			[-.55, .47]	[-.34, .65]	[-.53, .50]	[-.68, .29]	[-.62, .39]	[-.58, .44]	[-.67, .31]	[-.50, .52]	[-.57, .45]	[-.28, .69]	[-.15, .75]			
13. CS657Q01S	0.16	0.24	-.41	-.17	-.23	-.30	-.11	-.27	-.34	-.17	-.34	-.47	-.14	-.17		
			[-.76, .13]	[-.63, .37]	[-.67, .32]	[-.71, .25]	[-.59, .43]	[-.69, .28]	[-.73, .20]	[-.63, .37]	[-.73, .21]	[-.79, .06]	[-.61, .40]	[-.63, .38]		
14. CS657Q02S	0.32	0.21	.03	.28	.17	.08	-.10	.07	-.27	.12	.22	.04	.12	-.02	-.24	
			[-.49, .53]	[-.27, .69]	[-.38, .63]	[-.45, .57]	[-.58, .43]	[-.46, .56]	[-.69, .28]	[-.42, .60]	[-.33, .66]	[-.48, .54]	[-.42, .60]	[-.53, .50]	[-.67, .31]	
15. CS657Q03S	0.37	0.20	.17	.11	.27	.20	.09	.26	.12	-.02	.35	.15	.25	.09	-.58*	.42
			[-.37, .63]	[-.43, .59]	[-.28, .69]	[-.34, .65]	[-.44, .58]	[-.29, .68]	[-.42, .59]	[-.53, .50]	[-.19, .73]	[-.39, .61]	[-.30, .67]	[-.45, .57]	[-.84, -.10]	[-.11, .77]

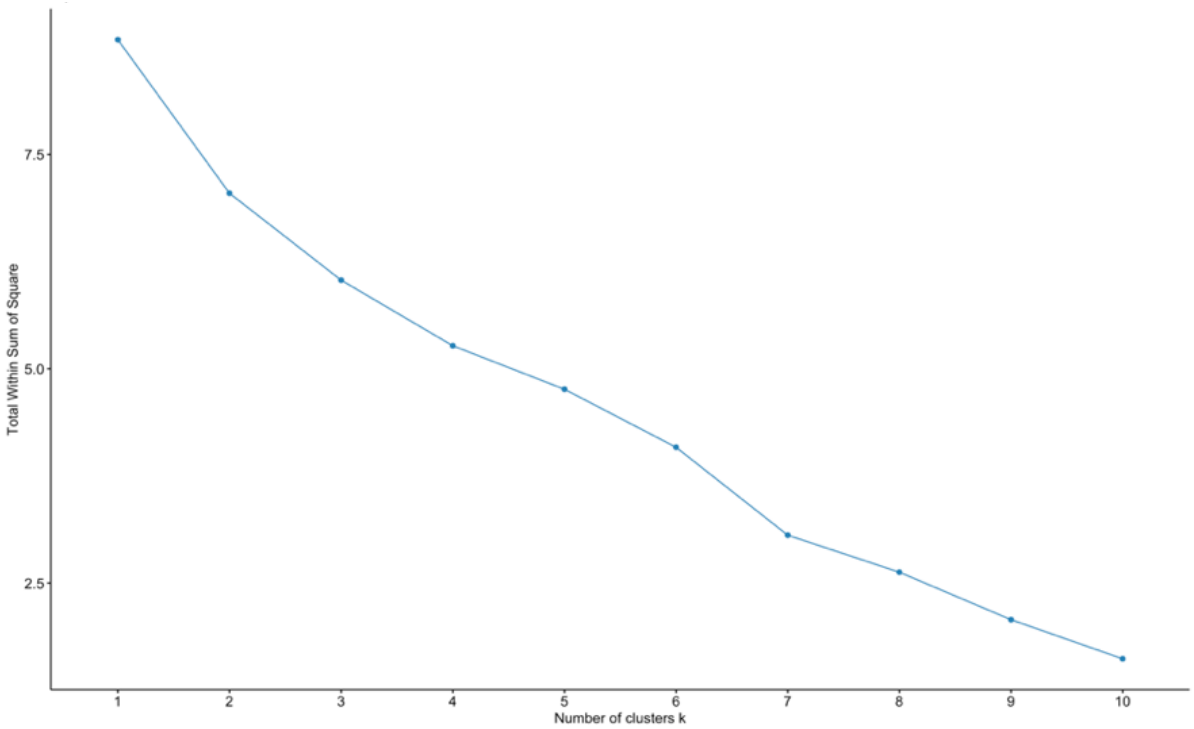
Note. Values in square brackets indicate the 95% CI for each correlation. The CI provides a range of population correlations that describe where the sample correlation may truly lay (Cumming, 2014). * Indicates $p < .05$ and ** indicate $p < .01$.

RQ2A: Cluster Analyses Results

The following scree plot in Figure 17 does not show a distinct elbow bend at any number of clusters because the slope is constantly decreasing without leveling off. There is no indication of multidimensionality from this analysis.

Figure 17

Scree Plot for Item Cluster S10 with Full Subsample



A second scree plot, shown in Figure 18, was obtained, as described in Chapter 2. Methods: Conducting Quantitative Analyses – Cluster Analyses, from the randomly chosen half of the cluster S10 student subsample. No elbow bend indicating optimal number of clusters is visible so there is no clear multidimensionality based on this analysis – no place where establishing clear dimensions would seem most appropriate.

Figure 18

Scree Plot for Item Cluster S10 with Random Half of Subsample

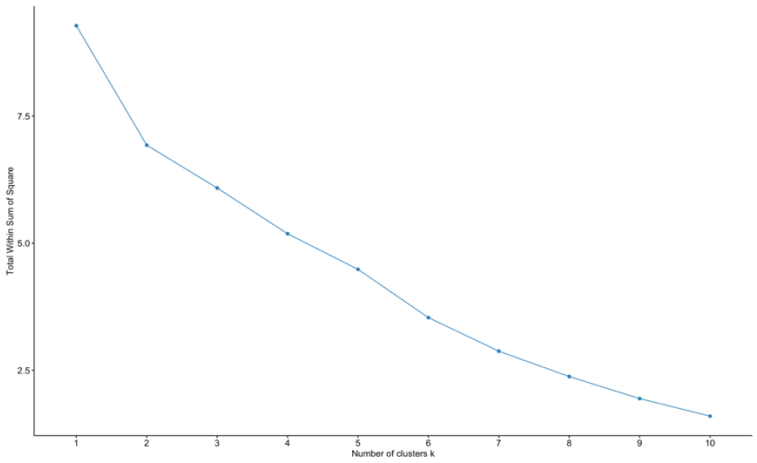
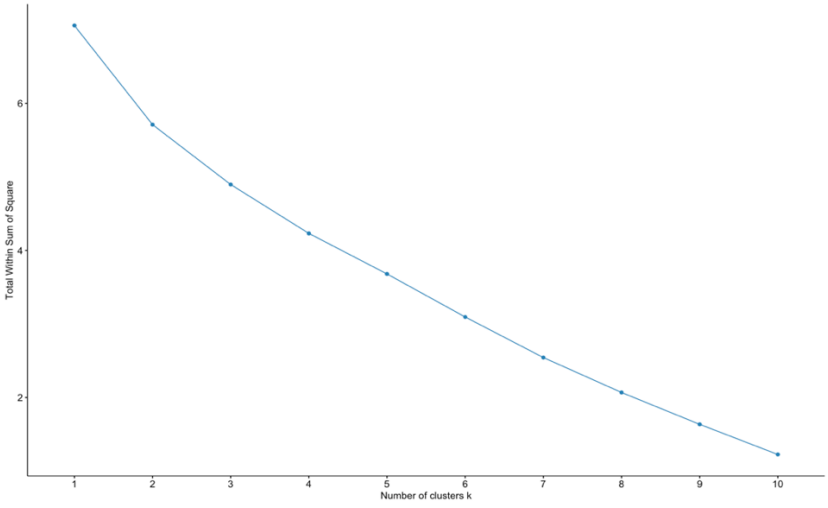


Figure 19 provides a scree plot that lacks a clear elbow bend, indicating no clusters for item cluster S11, thus revealing no clear multidimensionality by the above definition.

Figure 19

Scree Plot for Item Cluster S11



RQ2B: PCA Results

Figure 20 provides bar plots of the three PCA loadings (eigenvectors) explaining the most variation in the set of items in cluster S10. Principal components 4-15 were dropped as

they all explained 2.5% or less of the variation and were each far below eigenvalues of 1. The reader can note below that the only for PC 1 was the eigenvalue clearly 1 or greater, but since my research questions are about multidimensionality in three dimensions as compared to the usual unidimensionality with which PISA data are modeled, PC1-PC3 were kept. Per prcomp function documentation, “The signs of the columns of the rotation matrix (the loadings) are arbitrary, and so may differ between different programs for PCA, and even between different builds of R” (R Core Team, 2023). Only for PC 1 (2.01) was the eigenvalue greater than 1, PC 2 (0.1) and PC 3 (0.08) were much smaller. Eigenvalues greater than 1 indicate those components “account for more than the mean of the total variance in the items,” which follows the Kaiser-Guttman rule (Li, 2012). The vast majority of items load most strongly on PC 1, with only 1 item each loading most strongly for PC 2 and PC 3, as shown in the figures.

Figure 20

Loadings Bar Plots for Item Cluster S10 with Full Subsample

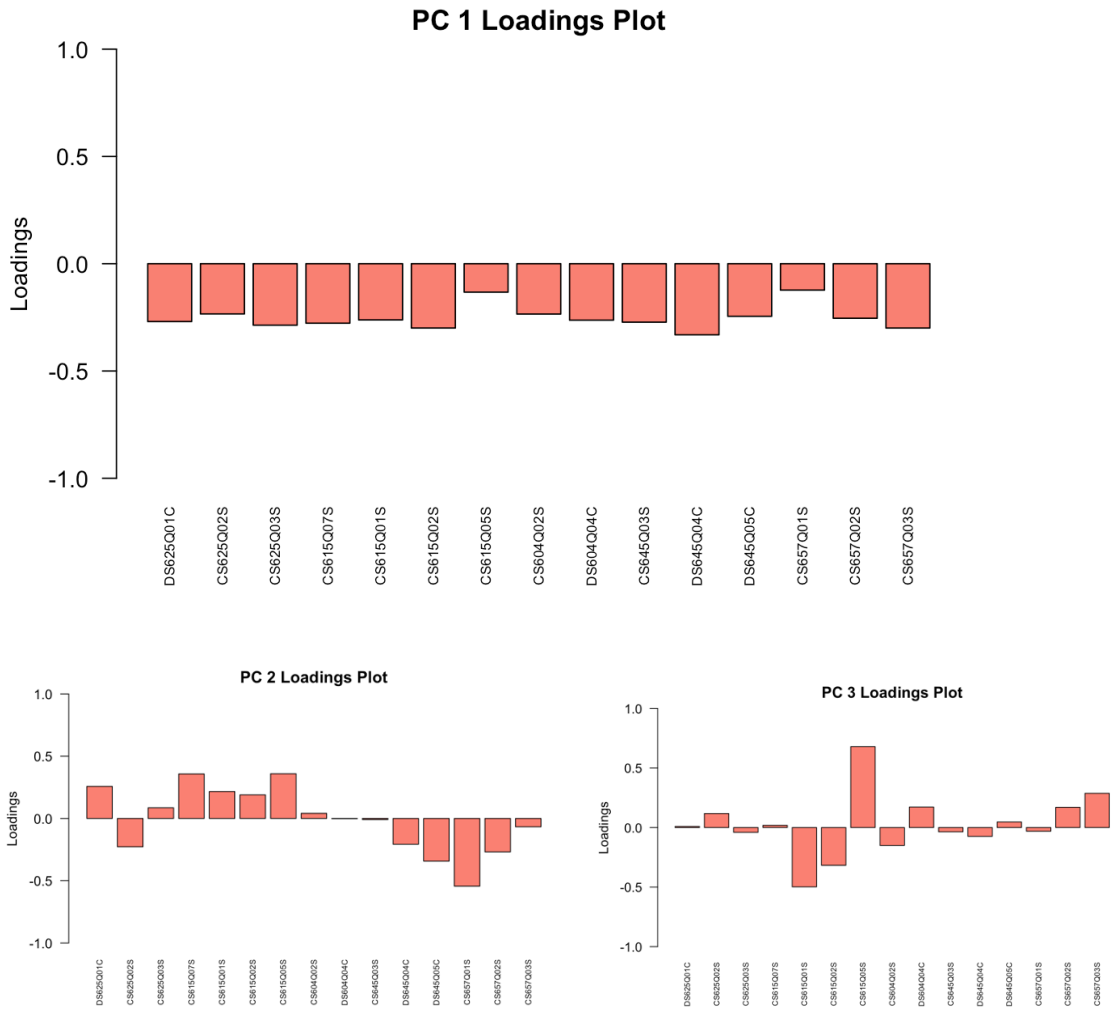
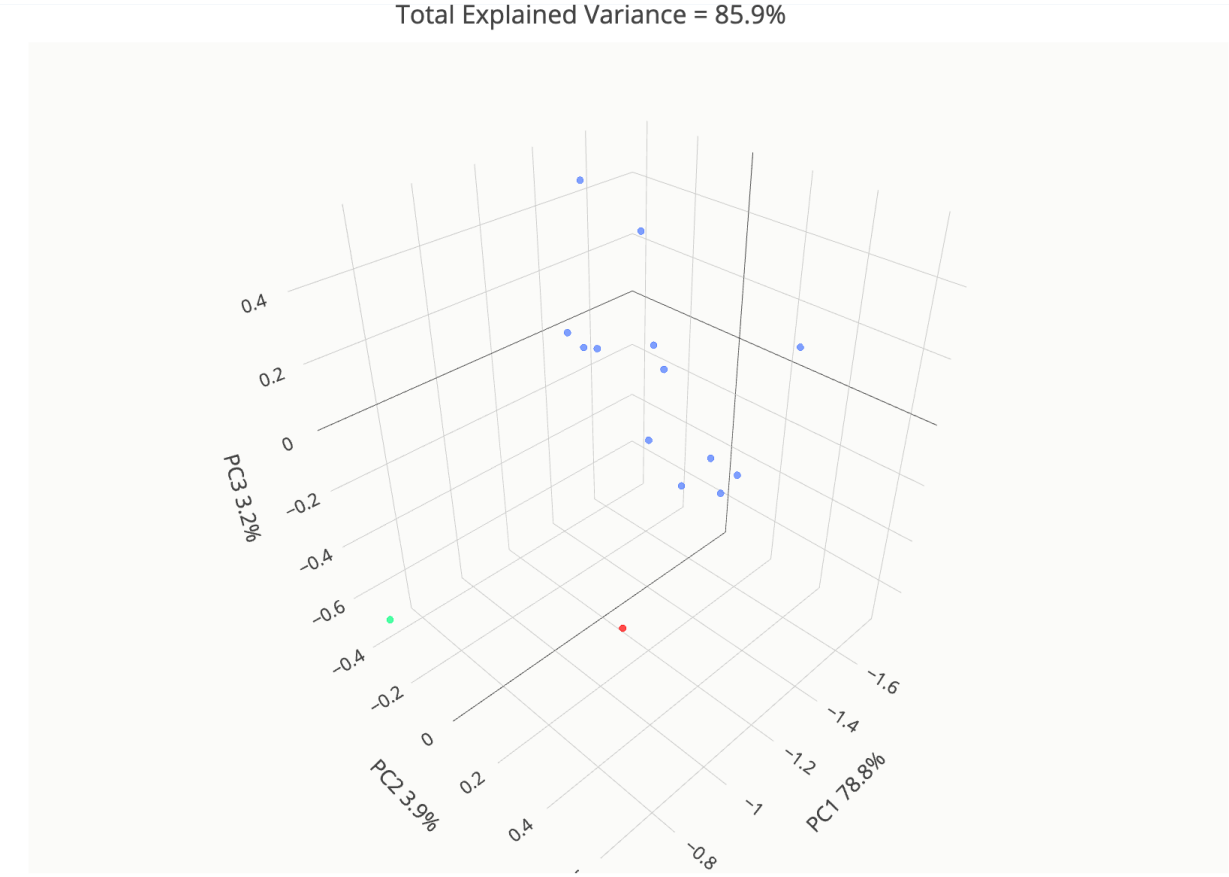


Figure 21 visualizes the three principal components in a 3D space based on the PCA scores. In general, PCA scores are developed by taking the original measurement and multiplying it by its eigenvector coefficient, which indicates the contribution of each measurement, and then additively summarizing all those values to get the score on any axis (Carr, 2001). However, Carr (2001) indicates that when a coefficient matrix, like this study uses, is used in place of a covariance matrix to run the PCA the calculated results will not match the results computed in a R program. Note, the low redundancy - data are spread apart and weakly

correlated. The two items isolated in PC 2 and PC 3 will be dropped as outliers in the second round of IRT analyses for item cluster S10, as no single item ever serves as a full measure in IRT. Multidimensionality does not seem to be indicated in this item cluster as the majority of variance (78.8%) is explained by PC 1. An interactive graph with rotation and data point hovering is available at https://chart-studio.plotly.com/~cnmalcom/3.

Figure 21

PCA Plot for Item Cluster S10 with Full Subsample



Note. Blue dots are items loading mainly on PC 1, while red dots indicate items loading mainly on PC 2, and green dots indicate items mainly loading on PC 3.

A second analysis for the random half subset of the set of items in cluster S10 was used to help confirm PCA results. Figure 22 provides bar plots of the three PCA loadings (eigenvectors) explaining the most variation. Principal components 4-15 were dropped as explained above. Only for PC 1 (2.18) was the eigenvalue greater than 1, PC 2 (0.11) and PC 3 (0.1) were much smaller. The majority of items load most strongly on PC 1.

Figure 22

Loadings Bar Plots for Item Cluster S10 with Random Half of Subsample

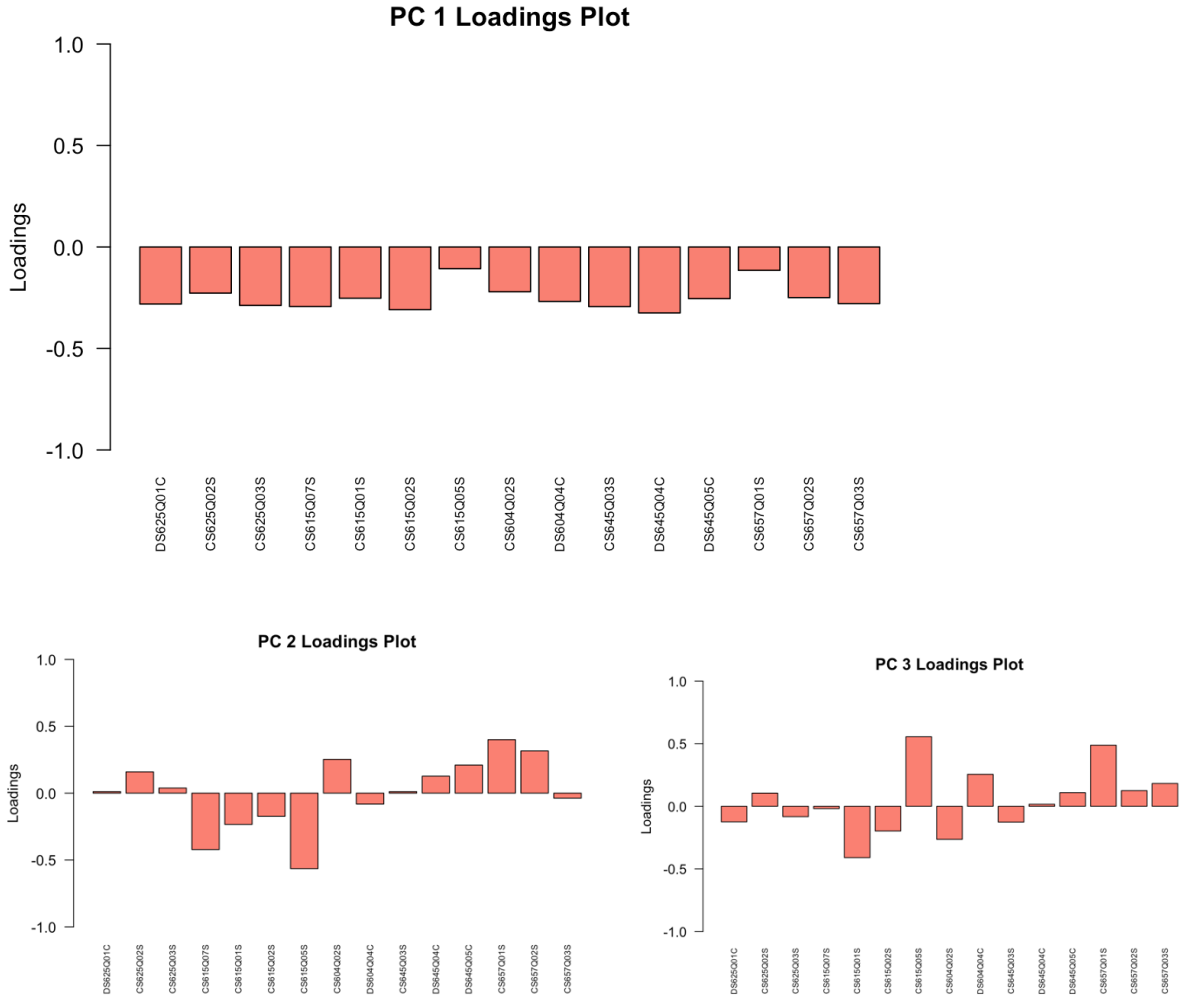
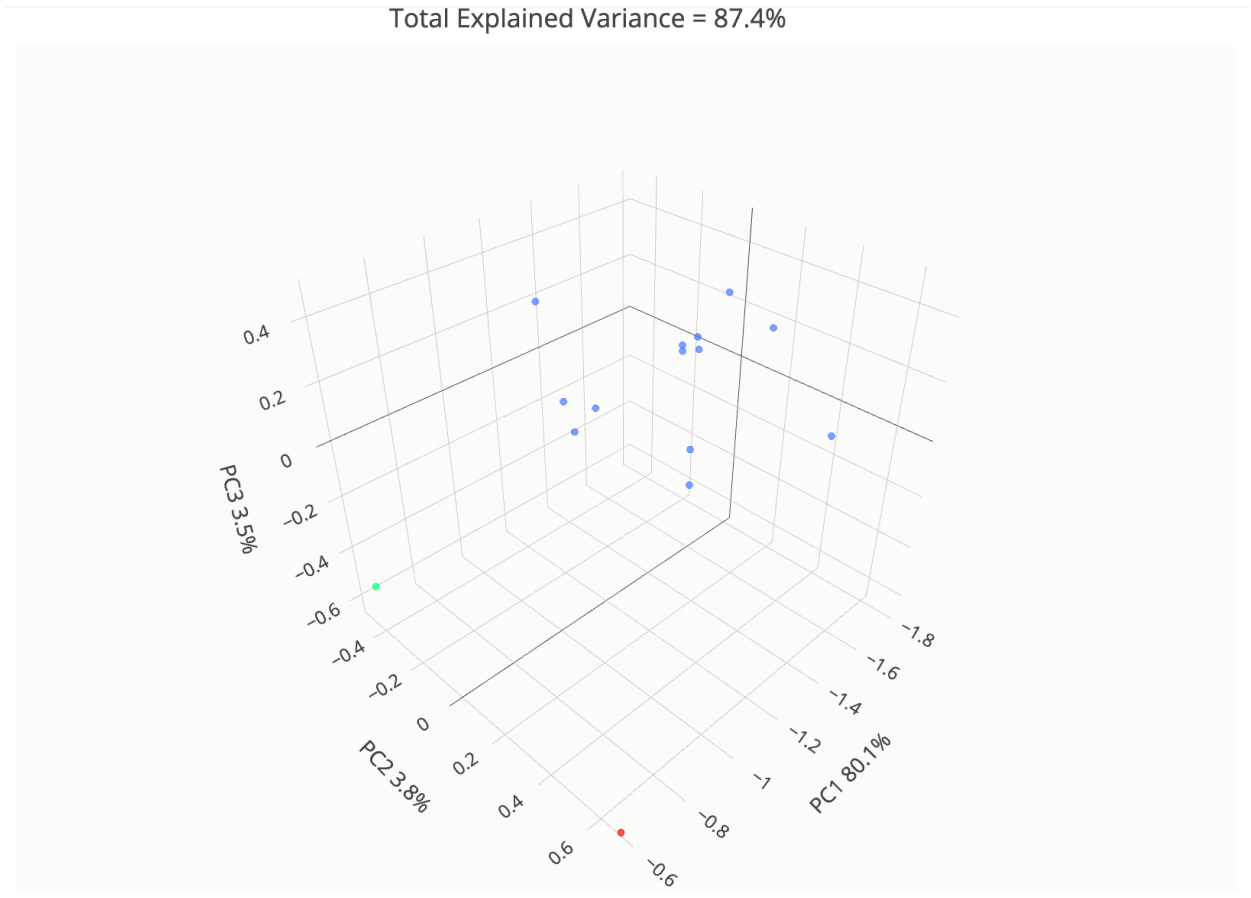


Figure 23 visualizes the three principal components in a 3D space based on the PCA scores. Again, the plot shows low redundancy - data are spread apart and weakly correlated. The same two items from the earlier PCA of the full subsample from item cluster S10 are still showing up independently on PC 2 and 3. Multidimensionality does not seem to be indicated as the majority of variance (80.1%) is explained by PC 1. This plot, as expected, is very similar in substance to Figure 21. An interactive graph with rotation and data point hovering is available at https://chart-studio.plotly.com/~cnmalcom/5.

Figure 23

Confirmation PCA Plot for Item Cluster S10 with Random Half of Subsample



A third PCA for the set of items in cluster S11 was run. Figure 24 provides bar plots of the three PCA loadings (eigenvectors) explaining the most variation. Principal components 4-15 were dropped as they all explained 1% or less of the variation. Only for PC 1 (2.86) was the eigenvalue greater than 1, PC 2 (0.1) and PC 3 (0.05) were much smaller. The majority of items load on PC 1.

Figure 24

Loadings Bar Plots for Item Cluster S11

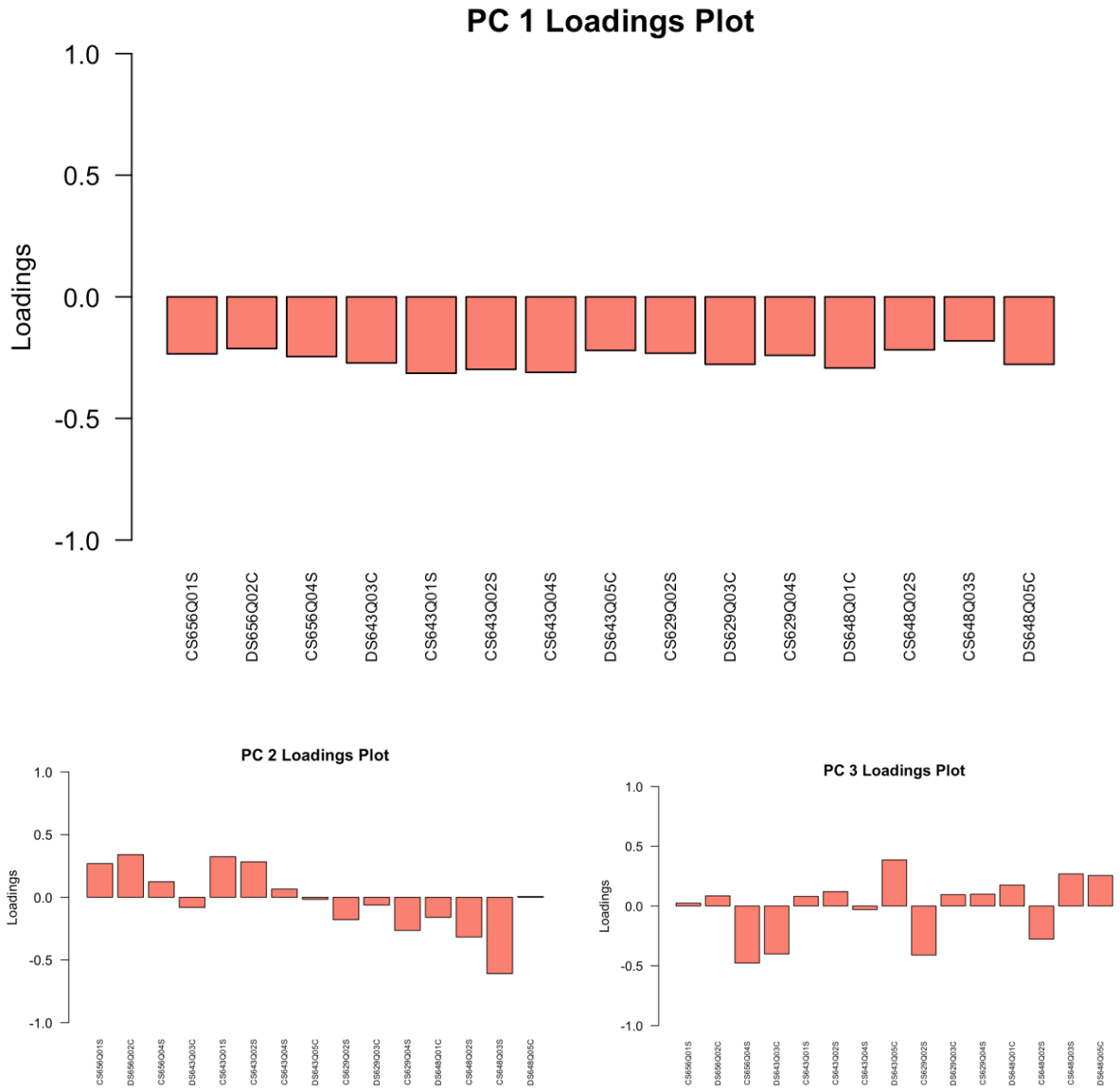
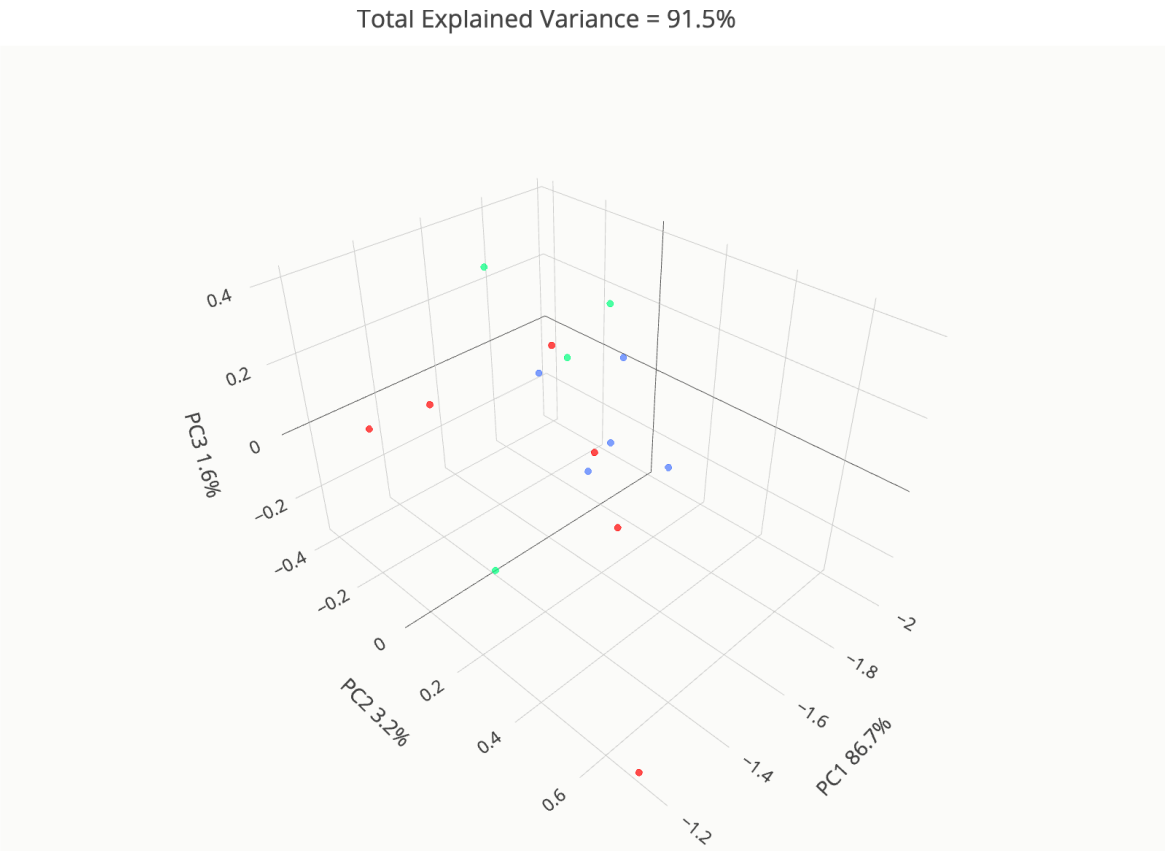


Figure 25 visualizes the three principal components in a 3D space based on the PCA scores. Again, the plot shows low redundancy - data are spread apart and weakly correlated. Multidimensionality does not seem to be indicated as the majority of variance (86.7%) is explained by PC 1. While more items are loading on PC 2 and PC 3 when only partial data are used, they do not appear to be in separate dimensions and do not align with science subdomains or with types of knowledge identified by OECD in the 2015 science framework. An interactive graph with rotation and data point hovering is available at https://chart-studio.plotly.com/~cnmalcom/7.

Figure 25

PCA Plot for Item Cluster S11



RQ2C: IRT Results

Both model fit and information from IRT analyses are provided in this section. Table 9 provides the subdomain groupings of items for each item cluster that were used in the multidimensional models. Physical System items were coded as Dimension 1, Earth and Space Systems as Dimension 2, and Living Systems as Dimension 3.

Table 9

Item Groupings for MIRT Models

Cluster S10				Cluster S11			
Item	Type of Knowledge	Subdomain	DOK	Item	Type of Knowledge	Subdomain	DOK
DS625Q01C	Content	Physical	L	CS643Q01S	Procedural	Physical	M
CS625Q02S	Content	Physical	L	CS643Q02S	Procedural	Physical	M
CS625Q03S	Content	Physical	M	DS643Q03C	Content	Physical	L
CS604Q02S	Content	Physical	M	CS643Q04S	Procedural	Physical	M
DS604Q04C	Epistemic	Physical	M	DS643Q05C	Epistemic	Physical	M
				CS629Q02S	Content	Physical	M
CS615Q07S	Procedural	Earth and Space	M	CS629Q04S	Epistemic	Physical	M
CS615Q01S	Procedural	Earth and Space	M				
CS615Q02S	Procedural	Earth and Space	M	DS648Q01C	Procedural	Earth and Space	M
CS615Q05S*	Epistemic	Earth and Space	M	CS648Q02S	Procedural	Earth and Space	M
CS645Q03S	Content	Earth and Space	M	CS648Q03S	Procedural	Earth and Space	M
DS645Q04C	Content	Earth and Space	M	DS648Q05C	Epistemic	Earth and Space	M
DS645Q05C	Content	Earth and Space	M	DS629Q03C	Procedural	Earth and Space	M
CS657Q01S*	Content	Living	L	CS656Q01S△	Content	Living	M
CS657Q02S	Content	Living	M	DS656Q02C△	Procedural	Living	H
CS657Q03S	Procedural	Living	M	CS656Q04S△	Procedural	Living	M

Note. *Items that were dropped from Models 3B and 4b. △Released item set. DOK = Depth of Knowledge: Low (L), Medium (M), High (H), developed by Webb in 1997 per OECD (2017a).

Cluster S10. Importantly, since the two items CS657Q01S and CS615Q05S each loaded in a separate dimension by themselves, the model fit analyses were conducted both with and

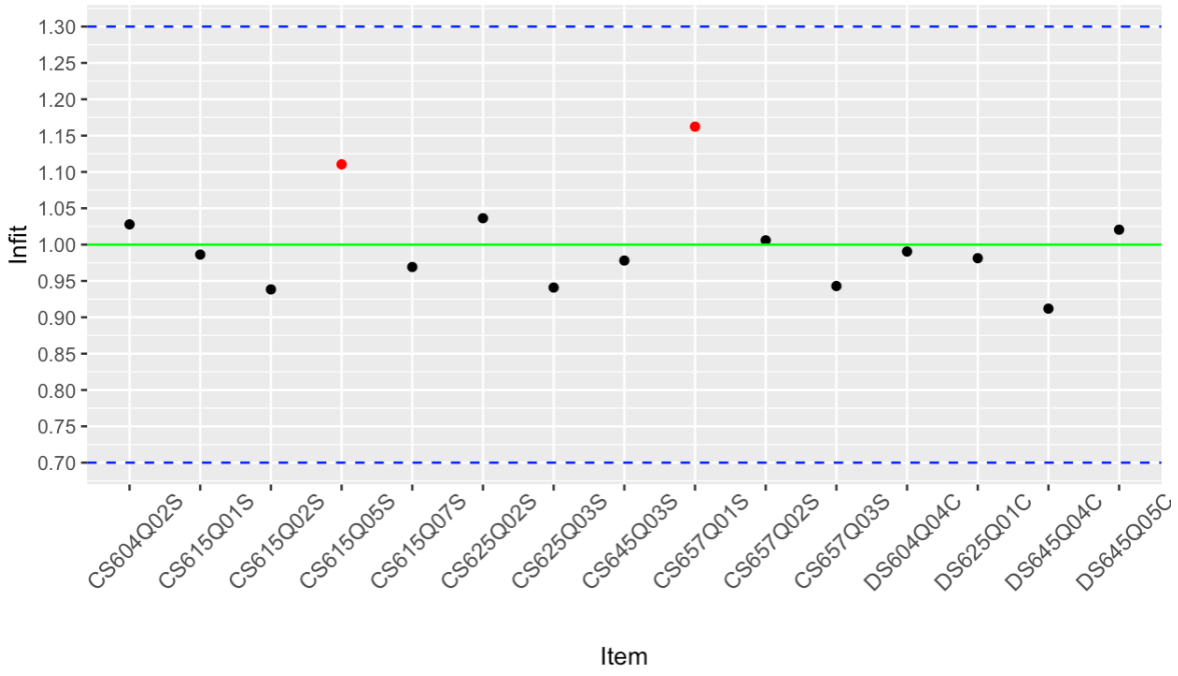
without these items in the item cluster S10 as shown in Table 10. Based on item difficulties for Model 1, these unusual items also have unusual characteristics: item CS615Q05S (xsi = 1.79) seems to be the second hardest item, item DS645Q05C (xsi = 2.08) being the hardest, and item CS615Q01S (xsi = -2.01) ranked the easiest while item CS657Q01S (xsi = -1.34) was the second easiest. This holds true for Models 2, 3, and 4 while the difficulty levels vary slightly.

Item infit (information-weighted fit) statistics for Model 1 showed that some items were slightly underfit and some slightly overfit when compared to the desired value of 1, but all were well within usual tolerances[67] of 0.70 to 1.30 (blue dashed lines in Figure 26). Figure 26 shows the infit statistic for each item with the desired expected value of 1.0 (green line). The two outlier items are marked in red. Note that items below 1.0 may have too predictable of responses and those greater than 1.0 may have responses that are too noisy, i.e., the excess variation may be masking what is a good model of the response pattern (Wind & Hua, 2021).

[67] These tolerances were adopted based off usage in a similar study by Pensavalle and Solinas (2013).

Figure 26

Infit Statistics for 1 PL UIRT Model of Item Cluster S10 with Full Subsample



There were 6 items whose infit statistic was significant at $p < 0.05$ for Models 1 and 3. Infit statistics remained close to the desired value of 1 for Models 2 and 4, and none were statistically significantly different. Models with items dropped have 3 items (Model 1b) and 2 items (Model 3b) that had a significant infit statistic respectively.

Table 10

Model Fit Indices for Comparison of Relative Model Fit – Item Cluster S10 Subsample

	Full Set of 15 Items				Without Items CS657Q01S and CS615Q05S			
	Model 1 (1PL UIRT)	Model 2 (2PL UIRT)	Model 3 (1PL MIRT)	Model 4 (2PL MIRT)	Model 1b (1PL UIRT)	Model 2b (2PL UIRT)	Model 3b (1PL MIRT)	Model 4b (2PL MIRT)
Deviance (-2 log-likelihood)	21,630.16	21,456.08	21,561.67	21,426.74	18,948.83	Not Run	18,925.41	Not Run
Number of Estimated Parameters	16	30	21	33	14		19	

AIC (constraint on students)	21,662	21,516	21,604	21,493	18,977	18,963
BIC	21,745	21,671	21,712	21,664	19,049	19,062
Iterations	15	21	428	151	17	389
EAP Reliabilities	0.74	0.76			0.75	
Dim 1	NA	NA	0.725	0.735	NA	0.736
Dim 2	NA	NA	0.729	0.743	NA	0.728
Dim 3	NA	NA	0.715	0.657	NA	0.635
Infit Range	0.915 to 1.159	0.989 to 1.015	0.905 to 1.146	0.985 to 1.011	0.929 to 1.001	0.933 to 1.081

Note. The green highlighted model is the most parsimonious when considering model statistics and guidelines.

Improvement of fit was determined for each model pair and is shown in Table 11, which can be interpreted for each pair of models as "from Model X to Model Y there is a Z% increase in improvement of model fit. Improvement of fit was determined with the following formula ((Model X Deviance – Model Y Deviance)/Model Y Deviance) * 100. Overall, improvement of the fit statistic for Model 1 (1PL UIRT) compared to Model 3 (1PL MIRT) for the three-dimensional model by content area is only 0.3%, which may not be meaningful enough for country level results. In other words, at least reporting at the group level for which representative samples were drawn (country in this case), the difference made by carrying forward all the extra parameters of the more complex model would not make a practical difference. Note, models for the full set of items versus the models lacking 2 items were not compared due to the difference in items, which made the models not able to be nested.

Table 11

Comparison of Model Fit – Item Cluster S10 Subsample

	Model 1 to 2	Model 1 to 3	Model 1 to 4	Model 3 to 2	Model 2 to 4	Model 3 to 4	Model 1b to 3b
Improvement of Fit	0.8%	0.3%	0.9%	0.5%	0.1%	0.6%	0.1%

Goodness of fit was also determined using chi-square test to compare sets of models with a significance level α set to 0.05. The models are compared below.

- Model 1 and Model 2: X^2 (14, N = 1,306) = 174.09, p = 0) so null hypothesis that both models are the same can be rejected and Model 2 is a significantly better fit based on lower deviance, AIC, and BIC statistics.
- Model 1 and Model 3: X^2 (5, N = 1,306) = 68.50, p = 0) so null hypothesis that both models are the same can be rejected and Model 3 is a significantly better fit based on lower deviance, AIC, and BIC statistics.
- Model 1 and Model 4: X^2 (17, N = 1,306) = 203.43, p = 0) so null hypothesis that both models are the same can be rejected and Model 4 is a significantly better fit based on lower deviance, AIC, and BIC statistics.
- Model 2 and Model 3: X^2 (9, N = 1,306) = 1015.59, p = 0) so null hypothesis that both models are the same can be rejected and Model 2 is a significantly better fit based on lower deviance, AIC, and BIC statistics.
- Model 2 and Model 4: X^2 (3, N = 1,306) = 29.34, p < 0.0001) so null hypothesis that both models are the same can be rejected and Model 4 is a significantly better fit based on lower deviance, AIC, and BIC statistics.
- Model 3 and Model 4: X^2 (12, N = 1,306) = 134.93, p = 0) so null hypothesis that both models are the same can be rejected and Model 4 is a significantly better fit based on lower deviance, AIC, and BIC statistics.

- Model 1b and Model 3b: X^2 (5, N = 1,306) = 23.42, $p < 0.0001$) so null hypothesis that both models are the same can be rejected and Model 3b is a significantly better fit based on lower deviance and AIC statistics (BIC was actually higher for Model 3b).

Cluster S11. Importantly, since the 1PL UIRT model was found to be the better fit for item cluster S10, only it and the 1PL MIRT model were compared for this item cluster as confirmation of unidimensional model fit. Analysis on IRT model fit is shown in Table 12. While Model 6 appears to be the better fitting model based on the lower AIC and BIC statistics, it does require more iterations to converge, and substantially more parameters are employed without making much practical difference in results at the group reporting level for this assessment. Model 5 has 9 items and Model 5 has 8 items whose infit statistic was significant at $p < 0.05$.

Table 12

Model Fit Indices for Comparison of Relative Model Fit – Item Cluster S11 Subsample

	Full Set of 15 Items	
	Model 5 (1PL UIRT)	Model 6 (1PL MIRT)
Deviance (-2 log-likelihood)	21,530.46	21,467.69
Number of Estimated Parameters	16	21
AIC (constraint on students)	21,562	21,510
BIC	21,645	21,618
Iterations	25	474
EAP Reliabilities	0.8	
Dim 1	NA	0.727
Dim 2	NA	0.752
Dim 3	NA	0.794
Infit Range	0.888 to 1.152	0.912 to 1.132

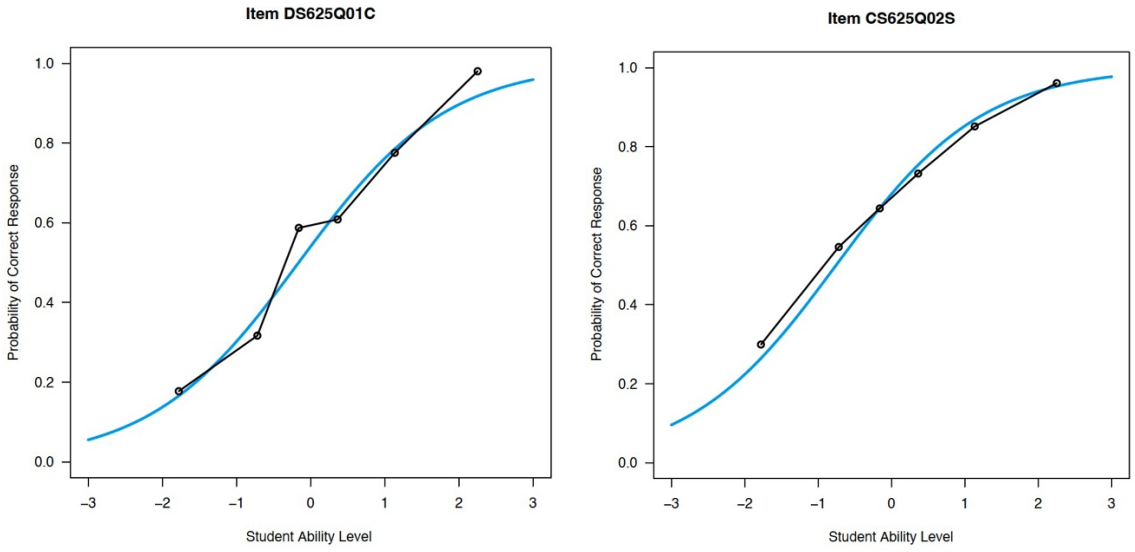
Note. The green highlighted model is the most parsimonious when considering model statistics and guidelines.

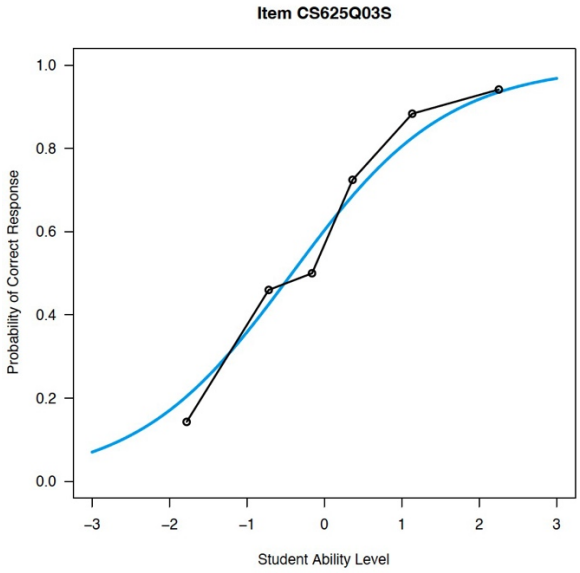
Improvement of fit was also determined for this model pair. Model 6 shows 0.3% improvement over the unidimensional Model 5. For goodness of fit chi-square test, Model 5 and Model 6: X^2 (5, N = 1,306) = 62.78, p = 0) so null hypothesis that both models are the same can be rejected and Model 6 is a significantly better fit based on lower deviance, AIC, and BIC statistics.

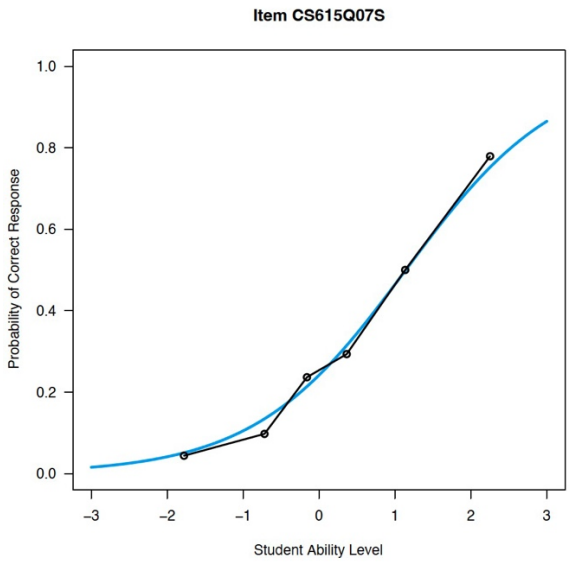
Item Fit Analyses. Using the best fitting model (Model 1b - 1PL UIRT) for item cluster S10 (full subsample) several analyses of item fit were completed. Figure 27 shows the ICCs for 13 items of item cluster S10 – the blue line is the model's expected scores curve and the black line is the actual scores curve. These plots showcase the relationship between the latent trait (student ability) and probability of an expected correct response (Wind & Hua, 2021).

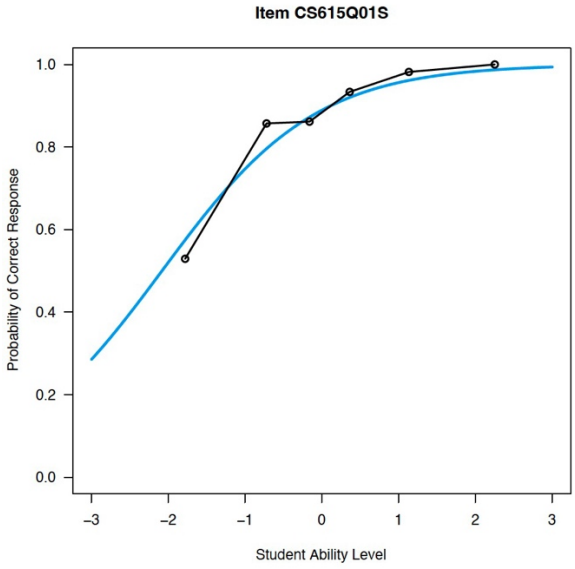
Figure 27

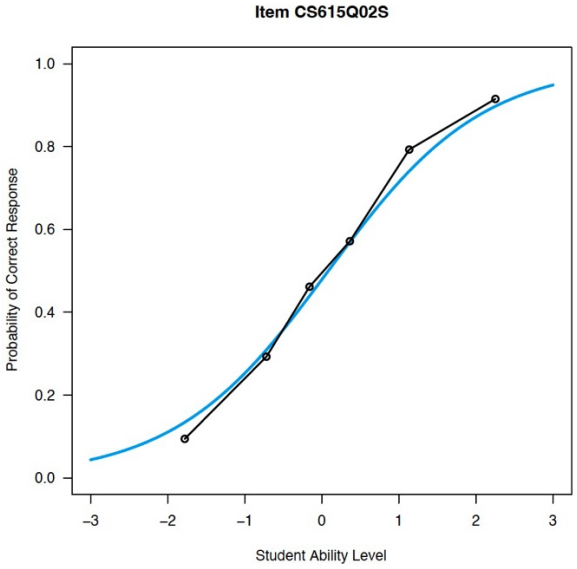
ICC Plots for Item Cluster S10 with Full Subsample

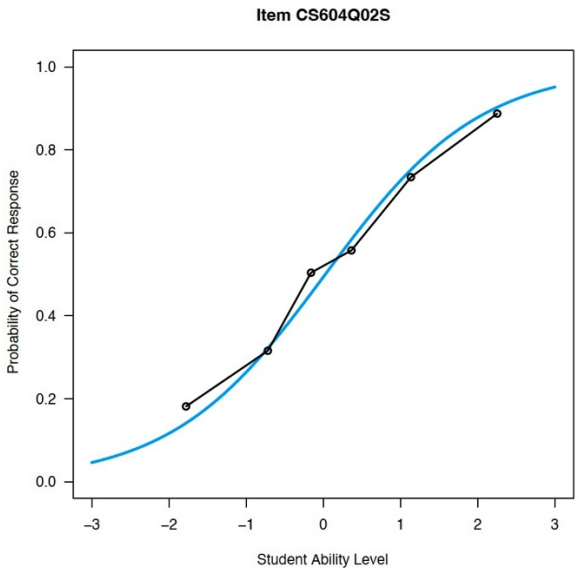


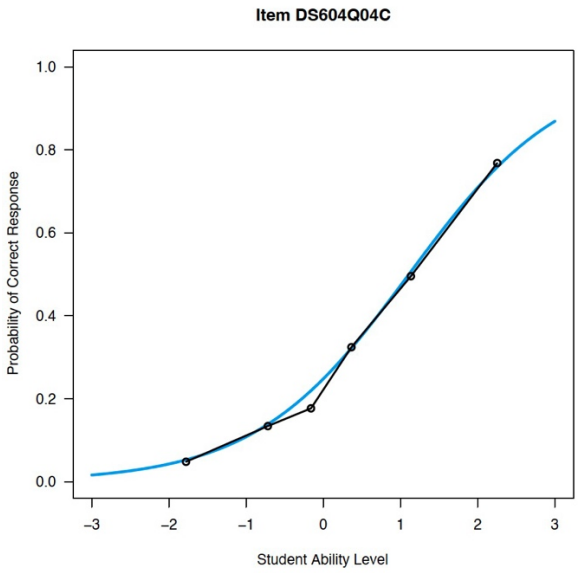

Item CS625Q03S
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

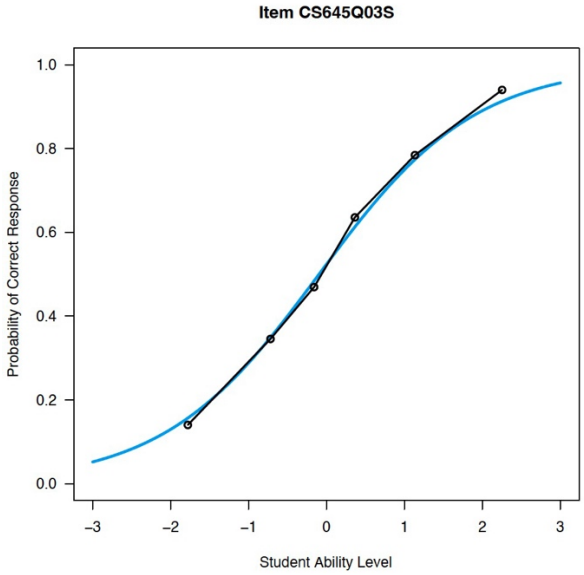

Item CS615Q07S
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

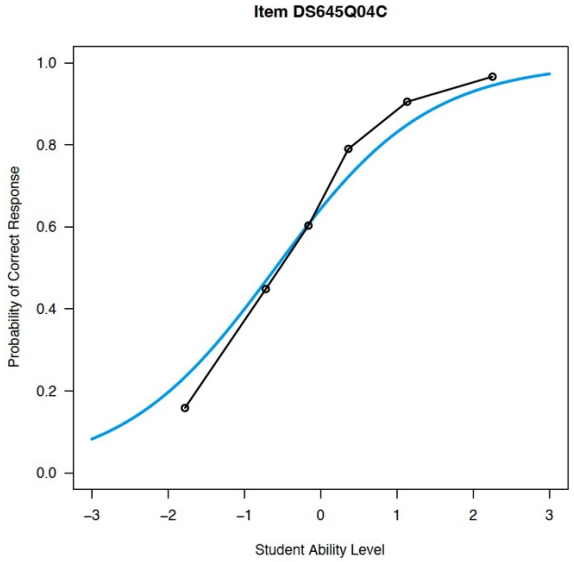

Item CS615Q01S
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

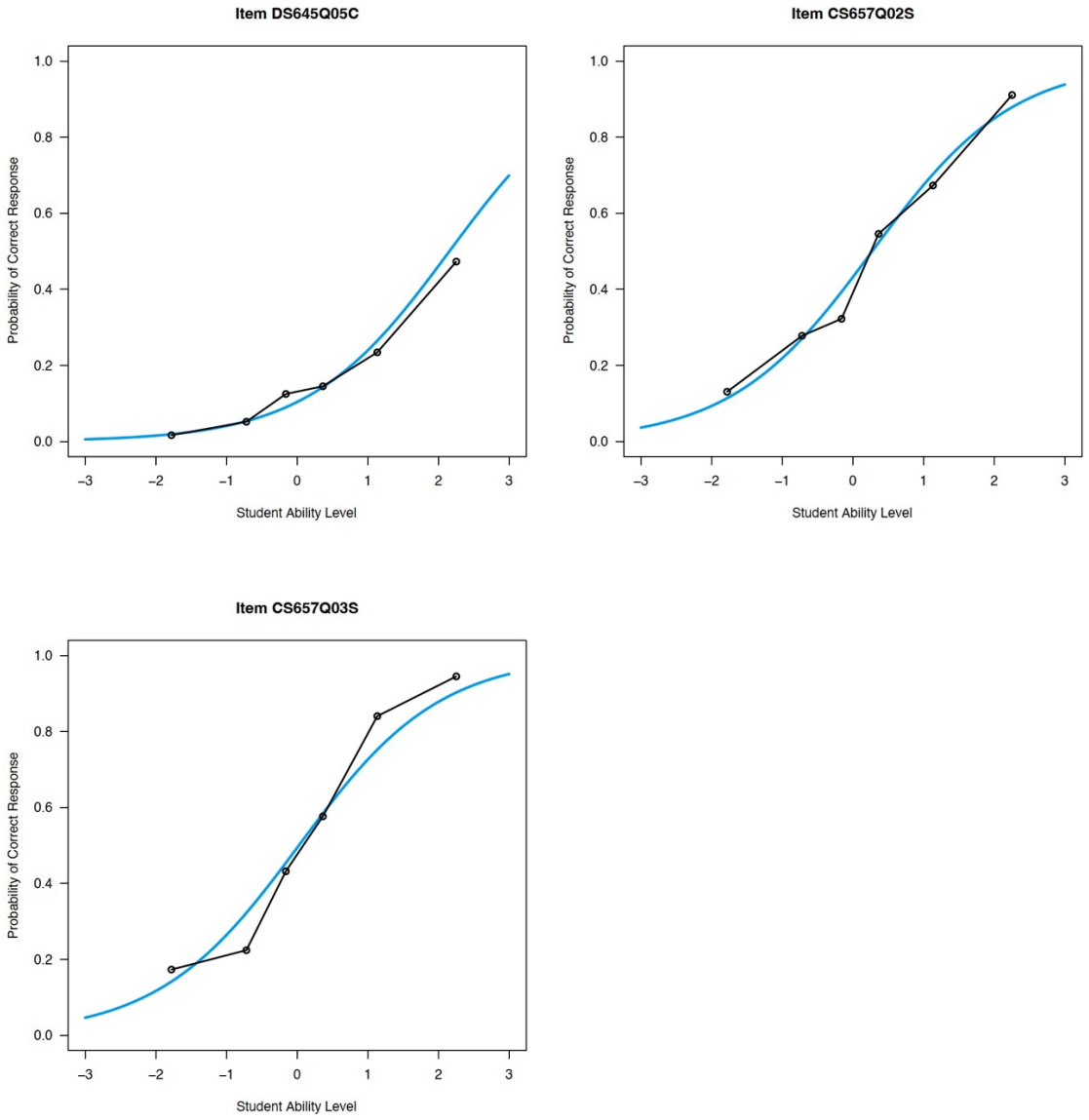

Item CS615Q02S
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


Item CS604Q02S
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


Item DS604Q04C
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


Item CS645Q03S
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


Item DS645Q04C
Probability of Correct Response
Student Ability Level
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

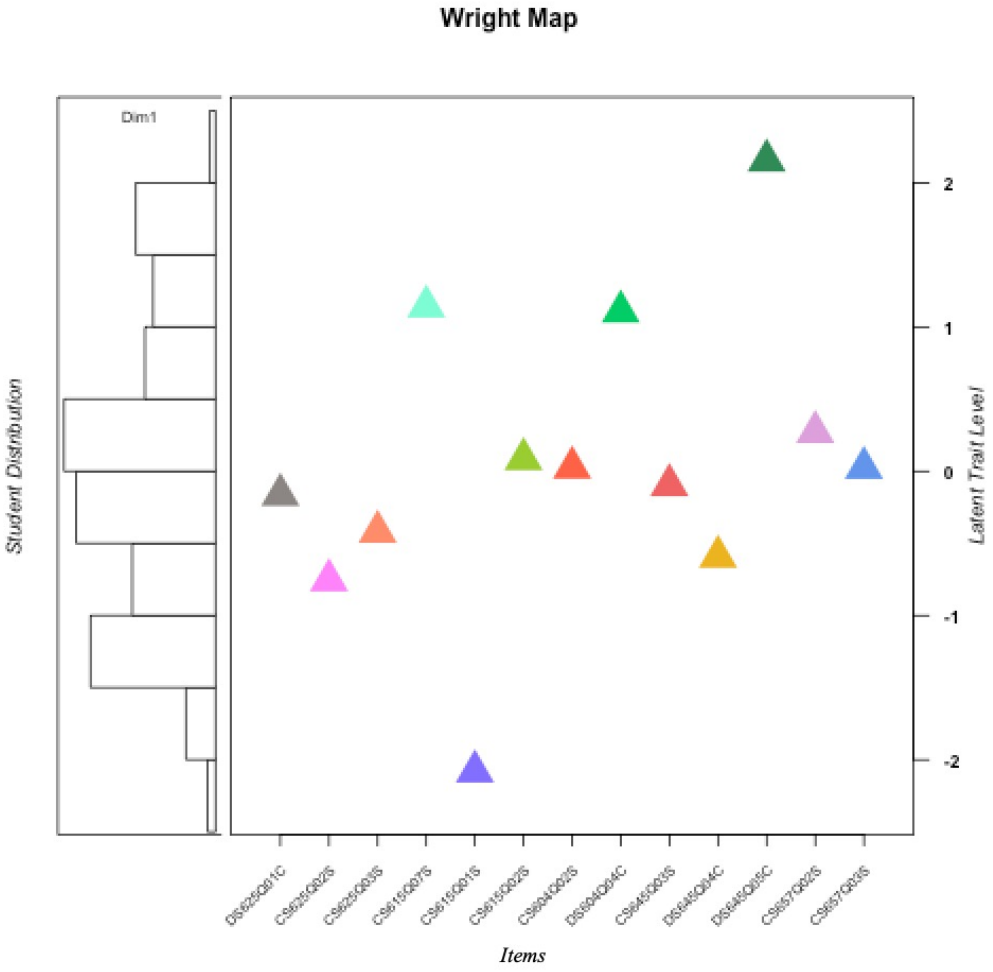


Note. Based off Model 1b and items that were dropped are not shown.

In Figure 28 the left side of the map provides a histogram of student ability (latent trait). The right side of the map provides item difficulty with harder items near the top of the map and easier items near the bottom. The distribution of difficulty shows a reasonably good spread given the location of the person respondents, although more very easy items might be helpful given the large number of respondents performing below the level of the bulk of the items.

Figure 28

Wright Map for Item Cluster S10 with Full Subsample



Note. Based off Model 1b and items that were dropped are not shown.

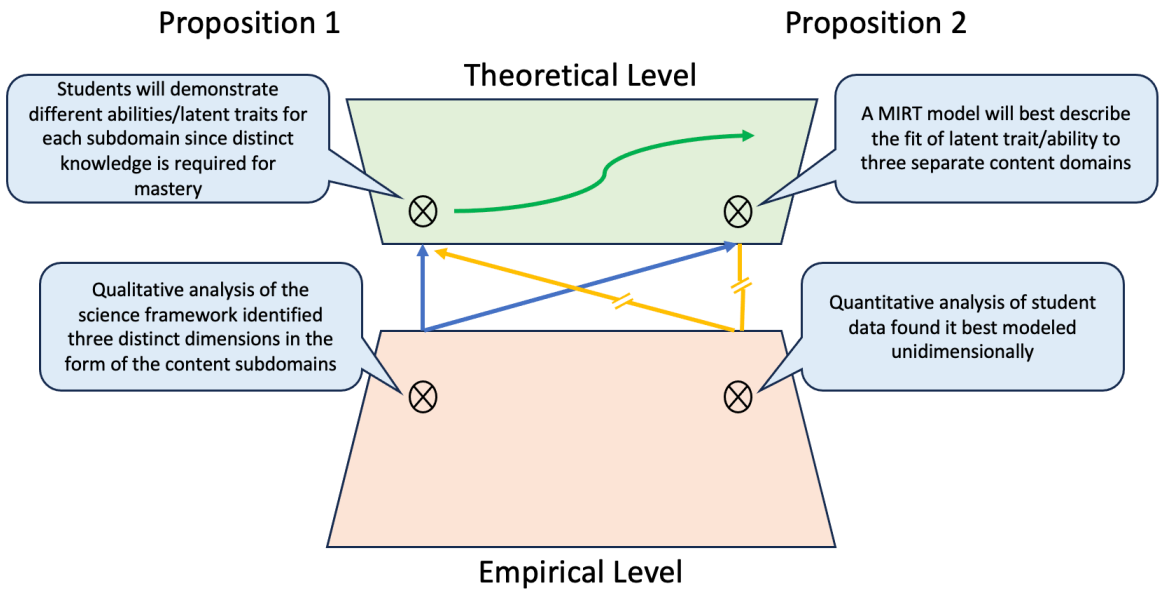
Triangulation

Extending on the earlier proposed qualitative education theory: distinct science subdomains require students to have differentiated knowledge to demonstrate mastery of each subdomain, a multidimensional model should best fit student data to accurately portray student abilities. Since this did not occur, a divergent triangulation is shown in Figure 29 as

qualitative analysis did reveal multiple content dimensions. Broken orange arrows indicate propositions that did not get proven by empirical evidence in the theoretical level. The curvy green arrow indicates that Proposition 1 should have directly led to Proposition 2 but was not proven by findings from the quantitative analysis. Blue arrows indicate empirical findings that did support each proposition. Note that the two broken arrow claims both rely on little separation by the more complex models, relative to the reporting claims.

Figure 29

Triangulation of Results



Note. Adapted from "Combining qualitative and quantitative research within mixed method research designs: A methodological review," by U. Östlund, L. Kidd, Y. Wengström, and N. Rowa-Dewar, 2011, *International Journal of Nursing Studies*, *48*(2011), p. 371. Copyright 2010 by Elsevier.

The IRT analyses did statistically support Proposition 2 for both S10 and S11 subsamples but not with practical significance. The initial exploratory analyses offer some clues, such as many

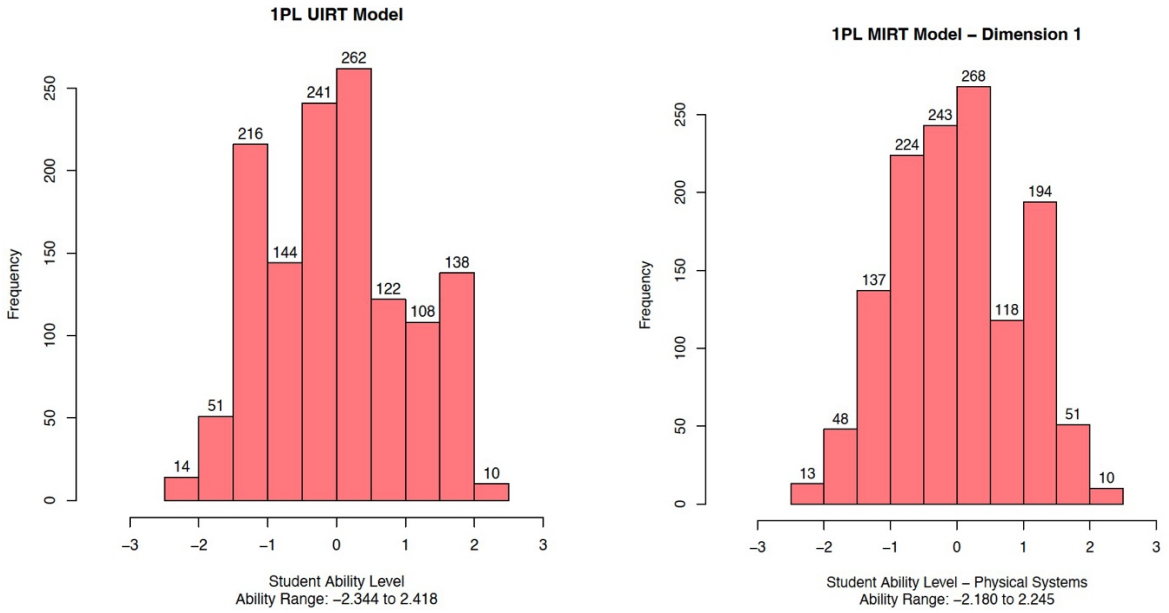
weakly correlated dimensions but not rising to the level of three-dimensional modeling or aligning with the theoretical level being assessed in this dissertation, for any of the subsamples.

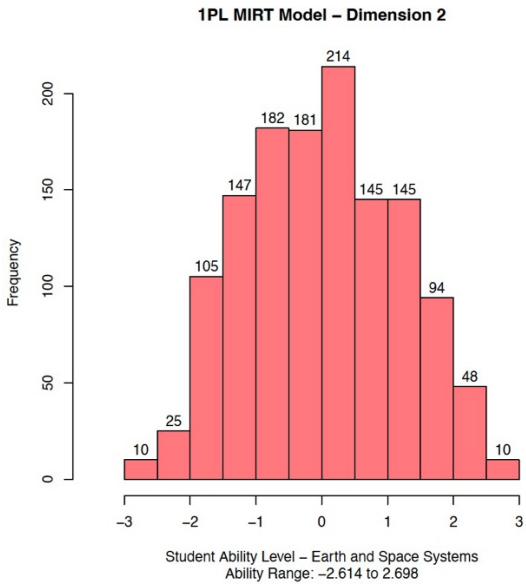
Results Relating to RQ3

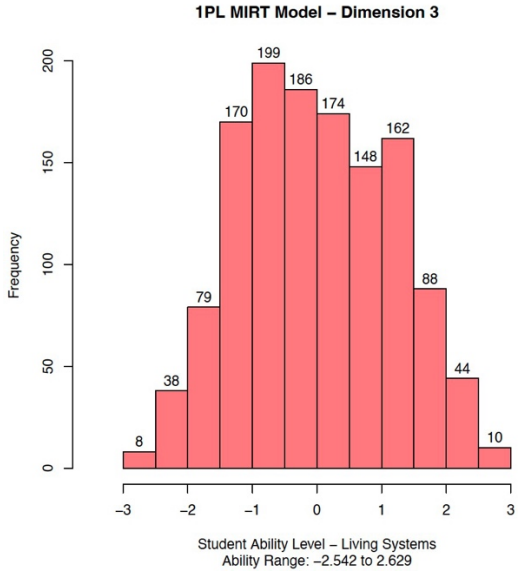
Figure 30 shows histograms for student ability level for both the 1PL UIRT and 1PL MIRT, Models 1b and 3b respectively, for item cluster S10. The number of students is greatest in the middle range of abilities for all models. For Models 1b and 3b dimension 1 the ability levels ranged from -2 to 2 while for Model 3b dimensions 2 and 3 they ranged from -3 to 3. Please note that the software does not allow dimensions to be directly compared as they are not aligned in this software, so do not make this comparison although the charts are located near each other.

Figure 30

Histograms of Student Ability Levels for Item Cluster S10







Note. Items CS657Q01S and CS615Q05S were not included in the models shown above.

CHAPTER 4: DISCUSSION

Analysis of educational data cannot exist in a silo apart from the education content being measured. What the construct is and how students are asked to learn and master it matter. Assessment data from students need to be modeled by researchers in a meaningful manner that incorporates these aspects. Often, assessment developers and educators start with an education standard and work towards developing an assessment/task to measure student performance rather than beginning with an understanding of the content framework and how students learn the subject (Claesgens et al., 2008; NRC, 2001). This study aimed to marry both a qualitative review of the 2015 science framework's content and how that content was learnt by students to a quantitative analysis of which model would be most appropriate for the assessment data. Major themes of the study's results are presented in this chapter, along with providing the study's limitations, threats to the study's validity and reliability, avenues for future research, and policy recommendations for stakeholders.

Study Overview

The literature synthesis revealed that accurately assessing science learning is still a relevant need in the U.S. In addition, the U.S. education system is still failing to improve gaps in science performance associated with a student's economic status and their race (Corcoran et al., 2009). Science education in U.S. schools continues to be centered around separate subject subdomains: life, physical, Earth and space, etc., which also drives how science assessments are formed. These science subdomains have been successfully modeled as multidimensional for other assessments, but some assessments like PISA still model science unidimensionally while reporting on student achievement in a multidimensional space – see Appendix C. UIRT models

are often more familiar and interpretable to large-scale assessment developers (Lang & Tay, 2021). This study was undertaken to determine if a mixed methods approach could reveal triangulated evidence of multidimensionality in the 2015 PISA science and how inaccurate modeling might impact equity issues relevant to students.

Key Takeaways

1. Qualitative analysis yielded a multidimensional view of science content in the 2015 PISA science framework by determining the science subdomains of living, physical, and Earth and space systems were developed as separate content dimensions.
2. Quantitative analysis yielded a 1PL UIRT model as the most practical fit for the 2015 PISA science assessment student data. This outcome was confirmed by PCA and cluster analysis results. Even though the 1PL MIRT model was statistically significant, it offered little improvement of fit for the inferences being made in PISA.
3. The equity investigation was limited by available data yet yielded one result of how using a unidimensional model instead of a MIRT model might negatively impact marginalized student groups by combining students into fewer ability levels, which leads to a loss of information about how students are performing in specific science subdomains.

A Lack of Synergy Between Results

Triangulation of results illustrated that the MIRT model advocated for by a qualitative analysis of the framework was not supported by quantitative analysis, which indicated the 1PL UIRT model was the most practical model in terms of using fewer parameters and viewed in the light of the MIRT model not improving fit substantially. Following is a discussion of what factors

might cause this disconnect between what is described in the framework and how the data end up being modeled. These factors are outside the scope of this study's research questions.

Reckase (1989, p. 9) stressed that "test items may require more than one cognitive skill for successful solution but still generate a statistically unidimensional data set through the interaction with a population that varies on many dimensions." This might be seen here in the weak correlations shown in the descriptive statistics.

One factor impacting multidimensionality of science could be that "most decisions about instruction and curriculum sequences in science have not been guided by a long-term understanding of learning progressions that are grounded in the findings of contemporary cognitive, developmental, education, and science studies research" (NRC, 2001; NRC, 2007, p. 214). Middle and high schools in the U.S. continue to offer science courses as distinct units and science subdomains are infrequently integrated (Enger & Yager, 2009). Up until NGSS was released in 2013 and advocated for crosscutting concepts and science practices that applied across science subdomains (NGSS Lead States, 2013), science learning focused mainly on recall of discrete facts. In 2015, when PISA last administered science as a major domain, the impact of NGSS on curriculum would not have been as far spread throughout the different states. This means students may still be responding to items on the science assessment via recall of facts rather than actually demonstrating mastery of science concepts. Use of memory to recall of facts might be a separate dimension (Leigh et al., 2006) from the science subdomains, but is outside the scope of the investigation here.

Another factor is hinted at by findings from a research study set in Brazil, which noted for that country that PISA scores could be impacted by higher participation rates of students

who had more school years completed (Gomes et al., 2020). If students of a similar age/grade level all have a similar ability level as evidenced by their scores that might also impact a construct showing up as multidimensional. One study showed that even though IRT analysis should be sample independent for the estimations of item characteristics “the stability of these estimations is enhanced when the sample is heterogeneous with regard to the latent trait” (Osteen, 2010, p. 79), which is less likely in a student sample whose members are homogenous in coursework taken and age.

A third factor that could impact or mask multidimensionality is general ability. Pokropek et al. (2022) found that only 17% of variance in science items is attributed to a specific science ability latent trait while the rest can be contributed to general ability (sometimes considered as working memory or another cognitive trait) based on a study of 33 OECD countries taking the 2018 PISA. A strong correlation of 0.88 exists between standard achievement tests that measure intelligence and PISA (Pokropek et al., 2022). This could indicate that PISA is measuring a student’s general intelligence or even their test taking skills rather than the subdomains outlined in the 2015 science framework. While these are outside of the scope of the research questions here, it is known that aspects such as test taking skills can present as an “unwanted dimension of performance” (Scalise & Gifford, 2006). Even the socioeconomic factors, such as number of books owned by the family, which are measured by OECD, may correlate more strongly with general ability (Pokropek et al., 2022).

Overview of Released Item Set

Since qualitative analysis of the framework indicates the science subdomains should be separate dimensions, a deeper look into what a released item set (see Figures 31-33)[68] reveals about the science content versus general ability requirements of the items being assessed in 2015 PISA science is warranted. If items are only asking students to perform recall of science facts, eliminate options based on test taking skills, or use their general intelligence to respond then this could eliminate the content-related multidimensionality being explored here that seems required by the framework. Note that this qualitative review of item content is based on one set of items as the other science items analyzed in this study were not released, leading to the caveat that other items may have more rigorous ties to science subdomains.

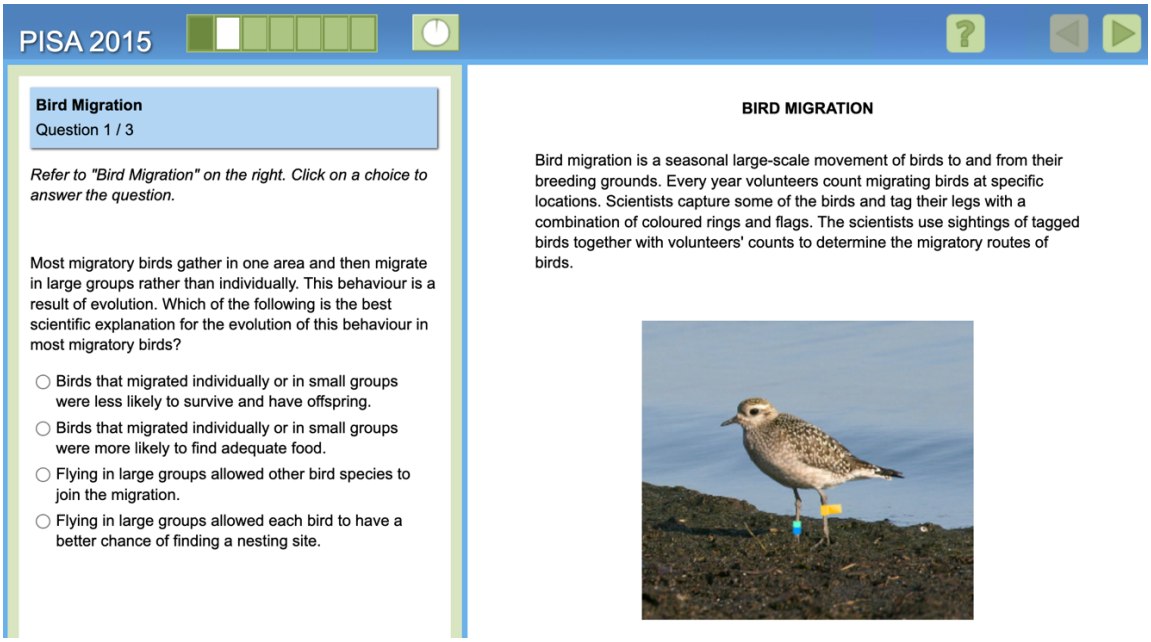
The first item of the Bird Migration set shown in Figure 31 (OECD scoring information provided directly below) seems to be a simple recall of the definition of natural selection, a mechanism of evolution. The item stimulus provides the key term evolution to solicit recall. The item also requires a level of detail about bird migration that is not normally taught in high school since option D (a second correct response) expects students to know if a species of bird can have a better chance at finding nesting sites, which Robins[69] do by having more experienced flock members lead less experienced to good nesting sites. OECD (n.d.-c) has tied this item to content knowledge of living systems, most likely to the standard "Populations (e.g. species, evolution, biodiversity, genetic variation)" – more detail on what aspect/s of evolution can be included in the 2015 PISA science are not provided by OECD. This item is also coded as a

[68] Images from OECD (n.d.-c).
[69] See Robin Migration website https://journeynorth.org/tm/robin/facts_migration.html with contributions by an ornithologist (a scientist who studies birds).

DOK of medium by OECD (n.d.-c), which typically requires more from a student than recall of facts. From an assessment developer perspective, I do not see knowledge of living systems being applied in this item.

Figure 31

Bird Migration Item 1 from Item Cluster S11



Competency	Explain Phenomena Scientifically
Knowledge System	Content - Living
Context	Global - Environmental Quality
Difficulty	501 - Level 3

For *full* credit the student selects "Birds that migrated individually or in small groups were less likely to survive and have offspring."

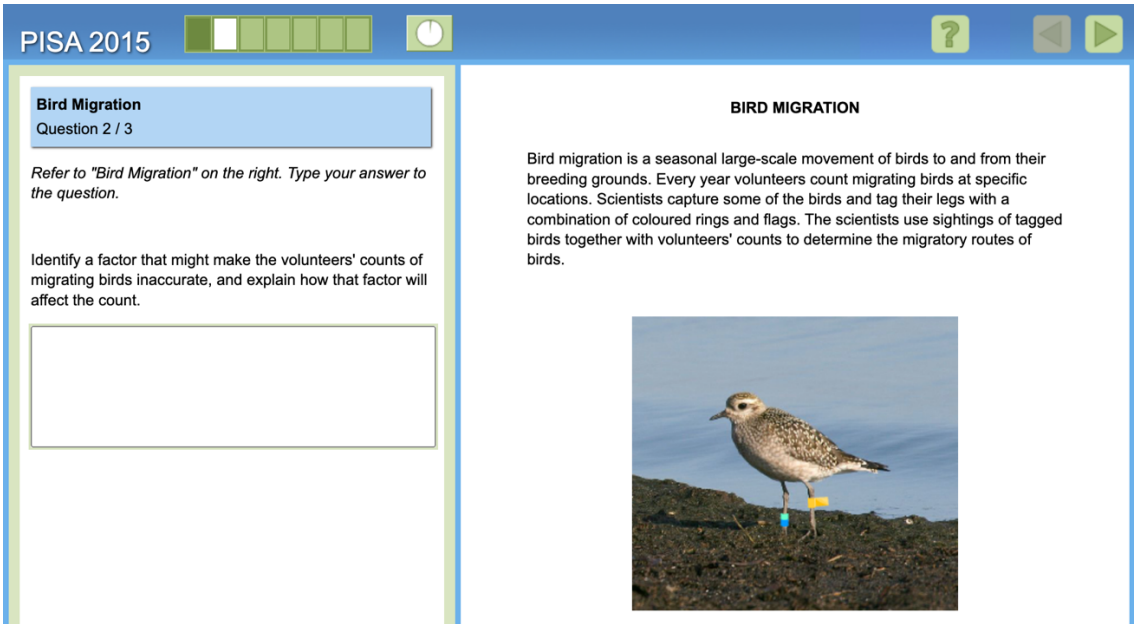
In question 1, students are asked to select an explanation for the specified phenomenon that birds migrate in large groups. This question, which is at the very low end of Level 3, requires that students identify an appropriate conclusion about the evolutionary benefit of this behavior.

OK

The second item of the Bird Migration set shown in Figure 32 (OECD scoring information provided directly below) is of the constructed response item type. The item expects students to draw on procedural knowledge of the correct process for scientific experiments that could apply to any science subdomain. OECD (n.d.-c) has tied this item to procedural knowledge of living systems, most likely to the same standard – a specific aspect of the "Populations" standard was not able to be teased out for this item. This item is also coded as a DOK of high by OECD (n.d.-c), which could be because the student has to make an analytical connection to factors that would invalidate a scientific investigation. From an assessment developer perspective, this item seems tied to the living systems subdomain merely because it mentions living animals, i.e., birds, but does seem to be a higher DOK, although the format effect between selected and constructed response might require a higher literacy DOK, which could be another source of dimensionality although not theoretically intended based on the PISA science framework analyzed.

Figure 32

Bird Migration Item 2 from Item Cluster S11



Competency	Evaluate and design scientific enquiry
Knowledge System	Procedural - Living
Context	Global - Environmental Quality
Difficulty	630 - Level 4

For *full* credit the student identifies at least one specific factor that can affect the accuracy of counts by observers for example:

- The observers may miss counting some birds because they fly high
- If the same birds are counted more than once, that can make the numbers too high
- For birds in a large group, volunteers can only estimate how many birds there are

To correctly answer this question, students must use procedural knowledge to identify a factor that might lead to inaccurate counts of migrating birds and explain how that could affect the data collected. Being able to identify and explain potential limitations in data sets is an important aspect of scientific literacy and locates this question at the top of Level 4.

OK

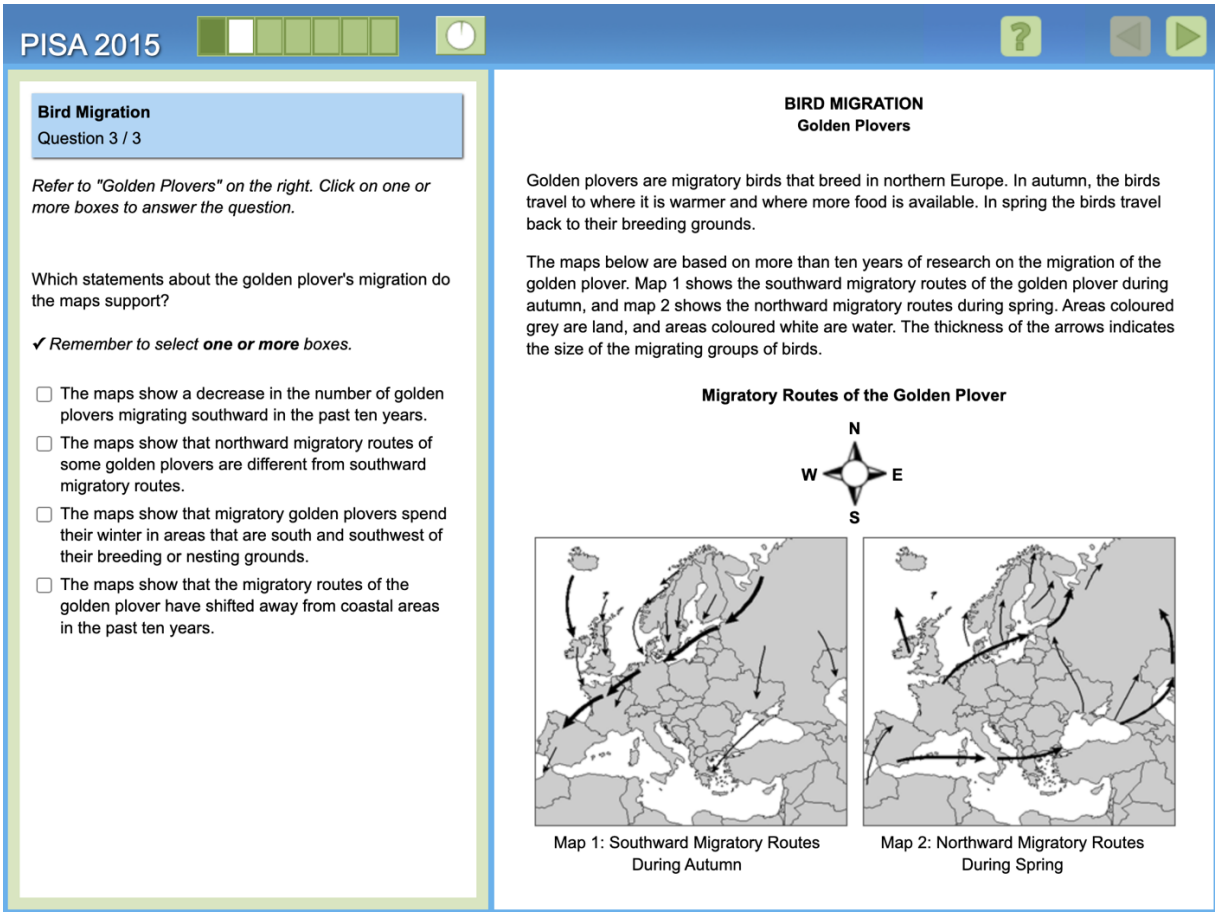
The third item of the Bird Migration set shown in Figure 33 (OECD scoring information provided directly below) is another multiple-choice item like Item 1 but with a multi-select option technology enhancement. Notice that the item's main stimulus has changed to a narrative about a specific bird species and the online version of this computer-based item set

does not seem to allow for students to return to the original stimulus. This could confuse some students and lead to a disconnect with the item. The item again expects students to draw on procedural knowledge, but this time tied to what seems to be actual research data from a living system. OECD (n.d.-c) has tied this item to procedural knowledge of living systems, most likely to the same standard – migrations of populations would fit into the "Populations" standard's example list. This item is also coded as a DOK of medium by OECD (n.d.-c), which this item seems to meet as the data must be interpreted by the student. From an assessment developer perspective, this item out of the set is most closely tied to the living systems subdomain yet focuses on a skill that could be used in any science subdomain – the interpretation of graphical data. Hence once again, subdomain multidimensionality could be masked in this example.

Based on just this item set, the application of general intelligence ability and perhaps a dimension of science procedural knowledge does seem to outweigh application of concepts required by each science subdomain in the 2015 PISA science framework. PISA's science procedural knowledge might be similar to "science practices" in the U.S. NGSS and PISA science epistemic knowledge similar to cross-cutting concepts, bringing these more unifying ideas into play and masking some subdomain multidimensionality. If other items are also disconnected from the subdomains this could help explain the practical unidimensionality of item clusters S10 and S11.

Figure 33

Bird Migration Item 3 from Item Cluster S11



Competency	Interpret data and evidence scientifically
Knowledge System	Procedural - Living
Context	Global - Environmental Quality
Difficulty	574 - Level 4

To get *full* credit the student selects **BOTH**:

- The maps show that northward migratory routes of some golden plovers are different from southward migratory route
- The maps show that migratory golden plovers spend their winter in areas that are south and southwest of their breeding or nesting grounds

Question 3 requires students to understand how data is represented in two maps and use that information to compare and contrast migration routes for the golden plover in the autumn and spring. This Level 4 interpretation task requires students to analyse the data and identify which of several provided conclusions are correct.

OK

Alternate Sources of Multidimensionality

- OECD (2017b) notes there is possible dimensionality between new and trend science items and claims that a UIRT model provided a better fit, yet the fit statistics (AIC and BIC) provided in the technical report show the MIRT model as statistically better with slightly more improvement of fit. This would be for the linked set of items across clusters, not possible to explore here. This is outside the themes being explored and was not codable but is consistent with the results found here for the clusters examined.
- In the 2015 science framework OECD (2017a) provides several types of knowledge, each of which could be a separate dimension. Only content knowledge was examined in this study, but procedural knowledge and/or epistemic knowledge may each be a dimension similar to the three-dimensional aspect of NGSS, as discussed earlier.
- Item format, specifically the item's stem (introductory sentence) has been found to impact multidimensionality of a math assessment (Kan et al., 2018). This could also be a source of multidimensionality in science assessments, as discussed earlier. Also, science content often uses math expressions in items, so the math issue like literacy discussed earlier could be involved in the weak correlations seen in the data exploration.
- Scientific inquiry requires a diverse skill set from students, such as creativity and ability to question. This is speculative based on data and research questions here, and is not explored in the literature review, but such broader issues could contribute if present so that many, but not separable, dimensions might be in play as discussed earlier. These do not seem aligned with the sources of dimensionality explored directly in the analysis.

Impact On Equity

Referring back to the findings and inferences drawn from the available data RQ3 could be better articulated as:

- What inferences can be made about equity in education based on model fit and range of student ability per different subdomains?

This rephrasing of the research question was necessitated by the lack of publicly available demographic data. Therefore, only inferences about how model fit might impact equity could be made.

The NRC (2012, p. 277) advocates that a "crucial role of a framework and its subject matter standards is to help ensure and evaluate educational equity" and "that all students should have adequate opportunities to learn". We should extend that to modeling data accurately with an eye towards the best model fit so all students have their performance modeled equitably. While PISA aims to impact policy at a broader national scale than the student level (Froese-Germain, 2010) the outcome of mismodelling student data can still be that stakeholders redirect resources to away from student groups and subdomains of science education that really need them.

As mentioned earlier, OECD did not make publicly available information on race/ethnicity of students for the 2015 PISA. This lack of data impacts stakeholder understanding of the diversity of the student population and disables researchers attempts to search for equity issues in student performance. We know that the U.S. sample was drawn to be representative, within the limitations for missing data described earlier, so subgroups must be present, and at representative rates, but they are not disaggregated in the PISA data shared.

Therefore, results from this investigation were limited to student ability level as evidenced by theta and how the range of ability levels changed between the UIRT and MIRT models, for the full sample. Future work should explore educational technology data sets or others internal to the U.S. where individual inferences are made and disaggregated samples might be available in science, to see for whom the impact of practical significance might be most meaningful if statistically significant multidimensionality that is separable is present but neglected in the modeling.

One issue can be explored here. When looking at Figure 30 and comparing the UIRT model to the MIRT model, one can see how the students are now compacted into fewer bins for the MIRT model. This may be doing a disservice to those students who do not have access to the higher-level science classes or are challenged by economic constraints of where they live (e.g., lack of internet or science labs). These underrepresented groups tend to be minorities – see Figures 32-34 of Appendix A. This lack of access or differential access to educational content continues to be a concern regarding a lack of advancement in the U.S. on science achievement by students (Vasquez, 2006).

Limitations

There were numerous limitations to this study. No linking information was released to examine all the items together thus limiting the size of sample. OECD did not document why data were missing for U.S. cases in the released sample, or provide additional information about them, so there was no additional information to report. Only one set of items from a subsample was released so a thorough qualitative review of the items was not possible. A lack of publicly available data on the diversity of the sample inhibited the equity investigation

required to answer RQ3. This absence of this information was a policy decision by some participating countries – in particular, the NCES controls access to ethnicity data provided by OECD for the U.S. and only makes it available unlinked and through a restricted use license, which were not the focus of this dissertation. Response rates were low for trend items so only new items were analyzed. Due to the assessment cycle in 2015 being the first science fully digital assessment, this could be due to trend items being mainly paper based while new items were developed natively as computer based and more schools are expecting electronic delivery of assessments.

Threats to Validity and Reliability

The model's generalizability may be limited due to sample size, which may be restricted by elimination of students with missing data. Therefore, external validity may be impacted and the results may not be generalizable to various student populations with differing demographics in the U.S., especially minority groups that tend to have smaller population sizes. This brings to question if the U.S. sample design was adequate enough. Next, as with most assessments there is always the question if what is being measured, such as the subdomain of life science, remains the same across clusters of variable content. Without any linking information available and no common form among all students, instrumentation also threatens the internal validity of this study. "A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes (Lang & Tay, 2021, p. 328)." Relatedly, students had access to prior versions of science items in the form of released items, which introduces the threat of *testing*, where

exposure to a pre-test similar to the final test might influence their outcomes as reported on in the PISA results.

Finally, opportunity to learn threats to internal validity could exist since students who were assessed varied in their learning stages by having taken different science courses. An important note is that because there was not a control group and students were not divided into groups (only by country) for the analysis, threats to validity associated with potential differences in group membership /selection bias are not relevant to PISA's sampling. There are different global windows of testing for PISA which could indicate a maturation threat, but the window of testing for the U.S. was fairly stable. Changes to the environment in which the assessment was taken and changes in student behavior were not shared by OECD.

With regards to reliability, a key threat is researcher error with regards to the number and construct type of the dimensions identified in the qualitative analysis. This error was mitigated by having a reviewer familiar with the construct analyze results. There will not be any agreement indices collected for the qualitative analysis.

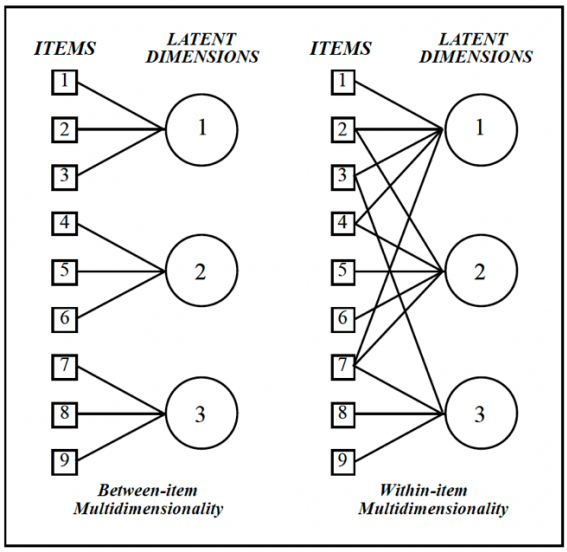
Future Research

Since the qualitative and quantitative analyses let to divergent results the question remains on how to accurately model the science content domain when its subdomains appear to be separate dimensions. Determining how dimensionality of science content affects assessments and their items is crucial to developing assessments that yield information equitably for diverse student subgroups. The following are recommendations for future research in this area:

- Since polytomous items were dropped from the study and only dichotomous items were analyzed, a future analysis needs to include a more complex model that covers both types of items.
- A structural equation model might be implemented if there more connections are found between items based on their subdomain content.
- The MIRT model in this study was developed for a between-item representation of multidimensionality – see Figure 6 (from Li et al., 2012). Baghaei (2012) provides an overview of a MIRT model for within-item dimensionality – see Figure 34 below. A within-item MIRT model should be tested to see if it provides better fit since some of the standards in the 2015 PISA science framework were found to load on each other during the qualitative analysis – see Figure 11. This relates to the expectation that students should be able to apply science content to either interdisciplinary or independent science items/tasks (Mostafa et al., 2018; OECD, 2017a).

Figure 34

Differences in Between-item and Within-item MIRT Models



Note. From "The application of multidimensional Rasch models in large-scale assessment and validation: An empirical example," by Purya Baghaei, 2012, *Electronic Journal of Research in Educational Psychology*, *10*(1), p. 239. Copyright 2012 by Electronic Journal of Research in Educational Psychology.

Policy Recommendations

The recommendations below are targeted towards several groups of science education stakeholders: researchers, learning scientists, policy makers, assessment and curriculum developers, and educators themselves. With the goal of clearer understanding of how science assessments should be modeled to increase equity for all students here are some crucial areas still needing attention:

- Prior researchers have advocated using an UIRT model simply because PISA has used this model in the past - see the Davier et al. (2019) study on model fit using PISA data. While following a similar methodology might increase reliability of results it does not

address equity concerns and new models should be analyzed to find the fit that provides accurate information on science performance for every student.

- At what point is an increase in fit too small if it can be justified by the help it provides the group existing in that range? Quantitative education researchers should review if model fit indices can be tailored to specific subgroups of the population.
- Developers of science standards for the U.S. should compare NGSS dimensionality and that found in PISA framework to identify areas where there is a content match or disconnect. These comparisons could guide how items are coded to a dimension in MIRT models
- For future PISA cycles, that more information such as linking be released, or that a similar study be conducted internally and a report released.
- OECD should consider making ethnicity data available publicly to increase transparency of results. This data release can be done on a nation-by-nation basis if requested.
- OECD assessment designers should release item specifications if available or build them to guide interpretation of future science framework standards.
- Learning scientists should identify any unique constructs for each science subdomain with the goal of describing how students develop mastery in each area.

Conclusions

In order to ensure we measure what is intended, both educators and researchers will benefit from digging deeper into what constitutes each science subdomain. If the science subdomains are truly individual constructs, then our scoring models should follow the framework's scaffolding, or we should identify what dimensionality aspect is more

predominantly being assessed. While large-scale assessments provide countries with a wealth of data, they may be doing harm to educational equity based on use inferences that routinely affect national educational objectives and school policies. In other words, we should not make claims about what is measured when we did not accurately differentiate the constructs or chose a model based on its usability only. As the U.S. moves forward with more three-dimensional science learning via NGSS how to model those constructs in a multidimensional space will become more critical.

To begin evaluating these aspects and fulfill the proposed policy recommendations above, a good start, to coin a phrase from Castillo & Gillborn (2023), is to "democratize evidence." Without complete sets of data, researchers are limited in verifying or building upon previous research and new discoveries cannot be made. Throughout this study it was clear that more complete student demographic data was needed to validate the equity inferences being drawn from the conclusion that there was model misfit based on the qualitative and quantitative findings.

REFERENCES

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, *8*(2), 155-194. https://www.nber.org/papers/w9411

American Educational Research Association. (2014). *Standards for Educational and Psychological Testing*.

Armstrong, C. (2021). Key methods used in qualitative document analysis. *SSRN eLibrary*. https://ssrn.com/abstract=3996213

Atkisson, M. (2010, 2010-10-15). *Social Negotiation as a Central Principle of Constructivism*. Ways of Knowing. https://woknowing.wordpress.com/2010/10/14/social-negotiation-as-a-central-principle-of-constructivism/

Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd edition). The Guilford Press.

Baghaei, P. (2012). The application of multidimensional rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology*, *10*(1), 233-252.

Baškarada, S. & Koronios, A., (2018). A philosophical discussion of qualitative, quantitative, and mixed methods research in social science. *Qualitative Research Journal*, https://doi.org/10.1108/QRJ-D-17-00042

Boon, M., Orozco, M., & Sivakumar, K. (2022). Epistemological and educational issues in teaching practice-oriented scientific research: Roles for philosophers of science. *European Journal for Philosophy of Science*, *12*(16), 1-23.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative*

Research Journal, *9*(2), 27-40.

Brandt, S. (2015). *Unidimensional interpretation of multidimensional tests*. [Doctoral dissertation, Christian-Albrechts-Universität zu Kiel].

Briggs, D. C. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87-100.

Broesch, T., Crittenden, A. N., Beheim, B. A., Blackwell, A. D., Bunce, J. A., Colleran, H., Hagel, K., Kline, M., McElreath, R., Nelson, R. G., Pisor, A. C., Prall, S., Pretelli, I., Purzycki, B., Quinn, E. A., Ross, C., Scelza, B., Starkweather, K., & Stieglitz, J. (2020). Navigating cross-cultural research: methodological and ethical considerations. *Proceedings B The Royal Society*, *287*(20201245), 1-7.

Brooks-Bartlett, J. (2018). Probability concepts explained:Maximum likelihood estimation. *Towards Data Science*.

Carpiano, R. M., & Daley, D. M. (2006). A guide and glossary on postpositivist theory building for population health. *Journal of Epidemiology and Community Health*, *60*, 564-570. doi: 10.1136/jech.2004.031534

Carr, S. M. (2001). *Interpreting a principal components analysis - Theory & practice*. Memorial University: Biology – Faculty of Science. https://www.mun.ca/biology/scarr/2900_PCA_Analysis.htm

Castillo, W. & Gillborn, D. (2023, September). *How to "QuantCrit:" Practices and Questions for Education Data Researchers and Users*. (EdWorkingPaper No. 22-546). https://doi.org/10.26300/v5kh-dd65

Center for Professional Education of Teachers. (n.d.). *Equity and Assessment*.

https://cpet.tc.columbia.edu/news-press/equity-and-assessment

Civil Rights Data Collection. (2023, November 20). *Data on Equal Access to Education*. Office for Civil Rights, U.S. Department of Education. https://ocrdata.ed.gov/

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2008). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education*, *93*, 56-85.

College Board. (2009). *Science: College board standards for college success*. The College Board.

Connected Papers | Find and explore academic papers. (2021). https://www.connectedpapers.com/

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (Report No. RR-63). Consortium for Policy Research in Education (CPRE). www.cpre.org

Csapó, B., & Funke, J. (Eds.). (2017). *The Nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing. https://read.oecd-ilibrary.org/education/the-nature-of-problem-solving_9789264273955-en#page5

Cummings, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16-29.

Davier, M. V., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 466-488.

DeMars, C. E. (2016). Partially compensatory multidimensional item response theory models: Two alternate model forms. *Educational and Psychological Measurement*, *76*(2), 231-257.

Duran, V. (2014, March 17). *Multidimensional item response theory: What have we learned thus far* [PowerPoint slides]. The Psychometrics Centre, University of Cambridge. https://www.psychometrics.cam.ac.uk/system/files/documents/multidimensional-item-response-theory.pdf

Enger, S. K., & Yager, R. E. (2009). Chapter 1: A framework for assessing student understanding in science: A standards-based K-12 handbook. In *Assessing student understanding in science* (2nd edition, pp.1-11). Sage.

Erzberger, C., & Kelle, U. (2003). Making inferences in mixed methods: The rules of integration. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of Mixed Methods in Social & Behavioral Research 1st edition* (pp. 457-488). Sage.

Froese-Germain, B. (2010). *The OECD, PISA and the impacts on educational policy* (Report). Canadian Teachers' Federation.

Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology*, *13*(117), 1-8.

Gao, N., Johnson, H., Lafortune, J., & Dalton, A. (2019). New eligibility rules for the University of California? The effects of new science requirements. Public Policy Institute of California. https://www.ppic.org/wp-content/uploads/new-eligibility-rules-for-university-of-california-the-effects-of-new-science-requirements.pdf

GEOstata. (2016). *PISA 2015 Results – Performance in Science*. OECD.

Godwin, A. (2017). Unpacking latent diversity. In *American Society for Engineering Education Annual Conference & Exposition*.

Gomes, M., Hirata, G., & Oliveira, J. B. A. E. (2020). Student composition in the PISA assessments: Evidence from Brazil. *International Journal of EducationalDevelopment*, *79*, 1-7.

Greene, J. C. & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. *New Directions for Evaluation*, *1997*(74), 5-17.

Greenwood, B. (2020). *Understanding Pedagogy - What is Social Constructivism?* Satchel. https://blog.teamsatchel.com/understanding-pedagogy-what-is-social-constructivism

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software*, *91*(1), 1-30. doi:10.18637/jss

Hanushek, E.A., Jamison, D.T., Jamison, E.A., & Woessmann, L. (2008). Education and economic growth: It's not just going to school, but learning something while there that matters. *Education Next*, *8*(2), 62-70.

Harris, D. (n.d.). Comparison of 1-, 2-, and 3-parameter IRT models. *Instructional Topics in Educational Measurement*, 157-163.

Ho, L., & Limpaecher, A. (2021, September 17). *The practical guide to grounded theory. practical guide to grounded theory research*. Delve. https://delvetool.com/groundedtheory

Iliescu, D. & Greiff, S. (2021). On consequential validity. *European Journal of Psychological Assessment*, *37*(3), 163–166.

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, *4*(45), 2-15.

Irribarra, D. T. & Arneson, A. E. (2023). The challenge of defining and interpreting dimensionality ineducational and psychological assessments. *Measurement*, *221*, 1-8.

Irribarra, D.T. & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration*. https://github.com/david-ti/wrightmap

Issayeva, L. (2022, December 18). *Multidimensional item response theory*. Assessment Systems Corporation (ASC). https://assess.com/multidimensional-item-response-theory/

Jerrim, J. (2016, November 1). *The design and use of test scores: Lecture 4* [PowerPoint slides]. Social Research Institute, University College London.

Jerrim, J., Micklewright, J., Heine, J., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476-493. https://doi.org/10.1080/03054985.2018.1430025

Johnson, R. B. & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*. *33*(7), 14-26.

Johnson, S. (2019, October 18). How one high school's dispute reflects the struggle to teach California's science standards. *EdSource*. https://edsource.org/2019/how-one-high-schools-dispute-reflects-the-struggle-to-teach-californias-science-standards/618752

Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: A review and recent Developments. *The Royal Publishing Society: Philosophical Transactions A*. *374*, 1-16.

Kaldaras, L., Akaeze, H., & Krajcik, J. (2021). A Methodology for Determining and Validating

Latent Factor Dimensionality of Complex Multi-Factor Science Constructs Measuring Knowledge-In-Use. *Educational Assessment, 26*(4), 241-263.

Kan, A., Bulut, O., & Cormier, D. C. (2018). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*, *24*(1), 13-32.

Kassambara, A. & Mundt, F. (2020) *Factoextra: Extract and visualize the results of multivariate data analyses*. (Version 1.0.7) [R program package]. https://CRAN.R-project.org/package=factoextra

Kelley, T. R. & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(11), 1-11.

Kiefer, T., Robitzsch, A., & Wu, M. (2015, July 2). *TAM: An R package for item response modelling* [PowerPoint slides]. https://user2015.math.aau.dk/presentations/205.pdf

Kose, I. A. & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia - Social and Behavioral Sciences*, *46*, 135 – 140.

Krutsch, E. & Roderick, V. (2022, November 4). STEM Day: Explore Growing Careers. *U.S. Department of Labor Blog*. https://blog.dol.gov/2022/11/04/stem-day-explore-growing-careers

Lang, J. W. B., & Tay, L. (2021). The Science and Practice of Item Response Theory in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 311-338.

Language Resource Center (LRC). (2022). Languages by countries. https://www.languagerc.net/languages-by-countries/

Learn more about PILA (n.d.). The Platform for Innovative Learning Assessments (PILA). https://pilaproject.org/.

Leigh, J. H., Zinkhan, G. M., & Swaminathan, V. (2006). Dimensional relationships of recall and recognition measures with selected cognitive and affective aspects of print ads. *Journal of Advertising*, *35*(1), 105-122.

Li, Y., Jiao, H., & Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large-scale science assessment. *Journal of Applied Testing Technology*, *13*(2), 1-27.

Lips, D. & Moritz, M. (2023). STEM and Computer Science Education: *Reforming Federal K-12 Education R&D Activities to Strengthen American Competitiveness: Prepared by Federation of American Scientists*. Lincoln Network, Foundation for American Innovation.

Mailman School of Public Health. (2023, November). *Item Response Theory*. Columbia University. https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory

Market Data Retrieval (MDR). (2024, March 26). *How many schools are in the U.S.?* MDR Education. https://mdreducation.com/how-many-schools-are-in-the-u-s/

Maxwell, J. A. & Mittapalli, K. (2010). Realism as a stance for mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of Mixed Methods in Social & Behavioral Research 2nd edition* (pp. 145-167). Sage.

Mazzei, L. A., & Jackson, A. Y. (Eds.). (2024). *Postfoundational approaches to qualitative inquiry*. Routledge. DOI: 10.4324/9781003298519

McLeod, S. (2019). *Constructivism as a Theory for Teaching and Learning | Simply Psychology*. https://www.simplypsychology.org/constructivism.html

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment* (Report No. RR-93-51). Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/j.2333-8504.1993.tb01562.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Mostafa, T., Echazarra, A., & Guillou, H. (2018). The science of teaching science: An exploration of science teaching practices in PISA 2015 (OECD Education Working Papers No. 188). www.oecd.org/edu/workingpapers

National Center for Education Statistics. (n.d.-a). *Frequently asked questions*. PISA resources. https://nces.ed.gov/surveys/pisa/faq.asp

National Center for Education Statistics. (n.d.-b). *Science literacy: Average scores*. Program for International Student Assessment (PISA). https://nces.ed.gov/surveys/pisa/pisa2015/pisa2015highlights_3.asp

National Center for Education Statistics. (2022). *High school mathematics and science course completion: Condition of education*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/sod

National Research Council. (2001). *Knowing what students know: The science and*

design of educational assessment. Washington, DC: The National Academies Press. https://doi.org/10.17226/10019

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11625

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states: Three dimensional learning*. https://www.nextgenscience.org/three-dimensional-learning

Organisation for Economic Co-operation and Development. (n.d.-a). *About PISA*. The Programme for International Student Assessment (PISA). https://www.oecd.org/pisa/aboutpisa/

Organisation for Economic Co-operation and Development. (n.d.-b). *FAQ*. The Programme for International Student Assessment (PISA). https://www.oecd.org/pisa/pisafaq/

Organisation for Economic Co-operation and Development. (n.d.-c). *PISA 2015 released field trial cognitive items*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2016a). *Country note: Key findings from PISA 2015 for the United States*. OECD Publishing. https://www.oecd.org/pisa/PISA-2015-United-States.pdf

Organisation for Economic Co-operation and Development. (2016b). *PISA 2015 results (volume I): Excellence and equity in education*. OECD Publishing. http://dx.doi.org/10.1787/9789264266490-en

Organisation for Economic Co-operation and Development. (2017a). *PISA 2015 Assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving, revised edition*. OECD Publishing. http://dx.doi.org/10.1787/9789264281820-en

Organisation for Economic Co-operation and Development. (2017b). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/

Organisation for Economic Co-operation and Development. (2018). *PISA 2015: Results in Focus*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2023). *Working draft: PISA learning in the digital world assessment framework*.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, *1*(2), 66-82.

Östlund, U., Kidd, L., Wengström, Y., and Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: A methodological review. *International Journal of Nursing Studies*. *48*(2011), 369-383.

Otarigho, M. D. & Oruese, D. D. (2013). Problems and prospects of teaching integrated science in secondary schools in Warri, Delta State, Nigeria. *Techno LEARN: An International Journal of Educational Technology*, *3*(1), 19-26.

Park, S., Reeger, A., & Aloe, A. M. (2020). Technically speaking: Determining test effectiveness with item response theory. *Iowa Reading Research Center, University of Iowa*. https://irrc.education.uiowa.edu/blog/2020/09/technically-speaking-determining-test-effectiveness-item-response-theory

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work developing transferable knowledge and skills in the 21st Century*. National Academies Press. DOI 10.17226/13398

Pensavalle, C. A. & Solinas, G. (2013). The Rasch model analysis for understanding mathematics proficiency—A case study: Senior high school Sardinian students. *Creative Education*, *4*(12), 767-773.

Pierson, A. E., Clark, D. B., & Kelly, G. J. (2019). Learning Progressions and Science Practices Tensions in Prioritizing Content, Epistemic Practices, and Social Dimensions of Learning. *Science & Education*, *28*, 833-841.

PISA USA. (2015). *Program for international student assessment: Frequently asked questions – Information for Students* [Brochure]. https://www.fldoe.org/core/fileparse.php/5389/urlt/PISA2015_FAQ_Student_Information.pdf

Pokropek, A., Marks, G. N., Borgonovi, F., Koc, P., & Greiff, S. (2022). General or specific abilities? Evidence from 33 countries participating in the PISA assessments. *Intelligence*, *92*.

Polites, G. L., Roberts, N., and Thatcher, J. (2012). Conceptualizing models using multidimensional constructs: A review and guidelines for their use. *European Journal of Information Systems*, *21*, 22-48.

Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC. https://plot.ly

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401-412.

Reckase, M. D. (1989). *The interpretation and application of multidimensional item response theory models; and computerized testing in the instructional environment: Final Report* (Report No. AD-A214109). The American College Testing (ACT) Program.

Reckase, M. D. (1990). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests* [Paper presentation]. Annual Meeting of American Educational Research Association, Boston.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research*. R package version 2.4.1. https://CRAN.R-project.org/package=psych

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules*. R package version 4.1-4, https://CRAN.R-project.org/package=TAM

RStudio Team. (2021). *RStudio: Integrated development environment for R*. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, *52*, 1893-1907.

Scalise, K. & Clarke-Midura, J. (2018). The many faces of scientific inquiry: Effectively measuring what students do and not only what they say? *Journal of Research in Science Teaching*, *55*, 1469-1496.

Scalise, K. & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment*, *4*(6), 1-45.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com

Singer, J. D. & Braun, H. I. (2018). Testing international education assessments: Rankings get headlines, but often mislead. *Science*, *360*(6384), 38-40.

Socha, A. (n.d.) *Multidimensional item response theory*. [Unpublished article – James Madison University]. https://educ.jmu.edu/~sochaab/index_files/Showcase/MIRT.pdf

Spencer, S. G. (2004). *The strength of multidimensional item response theory in exploring construct space that is multidimensional and correlated*. [Doctoral dissertation, Brigham Young University]. BYU ScholarsArchive.

Stehle, S. M., & Peters-Burton, E. E. (2019). Developing student 21st century skills in selected exemplary inclusive STEM high schools. *International Journal of STEM Education*, *6*(1), 1-15. https://doi.org/10.1186/s40594-019-0192-1

Strauss, V. (2019, December 3). Expert: How PISA created an illusion of education quality and marketed it to the world. *The Washington Post*.

The World Bank Group. (2023). *Compulsory education, duration (years)*. Data. https://data.worldbank.org/indicator/SE.COM.DURS

The World Bank Group. (2023). *World development indicators*. https://datatopics.worldbank.org/world-development-indicators/

Uesaka, Y., Suzuki, M., & Ichikawa, S. (2022). Analyzing students’ learning strategies using item

response theory: Toward assessment and instruction for self-regulated learning. *Frontiers in Education*, *7*(921844), 1-16.

USAGov. (2023, December 27). *Official language of the United States*. About the U.S. and its government. https://www.usa.gov/official-language-of-us

Vasquez, J. (2006). High school biology today: What the committee of ten did not anticipate. *CBE—Life Sciences Education: High School Biology Today*, *5*, 29-33.

Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS Quarterly*, *37*(1), 21-54.

Venkatesh, V., Brown, S. A., & Sullivan, Y. W. (2016). Guidelines for conducting mixed methods research: An extension and illustration. *Journal of AIS*, *17*(7), 435-495.

Voogt, J. & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299-321.

Wach, E., Ward, R., & Jacimovic, R. (2013). Learning about Qualitative Document Analysis. *IDS Practice Papers in Brief*, 1-10.

Wang, C. (2021). A brief history and next stage of multidimensional item response theory. *Quantitative and Qualitative Methods*.

Wang, C. & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*(2), 119-134.

Welch, W. W. (1977). Chapter 3: Evaluation and decision-making in integrated science. In D.

Cohen (Ed.), *Volume IV: New trends in integrated science teaching: evaluation of integrated science education* (pp. 26-36). United Nations Educational, Scientific, and Cultural Organization.

Western Governors University. (2020). *What is constructivism*? https://www.wgu.edu/blog/what-constructivism2005.html#close

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Version 3.4.4) [R program package]. Springer-Verlag, New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation* (Version 1.1.4) [R program package]. https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, *46*(9), 3766-3774.

Winarno, N., Rusdiana, D., Riandi, R., Susilowati, E., & Afifah, R. M. A. (2020). Implementation of Integrated Science Curriculum: A Critical Review of the Literature. *Journal for the Education of Gifted Young Scientists*, *8*(2), 795-817. DOI: http://dx.doi.org/10.17478/jegys.675722 https://www.wgu.edu/blog/what-constructivism2005.html#close

Wind, S. & Hua, C. (2021). *Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners*. Bookdown. https://bookdown.org/chua/new_rasch_demo2/

World Population Review. (2023). *Most racially diverse countries 2023*. https://datatopics.worldbank.org/world-development-indicators/

Yamamoto, K. (1995). *TOEFL technical report: Estimating the effects of test length and test time on parameter estimation using the hybrid model* (Report No. ETS-RR-95-2; TOEFL-TR-10). Educational Testing Service. https://files.eric.ed.gov/fulltext/ED395035.pdf

Yen, S. J. & Leah, W. (2007). *Multidimensional IRT models for Composite Scores* [Paper presentation]. 2007 Annual Meeting of the National Council of Measurement in Education, Chicago, IL, United States.

APPENDIX A: STUDENT ENROLLMENT IN SCIENCE COURSES BY ETHNICITY

The following bubble graphs showcase student enrollment for U.S. high school science courses based on data from Civil Rights Data Collection (CRDC, 2023) collected by the Office for Civil Rights (OCR). The data are from the 2020-21 school year, which was when OCR could restart data collection after a delay due to COVID-19. Student data is from all school districts and public schools, "as well as long-term secure juvenile justice facilities, charter schools, alternative schools, and special education schools that focus primarily on serving the educational needs of students with disabilities under IDEA or section 504 of the Rehabilitation Act (CRDC, 2023)." Administration of the CRDC occurs every two years in the 50 states, Washington, D.C., and the Commonwealth of Puerto Rico (CRDC, 2023). Enrollment was less than 1% for the American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander student populations, so those student populations are not depicted for all three bubble charts. The majority of students enrolled in high school physics (Figure 35), biology (Figure 36), and chemistry (Figure 37) courses are of White and Hispanic or Latino (of any race) ethnicities.

Figure 35

U.S. High School Physics Enrollment

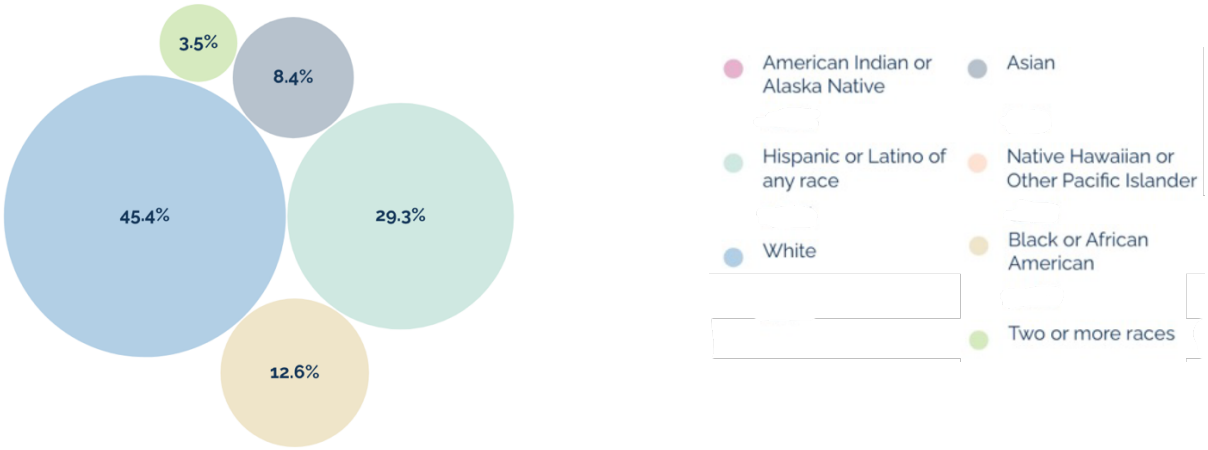


Figure 36

U.S. High School Biology Enrollment

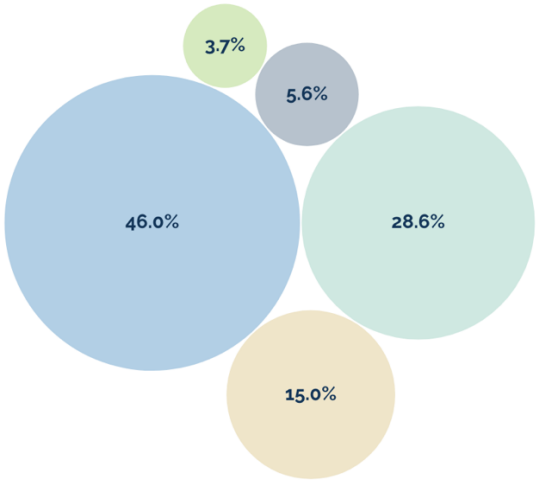
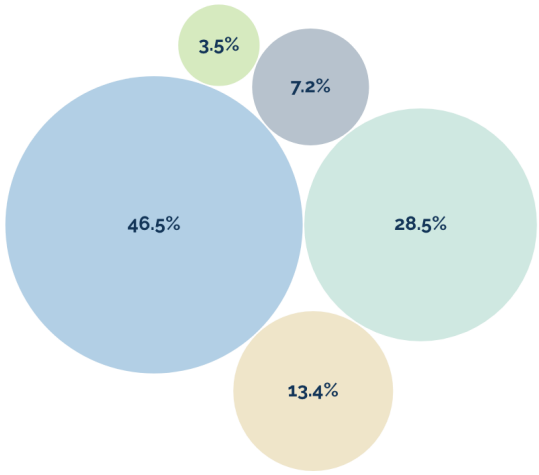


Figure 37

U.S. High School Chemistry Enrollment



Note. For Figures 35-37, from "Data on Equal Access to Education," 2023, Civil Rights Data Collection. Copyright 2015 by Office for Civil Rights U.S. Department of Education. https://ocrdata.ed.gov/

APPENDIX B: 2015 PISA AVERAGE SCORES FOR SCIENCE

Table 13 provides the 2015 PISA mean scores for the science literacy scale for all countries with the U.S. data highlighted in blue. Also included are the standard errors (SE). With regards to the U.S., it is important to consider that only two states and a territory were sampled individually to get their mean score per state/territory (OECD, 2016a). Figure 38 shows the relative stability of U.S. science mean scores over time from 2006 to 2018 (OECD, 2016a).

Table 13

2015 PISA Country Rankings by Average Score in Science

Education System	Average Score		*SE*	Education System	Average Score		*SE*
OECD average	493		0.4	Iceland	473	▼	1.7
Singapore	556	▲	1.2	Israel	467	▼	3.4
Japan	538	▲	3.0	*Malta*	465	▼	1.6
Estonia	534	▲	2.1	Slovak Republic	461	▼	2.6
Chinese Taipei	532	▲	2.7	Greece	455	▼	3.9
Finland	531	▲	2.4	Chile	447	▼	2.4
Macau (China)	529	▲	1.1	*Bulgaria*	446	▼	4.4
Canada	528	▲	2.1	*United Arab Emirates*	437	▼	2.4
Vietnam	525	▲	3.9	*Uruguay*	435	▼	2.2
Hong Kong (China)	523	▲	2.5	*Romania*	435	▼	3.2
B-S-J-G (China)	518	▲	4.6	*Cyprus*	433	▼	1.4
Korea, Republic of	516	▲	3.1	*Moldova, Republic of*	428	▼	2.0
New Zealand	513	▲	2.4	*Albania*	427	▼	3.3
Slovenia	513	▲	1.3	Turkey	425	▼	3.9
Australia	510	▲	1.5	*Trinidad and Tobago*	425	▼	1.4
United Kingdom	509	▲	2.6	*Thailand*	421	▼	2.8
Germany	509	▲	2.7	*Costa Rica*	420	▼	2.1
Netherlands	509	▲	2.3	*Qatar*	418	▼	1.0
Switzerland	506	▲	2.9	*Colombia*	416	▼	2.4
Ireland	503		2.4	Mexico	416	▼	2.1
Belgium	502		2.3	*Montenegro, Republic of*	411	▼	1.0
Denmark	502		2.4	*Georgia*	411	▼	2.4
Poland	501		2.5	*Jordan*	409	▼	2.7
Portugal	501		2.4	*Indonesia*	403	▼	2.6
Norway	498		2.3	*Brazil*	401	▼	2.3
United States	**496**		**3.2**	*Peru*	397	▼	2.4
Austria	495		2.4	*Lebanon*	386	▼	3.4
France	495		2.1	*Tunisia*	386	▼	2.1
Sweden	493		3.6	*Macedonia, Republic of*	384	▼	1.2
Czech Republic	493		2.3	*Kosovo*	378	▼	1.7
Spain	493		2.1	*Algeria*	376	▼	2.6

Latvia	490		1.6	*Dominican Republic*	332	▼	2.6
Russian Federation	487	▼	2.9				
Luxembourg	483	▼	1.1				
Italy	481	▼	2.5				
Hungary	477	▼	2.4	**U.S. States and Territories**			
Lithuania	475	▼	2.7	*Massachusetts*	529	▲	6.6
Croatia	475	▼	2.5	*North Carolina*	502		4.9
Buenos Aires (Argentina)	475	▼	6.3	*Puerto Rico*	403	▼	6.1

Note. Adapted from "Science literacy: Average scores," n.d., National Center for Education Statistics. Copyright 2015 by OECD. https://nces.ed.gov/surveys/pisa/pisa2015/pisa2015highlights_3.asp

▲ "Average score is higher than U.S. average score at the .05 level of statistical significance (NCES, n.d.-b)."

▼ "Average score is lower than U.S. average score at the .05 level of statistical significance (NCES, n.d.-b)."

"Education systems are ordered by 2015 average score. The OECD average is the average of the national averages of the OECD member countries, with each country weighted equally. Scores are reported on a scale from 0 to 1,000. All average scores reported as higher or lower than the U.S. average score are different at the .05 level of statistical significance. Italics indicate non-OECD countries and education systems. B-S-J-G (China) refers to the four PISA participating China provinces: Beijing, Shanghai, Jiangsu, and Guangdong. Results for Massachusetts and North Carolina are for public school students only (NCES, n.d.-b)." While Argentina, Malaysia, and Kazakhstan participated in PISA 2015, Argentina only provided a reliable sample from Buenos Aires, Malaysia was unable to meet response rate standards, and Kazakhstan administered only multiple-choice items, which limited comparison by rank in Argentina's case and prevented comparison by rank in the case of Malaysia and Kazakhstan (OECD, 2018).

Figure 38

U.S. Mean Scores for Science Stable Over Time

APPENDIX C: 2015 PISA AVERAGE SCORES BY SCIENCE SUBDOMAIN

Table 14 provides the 2015 PISA mean scores for the three science subscales: physical, living, and Earth and space systems for all countries with the U.S. data highlighted in blue. Also included are the SE. With regards to the U.S., it is important to consider that only two states and a territory were sampled individually to get their mean subscale scores per state/territory (OECD, 2016a). Note, the science competency[70] subscales (i.e., explain phenomena, evaluate and design inquiry, and interpret data and evidence) are outside the scope of this study so are not provided, but are available at the below NCES website.

https://nces.ed.gov/surveys/pisa/pisa2015/pisa2015highlights_3_2.asp

Table 14

2015 PISA Country Rankings by Average Score in Science Subdomain

Physical Systems			Living Systems			Earth and Space Systems		
Education System	**Average Score**	***SE***	**Education System**	**Average Score**	***SE***	**Education System**	**Average Score**	***SE***
OECD average	493	0.5	OECD average	492	0.5	OECD average	494	0.5
Singapore	555	1.6	*Singapore*	558	1.4	*Singapore*	554	1.6
Japan	538	3.2	Japan	538	3.2	Japan	541	3.3
Estonia	535	2.3	*Chinese Taipei*	532	2.7	Estonia	539	2.3
Finland	534	2.6	Estonia	532	2.1	Finland	534	3.0
Macau (China)	533	1.4	Canada	528	2.4	*Chinese Taipei*	534	3.1
Chinese Taipei	531	3.0	Finland	527	2.5	*Macau (China)*	533	1.2
Canada	527	2.4	*Macau (China)*	524	1.4	Canada	529	2.5
Hong Kong (China)	523	2.9	*Hong Kong (China)*	523	2.7	*Hong Kong (China)*	523	2.5
B-S-J-G (China)	520	5.3	*B-S-J-G (China)*	517	4.5	Korea, Republic of	521	3.3
Korea, Republic of	517	3.6	New Zealand	512	2.8	*B-S-J-G (China)*	516	4.9
New Zealand	515	2.7	Slovenia	512	1.6	Slovenia	514	1.8
Slovenia	514	1.6	Korea, Republic of	511	3.2	New Zealand	513	2.7
Netherlands	511	2.6	Australia	510	1.8	Netherlands	513	2.8
Australia	511	1.8	Germany	509	2.9	Germany	512	2.9

[70] NCES refers to OECD's framework competency subscales as "process subscales" on their website.

United Kingdom	509	▲	2.9
Denmark	508	▲	2.7
Ireland	507	▲	2.8
Germany	505	▲	2.8
Switzerland	503	▲	3.1
Poland	503	▲	2.7
Norway	503	▲	2.5
Sweden	500		3.8
Belgium	499		2.4
Portugal	499		2.7
Austria	497		2.7
United States	**494**		**3.5**
France	492		2.4
Czech Republic	492		2.5
Latvia	490		1.7
Russian Federation	488		3.4
Spain	487		2.3
Hungary	481	▼	2.9
Italy	479	▼	2.8
Luxembourg	478	▼	1.4
Lithuania	478	▼	2.8
Iceland	472	▼	1.9
Croatia	472	▼	2.6
Israel	469	▼	3.8
Slovak Republic	466	▼	2.9
Greece	452	▼	4.0
Bulgaria	445	▼	4.4
Chile	439	▼	3.0
United Arab Emirates	434	▼	2.8
Cyprus	433	▼	1.6
Uruguay	432	▼	2.6
Turkey	429	▼	4.3
Thailand	423	▼	3.2
Costa Rica	417	▼	2.4
Qatar	415	▼	1.5
Colombia	414	▼	2.7
Mexico	411	▼	2.2

United Kingdom	509	▲	2.6
Switzerland	506		3.2
Netherlands	503		2.4
Belgium	503		2.4
Portugal	503		2.5
Poland	501		2.8
Ireland	500		2.5
United States	**498**		**3.4**
Denmark	496		2.6
France	496		2.3
Norway	494		2.5
Spain	493		2.3
Czech Republic	493		2.4
Austria	492		2.6
Latvia	489	▼	1.7
Sweden	488		3.7
Luxembourg	485	▼	1.2
Russian Federation	483	▼	2.8
Italy	479	▼	2.7
Croatia	476	▼	2.6
Iceland	476	▼	2.0
Lithuania	476	▼	2.7
Hungary	473	▼	2.6
Israel	469	▼	3.5
Slovak Republic	458	▼	2.8
Greece	456	▼	4.0
Chile	452	▼	2.7
Bulgaria	443	▼	4.5
Uruguay	438	▼	2.5
United Arab Emirates	438	▼	2.6
Cyprus	433	▼	1.5
Turkey	424	▼	3.9
Qatar	423	▼	1.1
Thailand	422	▼	3.2
Costa Rica	420	▼	2.4
Colombia	419	▼	2.5
Mexico	415	▼	2.4

United Kingdom	510	▲	2.8
Australia	509	▲	2.1
Switzerland	508	▲	3.1
Denmark	505		2.7
Belgium	503		2.6
Ireland	502		2.6
Poland	501		2.8
Portugal	500		2.9
Norway	499		2.6
Austria	497		2.9
Spain	496		2.3
United States	**496**		**3.4**
France	496		2.5
Sweden	495		4.1
Czech Republic	493		2.6
Latvia	493		1.9
Russian Federation	489		3.3
Italy	485	▼	2.7
Luxembourg	483	▼	1.6
Croatia	477	▼	2.7
Hungary	477	▼	2.8
Lithuania	471	▼	3.0
Iceland	469	▼	1.9
Slovak Republic	458	▼	2.8
Israel	457	▼	3.8
Greece	453	▼	4.3
Bulgaria	448	▼	4.8
Chile	446	▼	2.5
United Arab Emirates	435	▼	2.8
Uruguay	434	▼	2.6
Cyprus	430	▼	1.6
Turkey	421	▼	4.3
Mexico	419	▼	2.4
Costa Rica	418	▼	2.4
Thailand	416	▼	3.2
Colombia	411	▼	2.7
Montenegro, Republic of	410	▼	2.0

Montenegro, Republic of	407	▼	1.6	*Montenegro, Republic of*	413	▼	1.3	*Qatar*	409	▼	1.2
Brazil	396	▼	2.6	*Brazil*	404	▼	2.6	*Brazil*	395	▼	3.1
Peru	389	▼	2.7	*Peru*	402	▼	2.7	*Peru*	393	▼	3.1
Tunisia	379	▼	2.4	*Tunisia*	390	▼	2.4	*Tunisia*	387	▼	3.4
Dominican Republic	332	▼	3.0	*Dominican Republic*	332	▼	2.8	*Dominican Republic*	324	▼	3.4
Buenos Aires (Argentina)	—		†	*Buenos Aires (Argentina)*	—		†	*Buenos Aires (Argentina)*	—		†
Romania	—		†	*Romania*	—		†	*Romania*	—		†
Jordan	—		†	*Jordan*	—		†	*Jordan*	—		†
Vietnam	—		†	*Vietnam*	—		†	*Vietnam*	—		†
Georgia	—		†	*Georgia*	—		†	*Georgia*	—		†
Albania	—		†	*Albania*	—		†	*Albania*	—		†
Trinidad and Tobago	—		†	*Trinidad and Tobago*	—		†	*Trinidad and Tobago*	—		†
Macedonia, Republic of	—		†	*Macedonia, Republic of*	—		†	*Macedonia, Republic of*	—		†
Algeria	—		†	*Algeria*	—		†	*Algeria*	—		†
Indonesia	—		†	*Indonesia*	—		†	*Indonesia*	—		†
Malta	—		†	*Malta*	—		†	*Malta*	—		†
Lebanon	—		†	*Lebanon*	—		†	*Lebanon*	—		†
Kosovo	—		†	*Kosovo*	—		†	*Kosovo*	—		†
Moldova, Republic of	—		†	*Moldova, Republic of*	—		†	*Moldova, Republic of*	—		†
U.S. States and Territories				**U.S. States and Territories**				**U.S. States and Territories**			
Massachusetts	526	▲	6.7	*Massachusetts*	533		6.9	*Massachusetts*	528	▲	6.6
North Carolina	501		5.2	*North Carolina*	503		5.4	*North Carolina*	502		5.0
Puerto Rico	—		†	*Puerto Rico*	—		†	*Puerto Rico*	—		†

Note. Adapted from "Science literacy: Average scores," n.d., National Center for Education Statistics. Copyright 2015 by OECD. https://nces.ed.gov/surveys/pisa/pisa2015/pisa2015highlights_3.asp

▲ "Average score is higher than U.S. average score at the .05 level of statistical significance (NCES, n.d.-b)."

▼ "Average score is lower than U.S. average score at the .05 level of statistical significance (NCES, n.d.-b)."

— "Not available (NCES, n.d.-b)."

† "Not applicable (NCES, n.d.-b)."

"Education systems are ordered by 2015 average subscale score. The OECD average is the average of the national averages of the OECD member countries, with each country weighted equally. Scores are reported on a scale from 0 to 1,000. Albania, Algeria, Buenos Aires (Argentina), Georgia, Indonesia, Jordan, Kosovo, Lebanon, Malta, Republic of Macedonia, Republic of Moldova, Puerto Rico, Romania, Trinidad and Tobago, and Vietnam

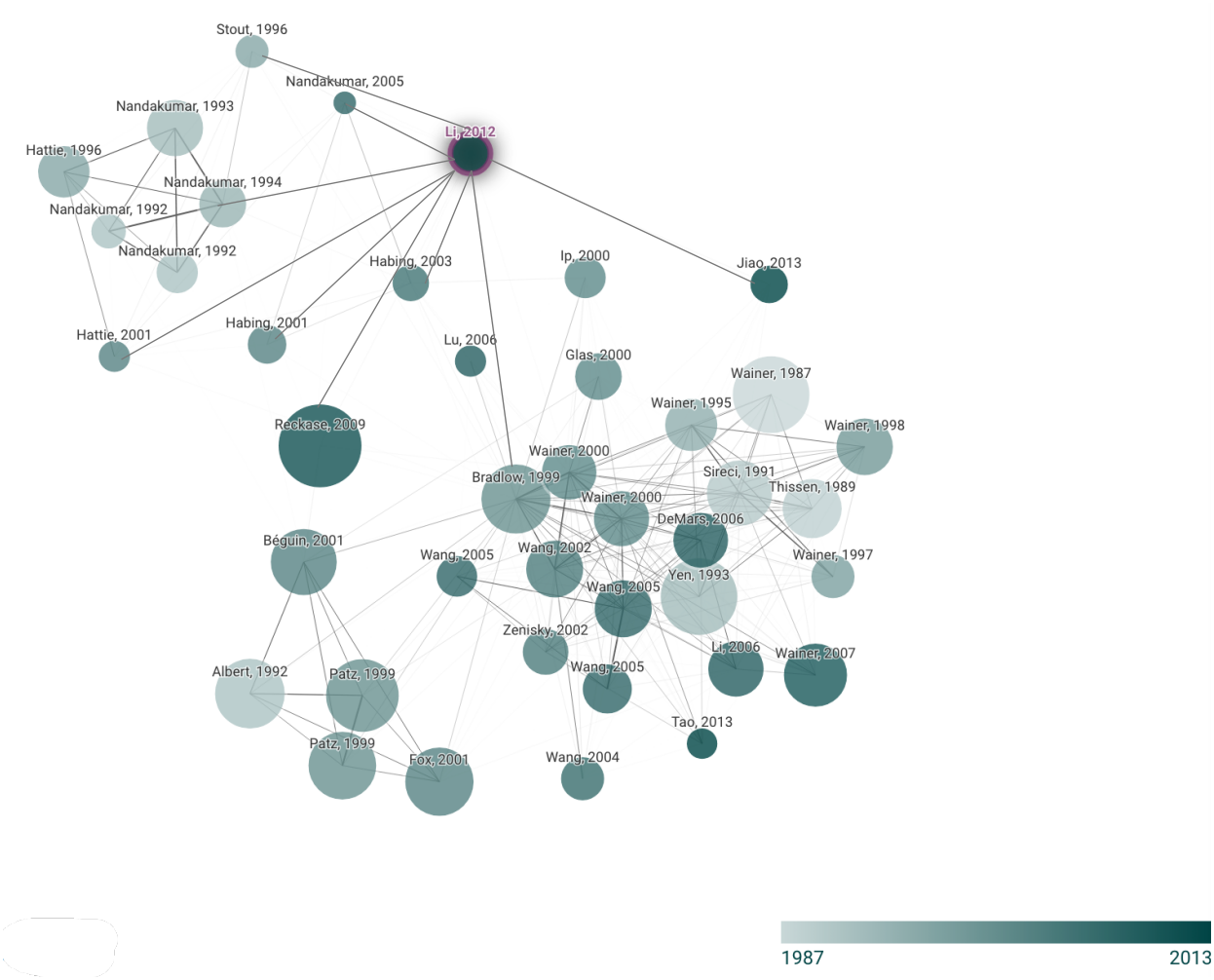
administered paper-based trend items and have no scores for science subscales. Italics indicate non-OECD countries and education systems. B-S-J-G (China) refers to the four PISA participating China provinces: Beijing, Shanghai, Jiangsu, and Guangdong. Results for Massachusetts and North Carolina are for public school students only (NCES, n.d.-b).” While Argentina, Malaysia, and Kazakhstan participated in PISA 2015, Argentina only provided a reliable sample from Buenos Aires, Malaysia was unable to meet response rate standards, and Kazakhstan administered only multiple-choice items, which limited comparison by rank in Argentina’s case and prevented comparison by rank in the case of Malaysia and Kazakhstan (OECD, 2018).

APPENDIX D: LITERATURE CONNECTIONS

Figure 39 details literature connections to the 2012 Li et al. article, which has similar methodology to my proposed study. Note, Figure 39 is hyperlinked to an interactive, larger version of the graphic, which was generated from https://www.connectedpapers.com/.

Figure 39

Connections to the 2012 Li Article



APPENDIX E: LITERATURE REVIEW MATRIX

Table 15 provides a detailed account of the literature review. The table is organized alphabetically with notes on various aspects of each piece of literature. The measurement type is color coded for ease of use – see key below. Also included are any barriers to using MIRT models to describe student performance in the science content subdomains. The big ideas driving the inclusion of the literature in this dissertation are summarized in the final column.

Color Key
Mixed Methods
Qualitative
Quantitative
Included in References

Table 15

Results of Literature Review

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
AERA (2014)	NA	NA	Book	Quantitative	NA	Education testing standards
Aktürk et al. (2017)	Turkey, Early Childhood	STEM	Article	Qualitative	NA	Example of curriculum document analysis
Armstrong (2021)	NA	NA	Article	Qualitative	NA	Document analysis; grounded theory; epistemology
Athalonz (2023)	NA	Life and Physical Sciences	Blog Post	NA	NA	Little curriculum overlap between life and physical science
Atkinsson (2010)	NA	NA	Online Article	Learning Theory	NA	Collaboration as part of constructivism

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Ayala (2022)	NA	NA	Book	Quantitative	Sources of indeterminacy (pg. 408)	Complete overview of IRT
Baghaei (2012)	Iran, HS	English	Article	Quantitative	Difficult to justify to test takers why scores on different dimensions depend on each other	MIRT in large-scale assessment, compares between- and within-item multidimensionality (see Figure 1) – hold for discussion
Baskarada & Koronios (2018)	NA	NA	Article	Mixed Methods	NA	Epistemology and philosophical considerations for mixed methods research
Berenzer & Adams (2017)	Global, 15-year-olds, 4th graders	Math, Reading, Science	Book	PISA, PIRLS, Quantitative, large-scale	NA	Use of scaling and IRT in large-scale assessments; IRT model choice
Beribisky & Hancock (2023)	NA	NA	Article	Quantitative	NA	RMSEA comparison
Binkley & Ma (2023)	U.S., HS	Advanced (AP) classes	Newspaper Articles	NA	NA	Inequity in advanced placement courses by student ethnicity (Black and Latino especially)
Boon et al. (2022)	NA	Science	Article	Epistemology	NA	Overview of constructivist approach to science education
Bowen (2009)	NA	NA	Article	Qualitative	NA	Document analysis as a research method – use in methods section
Brandt (2015)	Global, 15-year-olds, U.S.	NA	Dissertation	PISA, NAEP, Quantitative, large-scale,	Reliability of comprehensive scores	KEY! Large-scale assessment's unidimensional approach when really multidimensional;
Briggs & Wilson (2003)	NA	NA	Article	Quantitative	NA	Intro to multidimensional Rasch models; art and science of measurement
Broesch et al. (2020)	NA	NA	Article	Mixed Methods	NA	Research design that takes culture into consideration
Brooks-Bartlett (2017)	NA	NA	Online Article	Quantitative	NA	Introduction to probability

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Brooks-Bartlett (2018)	NA	NA	Online Article	Quantitative	NA	Maximum likelihood estimation
Camilleri (2023)	NA	NA	News Article	Quantitative	NA	Historical timeline of IRT
Carnoy et al. (2015)	U.S.	Math, Reading, Science	Briefing	PISA, NAEP, Quantitative	NA	State comparisons more useful than U.S. to international comparisons
Caro & Biecek (2017)	International	NA	Article	PISA, TIMSS, PIRLS, Quantitative	NA	An R package for analyzing large-scale assessment data
Carpiano & Daley (2006)	NA	NA	Article	Epistemology	NA	Figure 1 is a framework/theory/model view for philosophy that will work for qualitative review of content framework too – use in methods section; glossary for epistemology
CFPB (2019)	Global	Financial Literacy	Report	PISA, Mixed Methods	NA	Overview of PISA financial literacy results;
Claesgens et al. (2008)	U.S., HS and University	Chemistry	Article	Mixed Methods	NA	Uses IRT to match scores to a framework (pg. 8 for discussion chapter)
College Board (2009)	U.S.	Science	Book	NA	NA	Science standards for college success
Columbia (2023)	NA	NA	Webpage	Quantitative	NA	Item parameters and IRT
Corcoran et al. (2009)	U.S., K-12	Science	Report	NA	NA	Status of U.S. science education; role of science learning progressions (for possible use in my discussion chapter)
CPET (n.d.)	NA	NA	Webpage	NA	NA	Equity in assessment
CRCD (2023)	U.S., K-12	Math, Science, AP	Webpage	Quantitative	NA	Civil rights data for education in U.S. K-12

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Creswell (2015)	NA	NA	Article	Mixed Methods	NA	Approaches to and handbook for mixed methods
Crisan et al. (2017)	NA	NA	Article	Quantitative, Simulation	NA	Consequences of unidimensional IRT model misfit
Csapó (2017)	NA	NA	Online Book	NA	NA	Overview of 21st Century skills
Curran et al. (1996)	NA	NA	Article	Quantitative	NA	Skew and kurtosis acceptable ranges
Davier et al. (2019)	Global, 15-year-olds	Science, Math, and Reading	Article	PISA, Quantitative	MIRT was not used in original analysis so cannot be used in newer research in order to preserve trend and prior conclusions	Using multiple-group Rasch model rather than unidimensional IRT for linking in PISA to generate cross-country comparisons (for discussion chapter)
DeMars (2016)	NA	NA	Article	Quantitative, Simulation	Item difficulties can vary by dimension	Key! Non-compensatory MIRT equation (for methods and results chapters)
Dorans & Kingston (1985)	NA	NA	Article	Quantitative	NA	Violating unidimensionality
Duran (2014)	NA	NA	PowerPoint Presentation	Quantitative	Separate multi into unidimensional subtests	KEY! IRT and MIRT advantages overview
Duschl et al. (2007)	U.S., K-8	Science	Book	NA	NA	Science learning and learning progressions (for possible use in my discussion chapter)
El Masri & Andrich (2020)	UK, France, Jordan, 15-year-olds	Science	Article	PISA, Quantitative, IRT	NA	Model fit analysis in relation to invariance and validity with regards to DIF
Enger & Yager (2009)	U.S., K-12	Science	Book	NA	NA	Discussion on how concept subdomains of science should be taught; change to inquiry learning; background on other learning domains
EPI (2015)	Global, 15-year-olds	Science, Reading, Math	Report	PISA, Mixed Methods	NA	2012 PISA, NAEP, and TIMSS results comparison between

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
						U.S. states rather than international
Ercikan & Oliveri (2016)	NA	21st Century Skills	Article	NA	NA	Assessment design uses standards to measure desired traits; ECD; cognitive evidence needed for complex constructs rather than just expert reviews of items
Erzberger & Kelle (2003)	NA	NA	Book	Mixed Methods	NA	Mixed methods handbook
Fisher (2023)	Unspecified Children	Adverse Childhood Experiences	Dissertation	Mixed Methods	NA	Example of methods section split into two plans
Fu (2016)	U.S., Grade 8	Algebra	Article	Quantitative	Practical significance vs. statistical significance	MIRT models with covariates applied to longitudinal test data
Gale et al. (2013)	NA	NA	Article	Qualitative	NA	Framework method to compare/contrast qualitative data
Gao et al. (2019)	U.S., HS to College	Science	Report	Mixed Methods	NA	PPIC report on HS science course requirements for CA university admission – save for discussion as mentions racial disparity in meeting new requirements
Garnier-Villarreal et al. (2021)	NA	NA	Article	Quantitative	The number of factors to be evaluated	Estimation limits of between-item MIRT models
GEOstata (2016)	Global	Science	Webpage	PISA	NA	Map of PISA 2015 country/economy participants
Godwin (2017)	College	Engineering	Conference Article	Mixed Methods	NA	Latent diversity's impact on creative solutions to engineering problems

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Gomes et al. (2020)	Brazil, Student age varied	Math	Article	PISA, Quantitative	NA	Why scores may be impacted by student age due to taking more classes (hold for discussion)
Greene & Caracelli (1997)	NA	NA	Article	Mixed Methods	NA	Pragmatic paradigm in mixed methods design using different epistemologies
Greenwood (2020)	NA	NA	Blog Post	Learning Theory	NA	Social constructivism overview
Griffin & McGaw (2012)	NA	NA	Book	Quantitative	NA	Assessment of 21st century skills
Haksing (2010)	NA	NA	Article	Quantitative	Multidimensionality does not mean MIRT has to be used	Multidimension model indistinguishable from unidimensional model
Hanushek et al. (2008)	Global, 15-year-olds	Math	Article	PISA, Quantitative	NA	How cognitive growth as measured by PISA impacts the U.S. economy
Harris (n.d.)	NA	NA	Article	Quantitative	NA	Compares equations for UIRT models
Harrison et al. (2015)	U.S., HS, MS, Hawaii	Nature of Science	Article	NGSS, Mixed Methods	NA	Multidimensional nature of nature of science learning; MRCML model
Hartig & Hohler (2009)	NA	NA	Article	Quantitative	NA	KEY! SEM models for MIRT: between and within items;
Hebel et al. (2017)	Global, 15-year-olds, France	Science	Article	PISA, Mixed Methods	NA	Difficulty of PISA science items
Hoover et al. (2018)	Secondary	Earth and Space Sciences (ESS)	Report	NA	NA	Statistics on school requirements for ESS education
Hsu et al. (2023)	Undergraduate	Biology	Article	Quasi-random, Mixed Methods	NA	How item framing may impact student performance (for possible use in my discussion chapter)
IES (n.d.)	U.S., Grades 4 and 8	Science, Math	Webpage	TIMSS, Quantitative	NA	Overview of TIMSS science assessment

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Iliescu (2021)	NA	NA	Article	Validity	NA	Social consequence of testing;
Immekus (2019)	Undergraduate	Engineering	Article	Quantitative, MIRT, CFA	NA	KEY! Overview of IRT (2PL) and MIRT
Intasoi et al. 2020	Thailand, Grade 7	Science	Article	Quantitative	NA	Multidimensional MRCMLM model works better to measure scientific competency framework
Irribarra & Arneson (2023)	NA	Social Sciences, Education	Article	Quantitative	NA	Defining dimensional structure
Issayeva (2022)	NA	NA	Webpage	Quantitative	NA	Compensatory vs. non-compensatory MIRT models; MLE; guessing parameter
Jerrim (2016)	NA	NA	PowerPoint Presentation	PISA, Quantitative	NA	Types of weighting; handling missing data
Jerrim (2023)	Global	Science, Reading, Math	Article	PISA, TIMSS, PIRLS, Quantitative	NA	Interest in large-scale assessments by country
Jerrim et al. (2018)	Sweden, Germany, Ireland, 15-year-olds	Science, Reading, Math	Article	PISA, Quantitative	NA	Change of mode to computer-based assessment
Ji (2023)	China and Canada, Kindergarten	Self-regulation	Thesis	Qualitative	NA	Document analysis methodology
Johnson (2019)	U.S., HS	Integrated Science	Web article	NA	NA	Parent and student concerns over integrated science
Johnson & Onwuegbuzie (2004)	NA	NA	Article	Mixed Methods	NA	Pragmatic mixed methods – equal status design
Jolliffe & Cadima (2016)	NA	NA	Article	Quantitative	NA	A review of PCA
Kaldaras et al. (2021)	NA	Science	Article	NGSS, Quantitative,	NA	Validating latent multi-factor science constructs

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
				EFA, CFA, Invariance Analysis		
Kandanaarachchi & Smith-Miles (2023)	NA	NA	Article	Quantitative	NA	IRT used to evaluate machine learning algorithm (future research section)
Kaplan & Huang (2021)	U.S.	Math, Reading	Article	NAEP, Quantitative	NA	Bayesian probabilistic forecasting view; combining NSLP variable as a proxy for socio-economic status
Kapucu (2021)	Turkey, 9th grade	Science	Article	Mixed Methods	NA	Student perceptions of different science subdomains differentiate from one another
Kelley & Knowles (2016)	International	STEM	Article	NA	NA	Integrated STEM education
Kim & Wilson (2020)	NA	NA	Article	Quantitative	NA	Polytomous item explanatory item response theory models via MGLMM
Kim et al. (2019)	NA	NA	Article	Quantitative	NA	Overview of information criteria usage when comparing model fit
Kose & Demirtasli (2012)	Turkey, 8th grade	Language	Article	Quantitative	Longer tests and larger sample sizes are needed to increase model sensitivity and decrease error	Comparing sample size and test length for UIRT and MIRT models
Krutsch & Roderick (2022)	NA	STEM	Blog Post	Quantitative	NA	U.S. Department of Labor chart on STEM job growth
Kuo & Sheng (2016)	NA	NA	Article	Quantitative, Simulation	NA	Estimation methods for multi-unidimensional graded response

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Lang & Tay (2021)	NA	NA	Article	IRT, MIRT, Quantitative	Can be similar to other models such as CTT or CFA; unidimensional models are more familiar and easily interpreted (for discussion chapter)	KEY! Overview of IRT models; includes R code; history of MIRT development
Lau (2009)	Global	Scientific Literacy	Article	PISA, Qualitative	NA	Review of 2006 framework finds construct validity issues
Learn PILA (n.d.)	Global	Innovative	Webpage	Quantitative	NA	Platform for Innovative Learning Assessments
Lee & Tsai (2012)	College	Biology, Physics	Article	Qualitative	NA	Investigation across different domains of science is rare for student epistemological beliefs
Li et al. (2012)	U.S., K-12	Science	Article	Michigan Gr. 5, Quantitative, EFA, CFA	Unidimensionality and local item dependence assumptions	KEY! MIRT model used successfully for large-scale state assessment in science
Lin (1998)	NA	NA	Article	Qualitative	NA	Positivist vs. Interpretivist approaches
Lips & Moritz (2022)	U.S.	STEM	Report	NA	NA	Government spending on STEM
Liu et al. (2022)	U.S., Grade 8, Math	NA	Article	NAEP, MIRT, Quantitative	Scalability to big data	2PL IRT most commonly used; report RMSE
MacLeod & Nelson (1984)	U.S., Undergraduates	NA	Article	Quantitative	NA	Memory recall as a unidimensional construct
Mari et al. (2017)	NA	NA	Article	Quantitative	NA	Nature of measurement vs. measure
Marlowe (1986)	NA	NA	Article	Quantitative	NA	Multidimensionality in social intelligence
Masur (2022)	NA	NA	Webpage	Quantitative	NA	IRT models in MIRT R package
Maul (2019)	NA	NA	Article	NA	NA	Intersubjectivity of measurement
Maxwell & Mittapalli (2010)	NA	NA	Handbook	Mixed Methods	NA	Scientific realism

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Mazzei & Jackson (2024)	NA	NA	Book	Qualitative	NA	Re-animating documents in another form
McDonald (1999)	NA	NA	Book	CFA, IRT, MIRT, Quantitative	NA	Overview of multiple quantitative statistics
Mcleod (2023)	NA	NA	Online Article	Learning Theory	NA	Constructivism
Messick (1989)	NA	NA	Article	NA	NA	Use vs. interpretation inferences in assessment validity
Messick (1993)	NA	NA	Article	NA	NA	Consequential validity as an aspect of construct validity
Messick (1995)	NA	NA	Article	NA	NA	Construct validity
Monseur et al. (2011)	Global	Reading, Science, Math	Article	PISA, Quantitative	NA	Violation of independence assumption by items in a set
Moroi (2020)	NA	NA	Article	Mixed Methods	NA	Philosophies of research
Mostafa et al. (2018)	Global, 15-year-olds	Science	Report	PISA, Mixed Methods	NA	Student enjoyment of science linked to inquiry teaching; 2015 PISA science test design/scoring/scale
Müller (2020)	NA	NA	Article	Quantitative	NA	Infit item statistic acceptable bounds – save for results chapter
National Research Council (2012)	U.S., K-12	Science	Book	NA	NA	A framework for science education; chapter 11 discusses DEI in science education (for possible use in my discussion chapter)
NCES (2019)	U.S.	Math, Science	Report	PISA	NA	K-12 course completion statistics
NCES (2022)	U.S.	Science, Math	Report	NAEP, Quantitative	NA	Science course completion data
NCES (n.d.-a)	Global, 15-year-olds, U.S.	Science	Webpage	PISA	NA	Number of U.S. schools participating in 2015; implies

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
						students do not have to participate
NCES (n.d.-b)	Global, 15-year-olds, U.S.	Science	Webpage	PISA, Quantitative	NA	2015 PISA scores for all participating countries, also broken down by subdomain; U.S. sampling and data collection methods
NGSS (2013)	U.S.	STEM	Webpage	NA	NA	Next Generation Science Standards
Niiniluoto ed. et al. (2004)	NA	NA	Book	Qualitative	NA	Overview of different epistemologies and their origins
NWEA (2015)	NA	NA	Blog Post	Quantitative	NA	KEY! List of factors impacting use of MIRT
OCR (2023a)	U.S., K-12	Math, Science, Computer Science	Report	Mixed Methods	NA	Student access to education
OCR (2023b)	U.S., K-12	NA	Report	Quantitative	NA	Student enrollment
OECD (2016a)	Global, 15-year-olds	Science, Reading, Math	Report	PISA, Mixed Methods	NA	U.S.-specific report on 2015 data
OECD (2016b)	Global, 15-year-olds	Science, Reading, Math	Report	PISA, Mixed Methods	NA	Map of participating countries; 2015 results focusing on equity in education
OECD (2017a)	Global	Science	Framework	PISA	NA	2015 combined framework including science
OECD (2017b)	Global, 15-year-olds	Science, Reading, Math	Report	PISA, Mixed Methods	NA	Key! Technical report for PISA 2015 [Annex F – technical standards; Annex A – item codes and counts]
OECD (2018)	Global, 15-year-olds	Science, Reading, Math	Report	PISA, Mixed Methods	NA	Results in focus, data overview
OECD (2019)	Global	Science	Framework	PISA	NA	2018 science framework
OECD (2020)	Global	Science	Report	PISA	NA	Strategic vision for 2024 science assessment

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
OECD (2023)	Global	Learning in Digital World	Report	NA	NA	Digital assessment framework
OECD (n.d.-a)	Global, 15-year-olds	NA	Webpage	NA	NA	Overview of PISA
OECD (n.d.-b)	Global	NA	Webpage	NA	NA	Frequently asked questions about PISA
OECD (n.d.-c)	Global, 15-year-olds	Science, Collaborative Problem Solving	Report	PISA, Mixed Methods	NA	2015 released field trial items
Osteen (2010)	MSW students	NA	Article	Quantitative	NA	Integrating CFA and MIRT
Ostlund et al. (2011)	NA	NA	Article	Mixed Methods	NA	Triangulation of findings from mixed methods research
Otarigho & Oruese (2013)	Nigeria, Secondary Schools	Integrated Science	Article	Qualitative	NA	Using integrated science curriculum to provide students with an understanding of how science affects everyday lives
Park et al. (2019)	Belgium, 6- to 8-year-olds	Number Sense	Article	MIRT, Quantitative	Higher error for ability in unidimensional than multidimensional if item set is truly multidimensional	MIRT in adaptive learning systems
Park et al. (2020)	NA	NA	Blog Post	Quantitative	NA	ICC overview with item parameters
Pellegrino & Hilton (2012)	U.S., K-12	ELA, Math, and Science	Book	NA	NA	Developing transferable knowledge and skills in 21st century; emphasis on science inquiry (for possible use in my discussion chapter)
NRC - Pellegrino et al. (2001)	U.S., K-12	All	Book	NA	NA	Intersection of student learning and assessment (for possible use in my discussion chapter)
Pelz (n.d.)	NA	NA	Webpage	Social Science Research	NA	Takes at least two dimensions to be multidimensional

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Pierson et al. (2019)	U.S., K-12	Science	Article	NGSS	NA	Learning progressions; tension over teaching fact memorization vs. inquiry
PISA USA (2015)	U.S., 15-year-olds	Reading, Math, Science, Collaborative Problem Solving, Financial Literacy	Brochure	NA	NA	Students volunteer to take PISA if randomly selected by OECD.
Pokropek (2022)	Global, 15-year-olds	Science, Math, Reading	Article	PISA, Quantitative	NA	KEY! Only 17% for science is independent of the general ability factor and can be attributed to specific science ability factor (for possible use in my discussion chapter)
Polites et al. (2012)	NA	NA	Article	Mixed Methods	Conceptualizing dimension relationships using theory	Defining multidimensionality
Reckase (1985)	NA	NA	Article	Quantitative	NA	Multidimensional item difficulty
Reckase (1989)	NA	NA	Article	Quantitative	An item requiring two cognitive skills to solve may still be unidimensional	Application of MIRT – hold for discussion
Reckase (1990)	NA	NA	Paper Presentation	Quantitative	NA	Defining dimensionality
Reckase (1997)	NA	NA	Article	Quantitative	Multiple items can be selected based on MIRT, but modeled unidimensionally	Future directions for MIRT; early MIRT development
Reckase (2009)	NA	NA	Book	Quantitative	Small number of items on test	Psychological and educational context for MIRT

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Reed & Wolfson (2021)	U.S., HS, College	Chemistry	Article	Qualitative	NA	Learning progressions assume linear learning path; LPs not used by all
Reis (2016)	U.S., Massachusetts	Reading, Science, Math	Newspaper Article	PISA	NA	Massachusetts scores at a high level similar to country leaders of PISA 2015
Richman (2023)	K-12, Texas	Math	Newspaper Article	NA	NA	Using test scores to limit bias and determine which math class a student can take
Richman & Crain (2022)	K-12, U.S.	NA	Newspaper Article	NA	NA	Teacher shortages lead to accepting teachers with less training
Ruiz-Primo & Li (2015)	Global, 15-year-olds	Science		PISA 2006 and 2009, Quantitative	NA	How item context affects student performance
Saunders et al. (2018)	NA	NA	Article	Qualitative	NA	Saturation evaluation and grounded theory
Scalise (2017a)	U.S., MS	Science	Article	Quantitative	NA	MIRT model for tech-enhanced items
Scalise (2017b)	NA	Neuroscience	Book	NA	NA	Describes how students learn
Scalise & Clarke-Midura (2018)	U.S., MS	Science Inquiry	Article	Quantitative, MIRT	NA	KEY! Compares the fit of different IRT models to the data; Bayes net for process data
Scalise & Gifford (2006)	NA	NA	Article	NA	Can item type affect multidimensionality?	Overview of item types and constraints (for possible use in my discussion chapter)
Scalise & Wilson (2011)	NA	21st Century Learning	Article	Quantitative	NA	Multidimensionality in constructs
Scalise et al. (2018)	NA	STEM	Article	Literature Review and Analysis	NA	Digital accommodations for students
Scalise et al. (2021)	NA	NA	Article	Quantitative	NA	Learning analytics; figure 1

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Siegel (2006)	NA	NA	Article	Qualitative	NA	Epistemological diversity in education research
Socha (n.d.)	NA	NA	Unpublished article	Quantitative	Large sample size	MIRT assumptions; ICS analog to ICC
Spencer (2004)	NA	NA	Dissertation	Quantitative	NA	Comparing MIRT to UIRT when retrieving item difficulties and differentiation
Stehle & Peters-Burton (2019)	U.S., HS	STEM	Article	Quantitative	NA	U.S. students underperforming in science
Strauss (2019)	Global, 15-year-olds	NA	Newspaper Article	PISA	NA	Negative aspects of PISA scores
Taut & Palacios (2016)	Global, 15-year-olds	NA	Article	PISA	NA	Intended and unintended interpretations and uses of PISA results
Thomson et al. (2013)	Australia	Scientific Literacy	Report	PISA	NA	Overview of scientific literacy as assessed by PISA
Tucker (2016)	U.S.	Reading, Math, Science	Newspaper article	PISA	NA	No meaningful change in science scores; U.S. 2nd generation immigrants worst educated; math teachers lack appropriate training; poor recruitment strategy of teachers
Tulodziecki (2012)	NA	NA	Article	Qualitative	NA	Epistemic equivalence
Turcotte (2023)	HS	Math	Newspaper article	NA	NA	Detracking students to increase equity led to increased math performance
Tykoski (2017)	U.S., HS	ESS	Blog Post	NA	NA	Lack of ESS in IB and AP courses
Uesaka et al. (2022)	Japan, higher, middle, and lower-ranked universities	Self-regulated learning	Article	Quantitative	NA	Usefulness of IRT to classroom instruction

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Ulitzsch & Nestler (2022)	NA	NA	Article	Quantitative	NA	Bayesian IRT model
Vazquez (2006)	U.S., HS	Biology, Physics, Chemistry	Article	Qualitative	NA	Description of flow of course delivery in the sciences, i.e., biology to chemistry to physics, and if it should change to a physics first approach – save for discussion on inequity in course access
Venkatesh et al. (2013)	NA	Information Systems	Article	Mixed Methods	NA	Purposes for and guidelines of mixed methods research; developing meta-inferences; validity of mixed methods research
Venkatesh et al. (2016)	NA	NA	Article	Mixed Methods	NA	KEY! Extension of 2013 article with variations of mixed methods research; epistemology
Voogt & Roblin (2012)	International	21st Century Competencies	Article	Qualitative	NA	Document selection; screening framework for sub-themes (see Table 3)
Wach (2013)	NA	NA	Article	Qualitative	NA	Document analysis; set inclusion criteria; coding; validity
Walker (2016)	U.S., Global, 15-year-olds	Reading, Math, Science, Collaborative Problem Solving, Financial Literacy	News Article (web)	PISA, Mixed Methods	NA	U.S. scores remain in middle of other country averages for 2015
Wang (2021)	NA	NA	Article	Quantitative	Costs of large data sets	Unidimensionality assumption; history of MIRT

Author/s or Editor/s or Abbreviated Title (Date)	Student Demographics	Content Domain	Reference Type	Measurement Type	Barriers to Using MIRT Model	Big Idea/s
Wang & Nydick (2015)	NA	NA	Article	Quantitative, Simulation	Items on multiple dimensions leads to increased variability with regards to difficulty parameters on each dimension, which results in decreased information	Algorithms for non-compensatory MIRT models
Welch (1977)	U.S., K-12	Science	Book	NA	NA	Chapter 3 focuses on history of integration for teaching science subdomains
Wess et al. (2021)	NA	NA	Book	Quantitative	NA	Ch. 4: Test quality with regards to types of validity
Western Governors University (2020)	NA	NA	Blog Post	Epistemology	NA	Definitions and key characteristics/principles of social constructivism
Wilson (2013)	NA	NA	Article	Quantitative	NA	IRT overview; changes from CCT
Winarno et al. (2020)	International	Science	Article	Literature Review	NA	Problems with teaching integrated science classes
Yen & Leah (2007)	K-12	EL	Presentation Paper	Quantitative, EFA	Number of parameters to be estimated	KEY! Exploratory approach to MIRT model emphasizes finding the best fitting model
You et al. (2020)	Global, 15-year-olds	Science	Article	PISA, Quantitative	NA	School characteristics impact scientific literacy in students
Zakharov (2016)	NA	NA	Article	Quantitative	NA	Cluster analyses should be evaluated for reliability and validity – save for discussion
Zhao & Hambleton (2017)	NA	NA	Article	Quantitative, Simulation	NA	Consequences of IRT model misfit

APPENDIX F: PISA 2015 SCIENCE FRAMEWORK[71]

2

PISA 2015 science framework

Science is the main subject of assessment in the Programme for International Student Assessment (PISA) in 2015. This chapter defines "scientific literacy" as assessed in PISA. It describes the types of contexts, knowledge, competencies and attitudes towards science that are reflected in the assessment's science problems and provides several sample items. The chapter also discusses how student performance in science is measured and reported.

[71] From chapter 2 in *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, revised edition* (OECD, 2017).

This document provides a description of and rationale for the framework that forms the basis of the instrument to assess scientific literacy – the major domain in PISA 2015. Previous PISA frameworks for the science assessment (OECD, 2006, 2004, 1999) have elaborated a conception of scientific literacy as the central construct for science assessment. These documents have established a broad consensus among science educators of the concept of scientific literacy. This framework for PISA 2015 refines and extends the previous construct, in particular by drawing on the PISA 2006 framework that was used as the basis for assessment in 2006, 2009 and 2012.

Scientific literacy matters at both the national and international levels as humanity faces major challenges in providing sufficient water and food, controlling diseases, generating sufficient energy and adapting to climate change (UNEP, 2012). Many of these issues arise, however, at the local level where individuals may be faced with decisions about practices that affect their own health and food supplies, the appropriate use of materials and new technologies, and decisions about energy use. Dealing with all of these challenges will require a major contribution from science and technology. Yet, as argued by the European Commission, the solutions to political and ethical dilemmas involving science and technology "cannot be the subject of informed debate unless young people possess certain scientific awareness" (European Commission, 1995: 28). Moreover, "this does not mean turning everyone into a scientific expert, but enabling them to fulfil an enlightened role in making choices which affect their environment and to understand in broad terms the social implications of debates between experts" (ibid.: 28). Given that knowledge of science and science-based technology contributes significantly to individuals' personal, social, and professional lives, an understanding of science and technology is thus central to a young person's "preparedness for life".

The concept of scientific literacy in this framework *refers to a knowledge of both science and science-based technology*, even though science and technology do differ in their purposes, processes and products. Technology seeks the optimal solution to a human problem, and there may be more than one optimal solution. In contrast, science seeks the answer to a specific question about the natural, material world. Nevertheless, the two are closely related. For instance, new scientific knowledge leads to the development of new technologies (think of the advances in material science that led to the development of the transistor in 1948). Likewise, new technologies can lead to new scientific knowledge (think of how knowledge of the universe has been transformed through the development of better telescopes). Individuals make decisions and choices that influence the directions of new technologies (consider the decision to drive a smaller, more fuel-efficient car). Scientifically literate individuals should therefore be able to make more informed choices. They should also be able to recognise that, while science and technology are often a source of solutions, paradoxically, they can also be seen as a source of risk, generating new problems that can only be solved through the use of science and technology. Therefore, individuals need to be able to weigh the potential benefits and risks of applying scientific knowledge to themselves and society.

Scientific literacy also requires not just knowledge of the concepts and theories of science but also knowledge of the common procedures and practices associated with scientific enquiry and how these enable science to advance. Therefore, individuals who are scientifically literate have a knowledge of the major concepts and ideas that form the foundation of scientific and technological thought; how such knowledge has been derived; and the degree to which such knowledge is proved by evidence or theoretical explanations.

Undoubtedly, many of the challenges of the 21st century will require innovative solutions that have a basis in scientific thinking and scientific discovery. Societies will require a cadre of well-educated scientists to undertake the research and nurture the innovation that will be essential to meet the economic, social and environmental challenges that the world faces.

For all of these reasons, scientific literacy is perceived to be a key competency (Rychen and Salganik, 2003) and defined in terms of the ability to use knowledge and information interactively – that is "an understanding of how it [a knowledge of science] changes the way one can interact with the world and how it can be used to accomplish broader goals" (ibid.: 10). As such, it represents a major goal for science education for all students. Therefore, the view of scientific literacy that forms the basis for the 2015 international assessment of 15-year-old students is a response to the question: What is important for young people to know, value and be able to do in situations involving science and technology?

DEFINING SCIENTIFIC LITERACY

Current thinking about the desired outcomes of science education is rooted strongly in a belief that an understanding of science is so important that it should be a feature of every young person's education (American Association for the Advancement of Science, 1989; Confederacion de Sociedades Cientificas de España, 2011; Fensham, 1985; Millar and Osborne, 1998; National Research Council, 2012; Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2005; Taiwan Ministry of Education, 1999). Indeed, in many countries science is an obligatory element of the school curriculum from kindergarten until the completion of compulsory education.

Many of the documents and policy statements cited above give pre-eminence to an education for citizenship. However, many of the curricula for school science across the world are based on a view that the primary goal of science education should be the preparation of the next generation of scientists (Millar and Osborne, 1998). These two goals are not always compatible. Attempts to resolve the tension between the needs of the majority of students who will not become scientists and the needs of the minority who will have led to an emphasis on teaching science through enquiry (National Academy of Science, 1995; National Research Council, 2000), and new curriculum models (Millar, 2006) that address the needs of both groups. The emphasis in these frameworks and their associated curricula lies not on producing individuals who will be "producers" of scientific knowledge, i.e. the future scientists; rather, it is on educating all young people to become informed, critical users of scientific knowledge.

To understand and engage in critical discussions about issues that involve science and technology requires three domain-specific competencies. The first is the ability to provide explanatory accounts of natural phenomena, technical artefacts and technologies, and their implications for society. Such an ability requires a knowledge of the fundamental ideas of science and the questions that frame the practice and goals of science. The second is the knowledge and understanding of scientific enquiry to: identify questions that can be answered by scientific enquiry; identify whether appropriate procedures have been used; and propose ways in which such questions might be answered. The third is the competency to interpret and evaluate data and evidence scientifically and evaluate whether the conclusions are justified. Thus, scientific literacy in PISA 2015 is defined by the three competencies to:

- explain phenomena scientifically
- evaluate and design scientific enquiry
- interpret data and evidence scientifically.

All of these competencies require knowledge. Explaining scientific and technological phenomena, for instance, demands a knowledge of the content of science (hereafter, content knowledge). The second and third competencies, however, require more than a knowledge of what is known; they depend on an understanding of how scientific knowledge is established and the degree of confidence with which it is held. Some have argued for teaching what has variously been called "the nature of science" (Lederman, 2006), "ideas about science" (Millar and Osborne, 1998) or "scientific practices" (National Research Council, 2012). Recognising and identifying the features that characterise scientific enquiry requires a knowledge of the standard procedures that underlie the diverse methods and practices used to establish scientific knowledge (hereafter, procedural knowledge). Finally, the competencies require epistemic knowledge – an understanding of the rationale for the common practices of scientific enquiry, the status of the knowledge claims that are generated, and the meaning of foundational terms, such as theory, hypothesis and data.

Box 2.1 **Scientific knowledge: PISA 2015 terminology**

This document is based upon a view of scientific knowledge as consisting of three distinguishable but related elements. The first of these and the most familiar is a knowledge of the facts, concepts, ideas and theories about the natural world that science has established. For instance, how plants synthesise complex molecules using light and carbon dioxide or the particulate nature of matter. This kind of knowledge is referred to as "**content knowledge**" or "knowledge of the content of science".

Knowledge of the procedures that scientists use to establish scientific knowledge is referred to as "**procedural knowledge**". This is a knowledge of the practices and concepts on which empirical enquiry is based such as repeating measurements to minimise error and reduce uncertainty, the control of variables, and standard procedures for representing and communicating data (Millar, Lubben, Gott and Duggan, 1995). More recently these have been elaborated as a set of "concepts of evidence" (Gott, Duggan and Roberts, 2008).

Furthermore, understanding science as a practice also requires "**epistemic knowledge**" which refers to an understanding of the role of specific constructs and defining features essential to the process of knowledge-building in science (Duschl, 2007). Epistemic knowledge includes an understanding of the function that questions, observations, theories, hypotheses, models and arguments play in science; a recognition of the variety of forms of scientific enquiry; and the role peer review plays in establishing knowledge that can be trusted.

A more detailed discussion of these three forms of knowledge is provided in the later section on scientific knowledge and in Figures 2.5, 2.6 and 2.7.

Both procedural and epistemic knowledge are necessary to identify questions that are amenable to scientific enquiry, to judge whether appropriate procedures have been used to ensure that the claims are justified, and to distinguish scientific issues from matters of values or economic considerations. This definition of scientific literacy assumes that, throughout their lives, individuals will need to acquire knowledge, not through scientific investigations, but through the use of resources such as libraries and the Internet. Procedural and epistemic knowledge are essential to decide whether the many claims of knowledge and understanding that pervade contemporary media are based on the use of appropriate procedures and are justified.

People need all three forms of scientific knowledge to perform the three competencies of scientific literacy. PISA 2015 focuses on assessing the extent to which 15-year-olds are capable of displaying the three aforementioned competencies appropriately within in a range of personal, local/national (grouped in one category) and global contexts. (For the purposes of the PISA assessment, these competencies are only tested using the knowledge that 15-year-old students can reasonably be expected to have already acquired.) This perspective differs from that of many school science programmes that are dominated by content knowledge. Instead, the framework is based on a broader view of the kind of knowledge of science required of fully engaged citizens.

In addition, the competency-based perspective also recognises that there is an affective element to a student's display of these competencies: students' attitudes or disposition towards science will determine their level of interest, sustain their engagement, and may motivate them to take action (Schibeci, 1984). Thus, the scientifically literate person would typically have an interest in scientific topics; engage with science-related issues; have a concern for issues of technology, resources and the environment; and reflect on the importance of science from a personal and social perspective. This requirement does not mean that such individuals are necessarily disposed towards becoming scientists themselves, rather such individuals recognise that science, technology and research in this domain are an essential element of contemporary culture that frames much of our thinking.

These considerations led to the definition of scientific literacy used in PISA 2015 (see Box 2.2). The use of the term "scientific literacy", rather than "science", underscores the importance that the PISA science assessment places on the application of scientific knowledge in the context of real-life situations.

Box 2.2 **The 2015 definition of scientific literacy**

Scientific literacy is the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen.

A scientifically literate person is willing to engage in reasoned discourse about science and technology, which requires the competencies to:

- **Explain phenomena scientifically** – recognise, offer and evaluate explanations for a range of natural and technological phenomena.
- **Evaluate and design scientific enquiry** – describe and appraise scientific investigations and propose ways of addressing questions scientifically.
- **Interpret data and evidence scientifically** – analyse and evaluate data, claims and arguments in a variety of representations and draw appropriate scientific conclusions.

The competencies required for scientific literacy

Competency 1: Explain phenomena scientifically

The cultural achievement of science has been to develop a set of explanatory theories that have transformed our understanding of the natural world (in this document, "natural world" refers to phenomena associated with any object or activity occurring in the living or the material world), such as the idea that day and night is caused by a rotating Earth, or the idea that diseases can be caused by invisible micro-organisms. Moreover, such knowledge has enabled us to develop technologies that support human life by, for example, preventing disease or enabling rapid human communication across the globe. The competency to explain scientific and technological phenomena is thus dependent on a knowledge of these major explanatory ideas of science.

Explaining scientific phenomena, however, requires more than the ability to recall and use theories, explanatory ideas, information and facts (content knowledge). Offering scientific explanations also requires an understanding of how such knowledge has been derived and the level of confidence we might hold about any scientific claims. For this competency, the individual requires a knowledge of the standard forms and procedures used in scientific enquiry to obtain such knowledge (procedural knowledge) and an understanding of their role and function in justifying the knowledge produced by science (epistemic knowledge).

Competency 2: Evaluate and design scientific enquiry

Scientific literacy implies that students have some understanding of the goal of scientific enquiry, which is to generate reliable knowledge about the natural world (Ziman, 1979). Data collected and obtained by observation and experiment, either in the laboratory or in the field, lead to the development of models and explanatory hypotheses that enable predictions that can then be tested experimentally. New ideas, however, commonly build on previous knowledge. Scientists themselves rarely work in isolation; they are members of research groups or teams that engage, nationally and internationally, in extensive collaboration with colleagues. New knowledge claims are always perceived to be provisional and may lack justification when subjected to critical peer review – the mechanism through which the scientific community ensures the objectivity of scientific knowledge (Longino, 1990). Hence, scientists have a commitment to publish or report their findings and the methods used in obtaining their evidence. Doing so enables empirical studies, at least in principle, to be replicated and results confirmed or challenged. However, measurements can never be absolutely precise; they all contain a degree of error. Much of the work of the experimental scientist is thus devoted to resolving uncertainty by repeating measurements, collecting larger samples, building instruments that are more accurate and using statistical techniques that assess the degree of confidence in any result.

In addition, science has well-established procedures that are the foundations of any experiment to establish cause and effect. The use of controls enables the scientist to claim that any change in a perceived outcome can be attributed to a change in one specific feature. Failure to use such techniques leads to results where effects are confounded and cannot be trusted. Likewise, double-blind trials enable scientists to claim that the results have not been influenced either by the subjects of the experiment, or by the experimenter themselves. Other scientists, such as taxonomists and ecologists, are engaged in the process of identifying underlying patterns and interactions in the natural world that warrant a search for an explanation. In other cases, such as evolution, plate tectonics or climate change, scientists examine a range of hypotheses and eliminate those that do not fit with the evidence.

Facility with this competency draws on content knowledge, a knowledge of the common procedures used in science (procedural knowledge), and the function of these procedures in justifying any claims advanced by science (epistemic knowledge). Procedural and epistemic knowledge serve two functions. First, such knowledge is required by individuals to appraise scientific investigations and decide whether they have followed appropriate procedures and whether the conclusions are justified. Second, individuals who have this knowledge should be able to propose, at least in broad terms, how a scientific question might be investigated appropriately.

Competency 3: Interpret data and evidence scientifically

Interpreting data is such a core activity of all scientists that some rudimentary understanding of the process is essential for scientific literacy. Initially, data interpretation begins with looking for patterns, constructing simple tables and graphical visualisations, such as pie charts, bar graphs, scatterplots or Venn diagrams. At a higher level, it requires the use of more complex data sets and the use of the analytical tools offered by spreadsheets and statistical packages. It would be wrong, however, to look at this competency as merely an ability to use these tools. A substantial body of knowledge is required to recognise what constitutes reliable and valid evidence and how to present data appropriately.

Scientists make choices about how to represent the data in graphs, charts or, increasingly, in complex simulations or 3D visualisations. Any relationships or patterns must then be read using a knowledge of standard patterns. Whether uncertainty has been minimised by standard statistical techniques must also be considered. All of this draws on a body of procedural knowledge. The scientifically literate individual can also be expected to understand that uncertainty is an inherent feature of all measurement, and that one criterion for expressing confidence in a finding is determining the probability that the finding might have occurred by chance.

It is not sufficient, however, to understand the procedures that have been applied to obtain any data set. The scientifically literate individual needs to be able to judge whether they are appropriate and the ensuing claims are justified (epistemic knowledge). For instance, many sets of data can be interpreted in multiple ways. Argumentation and critique are essential to determining which is the most appropriate conclusion.

Whether it is new theories, novel ways of collecting data or fresh interpretations of old data, argumentation is the means that scientists and technologists use to make their case for new ideas. Disagreement among scientists is normal, not extraordinary. Determining which interpretation is the best requires a knowledge of science (content knowledge). Consensus on key scientific ideas and concepts has been achieved through this process of critique and argumentation (Longino, 1990). Indeed, it is a critical and sceptical disposition towards all empirical evidence that many would see as the hallmark of the professional scientist. The scientifically literate individual understands the function and purpose of argument and critique and why they are essential to the construction of knowledge in science. In addition, they should be able both to construct claims that are justified by data and to identify any flaws in the arguments of others.

The evolution of the definition of scientific literacy in PISA

In PISA 2000 and 2003, scientific literacy was defined as:

> "…the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity." (OECD, 2004, 2000)

In 2000 and 2003, the definition embedded knowledge *of* science and understandings *about* science within the one term "scientific knowledge". The 2006 definition separated and elaborated the term "scientific knowledge" by dividing it into two components: "knowledge *of* science" and "knowledge *about* science" (OECD, 2006). Both definitions referred to the application of scientific knowledge to understanding and making informed decisions about the natural world. In PISA 2006, the definition was enhanced by the addition of knowledge of the relationship between science and technology – an aspect that was assumed but not elaborated in the 2003 definition.

> "For the purposes of PISA, scientific literacy refers to an individual's:
>
> - Scientific knowledge and use of that knowledge to identify questions, acquire new knowledge, explain scientific phenomena and draw evidence-based conclusions about science-related issues.
> - Understanding of the characteristic features of science as a form of human knowledge and enquiry.
> - Awareness of how science and technology shape our material, intellectual and cultural environments.
> - Willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen." (OECD, 2006).

These ideas have evolved further in the PISA 2015 definition of scientific literacy. The major difference is that the notion of "knowledge *about* science" has been specified more clearly and split into two components – procedural knowledge and epistemic knowledge.

In 2006, the PISA framework was also expanded to include attitudinal aspects of students' responses to scientific and technological issues within the construct of scientific literacy. In 2006, attitudes were measured in two ways: through the student questionnaire and through items embedded in the student test. Discrepancies were found between the results from the embedded questions and those from the background questionnaire with respect to "interest in science" for all students and gender differences in these issues (OECD, 2009; see also Drechsel, Carstensen and Prenzel, 2011). More important, embedded items extended the length of the test. Hence, in PISA 2015, attitudinal aspects are only measured through the student questionnaire; there are no embedded attitudinal items.

As for the constructs measured within this domain, the first ("interest in science") and third ("environmental awareness") remain the same as in 2006. The second ("support for scientific enquiry") has been changed to a measure of "valuing scientific approaches to enquiry", which is essentially a change in terminology to better reflect what is measured.

In addition, the contexts in PISA 2015 have been changed from "personal, social and global" in the 2006 assessment to "personal, local/national and global" to make the headings more coherent.

A corrigendum has been issued for this page.
See: http://www.oecd.org/about/publishing/Corrigendum_PISA2015Revised_Framework.pdf

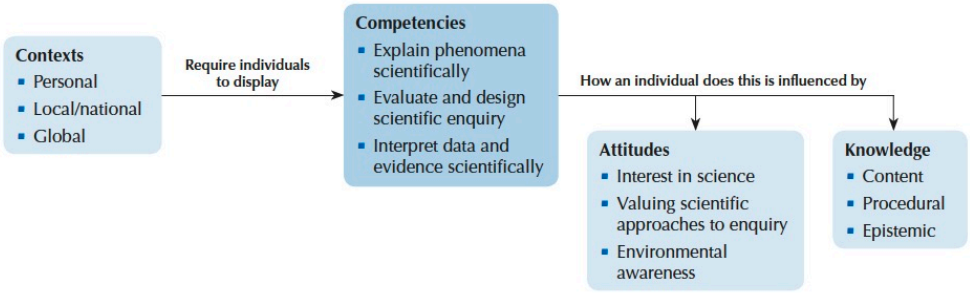
ORGANISING THE DOMAIN OF SCIENCE

The PISA 2015 definition of scientific literacy consists of four interrelated aspects (see Figures 2.1 and 2.2).

Figure 2.1 ▪ **Aspects of the scientific literacy assessment framework for PISA 2015**

Contexts	Personal, local/national and global issues, both current and historical, which demand some understanding of science and technology.
Knowledge	An understanding of the major facts, concepts and explanatory theories that form the basis of scientific knowledge. Such knowledge includes knowledge of both the natural world and technological artefacts (content knowledge), knowledge of how such ideas are produced (procedural knowledge), and an understanding of the underlying rationale for these procedures and the justification for their use (epistemic knowledge).
Competencies	The ability to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically.
Attitudes	A set of attitudes towards science indicated by an interest in science and technology, valuing scientific approaches to enquiry where appropriate, and a perception and awareness of environmental issues.

Figure 2.2 ▪ **Inter-relations between the four aspects**



Contexts of assessment items

PISA 2015 assesses scientific knowledge in contexts that are relevant to the science curricula of participating countries. Such contexts are not, however, restricted to the common aspects of participants' national curricula. Rather, the assessment requires evidence of the successful use of the three competencies required for scientific literacy in situations set in personal, local/national and global contexts.

Assessment items are not limited to school science contexts. In the PISA 2015 scientific literacy assessment, the items focus on situations relating to the self, family and peer groups (personal), to the community (local and national), and to life across the world (global). Technology-based topics may be used as a common context. Some topics may be set in historical contexts, which are used to assess students' understanding of the processes and practices involved in advancing scientific knowledge.

Figure 2.3 shows how science and technology issues are applied within personal, local/national and global settings. The contexts are chosen in light of their relevance to students' interests and lives. The areas of application are: health and disease, natural resources, environmental quality, hazards, and the frontiers of science and technology. They are the areas in which scientific literacy has particular value for individuals and communities in enhancing and sustaining quality of life, and in developing public policy.

The PISA science assessment is *not* an assessment of contexts. Rather, it assesses competencies and knowledge *in* specific contexts. These contexts are chosen on the basis of the knowledge and understanding that students are likely to have acquired by the age of 15.

Sensitivity to linguistic and cultural differences is a priority in item development and selection, not only for the sake of the validity of the assessment, but also to respect these differences among participating countries.

Figure 2.3 ▪ **Contexts in the PISA 2015 scientific literacy assessment**

	Personal	Local/National	Global
Health and disease	Maintenance of health, accidents, nutrition	Control of disease, social transmission, food choices, community health	Epidemics, spread of infectious diseases
Natural resources	Personal consumption of materials and energy	Maintenance of human populations, quality of life, security, production and distribution of food, energy supply	Renewable and non-renewable natural systems, population growth, sustainable use of species
Environmental quality	Environmentally friendly actions, use and disposal of materials and devices	Population distribution, disposal of waste, environmental impact	Biodiversity, ecological sustainability, control of pollution, production and loss of soil/biomass
Hazards	Risk assessments of lifestyle choices	Rapid changes (e.g. earthquakes, severe weather), slow and progressive changes (e.g. coastal erosion, sedimentation), risk assessment	Climate change, impact of modern communication
Frontiers of science and technology	Scientific aspects of hobbies, personal technology, music and sporting activities	New materials, devices and processes, genetic modifications, health technology, transport	Extinction of species, exploration of space, origin and structure of the universe

Scientific competencies

Figures 2.4a, 2.4b and 2.4c provide a detailed description of how students may display the three competencies required for scientific literacy. The set of scientific competencies in Figures 2.4a, 2.4b and 2.4c reflects a view that science is best seen as an ensemble of social and epistemic practices that are common across all sciences (National Research Council, 2012). Hence, all these competencies are framed as actions. They are written in this manner to convey the idea of what the scientifically literate person both understands and is capable of doing. Fluency with these practices is, in part, what distinguishes the expert scientist from the novice. While it would be unreasonable to expect a 15-year-old student to have the expertise of a scientist, a scientifically literate student can be expected to appreciate the role and significance of these practices and try to use them.

Figure 2.4a ▪ **PISA 2015 scientific competencies: Explain phenomena scientifically**

Explain phenomena scientifically

Recognise, offer and evaluate explanations for a range of natural and technological phenomena demonstrating the ability to:

- Recall and apply appropriate scientific knowledge.
- Identify, use and generate explanatory models and representations.
- Make and justify appropriate predictions.
- Offer explanatory hypotheses.
- Explain the potential implications of scientific knowledge for society.

Demonstrating the competency of *explaining phenomena scientifically* requires students to recall the appropriate content knowledge in a given situation and use it to interpret and explain the phenomenon of interest. Such knowledge can also be used to generate tentative explanatory hypotheses in contexts where there is a lack of knowledge or data. A scientifically literate person is expected to be able to draw on standard scientific models to construct simple representations to explain everyday phenomena, such as why antibiotics do not kill viruses, how a microwave oven works, or why gases are compressible but liquids are not, and use these to make predictions. This competency includes the ability to describe or interpret phenomena and predict possible changes. In addition, it may involve recognising or identifying appropriate descriptions, explanations and predictions.

Figure 2.4b ■ **PISA 2015 scientific competencies : Evaluate and design scientific enquiry**

Evaluate and design scientific enquiry
Describe and appraise scientific investigations and propose ways of addressing questions scientifically demonstrating the ability to: ▪ Identify the question explored in a given scientific study. ▪ Distinguish questions that could be investigated scientifically. ▪ Propose a way of exploring a given question scientifically. ▪ Evaluate ways of exploring a given question scientifically. ▪ Describe and evaluate how scientists ensure the reliability of data, and the objectivity and generalisability of explanations.

The competency of *evaluating and designing scientific enquiry* is required to evaluate reports of scientific findings and investigations critically. It relies on the ability to distinguish scientific questions from other forms of enquiry or recognise questions that could be investigated scientifically in a given context. This competency requires a knowledge of the key features of a scientific investigation – for example, what things should be measured, what variables should be changed or controlled, or what action should be taken so that accurate and precise data can be collected. It requires an ability to evaluate the quality of data, which, in turn, depends on recognising that data are not always completely accurate. It also requires the ability to determine whether an investigation is driven by an underlying theoretical premise or, alternatively, whether it seeks to determine patterns.

A scientifically literate person should also be able to recognise the significance of previous research when judging the value of any given scientific enquiry. Such knowledge is needed to situate the work and judge the importance of any possible outcomes. For example, knowing that the search for a malaria vaccine has been an ongoing programme of scientific research for several decades, and given the number of people who are killed by malarial infections, any findings that suggested a vaccine would be achievable would be of substantial significance.

Moreover, students need to understand the importance of developing a sceptical attitude towards all media reports in science. They need to recognise that all research builds on previous work, that the findings of any one study are always subject to uncertainty, and that the study may be biased by the sources of funding. This competency requires students to possess both procedural and epistemic knowledge but may also draw on their content knowledge of science, to varying degrees.

Figure 2.4c ■ **PISA 2015 scientific competencies: Interpret data and evidence scientifically**

Interpret data and evidence scientifically
Analyse and evaluate scientific data, claims and arguments in a variety of representations and draw appropriate conclusions, demonstrating the ability to: ▪ Transform data from one representation to another. ▪ Analyse and interpret data and draw appropriate conclusions. ▪ Identify the assumptions, evidence and reasoning in science-related texts. ▪ Distinguish between arguments that are based on scientific evidence and theory and those based on other considerations. ▪ Evaluate scientific arguments and evidence from different sources (e.g. newspapers, the Internet, journals).

A scientifically literate person should be able to interpret and make sense of basic forms of scientific data and evidence that are used to make claims and draw conclusions. Displaying this competency may require all three forms of scientific knowledge.

Those who possess this competency should be able to interpret the meaning of scientific evidence and its implications to a specified audience in their own words, using diagrams or other representations as appropriate. This competency requires the use of mathematical tools to analyse or summarise data, and the ability to use standard methods to transform data into different representations.

This competency also includes accessing scientific information and producing and evaluating arguments and conclusions based on scientific evidence (Kuhn, 2010; Osborne, 2010). It may also involve evaluating alternative conclusions using

evidence; giving reasons for or against a given conclusion using procedural or epistemic knowledge; and identifying the assumptions made in reaching a conclusion. In short, the scientifically literate individual should be able to identify logical or flawed connections between evidence and conclusions.

Table 2.1 shows the desired distribution of items, by competency, in the PISA 2015 science assessment.

Table 2.1 Desired distribution of items, by competency

Competency	Percentage of total items
Explain phenomena scientifically	40-50
Evaluate and design scientific enquiry	20-30
Interpret data and evidence scientifically	30-40

Scientific knowledge

Content knowledge

Given that only a sample of the content domain of science can be assessed in the PISA 2015 scientific literacy assessment, clear criteria are used to guide the selection of the knowledge that is assessed. The criteria are applied to knowledge from the major fields of physics, chemistry, biology, earth and space sciences, and require that the knowledge:

- has relevance to real-life situations
- represents an important scientific concept or major explanatory theory that has enduring utility
- is appropriate to the developmental level of 15-year-olds.

It is thus assumed that students have some knowledge and understanding of the major explanatory ideas and theories of science, including an understanding of the history and scale of the universe, the particle model of matter, and the theory of evolution by natural selection. These examples of major explanatory ideas are provided for illustrative purposes; there has been no attempt to list comprehensively all the ideas and theories that might be considered fundamental for a scientifically literate individual.

Figure 2.5 ▪ **Knowledge of the content of science**

Physical systems that require knowledge of:

- Structure of matter (e.g. particle model, bonds)
- Properties of matter (e.g. changes of state, thermal and electrical conductivity)
- Chemical changes of matter (e.g. chemical reactions, energy transfer, acids/bases)
- Motion and forces (e.g. velocity, friction) and action at a distance (e.g. magnetic, gravitational and electrostatic forces)
- Energy and its transformation (e.g. conservation, dissipation, chemical reactions)
- Interactions between energy and matter (e.g. light and radio waves, sound and seismic waves)

Living systems that require knowledge of:

- Cells (e.g. structures and function, DNA, plant and animal)
- The concept of an organism (e.g. unicellular and multicellular)
- Humans (e.g. health, nutrition, subsystems such as digestion, respiration, circulation, excretion, reproduction and their relationship)
- Populations (e.g. species, evolution, biodiversity, genetic variation)
- Ecosystems (e.g. food chains, matter and energy flow)
- Biosphere (e.g. ecosystem services, sustainability)

Earth and space systems that require knowledge of:

- Structures of the Earth systems (e.g. lithosphere, atmosphere, hydrosphere)
- Energy in the Earth systems (e.g. sources, global climate)
- Change in Earth systems (e.g. plate tectonics, geochemical cycles, constructive and destructive forces)
- Earth's history (e.g. fossils, origin and evolution)
- Earth in space (e.g. gravity, solar systems, galaxies)
- The history and scale of the universe and its history (e.g. light year, Big Bang theory)

Figure 2.5 shows the content knowledge categories and examples selected by applying these criteria. Such knowledge is required for understanding the natural world and for making sense of experiences in personal, local/national and global contexts. The framework uses the term "systems" instead of "sciences" in the descriptors of content knowledge. The intention is to convey the idea that citizens have to understand concepts from the physical and life sciences, and earth and space sciences, and how they apply in contexts where the elements of knowledge are interdependent or interdisciplinary. Things viewed as subsystems at one scale may be viewed as whole systems at a smaller scale. For example, the circulatory system can be seen as an entity in itself or as a subsystem of the human body; a molecule can be studied as a stable configuration of atoms but also as a subsystem of a cell or a gas. Thus, applying scientific knowledge and exhibiting scientific competencies requires a determination of which system and which boundaries apply in any particular context.

Table 2.2 shows the desired distribution of items, by content of science.

Table 2.2 Desired distribution of items, by content

System	Percentage of total items
Physical	36
Living	36
Earth and space	28
Total	100

Procedural knowledge

A fundamental goal of science is to generate explanatory accounts of the material world. Tentative explanatory accounts are first developed and then tested through empirical enquiry. Empirical enquiry relies on certain well-established concepts, such as the notion of dependent and independent variables, the control of variables, types of measurement, forms of error, methods of minimising error, common patterns observed in data, and methods of presenting data.

It is this knowledge of the concepts and procedures that are essential for scientific enquiry that underpins the collection, analysis and interpretation of scientific data. Such ideas form a body of procedural knowledge that has also been called "concepts of evidence" (Gott, Duggan and Roberts, 2008; Millar et al., 1995). One can think of procedural knowledge as knowledge of the standard procedures scientists use to obtain reliable and valid data. Such knowledge is needed both to undertake scientific enquiry and engage in critical reviews of the evidence that might be used to support particular claims. It is expected, for instance, that students will know that scientific knowledge has differing degrees of certainty associated with it, and so can explain why there is a difference between the confidence associated with measurements of the speed of light (which has been measured many times with ever more accurate instrumentation) and measurements of fish stocks in the North Atlantic or the mountain lion population in California. The examples listed in Figure 2.6 convey the general features of procedural knowledge that may be tested.

Figure 2.6 ▪ **PISA 2015 procedural knowledge**

Procedural knowledge

- The concept of variables, including dependent, independent and control variables.
- Concepts of measurement, e.g. quantitative (measurements), qualitative (observations), the use of a scale, categorical and continuous variables.
- Ways of assessing and minimising uncertainty, such as repeating and averaging measurements.
- Mechanisms to ensure the replicability (closeness of agreement between repeated measures of the same quantity) and accuracy of data (the closeness of agreement between a measured quantity and a true value of the measure).
- Common ways of abstracting and representing data using tables, graphs and charts, and using them appropriately.
- The control-of-variables strategy and its role in experimental design or the use of randomised controlled trials to avoid confounded findings and identify possible causal mechanisms.
- The nature of an appropriate design for a given scientific question, e.g. experimental, field-based or pattern-seeking.

Epistemic knowledge

Epistemic knowledge refers to an understanding of the role of specific constructs and defining features essential to the process of knowledge building in science (Duschl, 2007). Those who have such knowledge can explain, with examples, the distinction between a scientific theory and a hypothesis or a scientific fact and an observation. They know that models, whether representational, abstract or mathematical, are a key feature of science, and that such models are

like maps rather than accurate pictures of the material world. These students can recognise that any particle model of matter is an idealised representation of matter and can explain how the Bohr model is a limited model of what we know about the atom and its constituent parts. They recognise that the concept of a "theory" as used in science is not the same as the notion of a "theory" in everyday language, where it is used as a synonym for a "guess" or a "hunch". Procedural knowledge is required to explain what is meant by the control-of-variables strategy; epistemic knowledge is required to explain *why* the use of the control-of-variables strategy or the replication of measurements is central to establishing knowledge in science.

Scientifically literate individuals also understand that scientists draw on data to advance claims to knowledge, and that argument is a commonplace feature of science. In particular, they know that some arguments in science are hypothetico-deductive (e.g. Copernicus' argument for the heliocentric system), some are inductive (the conservation of energy), and some are an inference to the best explanation (Darwin's theory of evolution or Wegener's argument for moving continents). They also understand the role and significance of peer review as the mechanism that the scientific community has established for testing claims to new knowledge. As such, epistemic knowledge provides a rationale for the procedures and practices in which scientists engage, a knowledge of the structures and defining features that guide scientific enquiry, and the foundation for the basis of belief in the claims that science makes about the natural world.

Figure 2.7 represents what are considered to be the major features of epistemic knowledge necessary for scientific literacy.

Figure 2.7 ▪ **PISA 2015 epistemic knowledge**

Epistemic knowledge

The constructs and defining features of science. That is:

- The nature of scientific observations, facts, hypotheses, models and theories.
- The purpose and goals of science (to produce explanations of the natural world) as distinguished from technology (to produce an optimal solution to human need), and what constitutes a scientific or technological question and appropriate data.
- The values of science, e.g. a commitment to publication, objectivity and the elimination of bias.
- The nature of reasoning used in science, e.g. deductive, inductive, inference to the best explanation (abductive), analogical, and model-based.

The role of these constructs and features in justifying the knowledge produced by science. That is:

- How scientific claims are supported by data and reasoning in science.
- The function of different forms of empirical enquiry in establishing knowledge, their goal (to test explanatory hypotheses or identify patterns) and their design (observation, controlled experiments, correlational studies).
- How measurement error affects the degree of confidence in scientific knowledge.
- The use and role of physical, system and abstract models and their limits.
- The role of collaboration and critique, and how peer review helps to establish confidence in scientific claims.
- The role of scientific knowledge, along with other forms of knowledge, in identifying and addressing societal and technological issues.

Epistemic knowledge is most likely to be tested pragmatically in a context where a student is required to interpret and answer a question that requires some of this type of knowledge rather than assessing directly whether they understand the features detailed in Figure 2.7. For example, students may be asked to identify whether the conclusions are justified by the data, or what piece of evidence best supports the hypothesis advanced in an item and explain why.

Table 2.3 describes the desired distribution of items by type of knowledge.

Table 2.3 Desired distribution of items, by type of knowledge

Knowledge	Percentage of total items
Content	54-66
Procedural	19-31
Epistemic	10-22

The desired balance, by percentage of items, among the three knowledge components – content, procedural and epistemic – is shown in Table 2.4. These weightings are broadly consistent with the previous framework and reflect a consensus view among the experts consulted during the drafting of this framework.

Table 2.4 Desired distribution of items for knowledge

Knowledge types	Systems			
	Physical	Living	Earth and space	Total over systems
Content	20-24	20-24	14-18	54-66
Procedural	7-11	7-11	5-9	19- 31
Epistemic	4-8	4-8	2-6	10-22
Total over knowledge types	36	36	28	100

Sample items

In this section, three examples of science units are presented. The first is from PISA 2006 and is included to demonstrate the linkage between the 2006 and the 2015 frameworks. Questions from the unit are shown in the original paper-based format and also how they might be transposed and presented on screen. The second example is a new onscreen unit illustrating the 2015 scientific literacy framework. The third example illustrates an interactive, simulated scientific-enquiry environment which allows for assessing students' proficiency in science within a real world setting.

Other examples of science items are available on the PISA website (www.oecd.org/pisa/), including interactive examples (November 2016).

Science example 1: GREENHOUSE

Science example 1 is entitled GREENHOUSE and deals with the increase in the average temperature of the Earth's atmosphere. The stimulus material consists of a short text introducing the term "Greenhouse effect" and includes graphical information on the average temperature of the Earth's atmosphere and carbon dioxide emissions on Earth over time.

The area of application is Environment Quality within a global setting.

Read the texts and answer the questions that follow.

THE GREENHOUSE EFFECT: FACT OR FICTION?

Living things need energy to survive. The energy that sustains life on the Earth comes from the Sun, which radiates energy into space because it is so hot. A tiny proportion of this energy reaches the Earth.

The Earth's atmosphere acts like a protective blanket over the surface of our planet, preventing the variations in temperature that would exist in an airless world.

Most of the radiated energy coming from the Sun passes through the Earth's atmosphere. The Earth absorbs some of this energy, and some is reflected back from the Earth's surface. Part of this reflected energy is absorbed by the atmosphere.

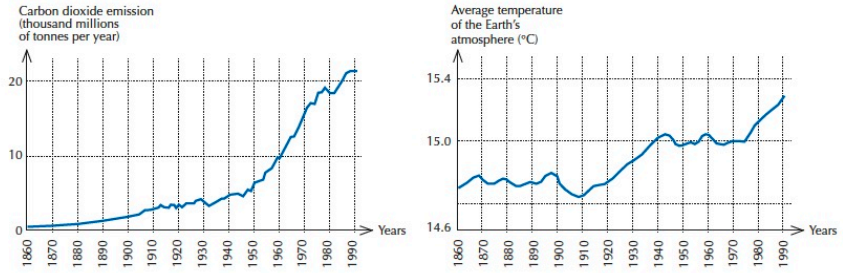
As a result of this the average temperature above the Earth's surface is higher than it would be if there were no atmosphere. The Earth's atmosphere has the same effect as a greenhouse, hence the term greenhouse effect.

The greenhouse effect is said to have become more pronounced during the twentieth century.

It is a fact that the average temperature of the Earth's atmosphere has increased. In newspapers and periodicals the increased carbon dioxide emission is often stated as the main source of the temperature rise in the twentieth century.

A student named André becomes interested in the possible relationship between the average temperature of the Earth's atmosphere and the carbon dioxide emission on the Earth.

In a library he comes across the following two graphs.



André concludes from these two graphs that it is certain that the increase in the average temperature of the Earth's atmosphere is due to the increase in the carbon dioxide emission.

GREENHOUSE – *QUESTION 1*

What is it about the graphs that supports André's conclusion?.....

Figure 2.8 ▪ **Framework categorisation for GREENHOUSE question 1**

Framework categories	2006 Framework	2015 Framework
Knowledge type	Knowledge about science	Epistemic
Competency	Explaining phenomena scientifically	Explaining phenomena scientifically
Context	Environmental, global	Environmental, global
Cognitive demand	Not applicable	Medium

Question 1 demonstrates how the 2015 framework largely maps onto the same categories as the 2006 framework, using the same competency and context categorisations. The 2006 framework included two categorisations of scientific knowledge: knowledge *of* science (referring to knowledge of the natural world across the major fields of science) and knowledge *about* science (referring to the means and goals of science). The 2015 framework elaborates on these two aspects, subdividing knowledge *about* science into procedural and epistemic knowledge. Question 1 requires students not only to understand how the data is represented in the two graphs, but also to consider whether this evidence scientifically justifies a given conclusion. This is one of the features of epistemic knowledge in the 2015 framework. The context categorisation is "environmental, global". A new feature of the 2015 framework is consideration of cognitive demand (see Figure 2.23). This question requires an interpretation of graphs involving a few linked steps; thus, according to the framework, it is categorised as medium cognitive demand.

GREENHOUSE – *QUESTION 2*

Another student, Jeanne, disagrees with André's conclusion. She compares the two graphs and says that some parts of the graphs do not support his conclusion.

Give an example of a part of the graphs that does not support André's conclusion. Explain your answer.

.....

Figure 2.9 ▪ **Framework categorisation for GREENHOUSE question 2**

Framework categories	2006 Framework	2015 Framework
Knowledge type	Knowledge about science	Epistemic
Competency	Explaining phenomena scientifically	Explaining phenomena scientifically
Context	Environmental, global	Environmental, global
Cognitive demand	Not applicable	Medium

Question 2 requires students to study the two graphs in detail. The knowledge, competency, context and cognitive demand are in the same categories as question 1.

GREENHOUSE – *QUESTION 3*

André persists in his conclusion that the average temperature rise of the Earth's atmosphere is caused by the increase in the carbon dioxide emission. But Jeanne thinks that his conclusion is premature. She says: "Before accepting this conclusion you must be sure that other factors that could influence the greenhouse effect are constant".

Name one of the factors that Jeanne means.

.....

Figure 2.10 ▪ **Framework categorisation for GREENHOUSE question 3**

Framework categories	2006 Framework	2015 Framework
Knowledge type	Knowledge about science	Procedural
Competency	Explaining phenomena scientifically	Explaining phenomena scientifically
Context	Environmental, global	Environmental, global
Cognitive demand	Not applicable	Medium

Question 3 requires students to consider control variables in terms of the critical review of evidence used to support claims. This is categorised as "procedural knowledge" in the 2015 framework.

The screenshots below illustrate how the GREENHOUSE question would be presented in an onscreen environment. The text and graphs are essentially unchanged, with students using page turners on the top right of the screen to view graphs and text as required. As the original questions were open responses, the onscreen version also necessitates an open-response format in order to replicate the paper version as closely as possible, ensuring comparability between delivery modes and therefore protecting comparability of data over time.

Figure 2.11 ▪ **GREENHOUSE presented onscreen: Stimulus page 1**

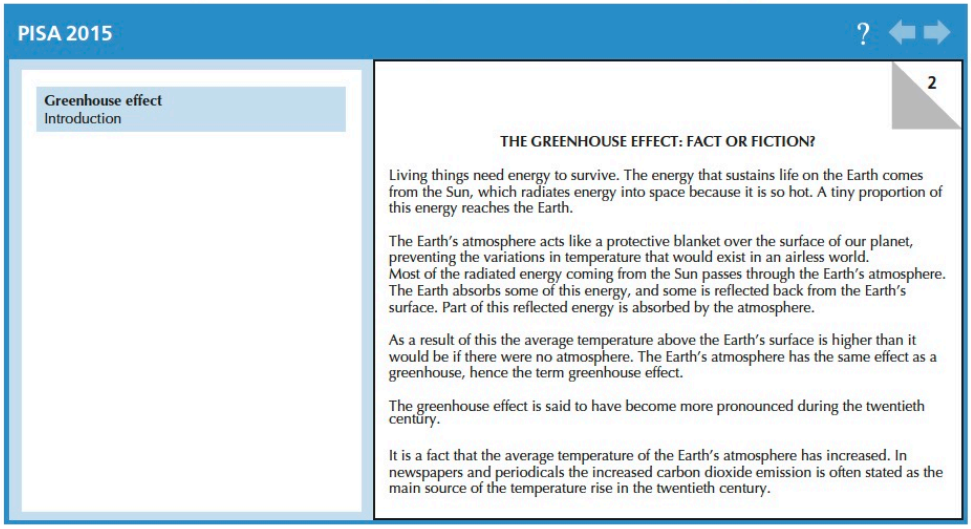


Figure 2.12 ▪ **GREENHOUSE presented onscreen: Stimulus page 2**

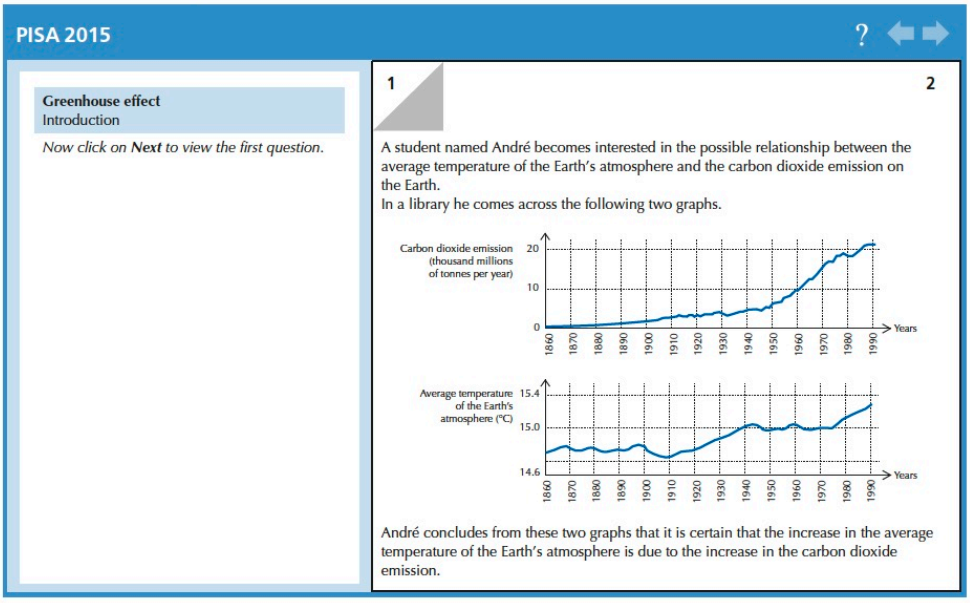


Figure 2.13 ▪ **GREENHOUSE presented onscreen: Question 1**

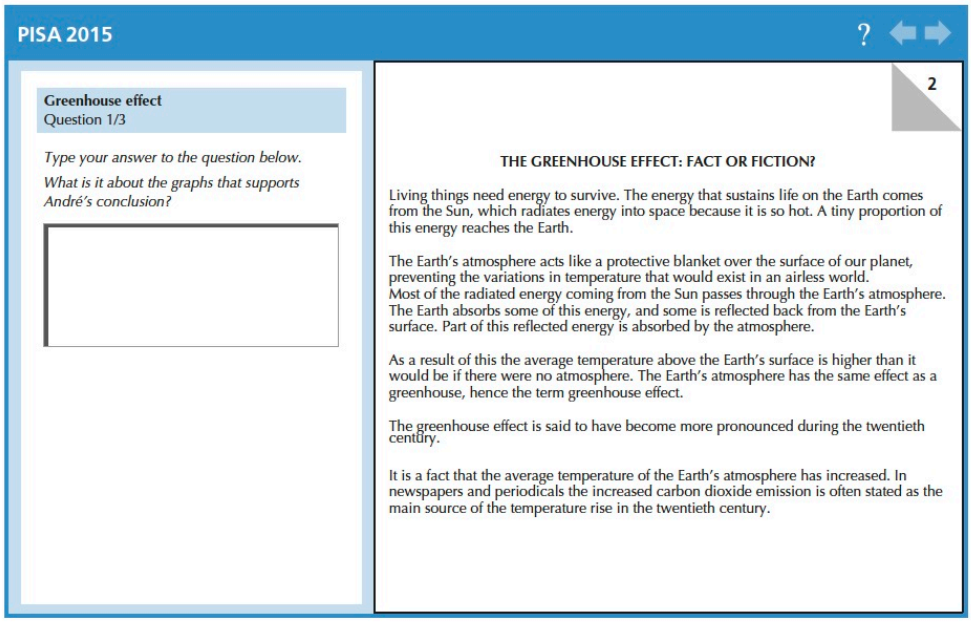


Figure 2.14 ▪ **GREENHOUSE presented onscreen: Question 2**

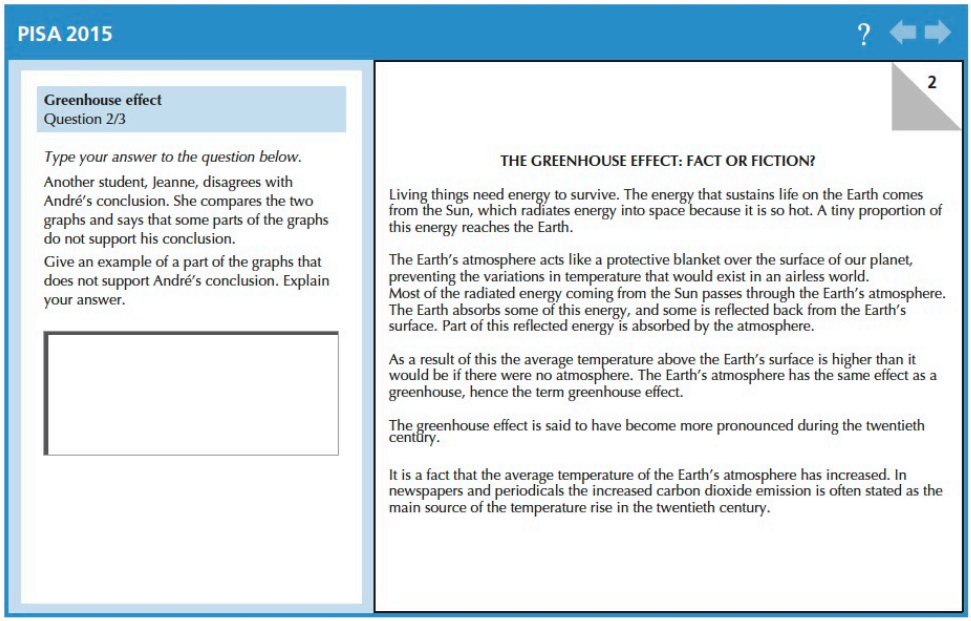
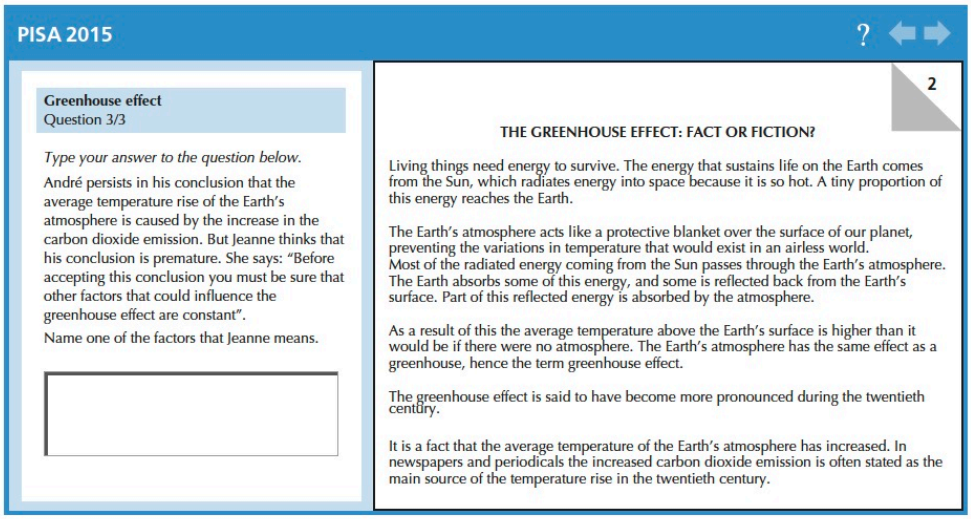


Figure 2.15 ▪ **GREENHOUSE presented onscreen: Question 3**

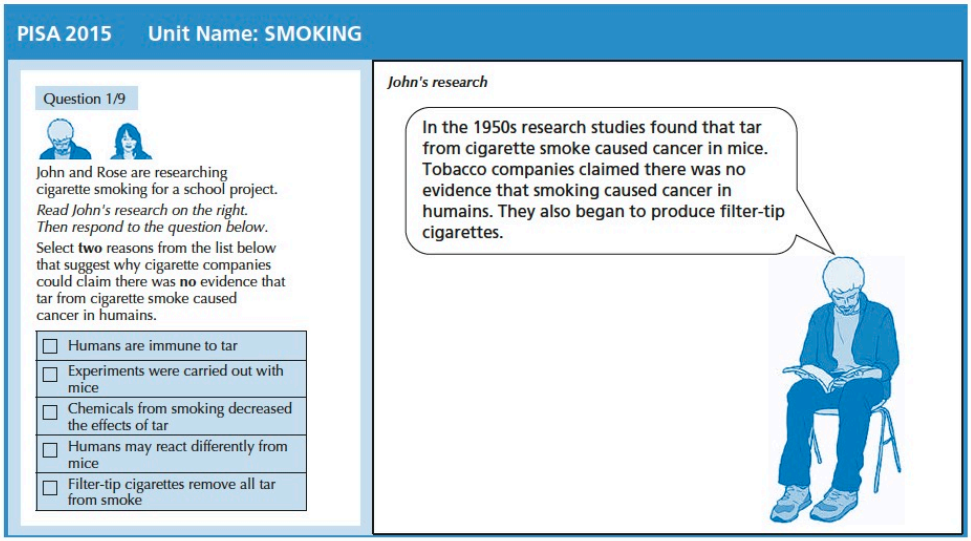


Science example 2: SMOKING

This new 2015 exemplar unit explores various forms of evidence linked to the harmful effects of smoking and the methods used to help people to stop smoking. New scientific literacy items for 2015 are only developed for computer-based delivery and therefore this exemplar is only shown in an onscreen format.

All onscreen standard question types in the PISA 2015 computer platform have a vertical split screen with the stimuli presented on the right side and the questions and answer mechanisms on the left side.

Figure 2.16 ▪ **SMOKING: Question 1**



SMOKING – *QUESTION 1*

This question requires students to interpret given evidence using their knowledge of scientific concepts. They need to read the information in the stimulus about early research into the potential harmful effects of smoking, and then select two options from the menu to answer the question.

In this question, students have to apply content knowledge using the competency of "explaining phenomena scientifically". The context is categorised as "health and disease" in a local/national setting. The cognitive demand requires the use and application of conceptual knowledge and is therefore categorised as a medium level of demand.

Figure 2.17 ▪ **Framework categorisation for SMOKING question 1**

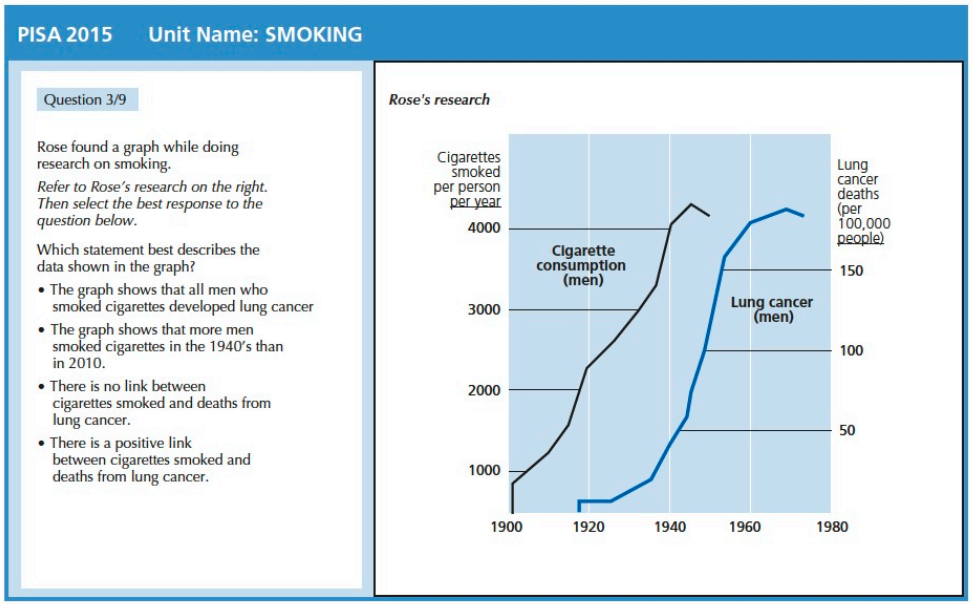
Framework categories	2015 Framework
Knowledge type	Content
Competency	Explaining phenomena scientifically
Context	Health and disease, local and national
Cognitive demand	Medium

SMOKING – *QUESTION 2*

This question explores students' understanding of data.

The right side of the screen shows authentic data of cigarette consumption and deaths from lung cancer in men over an extended period of time. Students are asked to select the best descriptor of the data by clicking on one of the radio buttons next to answer statements on the left side of the screen.

Figure 2.18 ▪ **SMOKING: Question 2**



This unit tests content knowledge using the competency of "interpreting data and evidence scientifically".

The context is "health and disease" applied to a local/national setting. As students need to interpret the relationship between two graphs, the cognitive demand is categorised as medium.

Figure 2.19 ▪ **Framework categorisation for SMOKING question 2**

Framework categories	2015 Framework
Knowledge type	Content
Competency	Interpret data and evidence scientifically
Context	Health and disease, local and national
Cognitive demand	Medium

Science example 3: ZEER POT

This new 2015 exemplar unit features the use of interactive tasks using simulations of scientific enquiry to explore and assess scientific literacy knowledge and competencies.

This unit focuses on an authentic low-cost cooling container called a Zeer pot, developed for local use in Africa, using readily available local resources. Cost and lack of electricity limit the use of refrigerators in these regions, even though the hot climate requires that people keep food cool so that it can be kept for a longer time before bacterial growth renders it a risk to health.

The first screen shot of this simulation introduces what a Zeer pot looks like and how it works. Students are not expected to have an understanding of how the process of evaporation causes cooling, just that it does.

Using this simulation, students are asked to investigate the conditions that will produce the most effective cooling effects (4°C) for keeping food fresh in the Zeer pot. The simulator keeps certain conditions constant (the air temperature and the humidity), but includes this information to enhance the authentic contextual setting. In the first question, students are asked to investigate the optimum conditions to keep the maximum amount of food fresh in the Zeer pot by altering the thickness of the sand layer and the moisture conditions.

Figure 2.20 ▪ **ZEER POT: Stimulus**

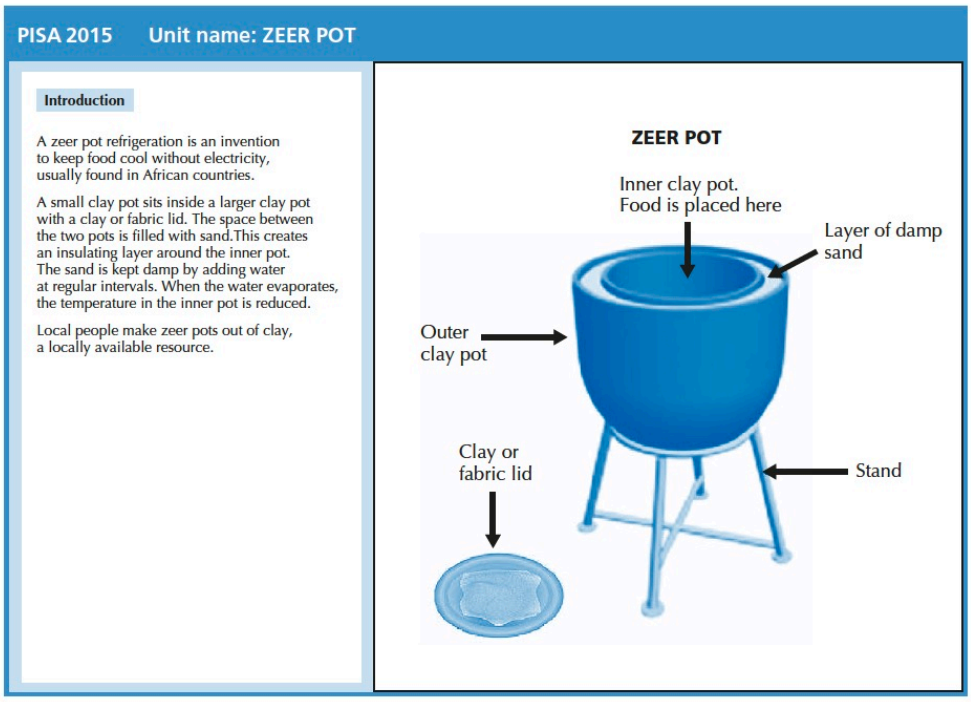
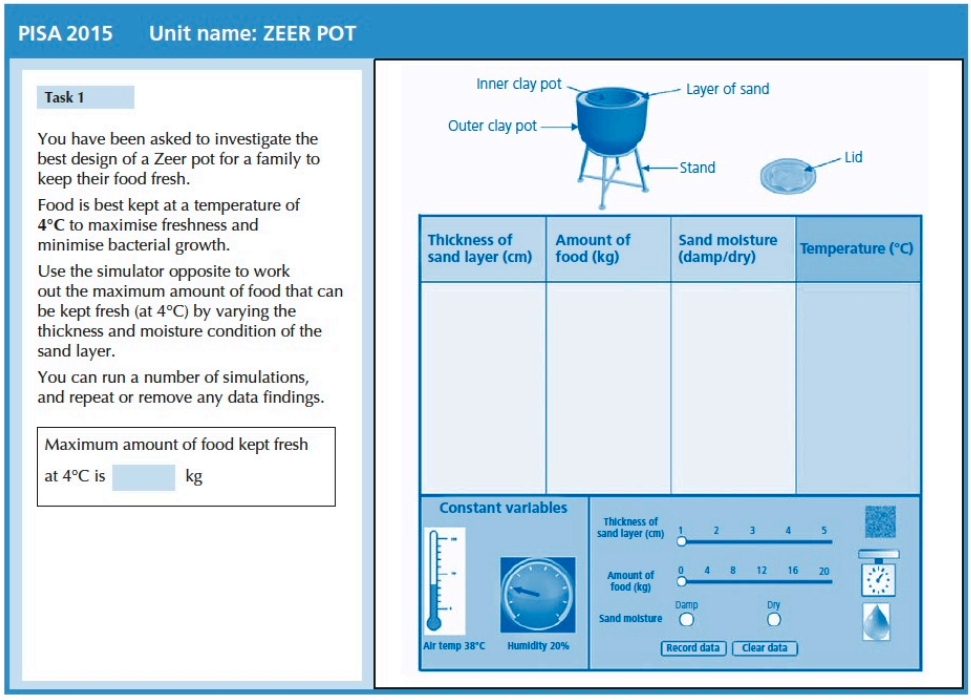


Figure 2.21 ▪ **ZEER POT: Question 1**



When students have set their conditions (which also alter the visual display of the on screen Zeer pot), they press the record-data button, which then runs the simulation and populates the data chart. They need to run a number of data simulations, and can remove data or repeat any simulations as required. This screen then records their response to the maximum amount of food kept fresh at 4°C. Their approaches to the design and evaluation of this form of scientific enquiry can be assessed in subsequent questions.

The knowledge categorisation for this item is "procedural", and the competence is "evaluate and design scientific enquiry". The context categorisation is "natural resources", although it also has links to "health and disease". The cognitive demand of this question is categorised as high because students are given a complex situation, and they need to develop a systematic sequence of investigations to answer the question.

Figure 2.22 ▪ **Framework categorisation for ZEERPOT question 1**

Framework categories	2015 Framework
Knowledge type	Procedural
Competency	Evaluate and design scientific enquiry
Context	Natural resources
Cognitive demand	High

Attitudes

Why attitudes matter

Peoples' attitudes towards science play a significant role in their interest, attention and response to science and technology, and to issues that affect them specifically. One goal of science education is to develop attitudes that lead students to engage with scientific issues. Such attitudes also support the subsequent acquisition and application of scientific and technological knowledge for personal, local/national and global benefit, and lead to the development of self-efficacy (Bandura, 1997).

Attitudes form part of the construct of scientific literacy. That is, a person's scientific literacy includes certain attitudes, beliefs, motivational orientations, self-efficacy and values. The construct of attitudes used in PISA draws upon Klopfer's (1976) structure for the affective domain in science education and reviews of attitudinal research (Gardner, 1975; Osborne, Simon and Collins, 2003; Schibeci, 1984). A major distinction made in these reviews is between attitudes towards science and scientific attitudes. While the former is measured by the level of interest displayed in scientific issues and activities, the latter is a measure of a disposition to value empirical evidence as the basis of belief.

Defining attitudes towards science in PISA 2015

The PISA 2015 assessment evaluates students' attitudes towards science in three areas: interest in science and technology, environmental awareness, and valuing scientific approaches to enquiry (see Figure 2.23), which are considered core to the construct of scientific literacy. These three areas were selected for measurement because a positive attitude towards science, a concern for the environment and an environmentally sustainable way of life, and a disposition to value the scientific approach to enquiry are characteristics of a scientifically literate individual. Thus, the extent to which individual students are, or are not interested in science and recognise its value and implications is considered an important measure of the outcome of compulsory education. Moreover, in 52 of the countries (including all OECD countries) that participated in PISA 2006, students with a higher general interest in science performed better in science (OECD, 2007:143).

Interest in science and technology was selected because of its established relationships with achievement, course selection, career choice and lifelong learning. For instance, there is a considerable body of literature which shows that interest in science is established by age 14 for the majority of students (Ormerod and Duckworth, 1975; Tai et al., 2006). Moreover, students with such an interest are more likely to pursue scientific careers. Policy concerns in many OECD countries about the number of students, particularly girls, who choose to pursue the study of science make the measurement of attitudes towards science an important aspect of the PISA assessment. The results may provide information about a perceived declining interest in the study of science among young people (Bøe et al., 2011). This measure, when correlated with the large body of other information collected by PISA through the student, teacher and school questionnaires, may provide insights into the causes of any decline in interest.

Valuing scientific approaches to enquiry was chosen because scientific approaches to enquiry have been highly successful at generating new knowledge – not only within science itself, but also in the social sciences, and even finance and sports. Moreover, the core value of scientific enquiry and the Enlightenment is the belief in empirical evidence as the basis of rational belief. Recognising the *value of the scientific approach to enquiry* is, therefore, widely regarded as a fundamental objective of science education that warrants assessing.

Appreciation of, and support for, scientific enquiry implies that students can identify and also value scientific ways of gathering evidence, thinking creatively, reasoning rationally, responding critically and communicating conclusions as they confront life situations related to science and technology. Students should understand how scientific approaches to enquiry function, and why they have been more successful than other methods in most cases. Valuing scientific approaches to enquiry, however, does not mean that an individual has to be positively disposed towards all aspects of science or even use such methods themselves. Thus, the construct is a measure of students' attitudes towards the use of a scientific method to investigate material and social phenomenon and the insights that are derived from such methods.

Environmental awareness is of international concern, as well as being of economic relevance. Attitudes in this area have been the subject of extensive research since the 1970s (see, for example, Bogner and Wiseman, 1999; Eagles and Demare, 1999; Rickinson, 2001; Weaver, 2002). In December 2002, the United Nations approved resolution 57/254 declaring the ten-year period beginning on 1 January 2005 to be the United Nations Decade of Education for Sustainable Development (UNESCO, 2003). The International Implementation Scheme (UNESCO, 2005) identifies the environment as one of the three spheres of sustainability (along with society, including culture, and economy) that should be included in all education programmes for sustainable development.

Given the importance of environmental issues to the continuation of life on Earth and the survival of humanity, young people today need to understand the basic principles of ecology and the need to organise their lives accordingly. This means that developing environmental awareness and a responsible disposition towards the environment is an important element of contemporary science education.

In PISA 2015 these specific attitudes towards science are measured through the student questionnaire. Further detail of these constructs can be found in the Questionnaire framework, Chapter 5.

ASSESSING SCIENTIFIC LITERACY

Cognitive demand

A key new feature of the PISA 2015 framework is the definition of levels of cognitive demand within the assessment of scientific literacy and across all three competencies of the framework. In assessment frameworks, item difficulty, which is empirically derived, is often confused with cognitive demand. Empirical item difficulty is estimated from the proportion of test-takers who solve the item correctly, and thus assesses the amount of knowledge held by the test-taker population, whereas cognitive demand refers to the type of mental processes required (Davis and Buckendahl, 2011). Care needs to be taken to ensure that the depth of knowledge required, i.e. the cognitive demand test items, is understood explicitly by the item developers and users of the PISA framework. For instance, an item can have high difficulty because the knowledge it is testing is not well known, but the cognitive demand is simply recall. Conversely, an item can be cognitively demanding because it requires the individual to relate and evaluate many items of knowledge – each of which is easily recalled. Thus, not only should the PISA test instrument discriminate in terms of performance between easier and harder test items, the test also needs to provide information on how students across the ability range can deal with problems at different levels of cognitive demand (Brookhart and Nitko, 2011).

The competencies are articulated using a range of terms defining cognitive demand through the use of verbs such as "recognise", "interpret", "analyse" and "evaluate". However, in themselves these verbs do not necessarily indicate a hierarchical order of difficulty that is dependent on the level of knowledge required to answer any item. Various classifications of cognitive demand schemes have been developed and evaluated since Bloom's Taxonomy was first published (Bloom, 1956). These have been largely based on categorisations of knowledge types and associated cognitive processes that are used to describe educational objectives or assessment tasks.

Bloom's revised Taxonomy (Anderson and Krathwohl, 2001) identifies four categories of knowledge – factual, conceptual, procedural and meta-cognitive. This categorisation considers these forms of knowledge to be hierarchical and distinct from the six categories of performance used in Bloom's first taxonomy – remembering, understanding, applying, analysing, evaluating and creating. In Anderson and Krathwohl's framework, these two dimensions are now seen to be independent of each other, allowing for lower levels of knowledge to be crossed with higher order skills, and vice versa.

A similar framework is offered by Marzano and Kendall's Taxonomy (2007), which also provides a two-dimensional framework based on the relationship between how mental processes are ordered and the type of knowledge required. The use of mental processes is seen as a consequence of a need to engage with a task with meta-cognitive strategies that define potential approaches to solving problems. The cognitive system then uses either retrieval, comprehension, analysis or knowledge utilisation. Marzano and Kendall divide the knowledge domain into three types of knowledge, information, mental procedures and psychomotor, compared to the four categories in Bloom's revised Taxonomy. Marzano and Kendall argue that their taxonomy is an improvement upon Bloom's Taxonomy because it offers a model of how humans actually think rather than simply an organising framework.

A different approach is offered by Ford and Wargo (2012), who offer a framework for scaffolding dialogue as a way of considering cognitive demand. Their framework uses four levels that build on each other: recall, explain, juxtapose and evaluate. Although this framework has not been specifically designed for assessment purposes, it has many similarities to the PISA 2015 definition of scientific literacy and the need to make more explicit references to such demands in the knowledge and competencies.

Another schema can be found in the framework based on Depth of Knowledge developed by Webb (1997) specifically to address the disparity between assessments and the expectations of student learning. For Webb, levels of depth can be determined by taking into account the complexity of both the content and the task required. His schema consists of four major categories: level 1 (recall), level 2 (using skills and/or conceptual knowledge), level 3 (strategic thinking) and level 4 (extended thinking). Each category is populated with a large number of verbs that can be used to describe cognitive processes. Some of these appear at more than one level. This framework offers a more holistic view of learning and assessment tasks, and requires an analysis of both the content and cognitive process demanded by any task. Webb's Depth of Knowledge (DOK) approach is a simpler but more operational version of the SOLO Taxonomy (Biggs and Collis, 1982) which describes a continuum of student understanding through five distinct stages of pre-structural, unistructural, multistructural, relational and extended abstract understanding.

All the frameworks described briefly above have served to develop the knowledge and competencies in the PISA 2015 Framework. In drawing up such a framework, it is recognised that there are challenges in developing test items based on a cognitive hierarchy. The three main challenges are that:

a) Too much effort is made to fit test items into particular cognitive frameworks, which can lead to poorly developed items.

b) Intended items (with frameworks defining rigorous, cognitively demanding goals) may differ from actual items (which may operationalise the standard in a much less cognitively demanding way).

c) Without a well-defined and understood cognitive framework, item writing and development often focuses on item difficulty and uses a limited range of cognitive processes and knowledge types, which are then only described and interpreted post hoc, rather than building from a theory of increasing competency.

The approach taken in this framework is to use an adapted version of Webb's Depth of Knowledge grid (Webb, 1997) alongside the desired knowledge and competencies. As the competencies are the central feature of the framework, the cognitive framework needs to assess and report on them across the student ability range. Webb's Depth of Knowledge Levels offer a taxonomy for cognitive demand that requires items to identify both the cognitive demand from the verbal cues that are used, e.g. analyse, arrange, compare, and the expectations of the depth of knowledge required.

Figure 2.23 ▪ **PISA 2015 Framework for Cognitive Demand**

		Competencies			Depth of Knowledge		
		Explain phenomena scientifically	Evaluate and design scientific enquiry	Interpret data and evidence scientifically	Low	Medium	High
Knowledge	Content knowledge						
	Procedural knowledge						
	Epistemic knowledge						

The grid in Figure 2.23 provides a framework for mapping items against the two dimensions of knowledge and competencies. In addition, each item can also be mapped using a third dimension based on a depth-of-knowledge taxonomy. This provides a means of operationalising cognitive demand as each item can be categorised as making demands that are:

- **Low**
  Carry out a one-step procedure, for example recall a fact, term, principle or concept, or locate a single point of information from a graph or table.
- **Medium**
  Use and apply conceptual knowledge to describe or explain phenomena, select appropriate procedures involving two or more steps, organise/display data, interpret or use simple data sets or graphs.
- **High**
  Analyse complex information or data; synthesise or evaluate evidence; justify; reason, given various sources; develop a plan or sequence of steps to approach a problem.

The distribution of items by depth of knowledge is described in Table 2.5.

Table 2.5 Distribution of items by depth of knowledge

Depth of knowledge	Percentage of items
Low	8
Medium	30
High	61
Total	**100**

Items that merely require recall of one piece of information make low cognitive demands, even if the knowledge itself might be quite complex. In contrast, items that require recall of more than one piece of knowledge, and require a comparison and evaluation of the competing merits of their relevance would be seen as having high cognitive demand. The difficulty of any item, therefore, is a combination both of the degree of complexity and range of knowledge it requires, and the cognitive operations that are required to process the item.

Therefore, the factors that determine the demand of items assessing science achievement include:

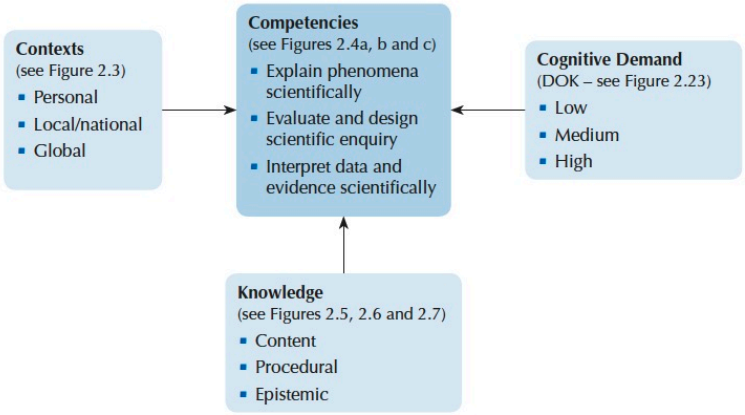
- The number and degree of complexity of elements of knowledge demanded by the item.
- The level of familiarity and prior knowledge that students may have of the content, procedural and epistemic knowledge involved.
- The cognitive operation required by the item, e.g. recall, analysis, evaluation.
- The extent to which forming a response is dependent on models or abstract scientific ideas.

This four-factor approach allows for a broader measure of scientific literacy across a wider range of student ability. Categorising the cognitive processes required for the competencies that form the basis of scientific literacy together with a consideration of the depth of knowledge required offers a model for assessing the level of demand of individual items. In addition, the relative simplicity of the approach offers a way to minimise the problems encountered in applying such frameworks. The use of this cognitive framework also facilitates the development of an a priori definition of the descriptive parameters of the reporting proficiency scales (see Figure 2.25).

Test characteristics

Figure 2.24 is a variation of Figure 2.2 that presents the basic components of the PISA framework for the 2015 scientific literacy assessment in a way that can be used to relate the framework with the structure and the content of assessment units. This may be used as a tool both to plan assessment exercises and to study the results of standard assessment exercises. As a starting point to construct assessment units, it shows the need to consider the contexts that will serve as stimulus material, the competencies required to respond to the questions or issues, the knowledge central to the exercise, and the cognitive demand.

Figure 2.24 ▪ **A tool for constructing and analysing assessment units and items**



A test unit is defined by specific stimulus material, which may be a brief written passage, or text accompanying a table, chart, graph or diagram. In units created for PISA 2015, the stimulus material may also include non-static stimulus material, such as animations and interactive simulations. The items are a set of independently scored questions of various types, as illustrated by the examples already discussed. Further examples can be found at the PISA website (www.oecd.org/pisa/) (November 2016).

PISA uses this unit structure to facilitate the use of contexts that are as realistic as possible, reflecting the complexity of real-life situations, while making efficient use of testing time. Using situations about which several questions can be posed, rather than asking separate questions about a larger number of different situations, reduces the overall time required for a student to become familiar with the material in each question. However, the need to make each score point independent of others within a unit needs to be taken into account. It is also necessary to recognise that, because this approach reduces the number of different assessment contexts, it is important to ensure that there is an adequate range of contexts so that bias due to the choice of contexts is minimised.

PISA 2015 test units require the use of all three scientific competencies and draw on all three forms of science knowledge. In most cases, each test unit assesses multiple competencies and knowledge categories. Individual items, however, assess only one form of knowledge and one competency.

The need for students to read texts in order to understand and answer written questions on scientific literacy raises an issue of the level of reading literacy that are required. Stimulus material and questions use language that is as clear, simple and brief, and as syntactically simplified as possible while still conveying the appropriate meaning. The number of concepts introduced per paragraph is limited. Questions within the domain of science that assess reading or mathematical literacy are avoided.

Response formats

Three classes of items are used to assess the competencies and scientific knowledge identified in the framework. About one-third of the items are in each of the three classes:

- simple multiple choice: items calling for
  - selection of a single response from four options
  - selection of a "hot spot", an answer that is a selectable element within a graphic or text
- complex multiple choice: items calling for
  - responses to a series of related "Yes/No" questions that are treated for scoring as a single item (the typical format in 2006)
  - selection of more than one response from a list
  - completion of a sentence by selecting drop-down choices to fill multiple blanks
  - "drag-and-drop" responses, allowing students to move elements on screen to complete a task of matching, ordering or categorising
- constructed response: items calling for written or drawn responses
  - Constructed-response items in scientific literacy typically call for a written response ranging from a phrase to a short paragraph (e.g. two to four sentences of explanation). A small number of constructed-response items call for drawing (e.g. a graph or diagram). In a computer-based assessment, any such item is supported by simple drawing editors that are specific to the response required.

In 2015, some responses are captured by interactive tasks, for example, a student's choices for manipulating variables in a simulated scientific enquiry. Responses to these interactive tasks are likely scored as complex multiple-choice items. Some kinds of responses to interactive tasks may be sufficiently open-ended that they are treated as constructed response.

Assessment structure

Computer is the primary mode of delivery for all domains, including scientific literacy, in PISA 2015. All new science literacy items are only available on computer. However a paper-based assessment instrument, consisting only of the trend items, is provided for countries choosing not to test their students by computer. (The PISA 2015 field trial studied the effect on student performance of the change in mode of delivery. For further details see Box 1.2.)

Scientific literacy items are organised into 30-minute sections called "clusters". Each cluster includes either only new units or only trend units. Overall for 2015, the target number of clusters included in the main survey is:

- six clusters of trend units in 2015 main survey
- six clusters of new units in 2015 main survey.

Each student is assigned one two-hour test form. A test form is composed of four clusters, with each cluster designed to occupy thirty minutes of testing time. The clusters are placed in multiple computer-based test forms, according to a rotated test design.

Each student spends one hour on scientific literacy, with the remaining time assigned to either one or two of the additional domains of reading, mathematics and collaborative problem solving. For any countries taking the paper-based assessment instrument, intact clusters of 2006 units are formed into a number of test booklets. The paper-based assessment is limited to trend items and does not include any newly developed material. In contrast, the computer-based instrument includes newly developed items as well as trend items. When transposing paper-based trend items to an onscreen format, the presentation, response format and cognitive demand remain comparable.

Item contexts are spread across personal, local/national and global settings roughly in the ratio 1:2:1, as was the case in 2006. A wide selection of areas of application are used for units, subject to satisfying as far as possible the various constraints imposed by the distribution of items shown in Tables 2.1 and 2.4.

Reporting proficiency in science

To achieve the aims of PISA, scales must be developed to measure student proficiency. A descriptive scale of levels of competence needs to be based on a theory of how the competence develops, not just on a post-hoc interpretation of what items of increasing difficulty seem to be measuring. The 2015 draft framework therefore defined explicitly the parameters of increasing competence and progression, allowing item developers to design items representing this growth in ability (Kane, 2006; Mislevy and Haertel, 2006). Initial draft descriptions of the scales are offered below, though it is recognised that these may need to be updated after the main survey. Although comparability with the 2006 scale descriptors (OECD, 2007) has been maximised in order to enable trend analyses, the new elements of the 2015 framework, such as depth of knowledge, have also been incorporated. The scales have also been extended by the addition of a level "1b" to specifically address and provide a description of students at the lowest level of ability who demonstrate minimal scientific literacy and would previously not have been included in the reporting scales. The initial draft scales for 2015 Framework therefore propose more detailed and more specific descriptors of the levels of scientific literacy, and not an entirely different model as shown in Figure 2.25.

Figure 2.25 ▪ **Initial draft of proficiency scale descriptions for science**

Level	Descriptor
6	At Level 6, students are able to use content, procedural and epistemic knowledge to consistently provide explanations, evaluate and design scientific enquiries, and interpret data in a variety of complex life situations that require a high level of cognitive demand. They can draw appropriate inferences from a range of different complex data sources, in a variety of contexts and provide explanations of multi-step causal relationships. They can consistently distinguish scientific and non-scientific questions, explain the purposes of enquiry, and control relevant variables in a given scientific enquiry or any experimental design of their own. They can transform data representations, interpret complex data and demonstrate an ability to make appropriate judgments about the reliability and accuracy of any scientific claims. Level 6 students consistently demonstrate advanced scientific thinking and reasoning requiring the use of models and abstract ideas and use such reasoning in unfamiliar and complex situations. They can develop arguments to critique and evaluate explanations, models, interpretations of data and proposed experimental designs in a range of personal, local and global contexts.
5	At Level 5, students are able to use content, procedural and epistemic knowledge to provide explanations, evaluate and design scientific enquiries and interpret data in a variety of life situations in some but not all cases of high cognitive demand. They draw inferences from complex data sources, in a variety of contexts and can explain some multi-step causal relationships. Generally, they can distinguish scientific and non-scientific questions, explain the purposes of enquiry, and control relevant variables in a given scientific enquiry or any experimental design of their own. They can transform some data representations, interpret complex data and demonstrate an ability to make appropriate judgments about the reliability and accuracy of any scientific claims. Level 5 students show evidence of advanced scientific thinking and reasoning requiring the use of models and abstract ideas and use such reasoning in unfamiliar and complex situations. They can develop arguments to critique and evaluate explanations, models, interpretations of data and proposed experimental designs in some but not all personal, local and global contexts.

...

Figure 2.25 [continued] ▪ **Initial draft of proficiency scale descriptions for science**

4	At Level 4, students are able to use content, procedural and epistemic knowledge to provide explanations, evaluate and design scientific enquiries and interpret data in a variety of given life situations that require mostly a medium level of cognitive demand. They can draw inferences from different data sources, in a variety of contexts and can explain causal relationships. They can distinguish scientific and non-scientific questions, and control variables in some but not all scientific enquiry or in an experimental design of their own. They can transform and interpret data and have some understanding about the confidence held about any scientific claims. Level 4 students show evidence of linked scientific thinking and reasoning and can apply this to unfamiliar situations. Students can also develop simple arguments to question and critically analyse explanations, models, interpretations of data and proposed experimental designs in some personal, local and global contexts.
3	At Level 3, students are able to use content, procedural and epistemic knowledge to provide explanations, evaluate and design scientific enquiries and interpret data in some given life situations that require at most a medium level of cognitive demand. They are able to draw a few inferences from different data sources, in a variety of contexts, and can describe and partially explain simple causal relationships. They can distinguish some scientific and non-scientific questions, and control some variables in a given scientific enquiry or in an experimental design of their own. They can transform and interpret simple data and are able to comment on the confidence of scientific claims. Level 3 students show some evidence of linked scientific thinking and reasoning, usually applied to familiar situations. Students can develop partial arguments to question and critically analyse explanations, models, interpretations of data and proposed experimental designs in some personal, local and global contexts.
2	At Level 2, students are able to use content, procedural and epistemic knowledge to provide explanations, evaluate and design scientific enquiries and interpret data in some given familiar life situations that require mostly a low level of cognitive demand. They are able to make a few inferences from different sources of data, in few contexts, and can describe simple causal relationships. They can distinguish some simple scientific and non-scientific questions, and distinguish between independent and dependent variables in a given scientific enquiry or in a simple experimental design of their own. They can transform and describe simple data, identify straightforward errors, and make some valid comments on the trustworthiness of scientific claims. Students can develop partial arguments to question and comment on the merits of competing explanations, interpretations of data and proposed experimental designs in some personal, local and global contexts.
1a	At Level 1a, students are able to use a little content, procedural and epistemic knowledge to provide explanations, evaluate and design scientific enquiries and interpret data in a few familiar life situations that require a low level of cognitive demand. They are able to use a few simple sources of data, in a few contexts and can describe some very simple causal relationships. They can distinguish some simple scientific and non-scientific questions, and identify the independent variable in a given scientific enquiry or in a simple experimental design of their own. They can partially transform and describe simple data and apply them directly to a few familiar situations. Students can comment on the merits of competing explanations, interpretations of data and proposed experimental designs in some very familiar personal, local and global contexts.
1b	At Level 1b, students demonstrate a little evidence to use content, procedural and epistemic knowledge to provide explanations, evaluate and design scientific enquiries and interpret data in a few familiar life situations that require a low level of cognitive demand. They are able to identify straightforward patterns in simple sources of data in a few familiar contexts and can offer attempts at describing simple causal relationships. They can identify the independent variable in a given scientific enquiry or in a simple design of their own. They attempt to transform and describe simple data and apply them directly to a few familiar situations.

The proposed level descriptors are based on the 2015 Framework described in this document and offer a qualitative description of the differences between levels of performance. The factors used to determine the demand of items assessing science achievement that have been incorporated into this outline of the proficiency scales include:

- the number and degree of complexity of elements of knowledge demanded by the item
- the level of familiarity and prior knowledge that students may have of the content, procedural and epistemic knowledge involved
- the cognitive operation required by the item, e.g. recall, analysis, evaluation
- the extent to which forming a response is dependent on models or abstract scientific ideas.

References

American Association for the Advancement of Science (1989), *Science for all Americans: a Project 2061 Report on Literacy Goals in Science, Mathematics and Technology*, AAS Publishing, Washington, DC, www.project2061.org/publications/sfaa/online/sfaatoc.htm.

Anderson, L.W. and **D.R. Krathwohl** (2001), *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Longman Publishing, London.

Bandura, A. (1997), *Self-Efficacy: The Exercise of Control,* W.H. Freeman and Company, Macmillan Publishers, New York.

Biggs, J. and **K. Collis** (1982), *Evaluating the Quality of Learning: The SOLO Taxonomy*, Academic Press, New York.

Bloom, B.S. (eds.) (1956), *Taxonomy of Educational Objectives Book 1: Cognitive Domain*, Longmans Publishing, London.

Bøe, M.V. et al. (2011), "Participation in science and technology: Young people's achievement-related choices in late-modern societies", *Studies in Science Education*, Vol. 47/1, pp. 37-72, http://dx.doi.org/10.1080/03057267.2011.549621.

Bogner, F. and **M. Wiseman** (1999), "Toward measuring adolescent environmental perception", *European Psychologist*, Vol. 4/3, http://dx.doi.org/10.1027//1016-9040.4.3.139.

Brookhart, S.M. and **A.J. Nitko** (2011), "Strategies for constructing assessments of higher order thinking skills", in G. Schraw and D.R. Robinson (eds.), *Assessment of Higher Order Thinking Skills*, IAP, Charlotte, NC, pp. 327-359.

Confederacion de Sociedades Cientificas de España (2011), *Informe ENCIENDE, Enseñanza de las Ciencias en la Didáctica Escolar para edades tempranas en España*, Madrid.

Davis, S.L. and **C.W. Buckendahl** (2011), "Incorporating cognitive demand in credentialing examinations", in G. Schraw and D.R. Robinson (eds), *Assessment of Higher Order Thinking Skills*, IAP, Charlotte, NC, pp. 327-359.

Drechsel, B., C. Carstensen and **M. Prenzel** (2011), "The role of content and context in PISA interest scales: A study of the embedded interest items in the PISA 2006 science assessment", *International Journal of Science Education*, Vol. 33/1, pp. 73-95.

Duschl, R. (2007), "Science education in three-part harmony: Balancing conceptual, epistemic and social learning goals", *Review of Research in Education*, Vol. 32, pp. 268-291, http://dx.doi.org/10.3102/0091732X07309371.

Eagles, P.F.J. and **R. Demare** (1999), "Factors influencing children's environmental attitudes", *The Journal of Environmental Education*, Vol. 30/4, www.researchgate.net/profile/Paul_Eagles/publication/271994465_Factors_Influencing_Children's_Environmental_Attitudes/links/553e677b0cf20184050f83a6.pdf.

European Commission (1995), "Teaching and learning: Towards the learning society", *White Paper on Education and Training*, Office for Official Publications in European Countries, Luxembourg, http://europa.eu/documents/comm/white_papers/pdf/com95_590_en.pdf.

Fensham, P. (1985), "Science for all: A reflective essay", *Journal of Curriculum Studies*, Vol. 17/4, pp. 415-435, http://dx.doi.org/10.1080/0022027850170407.

Ford, M.J. and **B.M. Wargo** (2012), "Dialogic framing of scientific content for conceptual and epistemic understanding", *Science Education*, Vol. 96/3, pp. 369-391, http://dx.doi.org/10.1002/sce.20482.

Gardner, P.L. (1975), "Attitudes to Science", *Studies in Science Education*, Vol. 2, pp. 1-41.

Gott, R., S. Duggan and **R. Roberts** (2008)," Concepts of evidence", University of Durham, www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm, (accessed 23 September 2012).

Kane, M. (2006), "Validation", in R.L. Brennan (eds.), *Educational Measurement*, 4th ed., Praeger Publishers and the American Council on Education, Westport, CT, pp. 17-64.

Klopfer, L.E. (1971), "Evaluation of learning in science" in B.S. Bloom, J.T. Hastings and G.F. Madaus (eds.), *Handbook of Formative and Summative Evaluation of Student Learning*, McGraw-Hill Book Company, London.

Klopfer, L.E. (1976), "A structure for the affective domain in relation to science education", *Science Education*, Vol. 60/3, pp. 299-312, http://dx.doi.org/ 10.1002/sce.3730600304.

Kuhn, D. (2010), "Teaching and learning science as argument", *Science Education*, Vol. 94/5, pp. 810-824, http://dx.doi.org/10.1002/sce.20395.

Lederman, N.G. (2006), "Nature of science: Past, present and future", in S. Abell and N.G. Lederman (eds.), *Handbook of Research on Science Education*, Lawrence Erlbaum, Mawah, NJ, pp. 831-879.

Longino, H.E. (1990), *Science as Social Knowledge,* Princetown University Press, Princetown, NJ.

Marzano, R.J. and **J.S. Kendall** (2007), *The New Taxonomy of Educational Objectives*, Corwin Press, Thousand Oaks, CA.

Millar, R. (2006), "Twenty first century science: Insights from the design and implementation of a scientific literacy approach in school science", *International Journal of Science Education*, Vol. 28/13, pp. 1499-1521, http://dx.doi.org/ 10.1080/09500690600718344.

Millar, R. and **J.F. Osborn** (eds.) (1998), *Beyond 2000: Science Education for the Future*, School of Education, King's College, London, www.nuffieldfoundation.org/sites/default/files/Beyond%202000.pdf.

Millar, R. et al. (1995), "Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance", *Research Papers in Education*, Vol. 9/2, pp. 207-248, http://dx.doi.org/ 10.1080/0267152940090205.

Mislevy, R.J. and **G.D. Haertel** (2006), "Implications of evidence-centered design for educational testing", *Educational Measurement: Issues and Practice*, Vol. 25/4, pp. 6-20.

National Academy of Science (1995), *National Science Education Standards*, National Academy Press, Washington, DC.

National Research Council (2012), *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, Committee on a Conceptual Framework for New K-12 Science Education Standards, Board on Science Education, Division of Behavioral and Social Sciences and Education, Washington, DC.

National Research Council (2000), *Inquiry and the National Science Education Standards*, National Academy Press, Washington DC.

OECD (2012), "What kinds of careers do boys and girls expect for themselves?", *PISA in focus*, No 14, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9d417g2933-en.

OECD (2009), *PISA 2006 Technical Report*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264048096-en.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264040014-en.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264026407-en.

OECD (2004), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264101739-en.

OECD (2000), *Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264181564-en.

OECD (1999), *Measuring Student Knowledge and Skills: A New Framework for Assessment*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264173125-en.

Ormerod, M.B. and **D. Duckworth** (1975), *Pupils' Attitudes to Science*, National Foundation for Educational Reasearch, Slough, UK.

Osborne, J.F. (2010), "Arguing to learn in science: The role of collaborative, critical discourse", *Science*, Vol. 328/5977, pp. 463-466, http://dx.doi.org/ 10.1126/science.1183944.

Osborne, J.F. and **J. Dillon** (2008), *Science Education in Europe: Critical Reflections*, Nuffield Foundation, London.

Osborne, J.F., S. Simon and **S. Collins** (2003), "Attitudes towards science: A review of the literature and its implications", *International Journal of Science Education*, Vol. 25/9, pp. 1049–1079, http://dx.doi.org/10.1080/0950069032000032199.

Rickinson, M. (2001), "Learners and learning in environmental education: A critical review of the evidence", *Environmental Education Research*, Vol. 7/3, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.454.4637&rep=rep1&type=pdf.

Rychen, D.S. and **L. H. Salganik** (eds.) (2003), *Definition and Selection of Key Competencies: Executive Summary*, Hogrefe Publishing, Göttingen, Germany.

Schibeci, R.A. (1984), "Attitudes to science: An update", *Studies in Science Education*, Vol. 11, pp. 26-59.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) (2005), Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10).

Tai, R.H. et al. (2006), "Planning early for careers in science", *Science*, Vol. 312, pp. 1143-1145.

Taiwan Ministry of Education (1999), Curriculum outlines for "Nature science and living technology", Ministry of Education, Taipei, Taiwan.

UNEP (2012), *21 Issues for the 21st Century: Result of the UNEP Foresight Process on Emerging Environmental Issues*, United Nations Environment Programme (UNEP), Nairobi, Kenya, www.unep.org/pdf/Foresight_Report-21_Issues_for_the_21st_Century.pdf.

UNESCO (2003), "UNESCO and the international decade of education for sustainable development (2005-2015)", *UNESCO International Science, Technology and Environmental Education Newsletter*, Vol. XXVIII, No. 1-2, UNESCO, Paris.

UNESCO (2005), "International implementation scheme" *United Nations Decade of Education for Sustainable Development (2005-2014)*, UNESCO, Paris, www.bibb.de/dokumente/pdf/a33_unesco_international_implementation_scheme.pdf.

Weaver, A. (2002), "Determinants of environmental attitudes: A five-country comparison", *International Journal of Sociology*, Vol. 32/1.

Webb, N.L. (1997), "Criteria for alignment of expectations and assessments in mathematics and science education", *Council of Chief State School Officers and National Institute for Science Education Research Monograph*, National Institution for Science Education, Washington, DC.

Wiliam, D. (2010), "What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment", *Review of Research in Education*, Vol. 34, pp. 254-284.

Ziman, J. (1979), *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*, Cambridge University Press, Cambridge, UK.

APPENDIX G: DISSERTATION TIMELINE