Educational Design Research

Volume 5 | Issue 1 | 2021 | Article 34

Contribution Academic Article

Title

# Rethinking Educational Assessment from the Perspective of Design Thinking

Author **William P. Fisher, Jr.**
University of California, Berkeley
USA

**Emily Pey-Tee Oon**
University of Macau
China

**Spencer Benson**
Education Innovations International Consulting, LLC
University Park, MD
USA

Abstract Design-based research in education emphasizes the contexts in which learning takes place as essential to more fruitful dialogues with practice. Contextual issues of social dynamics, place, and applications have been extensively investigated globally in recent design-based research, but the informational aspects of infrastructural concern to experimentation in naturalistic settings have yet to be addressed. Design Thinking (DT) augments design research by offering paths toward coherently integrating assessment and instruction across multiple levels of complexity and different communities' varying epistemic practices. DT is nonlinear but encompasses elements of empathy, problem definition, ideation, prototyping, and testing that inform the development of boundary objects mediating developmental, horizontal, and vertical forms of coherence by simultaneously functioning across denotative, metalinguistic, and metacommunicative levels of complexity. Prototype reports illustrate how emergent measured constructs can serve disparate communities' epistemic needs for shared languages while also structuring formative approaches to individuals’ unique learning processes.

Keywords
Design Thinking
Information systems
Complex adaptive systems
Computer supported cooperative work
Formative feedback
Metrology

DOI dx.doi.org/10.15460/eder.5.1.1537

Citation Fisher, W. P., Jr./ Oon, E. P.-T./Benson, S. (2021). Rethinking Educational Assessment from the Perspective of Design Thinking. *EDeR – Educational Design Research,* 5(1), 1-33. dx.doi.org/10.15460/eder.5.1.1537

# Rethinking Educational Assessment from the Perspective of Design Thinking

William Fisher

## 1.0 Introduction

Building productive connections between theory and practice remains an ongoing challenge in education. Design-based research emerged in the 1990s (Brown, 1992; Collins, 1992) as a means of addressing problems of how to situate the results of systematic investigations of learning outcomes in their naturalistic contexts (Barab, 1999; Barab & Squire, 2004). The complexity of the problems encountered has led to the emergence of a number of methodological variations and shifts in perspective (Akkerman et al., 2013; Design-Based Research Collective, 2003). These specializations, despite the shared intention of fostering more intensive collaborative engagements, have unfortunately exacerbated the problem, disconnecting research from practice to an even greater extent than was previously the case (Penuel et al., 2020).

That said, values relating to collaboration, problem solving, and research shared by these approaches point toward a potentially productive but as yet unexplored path toward more satisfying fulfilments of the need for interrelated research and practice. This suggestion is also supported from complementary directions by (a) design-based research focused on boundary-crossing alliances of collaborators (Akkerman et al., 2013, Akkerman & Bruining, 2016; Kali et al., 2018; Roth & McGinn, 1998; Zitter et al., 2012), and (b) model-based research focused on defining, testing, estimating, calibrating, and reporting multilevel measurements of progress in learning (Black, Wilson, & Yao, 2011; Scalise, Douskey, & Stacy, 2018). Maps of measured constructs functioning as boundary objects (Bowker et al., 2016; Fisher & Wilson, 2015; Star & Griesemer, 1989) offer important opportunities for improving the developmental, horizontal, and vertical coherence of learning outcome communications across research and practice. This body of research recognizes the importance of the interplay between global abstract idealizations and locally situated sociocultural practices (Squire, MaKinster, Barnett, Luehmann, & Barab, 2003; Zuiker & Whitaker, 2014), but has not yet formatively integrated assessment and instruction in ways that systematically separate and balance these abstract and concrete levels of complexity. This raises the possibility that systematically methodical integrations of boundary crossings in participatory research designs would leverage the simultaneously abstract and concrete nature of boundary objects to facilitate infrastructure development. This would seem preferable to continuing to allow researchers' and practitioners' subjective perspectives and definitions to dominate the relations of locally customized specifics and generalized standards.

We propose using a Design Thinking (DT) (Miller, 2015; Spinelli & Nixon, 2015) approach to adaptively engage with these problems in terms of the concepts and practices of developmentally, horizontally, and vertically coherent assessment (Gorin & Mislevy, 2013; Wilson, 2004; National Research Council, 2006) and the levels of complexity characterizing natural language use essential in the design of information infrastructures (Bateson, 1972; Bélanger, Cefaratti, Carte, & Markham, 2014; Bowker, 2016; Star & Ruhleder, 1996). The goal of the work is to contextualize feedback tools within each level of complexity to effect qualitatively substantive, meaningful, and quantitatively rigorous integrated assessment and instruction. Another way of stating this goal is as a meta-design problem of supporting collaboration by means of shared tools externalizing locally situated cognition in distributed ecosystem environments (Akkerman et al., 2007; Fischer, Giaccardi, Eden, Sugimoto, & Ye, 2005; Fisher & Stenner, 2018; Morrison & Fisher, 2018, 2019, 2020). Yet another way, following the work of Latour (1987, 1998, 2004, 2005) and others, focuses on the circulation of inscriptions throughout networks of actors stabilized by means of adherence to standards implemented at obligatory passage points (Fisher & Wilson, 2015; Roth & McGinn, 1998). The critique of such networks is, of course, inherently itself also constituted as a network. The infrastructuring processes proposed here, then, follow through on that critique to apply the lessons learned in ways that do not render invisible the lives, work, and voices of those whose learning outcomes are inscribed and circulated. Instead, we attend to the coherence of learning outcome measures over time and space, and to their multiple levels of semiotic complexity, with the aim of individualizing actionable information in concrete local circumstances that are not disconnected from the larger social context.

Design Thinking is a creativity process combining elements of engineering, empathy, and art for the development of practical solutions to pervasively "wicked" problems. Originating at Stanford University, DT advocates suggest it may replace the traditional liberal arts curriculum in higher education (Miller, 2015). DT may work to reorganize artificial, vertical bureaucracies and siloes into the lateral relationships of more horizontal ecosystems. Others suggest that DT will blossom across universities at systems levels, as liberal arts and professional education are integrated (Spinelli & Nixon, 2015). Yet unnoticed, however, are commonalities between DT's focus on enhanced communications and information flows, and similar efforts focused on (a) more coherent co-ordinations of learning outcome measurement and management (Gorin & Mislevy, 2013; Wilson, 2004; National Research Council, 2006), and (b) multilevel conceptions of information infrastructures (Bateson, 1972; Bélanger et al., 2014; Bowker, 2016; Fischer et al., 2005; Star & Ruhleder, 1996). These commonalities remain unnoticed and unleveraged even in those rare instances in which educational assessments are viewed from a DT perspective (Benson & Dresdow, 2014).

In this paper, we propose the use DT for the development of an educational measurement information infrastructure that connects the various coherence levels for formative and summative assessments with each of three levels of complexity occurring in everyday language. In particular, we provide new prototypes for developmentally, horizontally, and vertically coherent information and discuss how that information can be scaled and embedded in a multilevel assessment information framework for various stakeholders.

In this context, the five components of DT inform new approaches to educational assessment's needed for developmental coherence, horizontal coherence, and vertical coherence. Measuring and managing individual students' growth and progress in learning requires developmental coherence that support formative processes by relating what has been learned to objectives and also to what comes next, instructionally. Measuring and managing classroom- and school-level outcomes requires horizontal coherence and lateral expansions of the community: the comparability of learning progressions and outcomes across students, classrooms, and schools facilitates new conversations deepening, broadening, and intensifying collective efforts. Measuring and managing the accountability of the education system requires vertical coherence of classroom and high-stakes summative assessments at regional, national, and international levels. Figure 1 illustrates the relationships between these three forms of coherence.

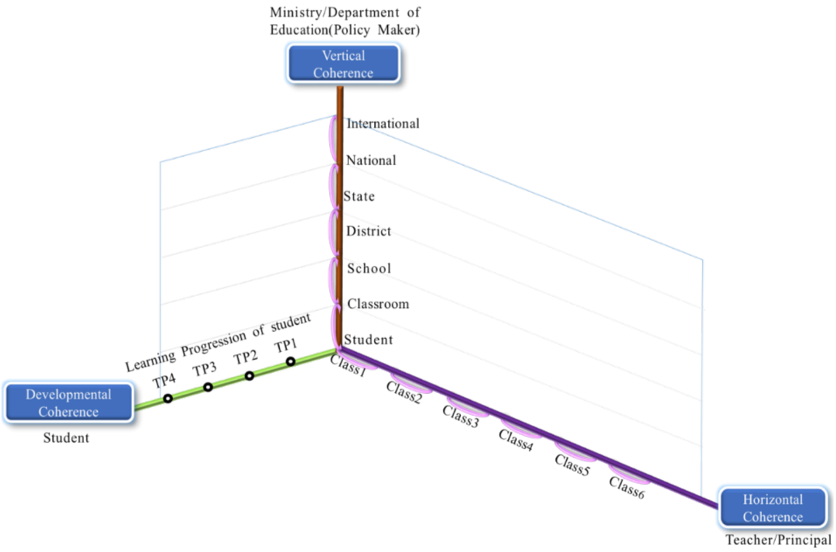

*Figure 1: Developmental, horizontal, and vertical coherence (Fisher, Oon, & Benson, 2018)*
*Note. TP = Time Point.*

Problems of coherence in educational measurement cannot be solved by tests with common items administered repeatedly over time and across classrooms and schools (Wilson, 2004; Moss, 2004). The three domains of coherence share a common focus on educational processes and outcomes, which suggests a realistic potential for improved alignments of information on students' learning progressions, designated learning outcomes, and the management and policy needs of individual students, classrooms, schools, districts, etc. Education as a systematic ecological institution is based on the conceptual and practical application of curricula to routinely observed patterns of physical, cognitive, and moral development (Dawson, 2002, 2004), even though the information applied and produced is represented incoherently across and within all of these contexts.

Successfully addressing problems of coherence in information infrastructures requires attention to the multilevel nature of meaning and its variation across local, midrange, and broader contexts. Philosophers have long noted the ironies and paradoxes of how words change their meaning across levels of complexity (Palmgren, 2018; Russell, 1908; Whitehead & Russell, 1910, 1912, 1913; Wittgenstein, 1922; Whorf, 1956; Bateson, 1972; Beckner, Holland, Ke, Larsen-Freeman, & Schoenemann, 2009). For example, referring to something in the world, like a cat or a tree, is very different from referring metalinguistically to the words 'cat' or 'tree,' and this, in turn, is qualitatively distinct from referring metacommunicatively to a statement containing a reference to a cat or a tree.

Three essential points are worth noting. The first is that everyday commonly used words are distributed throughout communities as the media enabling shared forms of social life. Coherent expressions (standardized words referring to real things in the world relative to a determined concept) describing learning outcomes must also be distributed in this way for shared participation to be realized. Second, these coherent expressions come into being via social processes not directed by any individual or group. The constructs of learning outcomes measured in formative and summative assessments must also cohere by means of collective processes. Third, the abstract metalinguistic and metacommunicative standards of orthography, grammar, semantics, syntax, dictionary definitions, etc. are not rigidly imposed in a mechanical, cookie-cutter kind of way, but instead serve as malleable media used in the moment to negotiate shared meanings. Similarly learning outcomes should not be mechanical, cookie-cutter statements but adaptable measurable descriptions of student knowledge, attitudes, or skills developing as a result of a learning activity or set of activities.

Though it has a long history, this latter point bears emphasis. Plato noted in *The Republic* (522c-527c, especially 523c) that "The experiences that do not provoke thought are those that do not at the same time issue in a contradictory perception" (also see Gadamer, 1989, p. 120). We tend to notice salient differences, but differences are inherently relational; things worth remarking on appear only in contrast with something else. Variation in unique local circumstances can be meaningfully experienced only against the background of a contextualizing invariance (Marton & Pang, 2013, p. 39). In his study of beauty and growth in nature, art, and architecture, Cook (1914/1979, p. 400) similarly argued that "Clearly it will be unprofitable to emphasize the value of variations unless we also suggest some standard by which those variations can be measured." Cook was also sensitive to the way measurement standards in science function to take advantage of the routine predictability of many kinds of events, foreshadowing Kuhn's (1961, p. 180; 1977, p. 205) point that the function of measurement in science is to reveal anomalies. Finally, Rasch (1960, p. 124) also says that "Once a law has been established within a certain field then the law itself may serve as a tool for deciding whether or not added stimuli and/or objects belong to the original group." This paper's argument and the prototypes presented are expansions on this theme, following the models and reporting formats pioneered by Wright (Wright, Mead, & Ludlow, 1980; Wright & Stone, 1979, 1999; Wright & Masters, 1982; Mead, 2009; Chien et al., 2018; Masters, Lokan, & Adams, 1994; Holster & Lake, 2016).

Taking the emphasis on this matter still further, the process of effecting the transition to contextualized boundary crossings can be described as systematizing the inclusion of subjective perspectives and the comparisons of them with each other and with objectively reproducible evidence (Wright, 1958; Fisher, 2017). Penuel et al. (2020), for instance, recount researchers' participation in negotiations concerned with identifying differences and similarities across four design-based approaches. The researchers' subjective analysis of common themes was aligned and resulted in the clarification of shared principles. This process entailed a mediation of objective distance and subjective empathy, neither of which defines well-articulated science, and both are needed to engage in a "joint epistemic project addressing the historically changing and mutually conditioning relation of 'inside' and 'outside' knowledge" (Galison, 2008, p. 293). To be useful, subjectivity and objectivity together must "maximize the occasion for the phenomenon at hand to raise its own questions against the original intentions of the investigator – including of course the generous 'empathic' intentions" (Latour, 2004, p. 219). In the manner exemplified by Penuel et al, (2020), instead of trying to identify subjectivity as factor to be removed, we instead put individual and collective biases on the table to see whether the phenomenon asserts itself against them as different.

This capacity of maximizing opportunities for phenomena to assert their own properties independent of the investigator's intentions defines a matter of central concern in the design of educational assessments and measurements. Traditional test scores protect researchers' subjective biases by not allowing the phenomenon to raise its own questions. The voices of those participating in the research are then, in effect, silenced (Roth & McGinn, 1998). Measured quantities, in contrast, are explicitly oriented toward framing the contexts in which anomalous observations can reveal themselves for further examination. Test scores, as opposed to measured quantities, indiscriminately lump diverse performances into a single homogenized numeric representation. Test scores thereby protect researchers' subjective biases from risks by not compelling them to question their assumptions. Traditional test scoring methods fallaciously treating counts or percentages correct as measures accept as a methodological necessity the fact that these numbers mean different things depending on which questions were asked and answered. This dependency is not, however, a necessary limitation of quantitative methods in psychology, and has been understood as such since the work of Thurstone in the 1920s (Andrich, 1978b; Andrich & Marais, 2019, pp. 224–225).

Coherent meaningfulness that does not silence but celebrates the voices of those whose learning outcomes are measured requires close attention to the relational processes by which words and concepts come to represent things in the world (Overton, 2015). Like other complex ecosystems (Beckner et al. 2009), communication networks involve micro-level processes within individuals that are distinguished from meso-level processes between individuals, which in turn are distinct from macro-level group processes. Issues stemming from these varying levels of complexity affect the efficiency and effectiveness of efforts aiming to improve the developmental, horizontal, and vertical coherence in educational measurement information infrastructures. Design thinking provides a novel route to solving the complex problem of contextualization, integration, and interconnectivity of these three coherences.

Previous approaches to the problem of coherence in education (Fullan & Quinn, 2015; Fortus & Krajcik, 2012; McQuillan, Welch, & Barnatt, 2012) failed to recognize the important roles that creating and sustaining shared meanings play in each of the three forms of coherence as they emerge within each of the three levels of complexity. Though a clear emphasis on communication and collaboration points toward the need for attending to information infrastructures (Fullan & Quinn, 2015, p. 5), considerations of what coherence means for education generally have not taken up issues related either to assessment (Gorin & Mislevy, 2013; Wilson, 2004; National Research Council, 2006) or to the levels of complexity that must be addressed meta-systematically as distinct but related spheres of activity (Bateson, 1972; Bélanger et al., 2014; Bowker, 2016; Fischer et al., 2005; Star & Ruhleder, 1996). Education researchers addressing complexity (Mason, 2008; Crick, Barr, & Pedder, 2017) often fail to attend to assessment coherence, or the details of how variation in complexity can be measured and managed in practical terms.

DT attends not just to the technical feasibility of engineering solutions but also to the desirability end users may associate with particular features and to the viability of producing those solutions. The complexity of the problem of educational coherence follows from the seemingly insurmountable difficulties encountered in trying to establish meaningful comparisons using different assessments with separate formative or summative purposes administered to different students in widely separated locations and times. Decades of research and practice, however, support the feasibility of a variety of technical solutions, such as adaptively administered item banks (Barney & Fisher, 2016; Wright & Bell, 1984), theory-driven and automatic item generation (Dawson, 2002, 2004; Embretson, 2010; Stenner, Fisher, Stone, & Burdick, 2013; Fischer, 1973; Fisher & Stenner, 2016), metrological traceability (Fisher & Stenner, 2016; Pendrill, 2014, 2019; Mari & Wilson, 2014; Mari, Wilson, & Maul, 2020), instrument equating (Masters, 1985), multifaceted (Linacre, Engelhard, Tatum, & Myford, 1994), multilevel (Beretvas & Kamata, 2007), and multidimensional (Briggs & Wilson, 2003) methods and models able to place multiple approaches to assessment item formats (multiple choice, open response, judged performance, etc.) on the same scale (Öztürk-Gübes & Kelecioglu, 2016). All of these solutions integrate qualitative and quantitative methods and data in meaningful unit definitions with known uncertainties, and data consistency indexes (Fisher 2004; Fisher & Cavanagh 2016).

A model for viable, feasible, and desirable solutions to assessment coherence can be discerned in the history of the last 20 years of developments in public health and clinical medicine. In these fields, new interdisciplinary relationships based on practical applications of distinctions between micro, meso, and macro levels of complexity have advanced the state of the art in epidemiology (Susser & Susser, 1996; Bizouarn, 2016). An analogous ecosystems approach to the coherence of educational assessments starts with the intention to situate learning in its appropriate environmental context, following the parallel pattern also set over the last 20 years in design-based research (Akkerman & Bruining, 2016; Barab, 1999; Barab & Squire, 2004; Kali et al., 2018; Penuel et al., 2020). Close attention must be paid to the emergence of meaningful assessment information in individual student response data (the micro level), the aggregation of this data into composite scales that retain their identities across students and assessments (the meso level), and the mapping of these composite constructs (the macro level). Coherent communications integrating formative and summative concerns across the boundaries of research and practice are feasible, viable, and desirable only to the extent these levels of complexity are integrated into information infrastructures in ways that allow each of them to retain their characteristic properties in relation to each respective sector of end users.

Empathy with end users determines the desirability and utility of design solutions. Economically viable and technically feasible solutions lacking sensitivity to the individual micro level of complexity have long histories of failure (Star & Ruhleder, 1996; Moss, 2004; Scott, 1998; Vonderau, 2018). End users' opinions and input need not be taken literally, however. Henry Ford is reputed, perhaps apocryphally, to have said he would have tried to develop faster horses if he had listened strictly to what people expected in the way of better ways of getting around. The same kind of situation can be observed in the context of most game-changing major technical innovations. For example, the public was not clamoring – or even aware of the possibility – for electricity, automobiles, the telegraph, telephone, television, faxes, computers, the Internet or smartphones. The question then arises as to whether new educational assessment technologies might meet unstated but urgent needs in a way students and educators will find desirable, even though they are not seeking or expecting them.

## 2.0 Illustrative DT integrative process

DT is often described as interconnected applications of empathy, problem definition, ideation, prototyping, and testing to yield innovative solutions that take into account the needs of end users. While the steps are often presented in a linear fashion, one can engage DT in any order, with variation in emphasis across the five steps, both within and across iterations. The education example described here moves through a DT sequence solely for the purpose of orienting the reader to different aspects of DT integrated as a process.

## 2.1 *Empathy*

In educational cultures that prioritize high-stakes exams, assessment often does not simultaneously serve its formative and summative purposes (Ladd, 2017). Grades in this context become the focal interest, at the expense of learning. Summative assessment scores are assumed to document performance levels for both individuals and groups. Opportunities for classroom implementations of formative assessments capable of informing individualized instruction are often overlooked due to the inordinate focus on the total score/grade. In serving summative purposes, grades are expected to provide valid measures of learning and performance, and are not intended as a tool for judging personal worth. Summative assessments nonetheless still impact students' self-images in formative ways (Boaler, Wiliam, & Brown, 2000; Pilcher, 1994). In principle, low grades are supposed to provide information on areas in need of improvement and attention. Parents and teachers tend to use grades, however, as a means of rewarding and punishing students for their learning and performance (Pilcher, 1994) rather than identifying areas of need and helping students to improve their learning.

Another problem with grades is that the total correct score does not provide a clear indication of how much learning has occurred, but is rather an ordinal and assessment-dependent indicator of performance that artificially ranks students in an undefined unit of usually unstated ranges of uncertainty. As forerunner agents in assessment, teachers are often criticized, blamed, or rewarded based on student and class test scores, despite their lack of access to instructionally relevant formative information (Ladd, 2017; Wilson, 2004, 2018). This can have intended and unintended consequences for the teacher, school, and the students. Teachers often lack the knowledge, technology, and authority to formatively measure and manage learning outcomes, as these resources may be externally controlled. Though researchers have tested a plethora of relevant and often useful methods, many teachers do not know how to document the performances of students relative to learning progressions, and do not know where and how to obtain the information needed to address learning gaps with respect to the learning objectives.

Current assessment frameworks lack the information infrastructures needed to support teachers in these tasks. Researchers have repeatedly demonstrated the objective reproducibility of dozens of measured constructs, from reading comprehension (He & Kingsbury, 2016; Stenner et al., 2013) to mathematics (Fischer, 1973) and writing abilities (Engelhard, 1992) to socio-emotional outcomes (Crowder et al., 2019) and moral and cognitive development (Dawson, 2002, 2004). The capacity to connect formative classroom assessments with high-stakes accountability exams (Wilson, 2004, 2018) has been investigated ever since Wright (1977, p. 108; Wright & Bell, 1984) and his student, Choppin (1968, 1976), suggested how this could be done. Persistent misconceptions and unexamined assumptions about tests, assessments, and educational measurement have had enduring negative impacts on students, teachers, educational outcomes, and the culture at large because of the ways they limit capacities to envision new possibilities designed to better meet end-user needs. In these ways, education's information infrastructures fail to take into account the needs of learners, parents and the teacher end users and lack empathy.

As a consequence of these limitations, teachers are overburdened with repetitive tasks that follow largely from the need to micro-manage the details of individual student’s learning, and the assessment of that learning. Students and teachers alike find it difficult to know what lessons and pedagogies are most appropriate to the needs of individual learners.

The information on educational processes and outcomes that is available in principle from assessments far exceeds the information made available to, and actually used by, teachers. Low quality information, such as total correct scores and overall grades, offer few resources for matching readers' abilities to texts' difficulties, for instance, and makes virtually impossible more complex tasks like plotting student learning progressions, making meaningful comparisons over time, and setting achievable quality improvement and accountability goals. To a large degree, teachers are being held accountable for managing outcomes they cannot effectively analyze, visualize, and compare day after day, either within their own classes, or across classrooms. Viable solutions to effective uses of assessments for deeper understanding and improved student learning must “empathize” with frontline teachers charged with the internally contradictory tasks of fostering student learning while functioning in an accountability environment that systematically undercuts opportunities for learning.

## 2.2 *Define*

The challenge in empathizing with teachers concerns the incoherence of the available learning outcome management information systems and the expectations placed on both teachers and students. Though progress has been made in recent years with respect to the increased emphasis on accountability for classroom assessments (Wilson, 2017), infrastructure supporting the management of coherent information remains underdeveloped. What might a coherent education management information system look like? Developmentally, horizontally, and vertically coherent assessments are needed with respect to addressing individual students' learning progressions, comparisons over time and across classrooms, and long-term accountability goals. To be coherent in these ways, measures must be interpretable as evidence of progress in learning. What is needed are processes that allow teachers to easily extract information on student learning, to apply that information to individual and class learning, and to evaluate and refine learning outcomes within and across systems of instruction. Teachers' practical skills, knowledge, and understanding of the roles of assessment (realized and possible) must be cultivated and nurtured at the same time that overall continuity and navigability of assessment processes are supported.

## 2.3 *Ideate*

How do developmental, vertical, and horizontal forms of coherence fit into the larger context and goals of education? Education optimally provides students with experiences that enable them to successfully address real-life problems and create value for themselves and others. A good curriculum enables knowledge building and samples problems and lessons that represent discipline content and concepts through real-life examples.

Schools are expected to nourish students who can meet designated learning outcomes independent of the particular students in attendance. This expectation is, broadly and generally speaking, borne out when considering the full range of variation across the span of formal education. Schools, curricula, textbooks, assessments, pedagogies, etc., are all created in terms relevant to the population level of meta-communicative complexity. The unstated and poorly tested, but widely adopted, assumption is that the student learning potential is governed by the interplay of student abilities and the difficulties of the materials to be learned. This can be expressed at the metalinguistic level of complexity as an assertion that the probability of student success is a function of the difference between ability and difficulty (Rasch, 1960; Bond & Fox, 2015; Wilson, 2005; Andrich, 1988; Fisher & Wright, 1994; Wright, 1999). This expression of a model for meso-level instrument calibration and measurement articulates the pre-existing macro-level theory of variation and sets the stage for contextualized, instructionally relevant micro-level individual reports of response data.

In this context, how might we formulate the problem of coherent assessment information infrastructures? One pathway is to brainstorm "possibly impossible" solutions for trial and error applications. Are there current and past situations in science and engineering where DT has been applied in situations that involve similar kinds of mutual, interdependent relations? For instance, might the theoretical and empirical convergence of psychometrics and metrology (Mari & Wilson, 2014; Pendrill & Fisher, 2015; Wilson & Fisher, 2016; Wilson & Fisher, 2018) provide a basis for an effective model of the problem of coherence in educational assessments? Metrology has a long history of addressing comparisons via common languages distributed throughout distributed ecosystems that emerge from within locally situated forms of knowledge (Latour, 1998; Berg & Timmermans, 2000; Golinski, 2012; O'Connell, 1993; Fisher & Wilson, 2015).

It is important here to acknowledge and respond to those (Moss, 2004) who anticipate coherent assessments to be yet another in a long line of well-intended but harmful impositions of homogenous uniformity that aid bureaucratic management but stifle creativity, meaning, and innovation (Scott, 1998). The practical question is how to use language as a model: how to let things/objects come into words, to inscribe meanings, in ways that enable them to be locally situated and concrete at the same time that they are abstract and ideal (Star & Griesemer, 1989; Star & Ruhleder, 1996; Scott, 1998, p. 357). Might not psychometric metrology also yield ecological conceptualizations and organization, with overlapping formative and summative assessment niches nurturing teachers' locally situated knowledge at the same time that they facilitate navigable continuity over time and space (Fisher, Oon, & Benson, 2018; Fisher & Stenner, 2018; Fisher & Wilson, 2015)? In other words, how might boundary objects spanning multiple levels of complexity in research and practice domains be constituted via a simultaneous combination of formal mathematical theory, abstract instrument standards, and concrete data?

A formal theoretical model of measurement relating concrete individual student abilities and item difficulties with the aim of calibrating an abstract unit quantity takes the form

$$\ln[Pnij / (Pnij-1)] = Bn - Dij \tag{1}$$

This equation asserts that the log-odds (the natural logarithm of the response probability in one partial credit score category divided by the probability of being in the previous score category) of success for student n on item i at partial credit score j is equal to the difference between the estimate B of student n's ability and the difficulty estimate D of item i at rating j (Andrich, 1978a, 2010; Andrich & Styles, 2011; Masters, 1982; Bond & Fox, 2015).

A group- and meso-level construct emerges from within the micro-level individual responses as a self-organized pattern (Fisher, 2017a). The fit of data to a model of this kind requires statistically sufficient invariances (Andersen, 1977; Andrich, 2010) in the order and positions of student abilities on a scale relative to the questions asked, and vice versa. These invariances have been found to be repeatedly reproducible and theoretically explained for different constructs, and across decades and many millions of students (He & Kingsbury, 2016; Fisher & Stenner, 2016; Williamson, 2018), thus supporting the pre-calibration and adaptive administration of item banks and cognitive models (Barney & Fisher, 2016; Embretson, 2010; Wright & Bell, 1984).

Documenting individual student learning in this way leads to the identifications of patterns in that learning, such that we learn about learning at a higher order level of complexity. Repeated observations of reproducible patterns across students and assessments leads to the development of theories of learning capable of explaining meso-level variation from a macro level. When these levels are distinguished in practical terms, they can each be encapsulated in portable technologies that embody the available relevant information in different forms for developmentally, horizontally, and vertically coherent applications in instruction, quality improvement, and accountability. The invariant and reproducible scales calibrated from theory and data provide the basis for a new coherent multilevel language of learning outcomes. This new language differs from existing test scores in that the number words used are not tied to a single collection of assessment questions devised by individuals or groups. Instead, by modelling constructs that retain their properties across samples of students and assessment items, the door is opened to a wider world of applications and the creation of a language relationally capable of referring in principle to any student encountering any question. One way of describing this process (Ihde, 1991, pp. 132–135) is in terms of a model of something real in the world embodied in a portable and readable technology exported back into the world via distributed networks of metrologically traceable instruments. This is the point at which technical solutions are proposed as prototypes.

## 2.4 *Prototype*

Prototyping is the process in which possibly impossible solutions are proposed and developed. Developmental coherence deals with the problem of students knowing what they know, what to study, and the metacognitive issues of knowing and seeing their own learning progression, including special strengths and weaknesses in need of attention. In addition, students, teachers, and parents want to know where individuals stand relative to the class as a whole. Teachers need a means of mapping individual learning within the context of the class as a whole both to be better equipped to assist individual students and to be able to adjust pedagogies to meet the needs of the class. Each class, while similar to prior classes, is unique due to student composition, external factors, and the human nature of the teaching-learning endeavor.

One way of approaching this uniqueness coherently is to map individual student responses in terms of learning progressions using common metrics to measure student performance relevant to the desired learning outcomes. It is in this context that the capacity to identify variations in the context of an overall invariance, taking up the previously mentioned theme of anomalies, shows its value. Figure 2 shows a prototype form produced from standard educational measurement data analysis software (Andrich, Sheridan, & Luo, 2017; Linacre, 2020; Wu, Adams, & Wilson, 2015). Individual and class performances on a summative assessment are illustrated in terms of estimates based on the probabilistic model described above.

In the Figure 2 self-scoring form, the range of measurements for which incorrect responses are expected are scored 0 and are colored red. The range of measurements where correct responses are expected and are scored dichotomously (in two categories) are the 1s in the pink background. The fully correct responses scored in three categories accepting partial credit are shown as 2s in a green background. As the measures increase on the horizontal scale, the probability of a correct answer increases. The learning progression is mapped by the vertically increasing item difficulties.

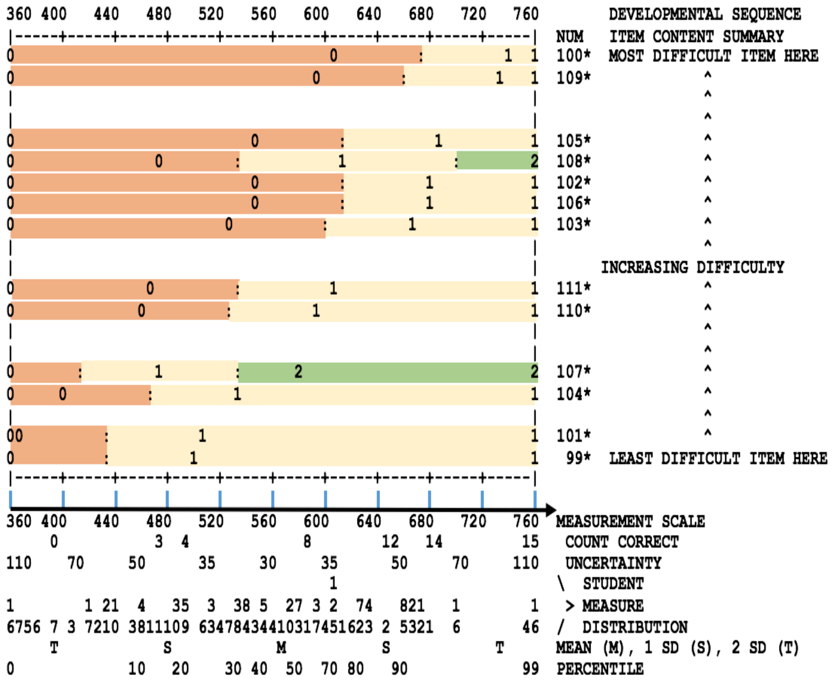

*Figure 2: Prototype individual student self-scoring form for implementing developmental coherence (see text for a description of the figure's components, and how the form is filled out and interpreted).*

The Figure 2 prototype for a self-scoring form provides information for the analysis of instructional value by mapping the learning progression in a developmentally coherent way. The measurements depicted in the figure are an instantiation of a boundary object. The formal mathematical model and construct theory have been used to design and calibrate an instrument measuring an abstract unit quantity on the basis of the evidence provided by concrete data.

The form in Figure 2 is filled out for an individual student by marking the score on each item that was obtained. Incorrect responses are recorded by circling the 0 in the item's horizontal row; correct responses, by circling the 1 for dichotomous items or the 2 for partial credit items. For these latter, partially correct responses are scored 1 by circling the middle option. No data ever fit a model perfectly, and so we do not expect the pattern of responses to conform exactly to the modeled expectations. However, given the overall fit of the data to the model, for a student with a measure of 600, for instance, we would expect the pattern of scores to be dominated by 1s at the bottom, up to about item 111, and by 0s at the top, starting from item 103. Further details on Figure 2 are given below.

Augmented with supplemental information on item content and a construct map illustrating the relevant learning progressions (Alonzo & Steedle, 2009; Black et al., 2011; Ketterlin-Geller et al., 2019; Wilson, 2005, 2009; di Uccio et al., 2020), the Figure 2 prototype informs the student and teacher as to what has been accomplished relative to the desired outcome, what comes next on the learning progression, what special weaknesses may need to be rectified, and what special strengths can be leveraged (Black & Wiliam, 1998, 2009). Items incorporating partial credit assignments, such as 107 and 108 in Figure 2, may offer added value by alerting the student and teacher to the presence of misconceptions (Andrich & Styles, 2011; Masters, 1982; Wind et al., 2019). Each student's pattern of responses may be unique without compromising fit of the data to the measurement model or the explanatory power of the predictive construct theory.

The information needed for horizontal and vertical coherence levels is of a different but related nature. Student performance data is needed but not at the level necessary for individual learning progressions. For horizontal coherence, teachers need information that identifies what is working pedagogically: what needs improvement, elimination, or replacement? Figure 3 illustrates a prototype form that provides information necessary for horizontal coherence. It allows for tracking at the student level, showing meaningful comparisons with other teachers' results and visualization on individual and class learning progressions. Reports of this kind can also be produced from the outputs of available psychometric software.

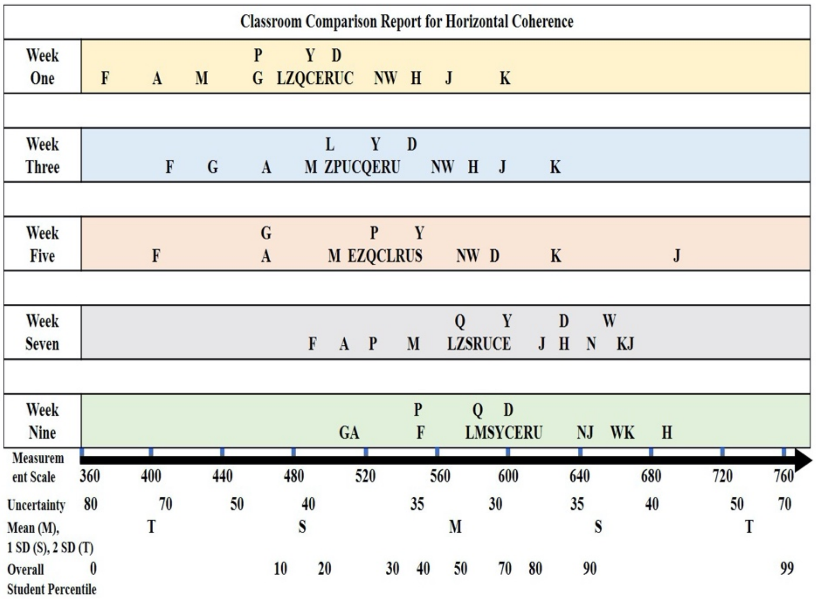

*Figure 3: Prototype comparison report for horizontal coherence across classrooms. Individual classrooms (student-teacher combinations) are represented by the same letters across weeks. By plotting the classroom performance (measurement scale) over time (week 1 to 9) one can discern the class's learning progression.*

Figure 4 illustrates a prototype for multilevel comparisons that allow vertical coherence across the hierarchy of accountability demands from the student to the institution, enabling information extraction and utilization from the micro to macro level extending to statewide, national, and international levels. Details of these and additional illustrations of coherent educational assessment information are discussed in the next section of this paper.

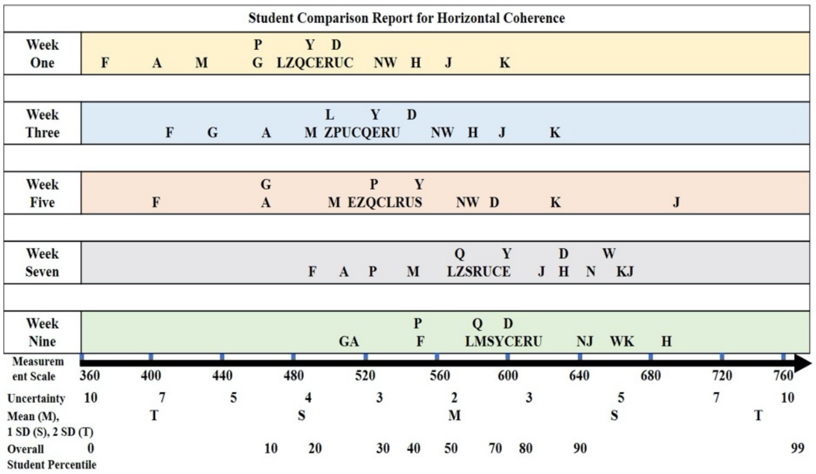

*Figure 4: Prototype student comparison report for horizontal coherence within a classroom. Individual students are represented by the same letters across weeks. By plotting the classroom performance (measurement scale) over time (week 1 to 9) one can discern the student's learning progression.*

## 2.4 Testing

Prototypes are tools for testing the proposed models, and for optimizing, refining, and possibly rejecting the results of empathizing, defining, and ideating. The DT approach is unique in that it seeks end user-centric innovations that have the potential to move the field forward in new and sometimes unforeseen ways. As such, the possibly impossible DT solution may fail to yield new information, which will drive reiteration of some or all of the other parts of the process to explore new developments and approaches. To test the prototype coherence tools proposed here, teachers will evaluate the strengths, weaknesses, opportunities, and limitations of prototypes in the context of their teaching practice. They will work with test data, students, and other teachers to assess the value of the prototypes relative to pre-existing methods and other tools they have previously used. The results from these tests will provide information necessary for refinement, optimization, and scalability of the prototype as testing is expanded to include additional classes, teachers, schools, and educational systems.

## 3.0 Developmental, horizontal and vertical coherence prototypes

Figure 2 shows the expected score on each item for a student responding to questions from a science assessment. The items (99-111) are listed on the right of the table in difficulty order, with the easiest item at the bottom. The items have been drawn from a pre-calibrated bank of hundreds or thousands of items. This bank measures a single unidimensional construct (intended learning outcome) that has been mapped conceptually in terms of the progression in learning that makes it possible for students to apply their knowledge to problems of greater difficulty. The construct map will be made available as a supplementary resource for teachers' use in interpreting the measures shown in these prototypes.

The 360-760 scale is shown horizontally, with the count of correct responses (0-15) shown just below it, and the uncertainty (standard error) indicating the 95% confidence ranges. The positions of the students on the scale are shown across the bottom in the three rows of numbers labelled Student Measure Distribution. These numbers are read vertically, such that the following:

```
 1
3 2
745
```

indicates 37, 4, and 125 students with measures at the respective horizontal positions. This vertical arrangement indicates how many students have measurements in a particular column at that horizontal point on the quantitative continuum, which is illustrated via the vertical hash marks on the 360-760 continuum. In Figure 2, the Mean (the M at about 570), first standard deviation (the S on the left at 480, and the other S on the right at 660), and second standard deviation (the T on the left at about 390, and the other T on the right at about 750) are indicated in the row of letters beneath the Student Measure Distribution. Finally, the percentile rankings of the student measurements are indicated in the bottom row of numbers; note the nonlinear nature of the percentile distribution along the horizontal axis.

To connect the measurements with expected performances, imagine a vertical line connecting the number 600 on the 360-760 scale at the bottom of Figure 2 with the 600 at the top of the table. Pick an item from those listed on the right (99-111) and read across the table until your line of sight intercepts the vertical line between the top and bottom 600s. That point of interception indicates the expected score on that item for a student with a performance measure of 600. The expected score (0, 1, 2) is the one to the left of the intercept, because the probability of a correct response increases the further to the right a measure is relative to an item. For example, a student with a measurement of 600 would be expected to get a score of 1 or 2 (item 107) for items 99, 101, 104, 107, and 111 and a score of 0 for items 103, 106, 102,108, 105,109, and 100. The colons between the scores show where transitions between scores are most likely to take place.

Reading up the page from the easiest to the hardest item along the vertical line indicates an expected score for each item. The highest probabilities of success are associated with items at the bottom of the scale and the lowest probabilities with items at the top. The digits in each row moving to the right as you read up the page represent the scores for each item. Items at the top of the page have scores further to the right since being to the right is associated with greater difficulty and higher measures, where higher measures mean greater ability, and greater ability is a capacity to succeed in relation to greater difficulties.

Figure 2 is then a snapshot of a moment in an individual student’s developmental trajectory. The self-scoring form shows the student's actual responses in the context of the learning progression. Special strengths are revealed as unexpected correct answers, and special weaknesses are shown as unexpected incorrect answers. Instead of concealing anomalous exceptions within a score assumed to always mean the same thing, the recording of students' concrete responses in the context of the abstract scale of measurement gives voice to unique individuals. Reporting the contextualized response data empowers students and teachers with instructionally actionable information that is otherwise unavailable.

The items are relatively on target, meaning they are of a difficulty level appropriate to the individual student and the class. The distribution of student measures at the bottom of Figure 2 is encompassed within the same range as the items' difficulties. That is, reading up from the bottom of Figure 2, only students with the lowest measures (furthest to the left) have little chance of answering any items correctly, and few students are located so far to the right that they are likely to succeed on all items. Targeting is substantively useful in instruction because it is only when students have some questions that they can answer correctly, and others they cannot, that we are able to state in concrete terms just what we expect them to be able to do, and what instructional intervention may be most likely to build on what they know.

That is, a test on which someone scores zero tells us only that all of the items were too difficult for that student; we have no idea how much too difficult. Conversely, a test on which a student scores the maximum possible says only that the questions asked were too easy, and we do not know how much too easy.

Additional information on learning objectives, or on vertically coherent proficiency standards or SAT equivalents, could also be added to Figure 2.

Figure 3 shows the same scale as Figure 2, but now with classroom-level uncertainty and percentile information. The same construct map and other supplementary information guides interpretation here. The difference in Figure 3 is that, instead of one student’s individual scored responses, average classroom measures are plotted over time. The same group of classrooms are plotted in two rows for each of several weeks (One, Three, Five, Seven, and Nine). Each classroom has the same letter designation in each week. Uncertainties are smaller in Figure 3 because of the larger sample sizes associated with the classroom-level focus. Mean model fit statistics could possibly be reported here so as to flag situations in which some classrooms or schools are consistently inconsistent, with students exhibiting repeated patterns of special strengths or weaknesses. This figure coherently combines developmental and horizontal information to make status and change visible and actionable across classrooms.

Figure 4 shows the same scale as the prior figures, but illustrating within-classroom student-level uncertainty and percentile information over time. The same construct map and other supplementary information again guides interpretation here. The difference between Figures 3 and 4 is that, instead of average classroom measures plotted over time, Figure 4 shows individual student measures within a classroom over time. The same group of students are plotted in two rows for each of several weeks (One, Three, Five, Seven, and Nine). This figure coherently combines developmental and horizontal information to make status and change visible and actionable across students within a classroom.

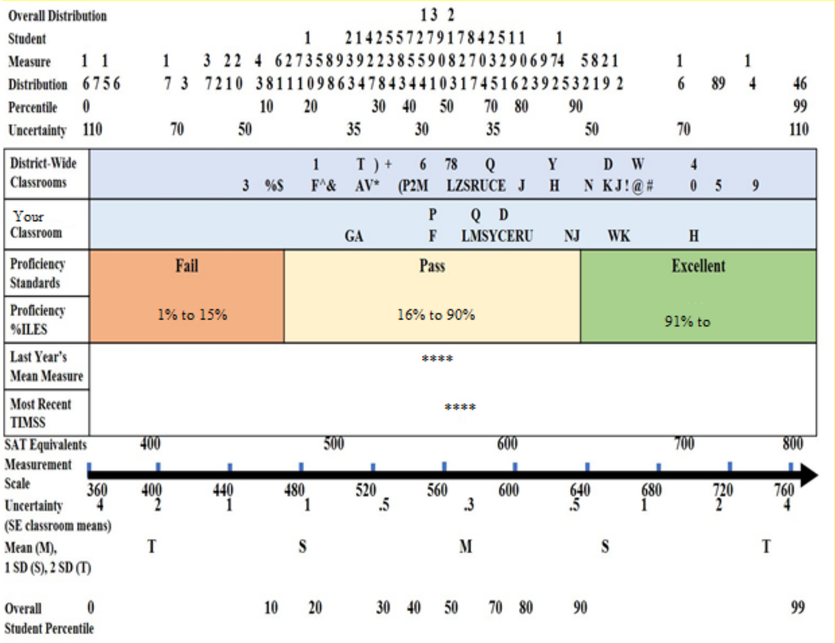
Overall Distribution 1 3 2
Student 1 2 1 4 2 5 5 7 2 7 9 1 7 8 4 2 5 1 1 1
Measure 1 1 1 3 2 2 4 6 2 7 3 5 8 9 3 9 2 2 3 8 5 5 9 0 8 2 7 0 3 2 9 0 6 9 74 5 8 2 1 1 1
Distribution 6 7 5 6 7 3 7 2 1 0 3 8 1 1 1 0 9 8 6 3 4 7 8 4 3 4 4 1 0 3 1 7 4 5 1 6 2 3 9 2 5 3 2 1 9 2 6 89 4 46
Percentile 0 10 20 30 40 50 70 80 90 99
Uncertainty 110 70 50 35 30 35 50 70 110
District-Wide Classrooms
1 T ) + 6 78 Q Y D W 4
3 %$ F^& AV* (P2M LZSRUCE J H N KJ!@# 0 5 9
Your Classroom
P Q D
GA F LMSYCERU NJ WK H
Proficiency Standards
Fail
Pass
Excellent
Proficiency %ILES
1% to 15%
16% to 90%
91% to
Last Year's Mean Measure
****
Most Recent TIMSS
****
SAT Equivalents 400 500 600 700 800
Measurement Scale 360 400 440 480 520 560 600 640 680 720 760
Uncertainty (SE classroom means) 4 2 1 1 .5 .3 .5 1 2 4
Mean (M), 1 SD (S), 2 SD (T)
T S M S T
Overall Student Percentile 0 10 20 30 40 50 70 80 90 99

*Figure 5: Prototype multilevel comparison report for vertical coherence (see text for a description of the figure's components).*

Figure 5 again repeats the same scale as the prior figures, but now illustrates district-, school-, classroom-, and student-level information in a common frame of reference, alongside proficiency standards and other accountability metrics (TIMSS, PISA, and SAT equivalents expressed in the shared common metric, etc.) in a single snapshot taken at a particular time. This figure coherently combines horizontal and vertical information to make status and change visible and actionable across students within a classroom.

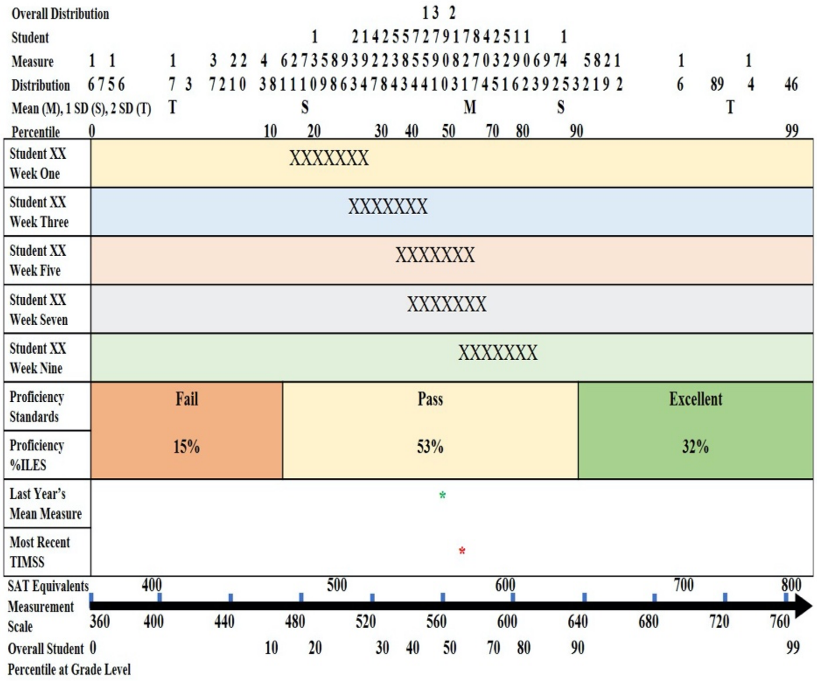

*Figure 6: Prototype individual student progress multilevel comparison report for vertical coherence (see text for a description of the figure's components).*

Figure 6 again repeats the same scale as the prior figures in the context of what might come to be a learning-oriented replacement for report cards, grades, and test scores. An individual student's measures in one subject are plotted over the course of a semester in association with the district-wide measure distribution and percentiles in a single end-of-semester snapshot. This figure coherently combines horizontal and vertical information to make status and change visible and actionable for a single student.

## 4.0 Discussion: Information infrastructures for developmental, horizontal, and vertical coherence

The components of an information infrastructure that coordinates developmental, horizontal, and vertical forms of coherence is illustrated in Figure 1. The prototypes presented in Figures 2-6 represent various segments within the x-, y-, and z-axes of Figure 1. This spatial organization assists in ensuring that each kind of coherence is coordinated with the other two. Within the developmental and horizontal axes, for instance, comparison on learning outcomes can be made across classes either at a given time point or longitudinally, across time points. Similarly, developmentally coherent comparisons of learning outcomes can be made across time points for individual students, within a class of students, or across classes of students. A time point can be taken as a collection of learning evidences over a period of instruction. With this infrastructure, information about learning can be coherently and systematically coordinated across contexts and aspects.

Figure 7 illustrates an example of how learning outcomes can be documented developmentally and horizontally. Selecting a two-dimensional space within Figure 1, individual student scores from a series of tests administered at different times (after different periods of instruction) pertaining to a common set of learning objectives can be scaled and compared, providing developmental coherence. Similarly, the learning outcomes can also be scaled and compared across classrooms, providing horizontal coherence.

Each of the three levels of complexity are accessible from within each form of coherence. Denotative information on individual students' correct and incorrect responses (Figure 2), for instance, is consolidated as metalinguistic information on measured performance in classroom-level documentation and comparisons. Metacommunicative justifications for instructional decisions, course placements, graduation, admissions, etc. will be supported not only by theoretical models capable of explaining variation, but also by the repeated instructional value obtained from the denotative facts of students' individual responses to questions, and by the capacity to calibrate an instrument from repeated experimental tests across varying samples of students and items of the hypothesis that a metalinguistic quantitative unit can be identified.

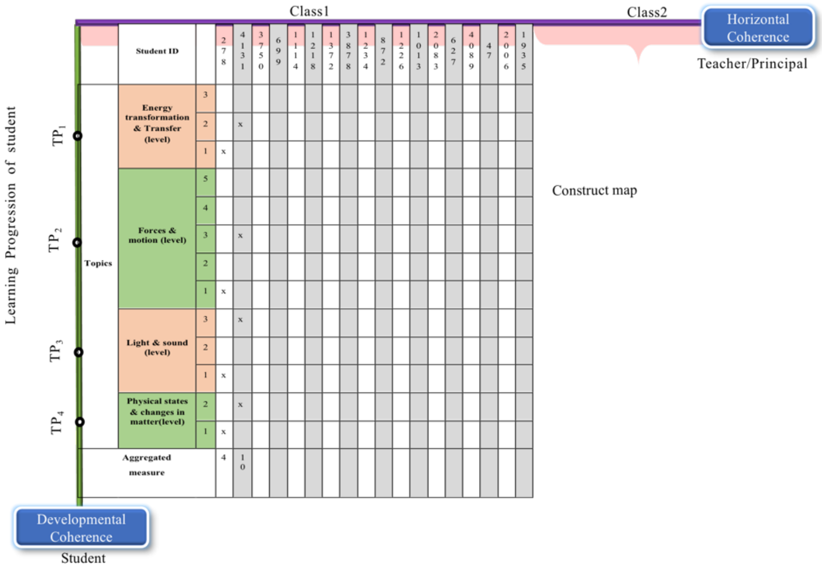

*Figure 7: Horizontal coherence in relation to developmental coherence (see text for a description of the figure's components).*

Figure 7 shows four formative assessments conducted at four different time points (TP1-TP4) in a semester covering content areas under four topics: (a) Energy transformation and transfer, (b) Forces and motion, (c) Light and sound, and (d) Physical states and changes in matter. The x under the student ID shows the relative understanding level of the student for each topic ranging from low/novice to high/expert levels (for illustrative purposes we have shown data for two students, 278 and 4131). Information at this level is contextualised based on learning theory, from which metacommunicative justifications are provided by explanatory models and demonstrated predictive theory.

On 'Force and motion', for example, learning outcomes are interpreted based on a metacommunicative account of various levels of understanding (Alonzo & Steedle, 2009, p. 39; cf. di Uccio et al., 2020; Wilson, 2009; Wind et al., 2019):

Level 0: No evidence or way off-track

Level 1: Student understands forces as a push or pull, but believes that only living or supernatural things can cause forces.

Level 2: Student recognizes that forces can be caused by nonliving things; however, student may believe that forces reside within moving objects.

Level 3: Student recognizes that forces are not contained within moving objects; however, student believes that motion implies a force in the direction of motion and that non-motion implies no force.

Level 4: Student understands that an object is stationary either because there are no forces acting on it or because there is no net force acting on it. However, student may have misconceptions related to a belief that the applied force is proportional to an object's speed or motion (rather than its acceleration). Student can use phrases such as "equal and opposite reaction" to justify the existence of no net forces but may not understand this as an interaction.

Level 5: Student understands that the net force applied to an object is proportional to its resulting acceleration (change in speed or direction), and that this force may not be in the direction of motion. Student understands forces as an interaction between two objects.

For other concepts the expected level of cognitive understanding based on the intended learning outcomes may include few or lower expectations. These abstract conceptualizations of understandings about force and motion provide information on where a student stands relative to intended learning outcomes. This construct map (Wilson, 2005, 2009) illustrates progress in learning over time (developmental coherence) and across classrooms (horizontal coherence). In addition, construct maps set up information sources for teachers to use in formulating useful and timely feedback for students. This improved coherence in documenting learning enhances classroom feedback and shifts the focus away from grades to more authentically serve the purposes of both formative assessment (to facilitate learning) and summative assessment (to provide information on where students stand relative to learning outcomes), at individual, class, and curriculum levels.

## 5.0 Conclusion

This application of Design Thinking began by empathizing with students and teachers as to the disempowered positions in which they are placed by assessments exclusively focused on summative and accountability applications. Seeing assessment from their point of view led to the identification of the possibly insoluble problems of coherence and complexity. Study and consideration of these problems focused attention on distinctions between developmental, horizontal, and vertical forms of coherence (Wilson, 2004; National Research Council, 2006), and between denotative, metalinguistic, and metacommunicative levels of complexity (Bateson, 1972; Star & Ruhleder, 1996).

Examples of prototypes addressing each form of coherence at each level of complexity illustrate practical means by which the dissonant conflict of competing demands on students and teachers might be resolved. By not treating ordinal test scores as though they are measured quantities, we open up rarely explored possibilities for reporting assessment responses in authentic contexts that give them otherwise inaccessible meaning. The prototypes suggest the viability of an experimental learning science focused on the evidence of trial and error assays mediated by instruments calibrated to unit quantity standards explained by predictive theory.

Testing of these prototypes is proceeding in conjunction with development of a computerized network providing access to item banks, assessment assembly and administration, response scoring and analysis, and reporting, extending previously described systems of this kind (Wilson & Scalise, 2015; Wilson & Sloane, 2000; Fisher & Wilson, 2015, 2020; Torres Irribarra, Freund, Fisher, & Wilson, 2015).

Resolving problems of coherence in ways sensitive to meaningfulness and to locally situated forms of knowledge demands close attention to varying levels of complexity in language. Historically, information infrastructures bureaucratically impose homogeneous uniformities from the top down and are insensitive to (a) the relational structures through which words acquire general meanings, and (b) the creative improvisations of practitioners adapting to the demands of the lived moment (Scott, 1998). These insensitivities render classroom communities more fragmented and less effective in achieving the desired learning outcomes than they otherwise might be (Ladd, 2017).

An alternative approach to the design of information infrastructures begins from the bottom up as a model of individual response processes informing variations on invariant relationships that can be meaningfully mapped on a number line (Andrich, 1978a; Andrich & Marais, 2019; Andrich, 2010; Bond & Fox, 2015; Fisher & Wright, 1994; Wilson, 2004; 2005, 2009; Wright, 1999). The discontinuous transformation of discrete responses into a continuous scale contextualizes and characterizes denotative statements of concrete fact in metalinguistic statements expressing those facts in an abstract language meaningful at a general level. This generality provides a basis for an experimental learning science in which researchers' and practitioners' subjective impressions, biases, and hunches can be tested not only against each other but against objectively reproducible learning progressions and measuring instruments. The complexity of the boundary object as simultaneously concrete, abstract, and formal means that the testing of subjective biases is not reduced to a solely quantitative method or model. Meaningful exceptions can provide insights into how learning processes can assert themselves as active agents in their own right. A learning science incorporating the multilevel complexity of boundary objects may take as a primary goal the infrastructuring of new simultaneously concrete, abstract, and formal standards for research and practice (Fisher & Wilson, 2015). These standards could be defined via experimental tests as exhibiting the stability needed to serve as multilevel media for scaffolded cognition and externalized remembering.

Managing the discontinuities between concrete factual data and abstract collective response patterns, and between a third formal level and those two concrete and abstract levels, is the point of individual-level measurement models positing invariant relationships between inferentially separable parameters (Andersen, 1977; Andrich, 2010). Explanatory models (DeBoeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Stenner et al., 2013) predicting item calibrations take matters to a formal level of complexity sustaining metacommunicative statements about statements. When this bottom up process is completed, coherent but nonlinear relationships between the three levels of complexity can be discerned and used to guide curriculum and pedagogy in meaningful ways. Instead of framing research results and accountability standards in terms of concepts and representations wrongly assumed to always mean the same thing across levels of complexity, they can be framed so as to respect and leverage the opportunities for communicating and improving learning outcomes afforded within each of those levels. As systems supporting coherent information of this kind are put in place, it is reasonable to expect new levels of trust to emerge among teachers, students, parents, researchers, administrators, and the public as they experience repeatedly demonstrated, reliable associations between learning processes and outcomes.

## 6.0 References

Ackerman, M. S. (2000). The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. Human-Computer Interaction, 15(2-3), 179–203.

Akkerman, S. F., Bronkhorst, L. H., & Zitter, I. (2013). The complexity of educational design research. Quality & Quantity, 47(1), 421–439.

Akkerman, S., & Bruining, T. (2016). Multilevel boundary crossing in a professional development school partnership. Journal of the Learning Sciences, 25(2), 240–284.

Akkerman, S., van den Bossche, P., Admiraal, W., Gijselaers, W., Segers, M., Simons, R.-J. et al. (2007). Reconsidering group cognition: From conceptual confusion to a boundary area between cognitive and socio-cultural perspectives? Educational Research Review, 2, 39–63.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. Science Education, 93(3), 389–421. https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.20303

Andersen, E. B. (1977). Sufficient statistics and latent trait models. Psychometrika, 42(1), 69–81. https://link.springer.com/article/10.1007/BF02293746

Andrich, D. (1978a). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573. https://link.springer.com/article/10.1007/BF02293814

Andrich, D. (1978b). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, 2, 449–460.

Andrich, D. (1988). Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07-068: Rasch models for measurement. Beverly Hills, California: Sage Publications.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. Psychometrika, 75(2), 292–308. https://link.springer.com/article/10.1007/s11336-010-9154-8

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory: Measuring in the educational, social, and health sciences. Cham, Switzerland: Springer.

Andrich, D., Sheridan, B., & Luo, G. (2017). RUMM 2030: Rasch unidimensional models for measurement. Perth, Australia: RUMM Laboratory Pty Ltd [www.rummlab.com.au].

Andrich, D., & Styles, I. M. (2011). Distractors with information in multiple-choice items: A rationale based on the Rasch model. Journal of Applied Measurement, 12(1), 67–95.

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

Barney, M., & Fisher, W. P., Jr. (2016). Adaptive measurement and assessment. Annual Review of Organizational Psychology and Organizational Behavior, 3, 469–490. https://www.annualreviews.org/doi/pdf/10.1146/annurev-orgpsych-041015-062329

Bateson, G. (1972). Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology. Chicago: University of Chicago Press.

Beckner, C., Blythe, R., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. Language Learning, 59(s1), 1–26. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2009.00533.x

Benson, J., & Dresdow, S. (2014). Design thinking: A fresh approach for transformative assessment practice. Journal of Management Education, 38(3), 436–461.

Beretvas, N., & Kamata, A. (Guest editors) 2007. Part II. Multi-level measurement Rasch models. In E. V. Smith, Jr. & R. M. Smith (Eds.), Rasch measurement: Advanced and specialized applications (pp. 291–470). Maple Grove, MN: JAM Press.

Berg, M., & Timmermans, S. (2000). Order and their others: On the constitution of universalities in medical work. Configurations, 8(1), 31–61. https://muse.jhu.edu/article/8183/summary

Bélanger, F., Cefaratti, M., Carte, T., & Markham, S. E. (2014). Multilevel research in information systems: Concepts, strategies, problems, and pitfalls. Journal of the Association for Information Systems, 15(9), 614–650.

https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1686&context=jais

Bizouarn, P. (2016, May). L'éco-épidémiologie: Vers une épidémiologie de la complexité. Médicine/Sciences, 32(5), 500–505. https://www.medecinesciences.org/en/articles/medsci/full_html/2016/06/medsci20163205p500/medsci20163205p500.html

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7–74. https://www.tandfonline.com/doi/abs/10.1080/0969595980050102

Black, P., & Wiliam, D. (2003). In praise of educational research: Formative assessment. British Educational Research Journal, 29(5), 623–637. https://www.tandfonline.com/doi/abs/10.1080/0141192032000133721

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

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. Measurement: Interdisciplinary Research and Perspectives, 9, 1–52.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (Eds.). (1971). Handbook on the formative and summative evaluation of student learning. New York, NY: McGraw-Hill.

Boaler, J., Wiliam, D., & Brown, M. L. (2000). Students' experiences of ability grouping - disaffection, polarisation and the construction of failure. British Educational Research Journal, 27(5), 631–648. https://www.tandfonline.com/doi/abs/10.1080/713651583

Bond, T., & Fox, C. (2015). Applying the Rasch model: Fundamental measurement in the human sciences, 3rd edition. New York: Routledge.

Bowker, G. C. (2016). How knowledge infrastructures learn. In P. Harvey, C. B. Jensen, & A. Morita (Eds.), Infrastructures and social complexity: A companion (pp. 391–403). New York: Routledge.

Bowker, G., Timmermans, S., Clarke, A. E., & Balka, E. (Eds). (2015). Boundary objects and beyond: Working with Leigh Star. Cambridge, MA: MIT Press.

Briggs, D., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. Journal of Applied Measurement, 4(1), 87–100. https://nepc.colorado.edu/publication/an-introduction-multidimensional-measurement-using-rasch-models

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. The Journal of The Learning Sciences, 2(2), 141–178.

Chien, T.-W., Linacre, J. M., & Wang, W.-C. (2011). Examining student ability using KIDMAP fit statistics of Rasch analysis in Excel. In H. Tan & M. Zhou (Eds.), Communications in Computer and Information Science: Vol. 201. Advances in Information Technology and Education, CSE 2011 Qingdao, China Proceedings, Part I (pp. 578–585). Berlin: Springer.

Choppin, B. (1968). An item bank using sample-free calibration. Nature, 219, 870–872.

Choppin, B. (1976). Recent developments in item banking. In D. N. M. DeGruitjer & L. J. van der Kamp (Eds.), Advances in Psychological and Educational Measurement (pp. 233–245). New York: Wiley.

Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), New directions in educational technology (pp. 15–22). New York: Springer.

Cook, T. A. (1914/1979). The curves of life. New York: Dover.

Crick, R. D., Barr, S., Green, H., & Pedder, D. (2017). Evaluating the wider outcomes of schools: Complex systems modelling for leadership decisioning. Educational Management Administration & Leadership, 45(4), 719–743. https://journals.sagepub.com/doi/abs/10.1177/1741143215597233

Crowder, M. K., Gordon, R. A., Brown, R. D., Davidson, L. A., & Domitrovich, C. E. (2019). Linking social and emotional learning standards to the WCSD Social–Emotional Competency Assessment: A Rasch approach. School Psychology, 34(3), 281.

Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral reasoning stages revisited. International Journal of Behavioral Development, 26(2), 154–166.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. Journal of Adult Development, 11(2), 71–85. https://link.springer.com/article/10.1023/B:JADE.0000024541.84265.04

De Boeck, P., & Wilson, M. (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. Statistics for Social and Behavioral Sciences. New York: Springer.

Design-Based Research Collective. (2003). Design-Based Research: An emerging paradigm for educational inquiry. Educational Researcher, 32(1), 5–8. https://doi.org/10.3102/0013189X032001005

di Uccio, U. S., Colantonio, A., Galano, S., Marzoli, I., Trani, F., & Testa, I. (2020). Development of a construct map to describe students' reasoning about introductory quantum mechanics. Physical Review Physics Education Research, 16(1), 010144.

Embretson, S. E. (2010). Measuring psychological constructs: Advances in model-based approaches. Washington, DC: American Psychological Association.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5(3), 171–191.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359–374. https://www.sciencedirect.com/science/article/pii/0001691873900036

Fischer, G., Giaccardi, E., Eden, H., Sugimoto, M., & Ye, Y. (2005). Beyond binary choices: Integrating individual and social creativity. International Journal of Human-Computer Studies, 63, 482–512. https://www.sciencedirect.com/science/article/pii/S1071581905000479

Fisher, W. P., Jr. (2004). Proper measurement is universally reproducible. Rasch Measurement Transactions, 18(1), 967 [https://www.rasch.org/rmt/rmt181e.htm].

Fisher, W. P., Jr. (2017a). A practical approach to modeling complex adaptive flows in psychology and social science. Procedia Computer Science, 114, 165–174. https://doi.org/10.1016/j.procs.2017.09.027

Fisher, W. P., Jr. (2017b). Provoking professional identity development: The legacy of Benjamin Drake Wright. In M. Wilson & W. P. Fisher, Jr. (Eds.), Psychological and social measurement: The career and contributions of Benjamin D. Wright (pp. 135–162). New York: Springer.

Fisher, W. P., Jr., Oon, E. P.-T., & Benson, S. (2018). Applying Design Thinking to systemic problems in educational assessment information management. Journal of Physics Conference Series, 1044, 012012 [http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012012].

Fisher, W. P., Jr., & Stenner, A. J. (2016). Theory-based metrological traceability in education: A reading measurement network. Measurement, 92, 489–496. http://www.sciencedirect.com/science/article/pii/S0263224116303281

Fisher, W. P., Jr., & Stenner, A. J. (2018). Ecologizing vs modernizing in measurement and metrology. Journal of Physics Conference Series, 1044(012025), [http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012025].

Fisher, W. P., Jr., & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana, 52(2), 55–78. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2688260

Fisher, W. P., Jr., & Wilson, M. (2020). An online platform for sociocognitive metrology: The BEAR Assessment System Software. Measurement Science and Technology, 31(034006). https://iopscience.iop.org/article/10.1088/1361-6501/ab5397/meta

Fisher, W. P., Jr., & Wright, B. D. (1994). Introduction to probabilistic conjoint measurement theory and applications (W. P. Fisher, Jr., & B. D. Wright, Eds.) [Special issue]. International Journal of Educational Research, 21(6), 559–568. https://www.sciencedirect.com/science/article/pii/0883035594900108

Fortus, D., & Krajcik, J. (2012). Curriculum coherence and learning progressions. In D. Fortus & J. Krajcik (Eds.), Second international handbook of science education (pp. 783–798). Dordrecht, the Netherlands: Springer (pp. 783–798). Dordrecht, the Netherlands: Springer.

Fullan, M., & Quinn, J. (2015). Coherence: The right drivers in action for schools, districts, and systems. Thousand Oaks, California: Corwin Press.

Gadamer, H.-G. (1989). Truth and method (J. Weinsheimer & D. G. Marshall, Trans.) (Rev. ed.). New York: Crossroad.

Galison, P. (2008). Image of self. In L. Daston (Ed.), Things that talk: Object lessons from art and science (pp. 256–294). New York: Zone Books.

Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. Osiris, 27(1), 19–36. https://www.journals.uchicago.edu/doi/abs/10.1086/667821

Gorin, J. S., & Mislevy, R. J. (2013). Inherent measurement challenges in the next generation science standards for both formative and summative assessment (K-12 Center at Educational Testing Service. Invitational Research Symposium on Science Assessment). Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf

Greaves, D. (1999, September). Meeting the educational needs of students with learning difficulties: A sociological study of three schools in Victoria. Australian Journal of Learning Disabilities, 4(3), 12–20.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. New York: Routledge.

He, W., Li, S., & Kingsbury, G. G. (2016). A large-scale, long-term study of scale drift: The micro view and the macro view. Journal of Physics Conference Series, 772, 012022. http://adsabs.harvard.edu/abs/2016JPhCS.772a2022H

Holster, T. A., & Lake, J. W. (2015). From raw scores to Rasch in the classroom. Shiken, 19(1), 32–41. http://teval.jalt.org/sites/teval.jalt.org/files/19-01-32%20Holster%20Lake%20Rasch%20in%20Classroom.pdf

Ihde, D. (1991). Instrumental realism: The interface between philosophy of science and philosophy of technology. Bloomington, Indiana: Indiana University Press.

Kali, Y., Eylon, B. S., McKenney, S., & Kidron, A. (2018). Design-centric research-practice partnerships: Three key lenses for building productive bridges between theory and practice. In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), Learning, design, and technology: An international compendium of theory, research, practice, and policy (pp. 1–30). Cham: Springer. https://doi.org/10.1007/978-3-319-17727-4_122-1

Ketterlin-Geller, L. R., Shivraj, P., Basaraba, D., & Yovanoff, P. (2019). Considerations for using mathematical learning progressions to design diagnostic assessments. Measurement: Interdisciplinary Research and Perspectives, 17(1), 1–22.

Kuhn, T. S. (1961). The function of measurement in modern physical science. Isis, 52(168), 161–193. (Rpt. in T. S. Kuhn. (1977). The essential tension: Selected studies in scientific tradition and change (pp. 178–224). Chicago: University of Chicago Press.

Ladd, H. F. (2017). No Child Left Behind: A deeply flawed federal policy. Journal of Policy Analysis and Management, 36(2), 461–469. https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.21978

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. New York: Harvard University Press.

Latour, B. (1998). To modernise or ecologise? That is the question. In B. Braun & N. Castree (Eds.), Remaking reality: Nature at the millennium (pp. 221–242). London: Routledge.

Latour, B. (2004). How to talk about the body? The normative dimension of science studies. Body & Society, 10(2-3), 205–229.

Latour, B. (2005). Reassembling the social: An introduction to Actor-Network-Theory. Oxford, England: Oxford University Press.

Linacre, J. M. (2019). A user's guide to WINSTEPS Rasch-Model computer program, v. 4.3.3. Chicago, Illinois: Winsteps.com.

Linacre, J. M., Engelhard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. International Journal of Educational Research, 21(6), 569–577. https://www.sciencedirect.com/science/article/pii/0883035594900116

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. Measurement, 51, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2020). Measurement across the sciences (R. Morawski, G. Rossi et al., Eds.). Springer Series in Measurement Science and Technology. Cham: Springer.

Marton, F., & Pang, M. F. (2013). Meanings are acquired from experiencing differences against a background of sameness, rather than from experiencing sameness against a background of difference: Putting a conjecture to the test by embedding it in a pedagogical tool. Frontline Learning Research, 1(1), 24–41. https://doi.org/10.14786/flr.v1i1.16

Mason, M. (Ed.). (2008). Complexity theory and the philosophy of education. West Sussex, UK: Wiley Blackwell.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174. https://link.springer.com/article/10.1007/BF02296272

Masters, G. N. (1985). Common-person equating with the Rasch model. Applied Psychological Measurement, 9(1), 73–82. https://journals.sagepub.com/doi/abs/10.1177/014662168500900107

Masters, G. N., Adams, R. J., & Lokan, J. (1994). Mapping student achievement. International Journal of Educational Research, 21(6), 595–610.

McQuillan, P. J., Welch, M. J., & Barnatt, J. (2012). In search of coherence: "Inquiring" at multiple levels of a teacher education system. Educational Action Research, 20, 535–551. https://www.tandfonline.com/doi/abs/10.1080/09650792.2012.727640

Mead, R. J. (2009). The ISR: Intelligent Student Reports. Journal of Applied Measurement, 10(2), 208–224.

Miller, P. N. (2015, April 3). Is 'Design Thinking' the new liberal arts? The Chronicle of Higher Education, 61(29). Retrieved from http://www.chronicle.com/article/Is-Design-Thinking-the-New/228779/

Morrison, J., & Fisher, W. P., Jr. (2018). Connecting learning opportunities in STEM education: Ecosystem collaborations across schools, museums, libraries, employers, and communities. Journal of Physics: Conference Series, 1065(022009). https://doi.org/10.1088/1742-6596/1065/2/022009

Morrison, J., & Fisher, W. P., Jr. (2019). Measuring for management in Science, Technology, Engineering, and Mathematics learning ecosystems. Journal of Physics: Conference Series, 1379(012042). https://doi.org/10.1088/1742-6596/1379/1/012042

Morrison, J., & Fisher, W. P., Jr. (2020, September 1). The Measure STEM Caliper Development Initiative [Online]. BEAR Seminar Series. BEAR Center, Graduate School of Education: University of California, Berkeley. http://bearcenter.berkeley.edu/seminar/measure-stem-caliper-development-initiative-online

Moss, P. (2004). The risks of coherence. In M. Wilson (Ed.), Towards coherence between classroom assessment and accountability (pp. 217–238). Chicago: University of Chicago Press.

National Research Council. (2006). Systems for state science assessment (M. R. Wilson & M. W. Bertenthal, Eds.). Washington, DC: The National Academies Press.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. Social Studies of Science, 23, 129–173. https://journals.sagepub.com/doi/abs/10.1177/030631293023001005

Overton, W. F. (2015). Processes, relations and Relational-Developmental-Systems. In W. F. Overton & P. C. M. Molenaar (Eds.), Theory and Method. Volume 1 of the Handbook of child psychology and developmental science (7th Ed.) (pp. 9–62). Hoboken, NJ: Wiley.

Öztürk-Gübes, N., & Kelecioglu, H. (2016, June). The impact of test dimensionality, common-item set format, and scale linking methods on mixed-format test equating. Educational Sciences: Theory & Practice, 16, 715–734. https://eric.ed.gov/?id=EJ1115021

Palmgren, E. (2018). A constructive examination of a Russell-style ramified type theory. Bulletin of Symbolic Logic, 24(1), 90–106. https://www.cambridge.org/core/journals/bulletin-of-symbolic-logic/article/constructive-examination-of-a-russellstyle-ramified-type-theory/51A828F64C72009D9AE1286FBA0CB0ED

Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. NCSLi Measure: The Journal of Measurement Science, 9(4), 22–33. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702

Pendrill, L. (2019). Quality assured measurement: Unification across social and physical sciences. Cham: Springer.

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. Measurement, 71, 46–55. http://dx.doi.org/10.1016/j.measurement.2015.04.010

Penuel, W. R., Riedy, R., Barber, M. S., Peurach, D. J., LeBouef, W. A., & Clark, T. (2020). Principles of collaborative education research with stakeholders: toward requirements for a new research and development infrastructure. Review of Educational Research, 90(5), 627–674.

Pilcher, J. K. (1994). The value-driven meaning of grades. Educational Assessment, 2(1), 69–88. https://www.tandfonline.com/doi/abs/10.1207/s15326977ea0201_4

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Ratnam-Lim, C. T. L., & Tan, K. H. K. (2015). Large-scale implementation of formative assessment practices in an examination-oriented culture. Assessment in Education: Principles, Policy & Practice, 22(1), 61–78. https://www.tandfonline.com/doi/abs/10.1080/0969594X.2014.1001319

Russell, B. (1908). Mathematical logic as based on the theory of types. American Journal of Mathematics, 30, 222–262. https://www.jstor.org/stable/2369948

Scalise, K., Douskey, M., & Stacy, A. (2018). Measuring learning gains and examining implications for student success in STEM. Higher Education Pedagogies, 3(1), 183–195.

Scott, J. C. (1998). Seeing like a state: How certain schemes to improve the human condition have failed. New Haven: Yale University Press.

Spinelli, S., Jr., & Nixon, N. (2015, April 17). Letters to the editor: Not 'the new liberal arts.' The Chronicle of Higher Education, 61(31). Retrieved from http://www.chronicle.com/article/Not-The-New-Liberal-Arts-/229225

Squire, K. D., MaKinster, J. G., Barnett, M., Luehmann, A. L., & Barab, S. L. (2003). Designed curriculum and local culture: Acknowledging the primacy of classroom culture. Science Education, 87(4), 468–489. https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.10084

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations,' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science, 19(3), 387–420. https://journals.sagepub.com/doi/abs/10.1177/030631289019003001

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1), 111–134. https://pubsonline.informs.org/doi/abs/10.1287/isre.7.1.111

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. Frontiers in Psychology: Quantitative Psychology and Measurement, 4(536), 1–14 [https://doi.org/10.3389/fpsyg.2013.00536].

Susser, M., & Susser, E. (1996). Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology [see comments] [published erratum appears in Am J Public Health 1996 Aug;86(8 Pt 1):1093]. American Journal of Public Health, 86(5), 674–677. https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.86.5.674

Tan, C. (2012). The culture of education policy making: Curriculum reform in Shanghai. Critical Studies in Education, 53(2), 153–167. https://www.tandfonline.com/doi/abs/10.1080/17508487.2012.672333

Tan, C., & Hairon, S. (2016). Education reform in China: Toward classroom communities. Action in Teacher Education, 38(4), 315–326. https://www.tandfonline.com/doi/abs/10.1080/01626620.2016.1226205

Torres Irribarra, D., Freund, R., Fisher, W. P., Jr., & Wilson, M. (2015). Metrological traceability in education: A practical online system for measuring and managing middle school mathematics instruction. Journal of Physics: Conference Series, 588, 012042. https://iopscience.iop.org/article/10.1088/1742-6596/588/1/012042/meta

Vonderau, A. (2018). Scaling the cloud: Making state and infrastructure in Sweden. Ethnos Journal of Anthropology. https://doi.org/10.1080/00141844.2018.1471513

Whitehead, A. N., & Russell, B. (1910, 1912, 1913). Principia Mathematica, 3 vols. Cambridge, England: Cambridge University Press.

Whorf, B. L. (1956). Language, thought, and reality: Selected writings of Benjamin Lee Whorf (J. B. Carroll, Ed.) (Foreword by Stuart Chase). Cambridge, Massachusetts, New York, and London: Published jointly by The Technology Press at MIT; John Wiley & Sons, Inc.; and Chapman & Hall, Ltd.

Williamson, G. (2018). Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data. Cogent Education, 5(1464424), 1–29. https://doi.org/10.1080/2331186X.2018.1464424

Wilson, M. R. (2017). From the president. NCME Newsletter, 25(1), 1–2. Retrieved from https://www.ncme.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=94b0192d-9c95-901b-c6b4-4ae9015f0feb&forceDialog=1

Wilson, M. (Ed.). (2004). National Society for the Study of Education Yearbooks. Vol. 103, Part II: Towards coherence between classroom assessment and accountability. Chicago, Illinois: University of Chicago Press.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilson, M. R. (2009). Measuring progressions: Assessment structures underlying a learning progression. Journal of Research in Science Teaching, 46, 716–730.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. Educational Measurement: Issues and Practice, 37(1), 5–20.

Wilson, M., & Fisher, W. (2016). Preface: 2016 IMEKO TC1-TC7-TC13 Joint Symposium: Metrology across the Sciences: Wishful Thinking? Journal of Physics Conference Series, 772(1), 011001.

Wilson, M., & Fisher, W. (2018). Preface of special issue Metrology across the Sciences: Wishful Thinking? Measurement, 127, 577.

Wilson, M., & Scalise, K. (2015). Assessment of learning in digital networks. In P. Griffin, E. Care (Eds). Assessment and Teaching of 21st Century Skills: Methods and Approach (pp. 57-81). Dordrecht: Springer Netherlands.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied Measurement in Education, 13, 181-208.

Wind, S. A., Alemdar, M., Lingle, J. A., Moore, R., & Asilkalkan, A. (2019). Exploring student understanding of the engineering design process using distractor analysis. International Journal of STEM Education, 6(1), 1–18.

Wittgenstein, L. (1922). Tractatus logico-philosophicus. London: Harcourt Brace.

Wright, B. D. (1958, July 1). On behalf of a personal approach to learning. The Elementary School Journal, 58, 365–375. (Rpt. in M. Wilson & W. P. Fisher, Jr., (Eds.). (2017). Psychological and social measurement: The career and contributions of Benjamin D. Wright (pp. 221–232). New York: Springer Nature.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14(2), 97–116 [http://www.rasch.org/memo42.htm].

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every educator and psychologist should know (pp. 65–104 [http://www.rasch.org/memo64.htm]). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. Journal of Educational Measurement, 21(4), 331–345 [http://www.rasch.org/memo43.htm].

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago, Illinois: MESA Press.

Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). KIDMAP: person-by-item interaction mapping (Tech. Rep. No. MESA Memorandum #29). Chicago: MESA Press [http://www.rasch.org/memo29.pdf].

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago, Illinois: MESA Press.

Wright, B. D., & Stone, M. H. (1999). Measurement essentials. Wilmington, DE: Wide Range, Inc. [http://www.rasch.org/measess/me-all.pdf].

Wu, M. L., Adams, R. J., & Wilson, M. R. (2015). ACER ConQuest Version 4: Generalized item response modelling software. Camberwell, Victoria, Australia: Australian Council for Educational Research.

Yan, C. (2015). 'We can't change much unless the exams change': Teachers' dilemmas in the curriculum reform in China. Improving Schools, 18(1), 5–19. https://journals.sagepub.com/doi/abs/10.1177/1365480214553744

Zitter, I., De Bruijn, E., Simons, R. J., & Ten Cate, O. (2012). The role of professional objects in technology-enhanced learning environments in higher education. Interactive Learning Environments, 20(2), 119–140.

Zuiker, S., & Whitaker, J. R. (2014). Refining inquiry with multi-form assessment: Formative and summative assessment functions for flexible inquiry. International Journal of Science Education, 36(6), 1037–1059. https://doi.org/10.1080/09500693.2013.834489

## Author Profile

**William P. Fisher, Jr.** is a Research Associate with the BEAR Center in the Graduate School of Education at the University of California, Berkeley, USA. He is also a Research Scientist with the Research Institute of Sweden, in Gothenburg, Sweden, and he is the Principal and Founder of Living Capital Metrics LLC in Sausalito, CA, USA.

**Emily Pey-Tee Oon** is an Associate Professor in the Faculty of Education at the University of Macau, an SAR of China, where she focuses on science education and advanced applications of measurement and assessment.

**Spencer Benson** advises educators globally via his Educational Innovations International Consulting practice, after past lives as Director of the Centre for Teaching and Learning Enhancement, Faculty of Education at the University of Macau, and as a faculty member of the School of Education at the University of Maryland, USA.

## Author Details

**William P. Fisher, Jr.**
BEAR Center, Graduate School of Education
University of California, Berkeley
CA 94720-5800
USA
+1-919-599-7245
wfisher@berkeley.edu

**Emily Pey-Tee Oon**
University of Macau
Avenida da Universidade, Taipa, Macau
China
+853 8822 8789 / 8776
PeyTeeOon@um.edu.mo

**Spencer Benson**
Educational Innovations International Consulting
University Park, MD 20782-1188
+1-240-676-8808
sbeson@eii-consulting.com

## Editor Details

**Prof. Dr. Tobias Jenert**
Chair of Higher education and Educational Development
University of Paderborn
Warburger Straße 100
Germany
+49 5251 60-2372
Tobias.Jenert@upb.de

Journal Details

EDeR – Educational Design Research
An International Journal for Design-Based Research in Education
ISSN: 2511-0667
uhh.de/EDeR
#EDeRJournal (our hashtag on social media services)

Published by

**Hamburg Center for University Teaching and Learning (HUL)**
University of Hamburg
Schlüterstraße 51
20146 Hamburg
Germany
+49 40 42838-9640
+49 40 42838-9650 (fax)
EDeR.HUL@uni-hamburg.de
hul.uni-hamburg.de

In collaboration with

**Hamburg University Press**
Verlag der Staats- und Universitätsbibliothek Hamburg –
Landesbetrieb
Von-Melle-Park 3
20146 Hamburg
Germany
+49 40 42838 7146
info.hup@sub.uni-hamburg.de
hup.sub.uni-hamburg.de