Establishing assessment scales using a novel disciplinary rationale for scientific reasoning

Per Morten Kind

School of Education, Durham University, Leazes Road, Durham DH1 1TA, United Kingdom

Published in: JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 50, NO. 5, PP. 530–560 (2013)

Abstract

The paper argues that science assessment should change from an item-driven to a construct-driven practice and pay more attention to disciplinary scientific reasoning. It investigates assessment scales developed from a novel theoretical rationale, describing scientific reasoning as three fundamental practices (hypothesising, experimenting and evidence evaluation) and building on three types of knowledge (science content knowledge, procedural knowledge and epistemic knowledge). The scale development follows a construct-driven approach by, first, detailing the knowledge involved and explaining progression; and second, operationalising the theoretical construct into items and score criteria. The scales are trialled in a small-scale study. The outcome is a coherent and supportive 'validity argument' for three sub-scales, but with a suggestion that merging these into one scale has higher validity. The main implication is rewriting rationales for many science assessments, including TIMSS, which emphasises domain-general reasoning, and NAEP and PISA, which pay attention to domain-specific reasoning but are unclear about the knowledge involved.

Key words: Assessment, Scientific reasoning, Scientific Argumentation, Science Inquiry.

Introduction

This paper presents a study into assessment of scientific reasoning motivated by two major challenges. These are first, the 'positivist' view of science dominating classroom practices, placing emphasis on factual recall and confirmatory experiments (Driver, Newton, & Osborne, 2000; Layton, 1973; Weiss, Pasely, Sean Smith, Banilower, & Heck, 2003) and second, the tendency in science assessment to use items making lower level cognitive demands of recall and comprehension as opposed to the higher level demands of synthesis and evaluation (J. Osborne & Ratcliff, 2002; Pellegrino, Chudowsky, & Glaser, 2001; Shavelson & Ruiz-Primo, 1999; Mark Wilson & Bertenhal, 2005). According to Au's (2007) meta-analysis these two challenges are related, because teachers match expectations of student performance to those of the test. Improving assessment of scientific reasoning is, therefore, an important step towards changing classroom practices.

The main argument of the paper is that improving assessment of scientific reasoning should start with, and work from, an improved construct definition. This is, in part, based on Wiliam's (2010) claim that assessors tend to mix up the two uses of the term *construct* identified by Wiley (2001, p. 212). These are, first, "to name the psychological characteristics actually estimated by an existing test

score or other measurement" and, second, "to name the psychological characteristics that a test score or other measurement is intended ('designed') to measure". By allowing the *actual* test score to define the construct, Wiliam claims, assessment becomes 'item-driven' rather than 'construct-driven'. The suggested solution is to lay out a formal construct definition that clarifies what should be included and excluded from an assessment, independently of the test, and then argue systematically how the construct can be operationalised into a measure. Wiliam's argument summarises a key development in assessment theory over the last decades, as can be observed in the 'validity chapters' written by Messick (1989) and Kane (2006). Kane, who introduced the phrase 'validity as argument', explains that the relationship between the intended and operationalised constructs cannot be absolutely determined and therefore requires informal reasoning as described by Toulmin (1958). Mislevy, Wilson, Erickan & Chudowsky (2003) demonstrate in more concrete terms how construct-driven assessment can be put into practice. They suggest constructing a *student model*, identifying and explaining the configuration of skills and knowledge students are intended to learn; a *task model*, describing and explaining the situations in which students should be able to act; a *scoring model*, explaining how students' performance to items can be graded; and a *measurement model*, explaining how student's performance can be transformed to a measure. Other examples working to implement similar perspectives are Wilson (2005), Rupp, Templin and Henson (2010) and Leighton, Gierl and Hunka (2004). All these use 'modern test theory' (Crocker & Algina, 1986), based on Item Response Theory (IRT) and Rasch models to support the construct-driven and evidence-centred assessment approach. They also include what Wilson (2005) terms a 'developmental perspective', namely that assessment models should be aligned with progression in teaching and students' learning.

Another reason for focusing on construct definition to improve assessment and classroom practices is the state of scientific reasoning research. Rather than being dominated by *one* rationale, scientific reasoning is characterised by a multitude of meanings, some conflicting. In particular, the influence of cognitive psychology gives impetus to the 'nothing special view' (Simon, 1966) – that general reasoning abilities account for the main characteristics of scientific reasoning. This contrasts with disciplinary perspectives, such as that of Ford (2008), in which scientific reasoning is seen as a domain-specific practice. A related issue is the knowledge-dependency of reasoning (Koslowski, 1996; Passmore & Stewart, 2002), including debate about what *type* of knowledge is involved (Li & Shavelson, 2001). Discrepancy also exists between psychologists' and philosophers' perspectives on reasoning. Psychologists commonly take a descriptive perspective, focusing on cognitive processes and abilities, while philosophers take a normative perspective, focusing on epistemological principles and values intrinsic to reasoning. Bailin and Siegel (2002) explain that the former group is concerned mainly with *how* someone reasons, while the latter seeks to understand *why* students reason as they do. In science education, these perspectives merge, since students learn to reason scientifically and understand how science works, however, with science educators often subscribing to rationales for scientific reasoning from *either* psychology *or* philosophy. An outcome is to persist with naïve perspectives about the other field. An example, discussed later in the paper, is 'process

science' in the 1970s and 80s, which found support in contemporary psychology but continued logical-empiricist views of science that philosophers had rejected. Hence, assessment needs better guidance about what scientific reasoning means, and also evidence to illustrate how it occurs in students' learning.

The current paper unites the two challenges above, that is, establishing construct-driven assessment with an improved rationale for scientific reasoning, because these are mutually dependent. Previously, the author and colleagues (paper in review) suggested a disciplinary reasoning rationale, aiming for consensus and improved teaching practice by synthesising perspectives in philosophy and psychology. The rationale has already demonstrated relevance through influencing the new K-12 Science Framework in the US (National Research Council, 2012), but has yet to be applied for scale development. The purpose here is, therefore, to establish assessment scales that operationalise the new rationale by following construct-driven ideals. The aimed-for outcome is a validity argument, as suggested by Kane (2006), and based on Mislevy et al.'s (2002) four models (student model, item model, scoring model and measurement model) to explain how scientific reasoning can be implemented in assessment. The aim is also to support this argument with empirical evidence from trialling. The research is significant in two respects: first, it contributes towards understanding implications of construct-driven assessment in science education. Few examples exist, so more research is needed to investigate how construct-driven ideals can be put into practice. Second, the research contributes towards understanding how scientific reasoning may be implemented into assessment to support teaching and learning directly. The research also promotes better understanding of scientific reasoning, by prompting 'dialog' between assessments and educational theory (Bond, 2003).

The paper begins by explaining the rationale for scientific reasoning and arguments behind its development, followed by a section defining scales. The empirical trialling of scales is introduced, and the validity argument is outlined – that is, how the intended construct is operationalised into item-, scoring- and measurement models. The discussion reflects on what has been learned from the trials for construct-driven assessment and development of reasoning scales.

Scientific reasoning

Definitions for scientific reasoning have been proposed, relying on differing understandings of *science* and *reasoning*, thereby drawing on developments in science studies and learning sciences. Prominent in 1960s-1980s science education was the *process approach,* which defined science inquiry as a list of processes and related reasoning to a matching list of cognitive skills. Gagne (1965), then prominent in the US, described his list of science processes as 'behaviours of scientists that could be learned by students'. His accompanying curriculum project, the 'Science – A Process Approach' (American Association for the Advancement of Science, 1965), became influential, demonstrating how processes could be taught in science classrooms. Piaget's (1954) 'stage theory' of cognitive development provided a second contribution to the process rationale, identifying reasoning strategies that children could or could not undertake at different stages. The outcome, a list of general, non-science specific 'cognitive operations', has similarities to Gagne's science processes: one common denominator is the claim that processes and operations

are *knowledge-independent* and therefore generalisable across content domains. Science educators linked these, arguing that learning scientific method developed science reasoning processes, thus improving general cognition and improving students' learning of science knowledge (Lawson, Karplus, & Adi, 1978; Shayer & Adey, 1981).

Dramatic changes in the conception of science method occurred in the 1960s -1980s due to critique of logical empiricism and positivism in science philosophy. This critique became general knowledge among science educators in this period, forming the basis for criticism of the process approach. For example, Finley (1983) attacked Gagne for inadequate understanding of scientific method, on the grounds that he failed to acknowledge Hempel (1966), Quine (1969) and Hanson (1958). These philosophers argue that scientific observations occur only in the context of a conceptual scheme, meaning that all observations are theory-laden, and suggest no algorithm makes it possible to generalize observations into scientific theories. Their arguments shaped a new science philosophy, in which scientific reasoning is *content*- rather than *process-led*. Finley concluded that "if science educators are to understand better the nature of science processes, the relationships between content and process must be understood" (p. 53). Similar critique of Piaget's psychology pointed out that students' abilities to use cognitive operations was strongly dependent on their understanding of content and contexts (Brown & Desforges, 1977; Donaldson, 1984). Millar and Driver (1987) summarised these philosophical and psychological critiques, showing that science education needed a better rationale for scientific reasoning to guide research and teaching practices.

Answering this critique became a central focus, resulting in three research strands. First, research focusing on science misconceptions and conceptual change (Driver & Easley, 1978; Hewson, 1981; R. J. Osborne, 1982) turned attention away from scientific reasoning and claimed, like Ausubel (Ausubel, 1968), that 'what students know' is more important than 'how they reason'. This strand evolved to claim that students' conceptual changes in science learning are similar to scientists' reasoning used to develop new ideas and theories (R.J. Osborne & Wittrock, 1985). Kelly's (1955) metaphor of *man-the-scientist* was an influential analogy, paralleling the way individuals construct personal representations of the world with scientists' construction of theories. Personal representations, Kelly argued, are subject to change over time from constant testing and modification permitting better predictions of real world behaviour. Constant questioning, exploring, revising and replacing in the light of predictive failure is symptomatic of scientific theorising, and is precisely how a person learns to anticipate life events. Many, including McCloskey, (1983) and Pope & Keen (1981) acknowledged the limitations of this analogy, but still supported the notion that students' reasoning when learning science content knowledge is a form of scientific reasoning.

Second, research on reasoning in scientific inquiry tried more directly to improve Gagne's process skills rationale. An important response was presented by Gott and Mashiter (1991) who suggested *procedural knowledge* was a "missing element" (p. 58) in the process approach. Based on Gott and Murphy (1987), they argued that students who failed to solve inquiry tasks often lacked understanding of experimental procedures and concepts. This *procedural knowledge*, together with science conceptual knowledge, they claimed, could

explain Gagne's process skills. Research proposed a framework, or 'taxonomy', for procedural knowledge, describing experimental concepts and ideas for the science curriculum (Duggan & Gott, 1995; Gott & Duggan, 1995, 1996). Millar, Lubben, Gott and Duggan (1994) also provided evidence that procedural understanding, when tested by questionnaire, accounted for 50% or more of the variance observed in students' performance on practical investigative tasks. These authors found that students hold misconceptions about scientific method that negatively influence practical task performance. Thus, this response also defines scientific reasoning as *knowledge-based,* but focuses on experimentation rather than theorisation, and explaining this as reasoning based on a combination of procedural and conceptual knowledge.

A third, more recent strand presents *scientific argumentation* (Driver et al., 2000) as the answer to the critic of the process approach. A rationale for science pedagogy "that is coherent and based on current scholarship and research in the field of science studies and the philosophy of science" (p. 290) should be the new focus for scientific reasoning activity. This meant accepting science knowledge as socially constructed, rather than relying on observations and experiments as the 'bedrock' of science. As replacements, scientific reasoning should rely on *evidence evaluation and coordination,* with *epistemic knowledge* as a pre-requisite knowledge-base (Duschl & Osborne, 2002; G. J. Kelly & Duschl, 2002). Attention, as in the assessment area mentioned earlier, has been drawn towards Toulmin's (1958) model for informal reasoning.

Together, these three strands focus on key scientific practices with which scientists handle different problems. First, they develop scientific theories; next, they gather empirical data used to test the theories; and third, they coordinate and evaluate evidence critically. These practices may occur in different orders but are complementary, as the outcome of one acts as the starting point for the next. An argument raised by this paper is, therefore, that the research strands *collectively* set a rationale for scientific reasoning by identifying three practices, each belonging to a different stage of the inquiry process.

This rationale is elaborated by a psychological and a philosophical model. From a psychological perspective, the individual enters a particular *mode of thinking* in each stage due to the problem they have to solve. This is described by Klahr and Dunbar's (1988) *Scientific Discovery as Dual Search* (SDDS) model, derived from results of psychological experiments (Klahr, 2000) in which candidates (university students and school children) solved simulated scientific inquiry tasks. The model describes the two first phases, *hypothesis generation* and *experimenting*, as 'problem spaces' and the third, *evidence evaluation*, as a coordination of outcomes from these. A problem space comprises states, operators, goals, and constraints, suggesting, in a similar way to Johnson-Laird's (1983) model-based reasoning, that an individual reasons with the help of personal conceptualisation of the problem to be solved. Klahr and Dunbar take a domain-general view, paying most attention to problem solving strategies (called *weak* methods) that apply across the three spaces, but admit that domain-specific knowledge of science phenomena and research techniques give rise to '*strong* methods'. In *hypothesis generation*, for example, the individual uses personal models of the physical phenomenon involved (Gentner & Stevens, 1983; Harrison & Treagust, 2000), and evaluate these using scientific epistemic criteria (Pluta, Chinn, & Duncan, 2011). The point to be made is that all three

types of knowledge identified earlier, i.e. conceptual-, procedural- and epistemic knowledge, play important roles in shaping the personal conceptualisations in students' scientific reasoning.

Gière, Bickle and Mauldin (2006) explain the three reasoning practices from a *normative*, philosophical perspective. This states that scientific reasoning is, fundamentally, critical reasoning, and that all three practices outlined above are linked to criteria explaining *how* problems should be solved. In other words, when scientists, or students, solve problems in each phase of scientific inquiry, adaptation to certain standards set by the science community is required. These standards are embedded in content, procedural and epistemic knowledge. When reasoning scientifically, an individual *should* use established science content knowledge, follow recognised experimental procedure(s) and adapt to certain epistemic criteria. Combining Gière et al.'s work with the SDDS model suggests that expert scientists are better at scientific reasoning because they have superior understanding of these three knowledge types. Novices may improve their reasoning ability by developing understanding of the same knowledge. These perspectives provide the foundation for the assessment scales developed in the current study.

Noticeably, the rationale takes a different knowledge-oriented perspective on scientific reasoning compared to psychologists such as Koslowski (1996), Klaczynski (2000), and Hogan and Maglienti (2001) and science educators such as Zeineddin and Abd-El-Khalick (2010), who also challenge the domain-general view. These authors focus on the role of knowledge in scientific reasoning, but focus mainly on science conceptual knowledge and take this to *influence* scientific reasoning. In the current rationale, knowledge is part of the definition. Hence, scientific reasoning is *defined* as reasoning with three types of knowledge. Examples elaborating this difference and the knowledge involved follow.

Assessment scales for scientific reasoning

Generally, assessment scales for scientific reasoning in extant literature reflect the trend Zimmerman (2000; 2007) observed - that researchers admit reasoning is knowledge-dependent, but has domain-general strategies as foci. In this perspective, content knowledge is a problem, solved by reduction or 'control'. In the UK, the Assessment of Performance Unit (APU) survey (Johnson, 1987), for example, "produc[ed] process questions with minimum content dependence" (p. 100) to overcome this problem; that is, 'knowledge-lean' items were used. The International Association for the Evaluation of Educational Achievement (IEA) adopted a different strategy for the Trends in International Mathematics and Science Study (TIMSS), utilising 'averaging' reasoning processes across items with different contents (Martin, Mullis, & Foy, 2008). Figure 1 shows how a *content scale* is made by adding up physics items across all 'behaviour domains' and a *reasoning scale* is made by adding up items with the same 'behaviour' across all content domains. The same is done in all four rows and three columns.

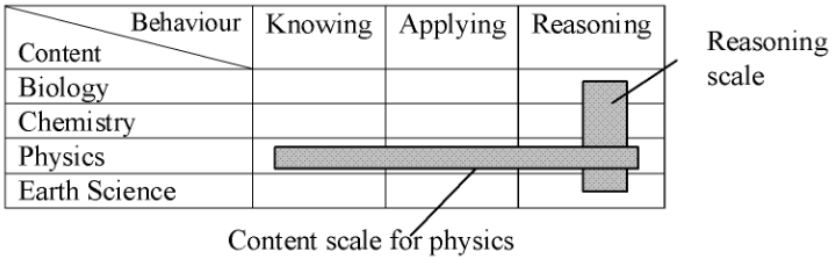


Figure 1. Item grid for Science in TIMSS 2007. Illustrating content and reasoning scales.

The scientific reasoning rationale applied here differs from earlier work in two ways. Firstly, reasoning and knowledge are *not* separated but instead treated as two integrated parts of scientific reasoning. The assessment regimes discussed above express concern about reasoning being knowledge-dependent but accept knowledge being reasoning-dependent: while trying to avoid knowledge in reasoning tasks it intentionally aims to include different types and levels of reasoning in knowledge tasks. Understanding *means* being able to reason with the knowledge. A determining point, however, is that no item can be placed in the grid in Figure 1 and classified along *one* axis only. *Any* item, whether in a content scale or a reasoning scale, tests a *combination* of 'behaviour' *and* 'content'. Thus, Figure 1 shows primarily that behaviour and content (i.e. reasoning and knowledge) have limited value on their own for constructing assessment scales. They must be seen together.

Secondly, the rationale for this study defines procedural and epistemic knowledge as well as science content knowledge. So far, this has been argued from a science education perspective, but has a wider meaning. Krathwohl (2002) describes how Anderson et al.'s (2001) revision of Bloom's taxonomy of the cognitive domain is based on an extended conceptualisation of knowledge. The new framework uses four categories to describe the knowledge dimension: factual knowledge, conceptual knowledge, procedural knowledge and metacognitive knowledge. Likewise, Li and Shavelson (2001) produce a similar framework splitting knowledge into declarative knowledge (knowing what), procedural knowledge (knowing how), schematic knowledge (knowing why), and strategic knowledge (knowing when, where, how knowledge applies). The common point in these, and other examples, is that they alter the relationship between knowledge and reasoning. What used to be explained as a reasoning 'skill' is now seen as application of a body of knowledge of 'what we know', 'how we know' and 'why it happens'. For example, the much investigated 'skill' of *controlling variables* (CoV) can be explained as applying *epistemic knowledge* about dependent, independent and control variables and criteria for defining their cause-effect relationships. It also involves understanding *procedural knowledge* about how to do 'fair testing' – that is, having a strategy for testing only one variable at the time and holding other variables the same.

On this basis, the format shown in Figure 2, which summarises the reasoning rationale from the previous section, has been used to specify assessment scales. It differs from Figure 1 by identifying three science practices rather than general cognitive levels. It also takes into consideration all three knowledge types. Together this creates various options for scale definition. Thus, scales may be 'narrow', measuring one or a few sub-constructs of scientific

reasoning, or 'wide', measuring all aspects involved (Andrich, 2007). The narrowest scale focuses on one type of knowledge used within one practice, for example, using procedural knowledge in experimenting. Most problems, however, will include a combination of knowledge types, for example, gathering data also means using conceptual knowledge to interpret the data and epistemic knowledge to evaluate the data. *Hypothesising*, *experimenting* and *evidence evaluation*, therefore, are three scales that are more meaningful: using all three types of knowledge to solve the problems of explaining a science phenomenon, of obtaining data to test explanations, and of coordinating explanations and evidence. The study will look at these scales separately, but also merge them into one combined reasoning scale.

Behaviour / Content	Science Practices		
	Hypothesising	*Experimenting*	*Evidence evaluation*
Content knowledge	**X**	x	x
Procedural knowledge	x	**X**	x
Epistemic knowledge	x	x	**X**

Figure 2. Construct framework for assessing scientific reasoning. Highlighted boxes are emphasised in scale development.

Figure 2 does not account for progression in the scales. Figure 1 accounts for this partly by a hierarchical relationship between the three behaviour domains, which TIMSS (Martin et al., 2008) refers to as 'cognitive demand'. The horizontal 'knowledge scales', therefore, involves answering questions at three cognitive levels (knowing, application and higher order reasoning). In Figure 2, the same would mean answering questions in three different practices (explaining phenomena, doing experiments and evaluating claims), which are not hierarchical. Comparing the two figures, however, suggests that cognitive level can be added to Figure 2 as a 'third dimension'. The authors of the National Assessment of Educational Progress (NAEP) science framework (National Assessment Governing Board, 2008) reached the same conclusion, naming the dimension 'performance expectation' and using three different levels 'basic', 'proficient' and 'advanced'. This progression dimension, however, also takes into consideration that knowledge can be hierarchically difficult, as illustrated in Figure 3. Item 1 assesses 'basic' performance, by including simple knowledge and low level general cognition, while Item 2 assesses 'advanced' performance, involving advanced ideas and high level general cognition. Any combination, of course, is possible, creating items between the two extremes. Using *three* levels is relevant only as a guideline.

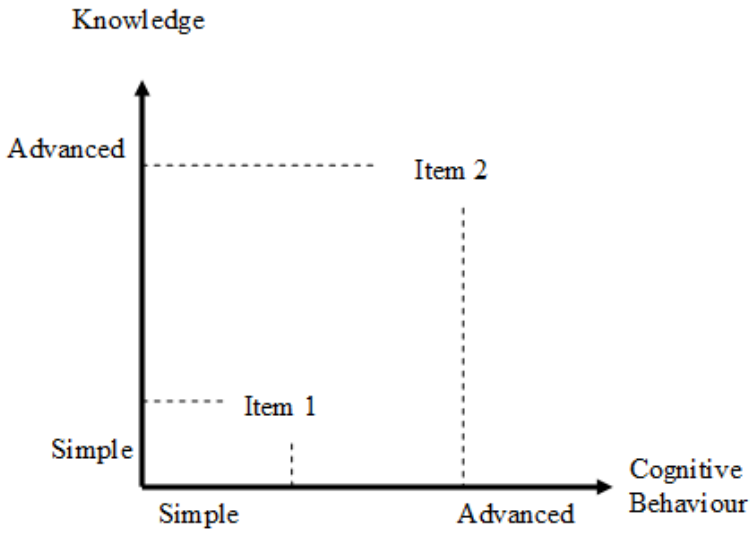


Figure 3. Progression in scales for scientific reasoning combining knowledge and cognitive behaviour. Item 2 aligns with higher progression level than Item 1.

The line of argument above leaves three scales, *hypothesising*, *experimenting* and *evidence evaluation*, based on understanding conceptual, procedural and epistemic knowledge and using cognitive demand together with complexity of the knowledge to set progression levels. This rationale differs theoretically but shares a common interest with NAEP and TIMSS in trying to integrate scientific reasoning into the science curriculum. It differs, however, substantially from Lawson's (1978; 2004) *Classroom Test of Scientific Reasoning* (LCTSR), which is one of the most commonly used test in scientific reasoning research (Bao et al., 2009) and based on Piaget's (1954) 'formal operational schemata'. The remainder of the paper develops the validity argument for the current rationale by operationalising the suggested scales and providing trial data. To help explore the knowledge dimensions, the scale development places increased focus on conceptual knowledge in *hypothesising*, procedural knowledge in *experimenting* and epistemic knowledge in *evidence evaluation* (as highlighted in Figure 2), although all three types of knowledge are involved in all three scales.

Methodology for scale development and trials

The study aimed to operationalise scales for *hypothesising*, *experimenting* and *evidence evaluation*, and to provide empirical evidence supporting the validity argument. For practical reasons, however, this full process was conducted for the two last scales only. Schools involved in trialling (see below) used teacher-designed tests for conceptual areas as a substitute for *hypothesising*, focused on explaining science phenomena. Although not ideal, this had the advantage that many content areas could be included and a series of subtests generated to produce the final scale. The scale was made with data from six content-led science tests: Human Biology, Evolution Theory, Geoscience (Rocks), Chemistry (Oil), Physics (Energy and Radiation). Each test was held

shortly after teaching and included items on a range of cognitive demands and conceptual difficulty. The two remaining scales were regarded as most interesting and challenging for the study. For these scales, the development process started with establishing Mislevy et al.'s (2002) 'student model' from research literature. Literature was used in identifying the procedural knowledge students use in *experimentation* and the epistemic knowledge used in *evidence evaluation*, and also, to establish the learning progression occuring in these. Next, items were selected from the same literature or developed to match identified knowledge and suggested progression. Third, scoring criteria were established for each item using the predicted progression in the student model. The study used a combination of ordered multiple choice items (OMC), as suggested by Briggs, Alonzo, Schwab, and Wilson (2006), ordinary MC items and open-ended (OE) items. This meant progression in a student's ability was reflected in the scale by overall difficulty of the item and difficulty of each response alternative/score category. Trialling was the last step in the development procedure. Items for both scales were placed in random order in a pencil-and-paper test for completion by students. Steps three and four, however, were repeated in four iterations to improve the face validity of the items and alignment with the student model. For example, some items were trialled in different formats (OE and MC) and MC items were adjusted to create better separation between response alternatives. These changes were based on statistical information from Rasch analysis (see below), but also 'expert comments' from teachers and fellow researchers.

The population in mind for the assessment is lower secondary education, which in the UK includes 12 to 16 year old students. The sample comprised mixed ability 14-15 year old students from six state-funded secondary schools in northern England, as shown in Table 1. The sample was not randomly selected, but from schools collaborating with the University. A random sample, however, was not regarded as a crucial since focus was on scale development. The main criteria were that students were drawn from a broad ability range and there was an 'opportunity to learn', that is, a match between taught curricula and test topic. The national science curriculum (Qualifications and Curriculum Authority, 2007) in England places much emphasis on experimenting and evidence evaluation in both laboratory and socio-scientific settings. Research studies (Abrahams & Millar, 2008) point out that laboratory experiments and teaching *about* science do not always match the ideals of the intended curriculum, yet items in the test were regarded as familiar problems to the students. The experiment and evidence evaluation test was presented as a test of 'How science works', a topic familiar to students.

Table 1

Reliability values (Cronbach's Alpha), number of items and sample sizes during four different iterations of test development. N is number of item in each scale.

	Test 1 *28 students*		***Test 2*** *38 students*		***Test 3*** *133 students*		***Test 4*** *139 students*	
	N	*α*	*N*	*α*	*N*	*α*	*N*	*α*
Evidence evaluation								
Item-level	26	0.14	23	0.85	20	0.80	19	0.83
Testlet-level	5	0.07	5	0.73	4	0.72	4	0.70
Experimenting								
Item-level	26	0.74	22	0.87	9	0.67	13	0.78
Testlet-level	5	0.58	4	0.59	2	0.62	3	0.61

.

Table 1 documents the iterations. The first two iterations tested items and scales crudely by groups of about thirty students in pencil-and-paper tests lasting about an hour each. Major changes were made between these iterations by using different item formats and improving response alternatives. Many multiple choice items in the first iteration were replaced with more open-ended items. Smaller changes, such as improving the text in line with comments from subject experts, were made between the third and fourth iterations Larger numbers of students took the third and fourth iteration tests. The number of items was reduced to reduce the time required to complete the test. Table 1 shows how these issues affected reliability. For example, although the *experimenting* scale showed good consistency in the first iteration, the scale comprised more than twenty questions in four items, requiring a long time to complete. By the third iteration, the scale was reduced to two items spread over nine individually scored questions. This resulted in a drop in Cronbach's alpha from 0.87 to 0.67. One previously deleted item was reinstated for the final iteration, making a scale with thirteen individually scored items, increasing alpha to 0.78. Reliability for *evidence evaluation* was very low in the first iteration, using multiple choice items, but improved when these were replaced by open-ended items.

Students' responses were coded and analysed quantitatively in two ways. Firstly, information about individual items from *raw scores* and item statistics, including distractor choice in MC items and response patterns in OMC/OE items, were examined. Secondly, scales were produced and examined using Rasch (1960) analysis in Winsteps (Linacre & Wright, 2001). The reason for using Rasch analysis is that it establishes a measurement model supporting construct-driven assessment and draws attention towards progression (Wilson, 2005). The Rasch model produces *measures*, or estimated score values, from raw data. Different Rasch models are used for dichotomous-, Likert- and rating-scored items, but all models estimate the probability of a student with a certain ability to get an item of a certain difficulty level correct. Here, analysis used the partial-credit model (Masters, 1982) with uni-dimensional analysis. The analysis produces separate measures for *item difficulty* and *student's ability*, presented in an item-person map (example shown in Figure 4). Where a student and an item

have the same value (aligned positions of the scale), the Rasch model estimates a 50% chance that the student will get that item correct. The student is *less* likely to get more difficult items correct and *more* likely to get easier items correct, a feature that can be read directly from the map placements. Comparing students' and items' measures thus permit some understanding of the level of performance associated with a specific ability level, thereby helping construct understanding. The item-person map can also analyse if progression aligns with theoretical expectations; in other words, if data confirm the predicted pattern of difficulty set out in the student model. This is important evidence for scale validity (J. Smith, E.V., 2004). Rasch analysis produces indicators of statistical quality (consistency) of the scale, including an indicator of 'fit' between raw data and estimated measures; analysis of the differentiated item functioning (DIF); category statistics; and measures of discrimination, separation and reliability. Detailed information about these can be obtained from Bond & Fox (2001).

A final step in the analysis examined 'dimensionality' of the scales using Principle Component Analysis (PCA) in Winsteps (Linacre & Wright, 2001). PCA 'tests' the Rasch model uni-dimensionality assumption by examining factors in the residual between the measures and the observed data. In the case of uni-dimensionality, no other significant factors should be found. The analysis looked separately at *hypothesising*, *experimenting* and *evidence evaluation*, but also at a *total scale* including all three subscales. Correlation values between the scales were also examined.

Outcomes from scale development and trialling

Separate outcomes and findings are presented for the *experimenting* and *evidence evaluation* scales. The 'student model' is explained for each scale with comments about the knowledge involved and progression. Examples follow showing how these student models were operationalised into items and scoring models, then empirical results emerging from the trials. After presenting the two scales, results from dimensionality analysis are presented. Unless mentioned otherwise, all data were obtained from the last iteration shown in Table 1.

Experimenting

Gott and Murphy's (1987) research on procedural knowledge was used as the basis for defining the scale on reasoning as e*xperimenting*. They pioneered work attributing students' failure or success in laboratory tasks to understanding 'strategies of scientific enquiry' (p13). Subsequently, Duggan and Gott (1995) established a procedural knowledge 'taxonomy', suggesting knowledge students should understand at various phases of a scientific investigation. These ideas were selected for our scale:

Knowledge about variables and their cause-effect relationships, e.g.
- identifying and explaining dependent and independent variables; and
- understanding the role of a confounding variable.

Knowledge about measurement of a single variable, e.g.
- understanding that a single measurement has uncertainty and that repeated measurements therefore are likely to vary;
- understanding strategies for making an accurate measurement,

such as doing repeated measurements and finding the mean; and
- knowing about anomalous data and having a strategy for handling these.

Understanding and using strategies for 'fair testing' of cause and effect among a set of variables

Lubben and Millar (1996) identified eight progression levels in students' understanding of procedural knowledge. Simplifying these into three fundamental levels suggests, at the lowest level, students may understand measurements as direct observations of 'true' values. This excludes uncertainty, making repeated measurements unnecessary. When two measurements of the same variable give different values, one must be wrong. Students, however, still understand the cause-effect relationship between variables and may be able to carry out 'fair testing' strategies. At the next level, students show more awareness of uncertainty and the importance of making several measurements, but may believe a *true* value is attainable. Commonly, for example, students will make measurements until two identical values are achieved, or claim more than one measurement is needed. At the most advanced level, students understand a *true* value is unattainable, and the best result is averaging several measurements. Students have strategies for handling this and evaluating a series of measurements, such as looking for anomalous data and evaluating variation in repeated measurements.

Research on 'control of variable' strategies (Kuhn, 2002; Zimmerman, 2007) gives more information about progression of the student model. As suggested, students at any level may understand cause and effect between variables and 'fair testing' strategies. The difficulty level, however, increases as more variables are included and when negative co-variation is involved. Control of variables is also more demanding if non-co-variation, that is, a dependent variable *not* influenced by an independent variable is included, rather than just co-variation (Kanari & Millar, 2004).

Procedural knowledge interacts with epistemic knowledge, firstly, because students need to understand the *purpose* of experimenting (Driver, Leach, Millar, & Scott, 1996), and secondly, because epistemological understanding of a measurement influences the strategies necessary for making accurate measurements. As shown later, the three levels of understanding measurements match levels of students' *personal epistemology* (Perry, 1970) from 'naïve positivism' towards 'informed relativism'. Procedural knowledge interacts also with science conceptual knowledge, because any measurement is of a science phenomenon, influencing both what measurement to make and what strategy to choose. Interactions cannot be avoided, but can be adapted to the purpose of the scale. For example, items may ask questions about the measurements and not explaining the phenomena, and include phenomena students are supposed to understand empirically. Students, for example, are supposed to learn to measure temperature and time. Importantly, however, the procedural knowledge listed above applies across phenomena and is not unique to any specific measurement.

Items were selected and adapted to fit the procedural knowledge and levels of reasoning described above. Examples of assessment items exist in Lubben and Millar (1996), J. Osborne and Ratcliffe (2002) and Roberts and Gott

(2006), and Goldsworthy, Watson and Wood-Robinson (2000) and Gott and Duggan (2003). Next, a scoring model was established for each item adapting the same rationale. Examples are shown below.

First, Table 2 illustrates a less successful OMC item from the first iteration, asking students to estimate the number of measurements needed in an experiment testing the time required to dissolve sugar in water. Response alternatives were provided with three levels of understanding measurements described above. At the lowest level (0 points) a response suggests one measurement is enough; at medium level (1 point) a response suggests more than one measurement (either a set number of measurements or until two the same are obtained); and at the highest level (2 points) a response takes into consideration the level of variation. The table presents the number of students (and percentages) giving each response with their average total score on the scale. The 11 students selecting alternative *d*, the highest scoring answer, had the lowest total score whereas the highest scoring students selected alternatives *b* and *e*. This response pattern caused negative discrimination, so despite being well aligned with the student model, the item is unsuitable for a scale. A likely reason the item failed is that high ability students learn measurement strategies (such as "always make a set number of measurements") without learning the epistemic knowledge; that is, they learn the strategy but not *why* its use is appropriate (Lubben and Millar, 1996). The problem, however, seemed to be brought out more severely because few students understand the importance of level of variation between measurements. Revising the item was considered, but with a conclusion that two items are needed to test the intended outcome: one testing 'the need for repeated measurements' and the other testing 'variation in measurements'. How 'variation' influences 'number of measurements' seems to be high level understanding beyond what the involved student group understands.

Table 2

Item in experimenting scale

Jasmine was asked to do an experiment to find how long it takes some sugar to dissolve in water. What advice should you give Jasmine to tell her how many repeated measurements she should make? (Choose one)

	N	*Percent*	*Ability measure*
a) Two or three measurements are always enough (1pt)	35	31.0	55.2
b) She should always make 5 measurements (1pt)	37	32.7	60.5
c) If she is accurate she only needs to measure once (0pt)	2	1.8	57.1
d) She should go on taking measurements until she knows how much they vary (2 pts)	11	9.7	54.7
e) She should go on making measurements until she gets two or more the same (1pt)	28	24.8	61.3
Total	113	100	

In contrast, Table 3 shows a successful item called 'EXP4'. This assessed students' reasoning using knowledge about uncertainty in measurement. A school experiment, with data, was presented as a stem, followed by three questions. Question A asked why measurements are not the same over four repetitions, offering four possible responses. Responses were designed to reflect the student model, with the correct alternative, *b*, placed at the second level of understanding (that measurements may have uncertainty). Table 3 (Question A) shows 56% of respondents got the item right, and that discrimination worked well. Question B asked for a strategy to handle uncertainty in measurements. Here, two response alternatives gave credit: students scored 2 points for alternative *b*, which suggested removing the irregular measurement before averaging the other measurements, and 1 point for alternative *a*, to average the measurements. These alternatives reflect different levels in progression of procedural understanding. Distractor *d* was not selected, but the overall discrimination pattern confirms good progression within the remaining responses.

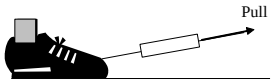
Students omitting the item had the lowest total scores. Question C asked students to select the conclusion they most agreed with. This anticipated higher cognitive demand, as using a strategy relating tabulated data and comparing results between fictitious students was expected. Responses scored 1 point for selecting a conclusion that matched the strategy suggested in question B, and 2 points for adding an appropriate reason. Results show good discrimination and indicate that the theoretical prediction of progression correlates well with student data.

Figure 4 presents the outcomes of applying the Rasch model and placing the item in Table 3 together with other items measuring *experimentation* into an *item-person map*. The scale is the vertical dotted line. Values at 10 point intervals are far left. Items, with each question marked separately are to the right of the line. Each is labelled 'EXP' for experimenting with the last letter indicating the question in each item (e.g. EXP5D is question *D* in item 5). The 'callouts' give the question topics. Each X to the left of the line represents a student. *Mean values* for item difficulty and students' abilities are marked with 'M' and one and two standard deviations with 'S' and 'T'.

Table 3

Item (EXP4) in experimenting scale. Stem above tables. Last column in tables show total score on the scale.

Daniel, Philip and Tom investigated how a trainer slips on different surfaces. They put a mass in the trainer and measured the pull (with a Newton-meter) needed to drag the trainer along. They tested each surface four times.



Here are their results:

Type of surface	Pull/force (Newtons) First time	Second time	Third time	Fourth time
Playground	11	12	13	21
Grass	14	13	13	14
Classroom carpet	8	9	8	9
Polished floor	8	7	7	8

A. They thought they had done everything the same. What is the most likely reason they didn't get the same measurement each time on the same surface? (Tick one box)

	Frequency	*Valid Percent*	*Ability measure*
a. They were not as accurate as they thought. Being more accurate they would have got it right	23	20.9	57.0
b. Measurements never are exactly the same, however hard you try get it accurate (1 pt)	62	56.4	62.9
c. The surfaces must have got slippier each time they did their test	13	11.8	49.2
d. There must have been something wrong with the Newton-meter as they repeated the measurement	12	10.9	55.8
Non-response	5	4.3	33.3
Total	115	100.0	

B. How should they decide which results they should use? (Choose one)

	Frequency	*Valid Percent*	*Ability measure*
a. Add up measurements from all trials and divide by 4 to get the average (1 pt)	58	50.4	56.1
b. Take away irregular (odd) measurements, then get average among the rest (2pt)	40	34.8	69.3
c. Choose measurements that are the same (occur several times)	9	7.8	47.2
d. Choose the lowest value, because this is the least force that was needed	0	0	
Non-response	8	7.0	29.2
Total	115	100.0	

C. The boys disagreed about the conclusion on which surface needed most force. Who do you agree with? Choose one and explain why.

	Frequency	*Valid Percent*	*Ability measure*
Select right person and give relevant reason (2 pt)	37	32.2	73.1
Select right person (1 pt)	47	40.9	55.6
Select wrong person (0 pt)	31	27.0	44.1
Total	115	100.0	

Questions A to E in Item EXP1 are lowest in the scale and the easiest questions. These asked students to identify variables. Item 1 is classified as 'low level' reasoning requiring simple use of the terms *dependent*, *independent* and *control* variables. Item EXP4 (see Table 3), is mid-scale. As this item required students to demonstrate understanding of strategies to handle uncertainty in data, it was classified as involving more complex knowledge and demanding higher level thinking than EXP 1. Questions A to D in EXP5 asked students to test hypotheses, use control of variable strategies, and reach the correct conclusion. Figure 4 shows EXP5 A and C were 'easier' than Band D, which were the two most difficult of all, as these appear towards the top of the scale. These data suggest 'control of variables' is not an advanced strategy, as predicted by the student model and supported by literature (Chen & Klahr, 1999), but using this strategy in advanced contexts makes an item more demanding. Overall, data demonstrated good alignment with the item difficulty pattern predicted by the student model.

The pattern of 'X's in Figure 4 shows how students spread along the scale. Some used simple concepts only, having a 50% chance of getting EXP1 questions correct. Most used some measurement strategies correctly and started to manage uncertainty. Item EXP4 in Table 3 therefore represents accurately the reasoning characterising a majority of students in this group. Some, shown by the Xs higher up the map, showed advanced reasoning using higher level strategies in items testing complex 'control of variables'. The highest scoring students found the most advanced question, EXP5_D, easy. The fact that mean value for students is higher than for items indicates that more difficult items in the scale were required. This would have helped characterise the most able students in more detail. The main outcome, however, is that the overall pattern is plausible compared to student and scoring models underpinning the scale, providing crucial evidence for the validity argument.

Table 4 presents consistency statistics for the scale. These demonstrate good reliability and 'fit statistics': Bond and Fox (2001) recommend 1 as a desired value for Infit and Outfit, with no single item larger than 1.3 or lower than 0.7. All items were within this range.

Table 4

Summary statistics for Rasch analysis of experimenting scale

	Person	*Item*	*Total*
Mean Measure[1]	58.43 (14.28)	49.8 (12.00)	
Model Error	7.04 (1.06)	2.54 (0.53)	
Infit Mnsq	0.99 (0.42)	0.98 (0.16)	
Outfit Mnsq	1.00 (0.82)	1.00 (0.34)	
Separation	1.75	3.38	
Reliability	0.78 (Alpha)	0.95	
Variance explained	23.4 %	22.7 %	46.2 %

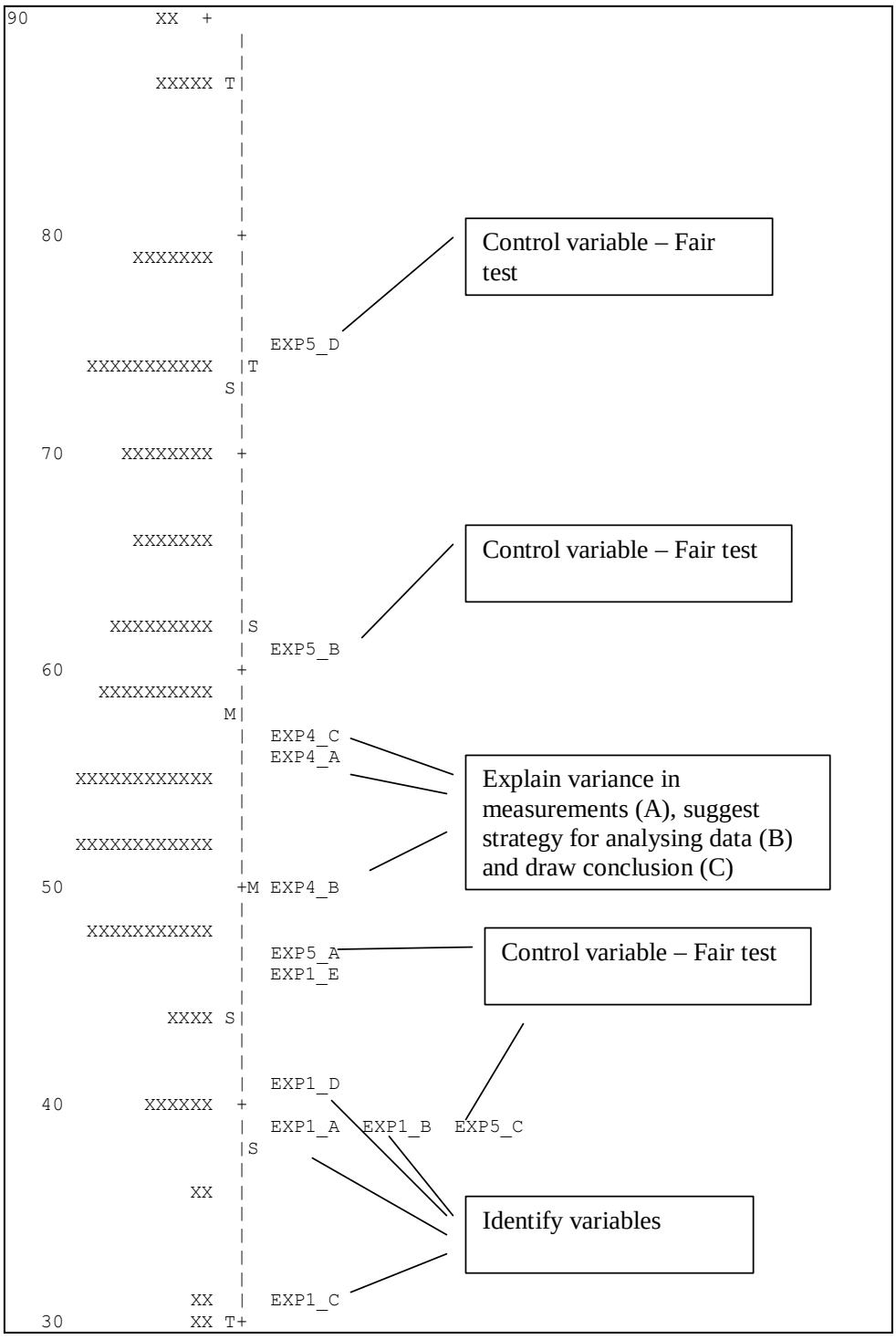


Figure 4. Item-person map for the *experimenting* scale. Each X to the left indicates the position of a student. Labels and callouts to the right indicate items.

Evidence evaluation

Research literature on nature of science (NOS) and scientific argumentation was examined to help define the student model of reasoning as *evidence evaluation* using epistemic knowledge. NOS literature reveals an ongoing debate about assessing *functional understanding* versus *declarative statements* (Allchin, 2011, p. 519). Allchin (ibid.) and Ford (2008) argue that expressing formal epistemology is insufficient for practical reasoning. Similarly, Sandoval (2005) points out that students may reason scientifically using different ideas from those they think scientists use, thus distinguishing between students' knowledge about formal epistemology, and the practical, personal epistemology built through participating in school science inquiry. This study utilised formal epistemology to identify *what* ideas to assess, but made use of functional understanding and argumentation literature in considering *how* to assess. Formal epistemology emerges from science philosophers', sociologists' and historians' attempts to establish normative models reflecting 'best practice' in science (Thagard, 1982), and are, therefore, the 'correct answers' for ideas educators *want* students to use in their reasoning. Applying these ideas meets G.J. Kelly and Duschl's (2002) definition of *epistemic practice*, and is what the current scale intends to measure.

Researchers have identified and agreeed on a set of formal epistemological ideas students ought to know (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; J. Osborne, Ratcliffe, Collins, Millar, & Duschl, 2003). Driver et al. (1996) suggest three main categories: the purpose of scientific work, the nature and status of scientific knowledge and science as a social enterprise. Similarly, Sandoval (2005) suggests students should understand four themes in order to evaluate scientific claims in the context of socio-scientific issues: scientific knowledge is constructed; scientific methods are diverse; there are different forms of scientific knowledge; and scientific knowledge varies in certainty. Ideas from this literature and the notion of understanding criteria scientists use when evaluating theories (Sampson & Clark, 2008) served as guidance, and following ideas were selected for the scale:

Understanding the claim-evidence relationship in science, e.g.
- claim and evidence are different epistemic entities;
- evidence is used in science to both support and reject claims;
- science value empirical evidence, but this is not the only evidence being used when evaluating claims;
- the credibility of the evidence is important to what conclusion can be drawn;
- the relationship between claim and evidence is not always final, uncertainty and disagreement about their relationship may occur.

Understanding criteria for evaluating and using evidence, e.g.
- the significance of how evidence is obtained (i.e. if sufficient method, context, sample etc. were used);
- the difference between anecdotal and scientific evidence;
- the relevance of a theoretical rationale to explain the evidence; and
- that an argument has to be presented for why an evidence is relevant to a claim.

Understanding the role of personal and social contexts in evidence evaluation

and coordination, e.g.

- that someone (person or group) may be biased towards their interests when using evidence;
- that science operates as a social enterprise and collectively try to evaluate and coordinate evidence; and
- that scientists may disagree in their interpretation of the claim-evidence relationship.

Information about progression in understanding and using evidence in science is sparse but growing. Research from NOS and scientific argumentation traditions were utilised. Erduran, Simon and J. Osborne (2004) suggest a five level scale for scoring quality of students' scientific argumentation. At the lowest level, students meet claims with counter-claims rather than using evidence, suggesting they do not separate claim and evidence as different epistemic entities. At the middle levels, students use evidence to support a claim, with or without providing *warrants* and *backings*. This suggests increasing understanding of the need for an argument and for the scientific argument to be valid. At the highest levels, students bring in rebuttals, showing understanding that evidence can support an argument by questioning the counterclaim. Berland and McNeill (2010) outline a learning progression for scientific argumentation with a similar and more detailed account. Their progression scale starts with students understanding the need for using evidence to defend claims and, at the next levels, include increasingly advanced reasoning with absence or presence of argumentation components (data, warrants, backings and rebuttals). The highest level in their account, similar to that of Erduran and colleagues, involves using rebuttals in an argument. Wilson (2005) offers a third approach, suggesting a 'using evidence' construct, describing progression from having *subjective perspective* – providing subjective opinions, not understanding the value of evidence, to an *scientific perspective* – 'questioning or justifying the source, validity, and/or quantity of evidence' (p. 33).

In NOS literature, general levels of epistemic development are identifiable. At the simplest level, students hold a naive positivist view that distinguishes little between describing and explaining nature, for example, an accurate observation is regarded as *true* knowledge. Driver et al. (1996) call this 'phenomenon-based reasoning'; Carey and C. Smith (1993) refer to 'level 1 epistemology', while Lederman et al. (2002) use the term 'naive epistemological views'. Next, 'intermediate' level individuals show awareness of the possibility of multiple explanations for science phenomena, but may take extreme 'positivist' or 'relativist' views. For example, although many explanations exist, *one* may be regarded as 'true', or opposite, on the grounds that 'no explanation is better than another'. Driver et al. call this stage 'relation-based reasoning'; to Carey and C. Smith (op cit) this is 'level 2' reasoning, while Lederman et al. describe this as 'in transition'. The highest, 'advanced', level of understanding shows individuals accepting the model-like nature of scientific knowledge, replacing a 'true – false' perspective with use of supporting evidence. Driver et al. describe this as 'model-based-reasoning', Lederman et al. call this 'informed' and Carey and C. Smith simply 'level 3'. This development theme is similar to psychological research on 'personal epistemology' (Hofer & Pintrich, 2002; Kuhn, Cheney, & Weinstock, 2000; Perry, 1970), which is found to be partly age related. Hence, general

development of students' personal epistemology occurs across subject domains, interacting with development of epistemic understanding and reasoning in science. This 'general' literature suggests large proportions of individuals in older age-groups (adolescents and young adults) only partly reach the third, most advanced stage.

Put together, literature suggests progression in students' reasoning combines 'epistemic growth', that is, a general development towards better understanding of knowledge, with better understanding of science formal epistemology, and use of specific science epistemic ideas. 'Naïve positivist' students do not understand the need for a scientific argument, because they think 'true' claims emerge automatically from science data. These students, are likely to have little or no understanding of the quality of scientific evidence. Next, 'intermediate' students accept multiple explanations and have some understanding about using evidence to support claims. These students are more likely than the positivists to understand that evidence has uncertainty and may or may not be a valid support to the claim. The most 'informed' students understand the modelling-like nature of science and the complicated relationship between claims and evidence. They understand the need for *rebuttals* when evaluating claims. Students, of course, are found along a continuum of this development.

As in the experimenting scale, cognitive demand was considered. Students may show recognition of an epistemic idea, for example, that evidence has uncertainty, but be unable to apply this to complex problems requiring higher order thinking. Berland and McNeil (2010), however, point towards an opposite, equally relevant problem; that students may be able to reason scientifically, for example, being able to identify relevant evidence for a claim, but be unable to present an explicit, reasoned explanation. They suggest giving an explicit reason appears higher in the learning progression than using the knowledge. This aligns with Allchin's (2011) suggestion that students develop functional epistemic understanding. Hence, this was used as a progression criterion in the student model and transferred to the item and scoring models.

Two items will be presented to illustrate operationalisation of the student model into item and scoring models, both based on J. Osborne and Ratcliffe (2002). The first item (EPI2 in Table5) had a socio-scientific setting and asked students to evaluate claims about the health risk of a chemical added to plastic. The item provided a list of seven pieces of information students could use as evidence, and asked specific questions about identifying evidence supporting or contradicting the claim, before asking for an overall conclusion. The 'evidence list' reflected the range of the ideas listed above about epistemic understanding, and included claims by the environmental lobbying group *Greenpeace*, technical information about tests and data, and outcomes from various research studies. From the student model, low-level reasoning was associated with responses selecting inappropriate information as evidence for the health risk, for example, using the *Greenpeace claim*. Medium-level reasoning related to understanding criteria for evaluating the evidence when identifying information contradicting the claim, for example, taking into consideration disagreement about research method. Two pieces of information supported the chemical *not* causing a health risk, giving an opportunity for two score levels. At the highest level, students synthesised all information to draw a conclusion. This took into consideration

synthesis of more information and weighing up of evidence on each side. Raw data are shown in Table 5.

Table 5

Item (EPI2) in evidence evaluation scale. Stem above tables.

"Phthalates" are chemicals added in small amounts to plastic to make it soft and flexible. Scientists disagree about the effects these chemicals may have on humans' health. Here is some information about phthalates: (list of information omitted)

Use the information to answer the questions:

A. What evidence supports that small amounts of phthalates could be dangerous to humans?

	Frequency	*Percent*	*Ability measure*
None correct	77	67.0	45.5
Identify correct evidence (1 pt)	38	33.0	55.4
Total	115	100.0	

B. What evidence supports that small amounts of phthalates should <u>not</u> be dangerous to humans?

	Frequency	*Percent*	*Ability measure*
None correct	41	35.7	40.2
One correct (1 pt)	61	53.0	52.7
Two correct (2 pt)	13	11.3	57.3
Total	115	100.0	

C. Which is the most correct conclusion, based on the information above?

	Frequency	*Percent*	*Ability measure*
Other conclusions (e.g. claim data with uncertainty should not be used)	56	48.7	42.3
Small amounts of phthalates are most likely not to cause a health problem (1 pt)	59	51.3	54.8
Total	115	100.0	

The data (in Table 5) support the student and scoring models by demonstrating discrimination as predicted, that is, students with the highest level of reasoning on the item have highest total score on the test. Selecting *Greenpeace*'s claim as scientific evidence was a common error in question A among the lowest scoring students. In question C, high scoring students explained that the total weight of evidence supported the chemical *not* causing a health risk, while lower level students selected the wrong conclusion or that 'you should never draw a conclusion when there is conflicting information'.

Table 6

Item (EPI5) in evidence evaluation scale. Stem above tables.

Three pupils are discussing how plants can increase in weight

Paul says:	**As a plant grows its extra weight comes from the soil**
Mary says:	**The extra weight comes from the air**
Laura says:	**The extra weight comes from the water the plant takes in through the roots**

They have 3 pieces of evidence for their argument. Say whose view(s) they support or contradict (there might be one or more). Explain why.

A. Evidence 1: Plants may grow and increase their weight in a glass of water.

	Frequency	*Percent*	*Ability measure*
None right (0 pt)	26	22.6	37.6
Supports Laura (1 pt)	34	29.6	47.1
Supports Laura contradicts Paul (2 pts)	49	42.6	54.0
Supports Laura and Mary, because plant is in air *and* water (3 pts)	6	5.2	63.6
Total	115	100.0	

B. Evidence 2: Pot plants grow and increase their weight when you add water, but no soil disappears.

	Frequency	*Percent*	*Ability measure*
Wrong or none-response (0 pt)	51	44.3	39.3
Supports Laura (1pt)	20	17.4	51.1
Supports Laura and contradicts Paul (2 pt)	41	35.7	58.2
Supports Laura and Mary, because plant is in air *and* is supplied with water (3 pt)	3	2.6	65.0
Total	115	100.0	

C. Evidence 3: If you weigh a growing pot plant you will find that the weight increases more than the water you add.

	Frequency	*Percent*	*Ability measure*
Wrong or none-response (0 pt)	58	50.4	42.7
Supports Paul or Laura, because plant is in soil and you add water (1 pt)	32	27.8	51.8
Supports Mary, because mass must come from somewhere else than soil and water (2 pt)	15	13.0	56.7
Supports Mary (as above), but some mass still could come from water or soil (3pt)	10	8.7	61.8
Total	115	100.0	

The second item (EP15 in Table 6) asked students to coordinate claims and evidence regarding plant growth. This item differs from EPI2 (above) in that there is no 'social' element and therefore fewer perspectives to consider when drawing conclusions. There was, however, in question B and C, a longer line of reasoning needed, because evidence supported several claims and information could be regarded as evidence that was not mentioned in the text. For example,

the plant is put in soil to which water is added, both variables being mentioned explicitly in the text, but the plant is also surrounded by air, which is implicit. The first trial showed some students answering what they thought was the 'right' claim without comparing claims and evidence. This risked getting the correct answer for the wrong reason. Asking students to explain their answer helped identify this problem, and brought the item into line with the student model. Including explanation significantly improved item discrimination and scale reliability between first and second iterations (see Table 1). The scoring model identified the *wrong* conclusion (students not connecting claim and evidence appropriately) as the lowest level, with a zero score. At the next level students connected a *mentioned* piece of evidence (e.g. water or soil) and the claim appropriately, scoring 1 point. A third level (2 points) included students mentioning 'rebuttals' (the claim a piece of evidence would reject). The highest level (3 points) of reasoning related to building an argument with rebuttals *and* mentioning the 'hidden' variable (air). As for the previous item, raw data in Table 6 gives positive evidence for this scoring model, showing that higher (theoretically) level responses relate to higher total scores on the test.

Table 7

Summary statistics for Rasch analysis of evidence evaluation scale

	Person	*Item*	*Total*
Mean Measure[1]	52.18 (11.70)	48.32 (13.70)	
Model Error	4.12 (0.94)	2.01 (0.74)	
Infit Mnsq	0.98 (0.17)	0.98 (0.17)	
Outfit Mnsq	0.97 (0.61)	0.97 (0.39)	
Separation	2.27	6.19	
Reliability	0.83 (Alpha)	0.97	
Variance explained	37.7 %	13.6 %	51.3 %

Further evidence is obtained from the item-person map for items in the evidence evaluation scale (Figure 5). An unmistakable pattern supporting the theoretical predictions emerges. EPI4 asked students to distinguish between items in a list of 'anecdotal' and 'scientific' evidence, to identify evidence relevant to the claim 'smoking is a health risk'. This was predicted as low level reasoning, because students had to evaluate the evidence only, not to co-ordinate this with a claim. Questions in the item occur at low end of Figure 5. Lowest ability students identified correctly that 'statistics show smokers on average die at younger ages than non-smokers' is valid evidence (EPI4C), but believed 'a close friend or relative of yours has been smoking for a long time and has got ill' (EPI4A) also counted as valid evidence. EPI1 asked students to evaluate if a claim could be tested scientifically, requiring understanding about the claim-evidence relationship. This is located in the mid area of the scale. All items including coordinating of claim and evidence (EPI2, EPI5 and EPI 7) are found towards the upper end of the scale. However, values on the left-hand scale in Figure 5 are average scores for each question. Threshold levels on the right of the item-

person map are used in the scoring model for question A in EPI5, demonstrating that the simplest co-variation between claim and evidence (scoring 1 point in question A in EPI5) was an easier question handled by much lower ability level students. Mid-level students align with the second threshold (2 points), while the highest threshold (3 points, requiring identification of all covariant and non-covariant variables) is placed higher than any student in the sample.

Table 7 provides Rasch item statistics, showing the scale has appropriate infit/outfit, reliability and separation. About half of the variance was explained, a figure similar to that for the *experimenting* scale.

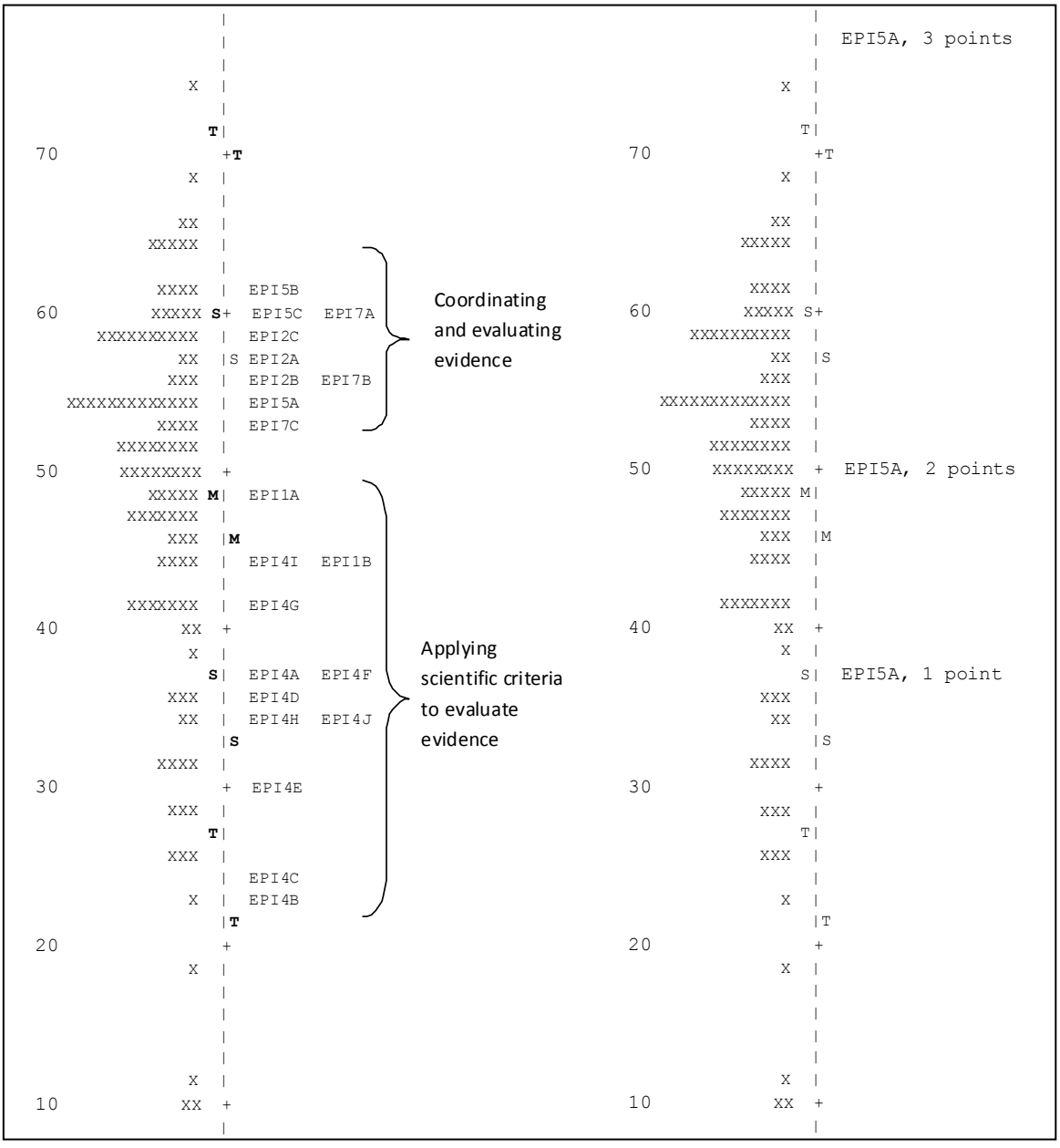


Figure 5. Item-person map for *evidence evaluation* scale. Scale to the left shows average difficulty for questions. Scale to the right shows data for question A in Item EPI5 only.

Dimensionality of the scales

Each of the three scales and the merged (total) scale were subjected to dimensionality analysis using 'PCA of residuals' in Winsteps.. The analysis examines if the scale is uni-dimensional by looking for patterns in variance *not* accounted for by Rasch measures. If truly uni-dimensional, no other significant factor in the unexplained variance should be present. Table 8 shows total variance in each scale (measured in Eigenvalue units and percent) and the explained and unexplained variances. If a second factor is significant, this should account for at least two eigenvalue units (Arrindell & van der Ende, 1985). In Table 8, three (highlighted) values in the first contrast are above this level. The factors are referred to as 'contrasts', because the software separates items with positive and negative loadings. Table 9 shows loadings for each highlighted value. In the *experimenting* and *evidence evaluation* scales, factors are related to a contrast between items with high and low difficulty. In the *total scale*, the factor relates to a contrast between *hypothesising* versus *experimentation* and *evidence evaluation*. In all cases, however, eigenvalues for the factors are small, suggesting the scales are near uni-dimensional. The measure explains 77.8% of variance in the *total scale*, while the second factor explains just 4.2%. The conclusion is that the three subscales fit well into a single scale. Table 10 presents correlation values between the subscales, confirming that *experimentation* and *evidence evaluation* are more closely related than *hypothesising* and any scale.

Table 8

Rasch dimensionality analysis in Winsteps

	Experimenting scale from items		***Evid.- Eval. scale from items***		***Combined scale from items***		***Combined scale from testlets***	
	Eigen-value	*Percent*	*Eigen-value*	*Percent*	*Eigen-value*	*Percent*	*Eigen-value*	*Percent*
Total variance	24.1	100	39.0	100	49.4	100	22.7	100
Explained by measures	11.1	46.2	20.0	51.3	23.3	47.4	15.7	69.2
Unexplained Variance	13.0	53.8	19.0	48.7	26.0	52.6	7.0	30.8
1st contrast	2.6	10.9	2.9	7.5	2.9	6.0	1.7	7.3
2nd contrast	1.8	7.4	1.9	5.4	2.5	5.2	-	-
3rd contrast	1.5	6.2	1.7	4.5	2.2	4.5	-	-

Table 9

Item loadings to secondary dimensions in Combined scale

1st contrast		***2nd Contrast***		***3RD Contrast***	
Item	*Loading*	*Item*	*Loading*	*Item*	*Loading*
Positive					
EXP1C	.58	EXP1A	.59	EPI5B	.62
EXP1A	.57	EPI4H	.51	EPI5C	.44
EXP1B	.56	EXP1C	.47	EPI5A	.35
EXP1F	.54	EXP1B	.46	EPI4E	.32
EXP1D	.49	EPI2B	.45	EPI4B	.28
Negative					
EPI4I	-.51	EPI5C	-.52	EXP5D	-.43
EPI4H	-.49	EPI2C	-.45	EXP1E	-.42
EPI7B	-.44	EPI5A	-.37	EXP1D	-.42
EPI4F	-.40	EXP5D	-.32	EXP4B	-.42
EPI4D	-.39	EPI2B	-.26	EXP5B	-.39

Table 10

Correlation table for the three scales

	Experimenting	*Evidence Evaluation*	*Science Conceptual Understanding*
Experimenting	1	0.69**	0.60**
Evidence evaluation		1	0.59**
Sc. Concept. Underst.			1

. ** Significant to 0.01 level.

Discussion

Two aims of the current paper have been to establish construct-driven assessment in science and to improve assessment of scientific reasoning. It commenced with a particular disciplinary rationale, suggesting scientific reasoning relates to the three practices, *hypothesising*, *experimenting* and *evidence evaluation*, and requires understanding of science conceptual, procedural and epistemic knowledge (Author & Colleagues, In review). From this rationale, the paper has followed Kane's (2006) suggestion of 'validity as argument', trying to build a line of reasoning that combines Mislevy et al.'s (2002) *student models*, *item models*, *scoring models* and *measurement models* and incorporates Wilson's (2005) 'developmental perspective'. Because assessment informs construct understanding (Bond, 2003), a third aim has been to explore and improve the theoretical rationale.

Looking at sub-scales for measuring *experimenting* and *evidence evaluation*, the analysis has placed *student models* in the context of epistemological growth from 'naïve positivism' towards 'informed relativism' (Perry, 1970). Students, it seems, become better positioned to learn science and to learn *about* science as their personal epistemology matures, but may also

develop their personal epistemology from science learning (C. L. Smith, Maclin, Houghton, & Hennessey, 2000). The development has importance to scientific reasoning, because science is an epistemic practice *requiring* an informed relativist epistemology (Kelly and Duschl, 2002). Science practice, however, and thereby the *experimenting* and *evidence evaluation* scales, relates to specific conceptual, procedural and epistemic ideas *within* the generic personal epistemological development. The epistemological growth described by Perry, therefore, serves as a frame that helps identify levels of progression in the scales, but does not inform about the ideas needed at each level. In *evidence evaluation*, for example, students have to understand particular criteria for evaluating claims and evidence, and in *experimenting* they have to learn measurements and strategies for testing hypotheses. This understanding is parallel to understanding science conceptual knowledge, for example, undergoing conceptual change (C. L. Smith et al., 2000) and including misconceptions (Lubben & Millar, 1996)

Following from these perspectives, the validity argument in the current study has been based on identifying relevant epistemic and procedural ideas, *and* demonstrating how these ideas relate to progression levels informed by the overall growth in students' personal epistemology. Thereafter, both of these, i.e. the epistemic and procedural ideas and the progression levels, have been related to assessment items and scoring criteria. Constructing the validity argument for *experimenting* and *evidence evaluation* has been possible because of the wide range of research literature studying teaching and learning of *the nature of science*, *scientific inquiry*, and *scientific argumentation*. The scale for *hypothesising* has not been outlined in the same way, but the idea that this practice is knowledge-based is already established and applied in content-led science tests. What the current rationale adds, however, is a stronger focus on progression and the involvement of epistemic and procedural knowledge. Learning progression research for conceptual knowledge is more comprehensive than for procedural and epistemic knowledge (e.g. Alonzo & Steedle, 2009; M. Wilson, 2009), and therefore suggests a similar argument as carried out for *experimenting* and *evidence evaluation* could be presented for *hypothesising*.

A challenging issue in the validity argument has been the role of domain-general reasoning abilities. On one hand, the current paper rests on a claim that domain-general cognition has been too dominant in teaching and assessment of scientific reasoning and seeks an alternative rationale related to the three types of domain-specific knowledge. On the other hand, general cognition plays an obvious role that has to be accounted for. The solution in the current study has been to relate domain-general reasoning to a third 'progression dimension', adding to the *knowledge* and *practices* dimensions in Figure 2. Progression, as shown in Figure 3, is explained as increasing complexity of domain-specific knowledge *combined* with increasing levels of domain-general reasoning (from lower to higher-order thinking). This has resemblance to the two-dimensional Tyler-Bloom 'content-behaviour' grid (Bloom, Hastings, & Madaus, 1971), commonly used in assessment. However, adding procedural and epistemic knowledge, rather than just science conceptual knowledge, *and* looking towards progression in understanding each of these make important differences. In the Tylor-Bloom rationale, for example as it is applied in the TIMSS science framework (Figure 1), *experimenting* and *evidence evaluation* become 'skills' and

related to 'cognitive demand'. The current rationale, in contrast, makes the same practices a matter of understanding of procedural and epistemic knowledge.

One outcome from the scale development that sheds light on the relationship between domain-general reasoning and domain-specific knowledge is that these tend to be more separated in lower than higher progression areas. In the higher areas, knowledge and general reasoning are mixed and make it harder to tell what exactly is being measured. For example, a low level item asking students to identify what information can be used as scientific evidence to support a claim tests specific epistemic knowledge about criteria for scientific evidence. In contrast, a higher order item asking students to draw a conclusion by synthesising and analysing a series of measurement, taking into consideration uncertainty and anomalous data, is much more complex. Students have to use epistemic and procedural knowledge, but the domain-general reasoning is more demanding and naturally will have stronger impact. It was also noticed during item development that 'advanced' items in *evidence evaluation*, asking students to coordinate claim and evidence, are similar to 'advanced' items in *experimenting*, asking students to draw a conclusion from interpreting data. In other words, the two scales seem to 'merge' in the higher progression areas *both* because of a stronger influence of domain-general higher order thinking (synthesis and analysis) *and* by items involving the same procedural, epistemic and conceptual knowledge. These issues have implication for dimensionality in the scales, and may partly explain the findings in Table 8. In a previous study, author and colleagues (Author & Colleagues, 2011) investigated students' reasoning in practical laboratory tasks and found *hypothesising*, *experimenting*, and *evidence evaluation* appearing as more distinct practices. It is therefore some reason to believe the test format in the current study (using written tests) has contributed to 'merging' the three types of reasoning and also made domain-general reasoning more dominant.

The current study has several implications. It underlines the importance of separating and clarifying the involved dimensions in any science assessment in order to increase the emphasis on scientific reasoning. Existing assessment practices tend to associate scientific reasoning with domain-general reasoning, and even if there has been a long call for making the reasoning 'domain-specific' and 'knowledge-based' (Zeineddin & Abd-El-Khalick, 2010) these aims have not been widely implemented. The TIMSS study has been used as an example, using a framework (Martin et al., 2008) that associates scientific reasoning with 'cognitive demand' rather than 'science content'. By producing scales for different levels of cognitive demands, TIMSS broadcasts a message that scientific reasoning *is* higher-order, domain-general reasoning. Introducing procedural and epistemic knowledge dimensions into the framework, in a similar way as science conceptual knowledge, therefore, makes a way forward for changing this framework. Similar notes can be made about the NAEP and Programme for International Student Assessment (PISA) science assessments. The NAEP science framework (National Assessment Governing Board, 2008) includes procedural and epistemic knowledge *embedded* in science performance. This is the line supported in the current study, but the dimensions have to be categorised explicitly, which is *not* a case in NAEP. The PISA framework (Organisation for Economic Co-operation and Development, 2006) separates between knowledge *of* science and knowledge *about* science, and also categorises both of these. The

framework, however, avoids explaining how procedural and epistemic knowledge link to specific elements of scientific reasoning. It is a categorisation of *formal epistemology* learned for 'scientific literacy' rather than epistemic knowledge used in reasoning. The current study suggests both frameworks have to identify particular procedural and epistemic ideas associated with scientific reasoning and also describe progression of such ideas.

The study also has implications for classroom teaching and assessment, again suggesting that all three practices and all three types of knowledge should be made explicit. As the study has illustrated, this does *not* mean that science teachers explicitly should teach science philosophy, but rather that students should learn to apply rules and criteria in the different science practices. The solution, in the current study, is to clarify *what* specific epistemic ideas students have to learn and then operationalise these into items and scoring criteria. The same applies to teaching; teachers have to understand the procedural and epistemic ideas and then 'operationalise' these into examples used for students. Assessment in this way can be applied as a tool illustrating to teachers and students what the ideas are and how they apply in scientific reasoning.

The study has limitations, particularly because of its small scale and for looking at a limited set of procedural and epistemic ideas. The main outcome is therefore establishing a line of reasoning that can be extended and investigated more thoroughly in further research. This line of reasoning fits well with existing construct-driven assessment in Milsevy et al. (2003) and Wilson (2005), and with the emphasise on *functional understanding* by Allchin (2011), but suggests more effort should be made to include and specify in more details all three practices and all three types of knowledge listed in the current rationale for scientific reasoning. Besides, the population discussed in the current paper is lower secondary, with little attention paid towards primary or upper secondary years. Expanding future research, within the same rationale, in both these directions seems plausible and may add valuable new information.

References

Abrahams, I., & Millar, R. (2008). Does Practical Work Really Work? A study of the effectiveness of practical work as a teaching and learning method in school science. International Journal of Science Education, 30(14), 1945-1969.

Allchin, D. (2011). Evaluating Knowledge of the Nature of (Whole) Science. Science Education, 95, 518-542.

Alonzo, A., & Steedle, J. (2009). Developing and Assessing a Force and Motion Learning Progressions. Science Education, 93(3), 389-421.

American Association for the Advancement of Science, A. (1965). Science - A process Approach. Commentary for teachers. USA: AAAS/Xerox Corporation.

Anderson, L. W., Krathwohle, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). A taxonomy for learning, teaching, and assessing: A revison of Bloom's Taxonomy of Educational Objectives. New York: Longman.

Andrich, D. (2007). On the fractal dimension of a social measurement: I Report No. 3 ARC Linkage Grant LP0454080: Maintaining Invariant Scales in State, National and International Level Assessments. D Andrich and G Luo Chief Investigators.: Graduate School of Education, The University of Western Australia.

Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. Applied Psychological Measurement, 9, 165 - 178.

Au, W. (2007). High Stakes Testing and Curricular Control: A Qualitative Metasynthesis. Educational Researcher, 36(5).

Ausubel, D. P. (1968). Educational Psychology: A cognitive View. New York: Holt, Rinehart and Winston.

Author, & Colleagues. (2011).

Author, & Colleagues. (In review). Towards a Model of Scientific Reasoning for Science Education.

Bailin, S., & Siegel, H. (2002). Critical thinking. London: Blackwell. In N. Blake, P. Smeyers, R. Smith & P. Standish (Eds.), The Blackwell guide to the philosophy of education (pp. 181-193). London: Blackwell.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., et al. (2009). Learning and Scientific Reasoning. Science, 323(5914), 586-587.

Berland, L., & McNeill, K. (2010). A Learning Progression for Scientific Argumentation: Understanding Student Work and Designing Supportive Instructional Contexts. Science Education, 94, 765-793.

Bloom, B. S., Hastings, J. T., & Madaus, G., E. (1971). Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book Company.

Bond, T. G. (2003). Validity and assessment: a Rasch measurement perspective. Metodología de las Ciencias del Comportamiento, 5(2), 179-194.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch Model. Fundamental Measurement in the Human Science. Mahwah, NJ: Lawrence Erlbaum Associates.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. R. (2006). Diagnostic assessment with ordered multiple-choice items. . Educational Assessment, 11(1), 33-64.

Brown, G., & Desforges, C. (1977). Piagetian Psychology and Education: Time for revision. British Journal of Educational Psychology, 47, 7-17.

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. Educational Psychologist, 28(3), 235-251.

Chen, Z., & Klahr, D. (1999). All other things equal: Children's acquisition of the control of variables strategy. Child Development, 70(5), 1098-1120.

Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. Fort Worth: Holt, Rinehart and Winston.

Donaldson, M. (1984). Children's Minds London Fontana

Driver, R., & Easley, J. (1978). Pupils and paradigms: a review of literature related to concept development in adolescent science students. Studies in Science Education, 5, 61-84.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). Young People's Image of Science. Buckingham: Open University Press.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. Science Education, 84(3), 287-312.

Duggan, S., & Gott, R. (1995). The place of investigations in practical work in the UK National Curriculum for Science. International Journal of Science Education, 17(2), 137 - 147.

Duschl, R. A., & Osborne, J. (2002). Supporting and Promoting Argumentation Discourse in Science Education. Studies in Science Education, 38, 39-72.

Erduran, S., Simon, S., & Osborne, J. (2004). Tapping into Argumentation: Developments in the Application of Toulmin's Argument Pattern for Studying Science Discourse. Science Education, 88, 915-933.

Finley, F. (1983). Science Processes. Journal of Research in Science Teaching, 20, 47-54.

Ford, M. (2008). Disciplinary Authority and Accountability in Scientific Practice and Learning. Science Education, 92, 404 – 423.

Gagne, R. M. (1965). The psychological basis of science - a process approach. AAAS miscellaneous publication, 65-68.

Gentner, D., & Stevens, A. L. (Eds.). (1983). Mental models. Hillsdale, NJ: Lawrence Erlbaum Associates.

Giere, R. N., Bickle, J., & Mauldin, R. F. (2006). Understanding scientific reasoning (6th ed.). CA: Belmont: Thomson Wadsworth.

Goldsworthy, A., Watson, J. R., & Wood-Robinson, V. (2000). Investigations: Developing Understanding. Hatfield: Association for Science Education.

Gott, R., & Duggan, S. (1995). Investigative work in science curriculum. Buckingham: Open University Press.

Gott, R., & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. International Journal of Science Education, 18(7), 791-806.

Gott, R., & Duggan, S. (2003). Understanding and using scientific evidence. How to critically evaluate data. London, UK: SAGE Publications.

Gott, R., & Mashiter, J. (1991). Practical work in science - a task-based approach? In B. Woolnough (Ed.), Practical science (pp. 53-66). Buckingham: Open University Press.

Gott, R., & Murphy, P. (1987). Assessing Investigation at Ages 13 and 15. Assessment of Performance Unit Science Report for Teachers: 9. London: Department of Education and Science, Welsh Office, Department of Education for Northern Ireland.

Hanson, N. R. (1958). Patterns of Discovery. Cambridge: Cambridge University Press.

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. International Journal of Science Education, 22(9), 1011-1026.

Hempel, C. G. (1966). Philosophy of natural science. Englewood Cliffs, NJ: Prentice-Hall.

Hewson, P. W. (1981). A conceptual change approach to learning science. European Journal of Science Education, 3(4), 383-396.

Hofer, B. K., & Pintrich, P. R. (Eds.). (2002). Personal Epistemology. Mahwah, NJ: Lawrence Erlbaum.

Hogan, K., & Maglienti, M. (2001). Comparing epistemological underpinnings of students' and scientists' reasoning about conclusions. Journal of Research in Science Teaching, 38, 663-687.

Johnson-Laird, P. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. . Cambridge, MA: Harvard University Press.

Johnson, S. (1987). Assessment in Science and Technology. Studies in Science Education, 14, 83-108.

Kanari, Z., & Millar, R. (2004). Reasoning from Data: How Students Collect and Interpret Data in Science Investigations. Journal of Research in Science Teaching, 41(7), 748-769.

Kane, M., T. (2006). Validation ., pp. 17-64). In L. Brennan (Ed.), Educational measurement, 4th ed. Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kelly, G. A. (1955). The Psychology of Personal Constructs. N.Y.: Norton.

Kelly, G. J., & Duschl, R. (2002). Toward a research agenda for epistemological studies in science education. Paper presented at the National Association for Research in Science Teaching.

Klaczynski, P. A. (2000). Motivated Scientific Reasoning Biases, Epistemological Beliefs, and the Polarization: A Two-Process Approach to Adolescent Cognition. Child Development, 71(5), 1347-1366.

Klahr, D. (2000). Exploring Science. The Cognition and Development of Discovery Processes. MA: Cambridge: Bradford.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-55.

Koslowski, B. (1996). Theory and evidence: The development of scientific reasoning. Cambridge, MA: MIT Press.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice, 41(2), 212-218.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), Handbook of childhood cognitive development. Oxford: Blackwell.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. Cognitive Development, 309-328.

Lawson, A. E. (1978). The development and validation of a classroom test for formal reasoning. Journal of Research in Science Teaching, 15(1), 11-24.

Lawson, A. E. (2004). The nature and development of scientific reasoning: A synthetic view. I. International Journal of Science and Mathematics Education, 2(4), 307-338.

Lawson, A. E., Karplus, R., & Adi, H. (1978). The acquisition of propositional logic and formal operational schema during secondary school years. Journal of Research in Science Teaching, 15(6), 461-466.

Layton, D. (1973). Science for the People: The Origins of the School Science Curriculum in England. London: Allen and Unwin.

Lederman, N. G., Abd-El-Khalick, F., Bell, R., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Towards valid and meaningful assessment of learners' conceptions of nature of science. Journal of Research in Science Teaching, 39(6), 497-521.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation of Tatsuoka's rule-space approach. Journal of Educational Measurement, 41(3), 205-237.

Li, M., & Shavelson, R. J. (2001). Validating the links between knowledge and test items from a protocol analysis. Paper presented at the Annual meeting of the American Educational Research Association.

Linacre, J. M., & Wright, B. D. (2001). Winsteps (Computer program). Chicago: MESA Press.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. International Journal of Science Education, 18, 955-968.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). TIMSS 2007 International Science Report. Findings from IEA's Trends in International Mathematics and Science Study at Fourth and Eight Grades. Boston, MA: TIMSS & PIRLS International Study Center.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & S. A. (Eds.), Mental Models (pp. 229–324). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement 3rd ed. New York: Macmillan.

Millar, R., & Driver, R. (1987). Beyond Process. Studies in Science Education, 14, 33-62.

Millar, R., Lubben, F., Gott, R., & Duggan, S. (1994). Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. Research Papers in Education, 9(2), 207 — 248.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. . In T. T. Kellaghan & D. Stufflebeam (Eds.), International Handbook of Educational Evaluation (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press.

National Assessment Governing Board. (2008). Science Framework for the 2009 National Assessment of Educational Progress. DC: Washington: Author.

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: The National Academies Press.

Organisation for Economic Co-operation and Development. (2006). PISA Released items - Science: OECD, PISA.

Osborne, J., & Ratcliff, M. (2002). Developing Effective Methods of Assessing "Ideas and Evidence". School Science Review, 83(305), 113-123.

Osborne, J., Ratcliffe, M., Collins, S., Millar, R., & Duschl, R. A. (2003). What 'ideas-about-science' should be taught in school science? A Delphi study of the expert community. Journal of Research in Science Teaching, 40(7), 692-720.

Osborne, R. J. (1982). Conceptual change - for pupils and teachers. Research in Science Education, 12, 25-31.

Osborne, R. J., & Wittrock, M. C. (1985). The generative learning model and its implications for science education. Studies in Science Education, 12, 59-87.

Passmore, C., & Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools. Journal of Research in Science Teaching, 39(3), 185-204.

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

Perry, W., G. Jr. (1970). Forms of Intellectual and Ethical Development in the College Years: A Scheme. New York: Holt, Inehart, and Winston.

Piaget, J. (1954). The Construction of Reality in the Child. New York: Basic Books.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' Epistemic Criteria for Good Scientific Models. Journal of Research in Science Teaching, 48(5), 486-511.

Pope, M. L., & Keen, T. R. (1981). Personal Construct Psychology and Education London: Academic Press.

Qualifications and Curriculum Authority. (2007). Science. Programme of study for key stage 3 and attainment targets. In The National Curriculum 2007 (pp. 206-219). London: QCA.

Quine, W. V. (1969). Natural kinds. In Q. W.V. (Ed.), Ontological relativity and other essays. New York: Columbia University Press.

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. : . Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980, Chicago: University Chicago Press).

Roberts, R., & Gott, R. (2006). Assessment of performance in practical science and pupil attributes. Assessment in Education, 13(1), 45-67.

Rupp, A. A., Templin, J., & Henson, R. J. (2010). Diagnostic measurement: Theory, methods, and applications. Mahwah, NJ: Lawrence Erlbaum.

Sampson, V., & Clark, D. B. (2008). Assessment of Ways Students Generate Arguments in Science Education: Current Perspectives and Recommendations for Future Directions. Science Education, 92(3), 447-472.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. Science Education, 89(4), 634-656.

Shavelson, R. J., & Ruiz-Primo, M. A. (1999). On the psychometrics of assessing science understanding. In J. Mintzes, J. H. Wamhersee & J. D. Novak (Eds.), Assessing science understanding: A human constructivist view (pp. 303–341). New York: Academic Press.

Shayer, M., & Adey, P. S. (1981). Towards a Science of Science Teaching. London: Heinemann Educational Books.

Simon, H. A. (1966). Scientific discovery and the psychology of problem solving. In R. Colony (Ed.), Mind and Cosmos (pp. 22-40). Pittsburgh: University of Pittsburgh Press.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, G. M. (2000). Sixth-Grade Students' Epistemologies of Science: The Impact of School Science Experiences on Epistemological Development. Cognition and Instruction, 18(3), 349-422.

Smith, J., E.V. (2004). Evidence for the Reliability of Measures and Validity of Measure Interpretations: A Rasch Measurement Perspective. In J. Smith, E.V. & R. M. Smith (Eds.), Introduction to Rasch Measurement (pp. 93-122). Minnesota: JAM Press.

Thagard, P. (1982). From the Descriptive to the Normative in Psychology and Logic. Philosophy of Science, 49(1), 24-42.

Toulmin, S. (1958). The uses of argument. Cambridge: Cambridge University Press.

Weiss, I. R., Pasely, J. D., Sean Smith, P., Banilower, E. R., & Heck, D. J. (2003). A Study of K-12 Mathematics and Science Education in the United States. Chapel Hill, NC: Horizon Research.

Wiley, D. E. (2001). Validity of constructs versus construct validity. In H. Braun, D. N. Jackson & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 207-227). Mahwah, NJ: Lawrence Erlbaum.

Wiliam, D. (2010). What Counts as Evidence of Educational Achievement? The Role of Constructs in the Pursuit of Equity in Assessment Review of Research in Education, 34, 254-284.

Wilson, M. (2005). Constructing Measures an Item Response Modeling Approach. New York: Psychology Press.

Wilson, M. (2009). Measuring Progressions: Assessment Structures Underlying a Learning Progression. Journal of Research in Science Teaching. Journal of Research in Science Teaching, 46(6), 716-730.

Wilson, M., & Bertenhal, M. (2005). Systems for State Assessment. Washington, DC: National Academic Press.

Zeineddin, A., & Abd-El-Khalick, F. (2010). Scientific Reasoning and Epistemological Commitments: Coordination of Theory and Evidence Among College Science Students. Journal of Research in Science Teaching, 47(9), 1064-1093.

Zimmerman, C. (2000). The Development of Scientific Reasoning Skills. Developmental Review, 20, 99-149.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. Developmental Review, 27, 172-223.