Review of Contemporary Research in Assessment and Reporting

# Robert Stevens, Liliana Ructtinger with Shantha Liyanage and Carole Crawford – The Research Team, School Policy and Information Management, NSW Department of Education March 2017

Contents

[Purpose statement 2](#_Toc476318260)

[Definition of ‘assessment’ 2](#_Toc476318261)

[Formative and Summative Assessment 2](#_Toc476318262)

[Integrating assessment into pedagogy 3](#_Toc476318263)

[The role of assessment in representing student progress 5](#_Toc476318264)

[Explicitly stating assessment criteria 7](#_Toc476318265)

[Visible Learning 7](#_Toc476318266)

[High Quality Tasks and High Expectations 8](#_Toc476318267)

[Reporting practices 9](#_Toc476318268)

[Reporting A-E 9](#_Toc476318269)

[Learning progressions 11](#_Toc476318270)

[Recent research trends in standardised assessment 12](#_Toc476318271)

[Evidence-centred Design 13](#_Toc476318272)

[ECD layers 13](#_Toc476318273)

[Design patterns 14](#_Toc476318274)

[Assessment that supports a diverse range of learners 16](#_Toc476318275)

[Universal Design for Learning 16](#_Toc476318276)

[Assessing students with disabilities 17](#_Toc476318277)

[Diagnosing students at risk of reading difficulty 18](#_Toc476318278)

[Teacher judgement and teacher agency in student assessment 19](#_Toc476318279)

[Innovative practice trends in assessment 20](#_Toc476318280)

[Theoretical issues 21](#_Toc476318281)

[References 23](#_Toc476318282)

# Purpose statement

This document has been produced by the NSW Department of Education Research team, Learning and Business at the request of Early Learning and Primary Education, Learning and Teaching. This work is intended to inform the revision of the policy on Curriculum planning and programming, assessing and reporting to parents K-12.

# Definition of ‘assessment’

The word ‘assessment’ has its roots in the Latin verb assidere meaning ‘to sit beside’. Mary James suggests that assessment has to do with the teacher sitting beside the student (literally or metaphorically) and gaining an understanding of what the student knows and can do in order to help them move on in their learning ([James, 2010](#_ENREF_21)).

The OECD defines assessment as “judgements on individual student progress and achievement of learning goals. It covers classroom-based assessments as well as large-scale, external assessments and examinations” ([OECD, 2013, p. 59](#_ENREF_30)).

Geoff Masters suggests that the fundamental purpose of assessment is to establish where learners are in their learning at the time of assessment ([Masters, 2013a](#_ENREF_22)). According to Steve Dinham, the purpose of assessment is to provide answers to four key questions:

* What can I do?
* What can’t I do?
* How does my work compare with that of others?
* How can I do better? ([Dinham, 2016, p. 39](#_ENREF_11))

Alternatively, assessment may answer the question “How does my work relate to expected standards?” and “What might be the next steps in my learning?”

# Formative and Summative Assessment

Summative assessment, or assessment of learning, aims to summarise learning that has taken place, in order to record, mark or certify achievements ([OECD, 2013, p. 140](#_ENREF_30)).

Formative assessment, or assessment for learning, aims to identify aspects of learning as it is developing in order to deepen and shape subsequent learning ([OECD, 2013, p. 140](#_ENREF_30)).

Assessment as learning occurs when students are their own assessors. Students monitor their own learning, ask questions and use a range of strategies to decide what they know and can do and how to use assessment for new learning. Black and Wiliam define formative assessment as:

all those activities undertaken by teachers – and by their students in assessing themselves – that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs ([Black & Wiliam, 1998, p. 161](#_ENREF_9)).

The term ‘feedback’ is a key term in Black and Wiliam’s account.

Dunn and Mulvenon define formative assessments, as “assessments designed to monitor student progress during the learning process” ([Dunn & Mulvenon, 2009, p. 3](#_ENREF_12)). But if these assessments are not used as feedback to modify teaching and learning activities, then they are indistinguishable from mere summative assessments or assessment of learning.

All formative assessments (assessment for and as learning) are also summative (assessment of learning). Not all summative assessments are formative (e.g. where students are not informed of the results of their assessment or are informed after a long delay). It all depends on the purpose of the assessments (not the timing of the assessments). Formative assessments are part of good instructional practice. Assessment is not an add-on to good teaching but a key part of it.

Hanover Research fairly summarises the research about the impact of formative assessment. Its findings indicate that students who receive formative assessment perform better on a variety of achievement indicators than their peers do. While many of these studies exhibit methodological limitations, researchers agree that the practice of assessment for learning shows promise in its ability to improve student performance ([Hanover, 2014, p. 3](#_ENREF_19)).

The effect size of teacher to student feedback is 0.75 – not the highest effect size on Hattie’s list but the most available for daily practice for the classroom teacher ([Hattie, 2012](#_ENREF_20)).

# Integrating assessment into pedagogy

Hattie summarises his approach to teaching as "know thy impact" ([Hattie, 2012, p. ix](#_ENREF_20)), but to know thy impact, thou must first assess it. Knowing one’s impact involves feedback. This might be through formal or informal assessments – sitting beside students, as it were, and listening to them. Hattie characterises his approach as "visible learning". For Hattie learning becomes visible only through assessment or feedback.

Hattie argues that expert teachers are aware of their impact by evaluating the effects that their teaching has on their students, rather than committing to particular teaching strategies. Student achievement increases when teaching and learning are “visible” to both teachers and students. More specifically:

1. Teachers are among the most powerful influences in learning and have the greatest impact when they can provide meaningful and appropriate feedback to each student
2. Teachers and students need to know the learning intentions and the criteria for success for their lessons, how well they are attaining these criteria for all students, and where to go next
3. Teachers need to move from the single idea to multiple ideas, and to relate and then extend these ideas such that the learners construct, and reconstruct knowledge and ideas; the learner’s construction of this knowledge and ideas is crucial ([Hattie, 2012, p. 22](#_ENREF_20)).

Assessment is an integral component of feedback to students, knowing learning intentions and the criteria for success, attainment of these criteria and where to go next.

Barron and Darling-Hammond stress the importance of assessment in successful inquiry-based pedagogies – though they apply to all types of pedagogy. This includes:

* the design of intellectually ambitious performance assessments that define the tasks students will undertake in ways that allow them to learn and apply the desired concepts and skills in authentic and disciplined ways
* the creation of guidance for students’ efforts in the form of evaluation tools such as assignment guidelines and rubrics that define what constitutes good work (and effective collaboration)
* using formative assessments to guide feedback to students and teachers’ instructional decisions throughout the process
* providing guidance to students about the quality of work and interactions they are aiming for
* providing students with frequent feedback about their learning, especially when that feedback takes the form of specific comments that can guide students’ ongoing efforts ([Barron & Darling-Hammond, 2010](#_ENREF_4)).

Barron and Darling Hammond suggest that successful inquiry-based approaches require planning and well thought out approaches to assessment. They argue that well-designed formative assessment and opportunities for revision support learning and well-designed summative assessments can be useful learning experiences. Formative assessment:

* provides guidance to students about the quality of work and interactions they are aiming for - clear criteria given in advance. The criteria used to assess performances should be multidimensional, representing the various aspects of a task rather than a single grade, and openly expressed to students and others in the learning community
* is a critical element in learning generally, and is especially important in the context of long-term collaborative work
* is designed to provide feedback to students that they can then use to revise their understanding and their work
* informs teaching so it can be adapted to meet students’ needs ([Barron & Darling-Hammond, 2010](#_ENREF_4)).

The use of particular modes of assessment has benefits.

* exhibitions, projects and portfolios provide occasions for review and revision to help students examine how they learn and how they can perform better
* student presentation of their work to an audience – teachers, visitors, parents, other students can be an excellent way of learning. This approach to assessment can be used to assess students' mastery. Presentations of work, particularly public presentations, can signal to students that their work is significant enough to be a source of public learning and celebration. It can provide opportunities for others in the learning community to engage with student work.
* performances can embody representations of school goals and standards so that they remain vital and energising, and develop important capabilities. Good performance tasks are complex intellectual, physical and social challenges stretch students’ thinking and planning abilities while also allowing student skills and interests to serve as a springboard for developing capabilities. Use of performance tasks is also important to assess the benefits of inquiry-based approaches ([Barron & Darling-Hammond, 2010](#_ENREF_4)).

# The role of assessment in representing student progress

Assessment is vital to knowing where each student has been, where they are and where they are going in terms of their learning ([Dinham, 2016, p. 20](#_ENREF_11)). To this end an assessment framework such as Bloom’s taxonomy or the SOLO taxonomy provides a valuable map that can guide this work.

The Structure of the Observed Learning Outcome (SOLO) taxonomy ([Biggs, 1995](#_ENREF_5); [Biggs & Collis, 1982](#_ENREF_6), [1991](#_ENREF_7)), provides a systematic way of describing how a learner’s performance grows in complexity when mastering varied tasks. The SOLO taxonomy postulates five levels of increasing complexity in growth or development of concepts or skills:

* Prestructural: The task is engaged, but the learner is distracted or misled by an irrelevant aspect belonging to a previous stage or mode.
* Unistructural: The learner focuses on the relevant domain and picks up one aspect to work with.
* Multistructural: The learner picks up more and more relevant and correct features, but does not integrate them.
* Relational: The learner now integrates the parts with each other, so that the whole has a coherent structure and meaning.
* Extended abstract: The learner now generalises the structures to take in new and more abstract features, representing a new and higher mode of operation ([Biggs & Collis, 1991, p. 65](#_ENREF_7)).

Hattie interprets the SOLO levels as meaning:

|  |  |  |
| --- | --- | --- |
| **SOLO Level** | **Hattie’s interpretation** | **Relationship Hattie’s level of understanding** |
| Pre-structural | No idea | No understanding |
| Unistructural | An idea  | Surface Understanding |
| Multistructural | Many ideas |
| Relational | Relating ideas  | Deep Understanding |
| Extended abstract | Extending ideas |

Hattie suggests that the Unistructural and Multistructural are about surface learning, and relational and extended abstract are about deeper processing. He comments “Together, surface and deep understanding lead to the student developing conceptual understanding.” ([Hattie, 2012, pp. 60-61](#_ENREF_20)).

Implicit in the SOLO model is a set of criteria for evaluating the quality of a response to (or outcome of) a task. The quality (or richness or complexity) of a response to a complex task varies with the relevance of the considerations brought to bear on the task, the range or plurality of those considerations, and the extent to which these considerations are integrated into a whole, and extended into broader contexts to create something new.

The SOLO framework can be used to assess the quality of a performance in a task. It can be used to assess the quality of an individual performance, the performance of a group working collaboratively on a task, and the contribution of an individual to a group performance. SOLO can be used to design learning and assessment tasks and sequencing of learning tasks from simpler to more complex. SOLO can be used to document a learning journey – identifying where a learner has been, where they are now, and where they might go next. It can also be said to make student thinking visible.

Recall Hattie’s claim that teachers [and implicitly students] need to move from the single idea to multiple ideas, and to relate and then extend these ideas such that the learners construct, and reconstruct knowledge and ideas; the learner’s construction of this knowledge and ideas is crucial. The SOLO framework provides a useful framework to represent student progress.

Building upon the developmental learning theories of Piaget and Bruner, Biggs and Collis recognised that learners move through distinct modes of functioning, generally corresponding to the following age periods:

|  |  |
| --- | --- |
| **Age** | **Stage** |
| From Birth | Sensori-Motor |
| From around 18 months | Ikonic |
| From around 6 years | Concrete-symbolic (approx. K-Year 10) |
| From around 16 years  | Formal (approx. Years 11 and 12+) |
| From around 20 years | Post-Formal (University/professional practice) |

([Biggs & Collis, 1991](#_ENREF_7)).

Differing from classical stage theory, it is not suggested that each stage replaces the previous one, but that each adds to the available cognitive repertoire. Much of schooling may be recognised as occurring within the concrete-symbolic mode, but the use of multi-modal strategies may be more extensive than previously realised ([Biggs & Collis, 1991](#_ENREF_7)).

An alternative taxonomy or framework to SOLO was developed by Benjamin Bloom and colleagues in 1956. ([Bloom & Krathwohl, 1956](#_ENREF_10)) Bloom’s original taxonomy was organised around six broad Levels: Knowledge; Comprehension; Application; Analysis; Synthesis and Evaluation. Bloom’s revised taxonomy is also organised around six levels: Remember; Understand; Apply; Analyse; Evaluate and Create ([Anderson, 2001](#_ENREF_3)).

The unistructural, multistructural and relational levels are recognised as the "target modes" for teaching; allowing for individual differences, it may be expected that all students should achieve one of these levels as a result of an effective learning experience. In the case of new work, it should be the teacher's objective to assist the students to move from a prestructural state (with no organised or coherent knowledge of the material) to one that is, ideally, relational.

Recent research suggests that the cycle of levels within the modes may be more complex than originally anticipated. In particular, at least two Unistructural-Multistructural-Relational cycles now appear to exist within the concrete-symbolic mode, as observed across a range of mathematical topics in the junior years of high school. ([Pegg, 2010](#_ENREF_31))

# Explicitly stating assessment criteria

Hanover Research states that the literature supports the efficacy of explicitly stated learning intentions and assessment criteria in improving student learning outcomes. Learning intentions, specifically as a part of guided instruction methods, positively impact student learning. Similarly, statement of learning objectives and assessment criteria improve students’ self‐assessment abilities and, as a result, improve learning outcomes ([Hanover, 2014, p. 3](#_ENREF_19)).

# Visible Learning

The conclusions of Visible Learning relate to these themes. He identifies a number of signposts:

Teachers are among the most powerful influences in learning and have the greatest impact when they can provide meaningful and appropriate feedback to each student [formative assessment]

Teachers and students need to know the learning intentions and the criteria for success for their lessons, how well they are attaining these criteria for all students, and where to go next. [Learning Intentions and Explicit Criteria]

Teachers need to move from the single idea to multiple ideas, and to relate and then extend these ideas such that the learners construct, and reconstruct knowledge and ideas; the learner’s construction of this knowledge and ideas is crucial. [The SOLO framework]

# High Quality Tasks and High Expectations

Amosa et al. reported students’ assessment tasks and their performances in response to these performances. The quality of these assessment tasks was based on a 14-item, three-dimensional Quality Teaching scale was used to code the degree to which the tasks exhibit high levels of quality pedagogy, within each of the three dimensions: Intellectual Quality, Quality Learning Environment, and Significance (see NSW DET, 2003). Each of the 14 items (classified as elements) in the guide for coding assessment practice is coded on a one to five scale. The 14 items are

**Intellectual Quality**: Deep Knowledge; Deep Understanding; Problematic Knowledge; Higher Order Thinking; Meta-Language; Substantive Communication

**Quality Learning Environment**: Explicit Quality Criteria; High Expectations; Student Direction

**Significance:** Background Knowledge; Cultural Knowledge; Knowledge Integration; Connectedness; Narrative.

Amosa et al found that “Increasing the rigour of intellectual demands of assignments significantly enhances student authentic performance and has the capacity to close the achievement gap between poor and wealthy students. Similarly, making more explicit high expectations about the quality of student work has a positive and significant effect on student authentic achievement, and can reverse traditional achievement patterns for students of Aboriginal and Torres Strait Islander descent” ([Amosa, Ladwig, Griffiths, & Gore, 2007, p. 6](#_ENREF_2)).

Rosenthal and Jacobson found that teacher expectations can significantly impact student achievement. The researchers gave teachers false information about the IQ scores of a group of students indicating that those students were on the brink of rapid intellectual growth. Those students whom teachers expected to perform well showed significantly higher gains in intellectual growth than their classmates at the end of the year ([Rosenthal & Jacobson, 1968](#_ENREF_34)). Many subsequent studies have since supported these findings ([Workman, 2012](#_ENREF_39)). Teacher expectations can be conveyed by the rigour of intellectual demand of assessment.

# Reporting practices

## Reporting A-E

The A-E Common Grade Scale reflects an international trend towards Standards Based Assessment and Reporting.

Research conducted in the United States presents a range of findings relating to reporting practices ([Hanover, 2011](#_ENREF_18)). One of the most contentious aspects of traditional grading practices is the incorporation of teacher judgements about student character and behaviour, including the amount of effort used in completing work, student attitude, timeliness, even bringing in classroom resources. Other traditional grading practices which are critiqued include the use of teacher expectations of students, awarding zeros as punishment, using a points systems and averages, grading homework and formative assessments, grading on a curve, and allowing students to compete for extra credit.

Many of these pitfalls could be avoided if teachers saw themselves as assessing student work rather than the students themselves.

The report identified that the best grading practices only address students’ academic performance. These grades are then able to provide specific and useful feedback. Understanding assessments as a tool to measure and improve student academic achievement entails the use of standard-based grading (also referred to as proficiency-based or criterion-referenced).

Standards-based assessment includes:

* Grading students entirely (or almost) on how well they progress toward learning objectives
* Measuring only a student’s most recent level of mastery over the course material
* Incorporating only summative assessment such as tests or essays, not formative assessments like homework
* Using formative assessments to provide qualitative feedback to students and parents
* Allowing students to redo summative assessments until demonstrating proficiency
* Using rubrics to define the specific learning criteria against which teachers will compare a student’s proficiency level (SOLO is an example of a framework that could be used to develop a rubric)
* Providing scores corresponding to performance standards rather than an aggregate value

It was found that the assumptions commonly-held by teachers about the connection between grading and student motivation were not well supported. Reducing the amount of work that was graded and avoiding the use of grades as rewards for good behaviour and completion of work, including homework, did not reduce the rates at which students completed homework if grades were replaced with descriptive feedback. Exemplar assessment systems are provided as part of the report.

When compared with traditional A-E grading in which teachers would assign a single overall grade for each subject, one study found that parents overwhelmingly preferred standards-based grading, in which teachers provided marks for individual standards within subjects ([Swan, Guskey, & Jung, 2014](#_ENREF_37)). Teachers surveyed as part of this study were similarly positive, but to a significantly lesser extent than the parents mostly due to the increase in workload.

As NESA explains “Achievement standards have two important components. These can be thought of in terms of **what** and **how well**

**what** students are expected to learn

and

**how well** they have achieved.

The NSW syllabuses state **what** students at each stage are expected to learn. The A to E grade scale describes **how well** students achieve ([NESA, 2016](#_ENREF_29)).

The Common Grade Scale describes performance at each of five grade levels.

|  |  |
| --- | --- |
| A | The student has an extensive knowledge and understanding of the content and can readily apply this knowledge. In addition, the student has achieved a very high level of competence in the processes and skills and can apply these skills to new situations. |
| B | The student has a thorough knowledge and understanding of the content and a high level of competence in the processes and skills. In addition, the student is able to apply this knowledge and these skills to most situations. |
| C | The student has a sound knowledge and understanding of the main areas of content and has achieved an adequate level of competence in the processes and skills. |
| D | The student has a basic knowledge and understanding of the content and has achieved a limited level of competence in the processes and skills. |
| E | The student has an elementary knowledge and understanding in few areas of the content and has achieved very limited competence in some of the processes and skills. |

([NESA, 2016](#_ENREF_29))

What are the criteria for Extensive; thorough; sound; basic; elementary?

What are the criteria for very high; high; adequate; limited and very limited?

There is scope for variation in teacher judgements about these evaluative terms – and what they mean in practical terms. For this reason it may be better to avoid the evaluative terms and adopt more objective developmental criteria such as those in the SOLO framework.

Recall that implicit in the SOLO model is a set of criteria for evaluating the quality of a response to (or outcome of) a task. The quality (or richness or complexity) of a response to a complex task varies with the relevance of the considerations brought to bear on the task, the range or plurality of those considerations, and the extent to which these considerations are integrated into a whole, and extended into broader contexts to create something new.

Using Meerbaum-Salants’ interpretation of the SOLO taxonomy, the following criteria could be used. ([Meerbaum-Salant, Armoni, & Ben-Ari, 2013](#_ENREF_26)):

|  |  |
| --- | --- |
| Grade | Criteria |
| E | The student mentions of uses unconnected and unorganized bits of information which make no sense. |
| D | A local perspective – The student uses or emphasises mainly one item or aspect. Others are missed, and no significant connections are made. |
| C | A multi-point perspective – the student uses or acknowledges several relevant items or aspects, but significant connections are missed and a whole picture is not yet formed. |
| B | A holistic perspective – the student grasps meta-connections. The significance of parts with respect to the whole is demonstrated and appreciated. |
| A | Generalization and transfer – the student sees the context as one instance of a general case. |

Geoff Masters critiques the A-E grading system for not providing sufficient information about the progress that individuals make. “Rather than recognising the progress individual students make, A to E grades judge all students against the same age-based expectations. A student who receives a D year after year is usually unable to see the progress they are making.” ([Masters, 2014](#_ENREF_24))The SOLO framework is not age based and could serve as a framework for identifying progress over time, as the richness of the student’s responses to a task grows.

Further information about student progress can be conveyed by a learning progression.

## Learning progressions

Learning progressions refers to sequences of how ideas and practices develop within domains. They were developed initially in the field of science education. In mathematics, the term Learning Trajectories was adopted, with similar intentions. They are not age-based but rather reflect the building of knowledge required to master a topic or domain.

Learning progressions are content-specific and usually relate to “big ideas” in a learning area, such as natural selection ([Furtak, Morrison, & Kroog, 2014](#_ENREF_16)), living organisms, or sinking and floating; and K-8 curricular specifications for, say, atomic structure ([examples taken from Shavelson & Kurpis, 2012](#_ENREF_35)). They are informed in a “top-down” direction, by seeking the views of content experts, and established curriculum standards, as well as “bottom-up” by seeking to understand how student learning actually intersects with this content.

 “Learning Progressions and Learning Trajectories shift the focus from endpoint mastery to understanding how ideas build upon one another as students develop desired knowledge, skills, and practices in a discipline. By providing a coherent description of how to build more sophisticated understanding of core ideas or skills of a discipline, Learning Progressions and Learning Trajectories provide a framework to align content (desired knowledge and skills), curriculum, instruction and assessment.” ([Solem, Huynh, & Boehm, 2015, p.2](#_ENREF_36)).

Researchers leading the learning progressions work are cognitive scientists – as such, the focus of learning progressions is about the development of abilities in the student, rather than the evaluation of work products. Although there is no reason to think that learning progressions could not be reflected in the development of students’ work over time.

Further information about learning progressions can be provided including:

* Application to a TESOL context: <http://onlinelibrary.wiley.com/doi/10.1002/tesq.176/full>)
* Detailed development of a learning progression ([Furtak et al., 2014](#_ENREF_16))
* Theoretical considerations about the criteria used to propose a hypothetical learning progression/trajectory and the measurements/evidence are being used to empirically define and refine it ([Duschl, Maeng, & Sezen, 2011](#_ENREF_13))

# Recent research trends in standardised assessment

The practice of standardised assessment is often considered in the same context as state-wide accountability and comparative education. Unlike school-developed classroom assessments, standardised assessments are often developed by an external body, such as a testing company or specialist arm of a jurisdictional entity.

The particular emphasis on efficiency is a standout feature of standardised assessment. This feature, in addition to the expected rigour in test reliability and validity – is suited to the use of several principled approaches to the design, development and implementation of standardised assessment.

Principled approaches involve six elements:

* clearly defined assessment targets
* statement of intended score interpretations and uses
* models of cognition, learning or performance
* aligned measurement models and reporting scales
* manipulation of assessment activities, and
* ongoing accumulation of evidence to support validity arguments ([Ferrara, Lai, Reilly, & Nichols, 2016](#_ENREF_14)).

Principled approaches include Evidence-centred Design, Cognitive design systems, Assessment engineering, Berkeley Evaluation and Assessment Research (BEAR) Centre assessment system, and Principles Design for Efficiency. The Evidence-centred Design approach is detailed below, as an exemplar principled approach.

## Evidence-centred Design

According to Masters ([2013b](#_ENREF_23)), the fourth principle of student assessment is that available assessment evidence should be used to draw a conclusion about where learners are in their progress within the learning domain.

The Evidence Centred Design (ECD) approach expands on Masters’ fourth principle. ECD is a systematic approach to designing assessment that uses the idea of constructing an evidentiary argument to make a claim about a student’s knowledge, skills and abilities.

There are different *layers* (or sequences of activities) which comprise ECD. For the purposes of this paper, the layers of *Domain analysis* and *Domain modelling* are of most relevance.

### ECD layers

ECD provides a means to identify the important things to be assessing in a learning area. This is done through *Domain analysis*, which involves gathering substantive information about the learning area (domain) to be assessed, including specific concepts, how knowledge is constructed, how it is acquired, how it is used, and modes of communication that are valued within that learning area. This process provides assessment designers with an understanding about the sorts of things that they should consider assessing in order to be able to make a valid claim about a student’s proficiencies.

The *Domain modelling* layer articulates the argument that connects observations of students’ actions to inferences about what they know or can do ([R. Mislevy & Riconscente, 2006](#_ENREF_27)). Information collected in the domain analysis layer is organised into a systematic format which can be used to build an assessment argument. Design patterns are an example of a tool that is used for the process of domain modelling.

### Design patterns

Design Patterns are a way to make explicit all decisions that an assessment designer can make about an item or a test. This evidentiary argument is made explicit because assessment design choices follow a consistent structure. This consistent structure is expressed through design pattern attributes (or categories) that provide a way to communicate information about assessment.

An important feature of Design Patterns is that they can be written to be content-neutral, so that, for example, a design pattern on critical and creative thinking can be applied to different learning areas. As such, ECD has the potential to provide an integrated approach to assessment across subjects.

The table below provides an example of how design pattern attributes are used to express the options available for consideration in a particular assessment domain – in this case, examples provided are from the [*Design Pattern for Model Revision in Model-Based Reasoning*](http://design-drk.padi.sri.com/padi/AddNodeAction.do?NODE_ID=2222&state=viewNode)*.*

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Definition** | **Example value(s)** |
| Rationale | A statement that describes the rationale or use of the design pattern. It explains a general situation in which it might be applied. | Model-based reasoning concerns making inferences about real-world situations through the entities and structures of a model. When the model is not appropriate for the job at hand, either because it does not fit or it does not adequately capture the salient aspects of the situation, it is necessary to be able to revise the model. |
| Focal Knowledge, Skills and Abilities | The primary knowledge, skills, and abilities (KSAs) targeted by the design pattern. These describe the learning outcomes that can be assessed through items that are informed by the design pattern. | * Ability, in a given situation, to modify a given model so that its features better match the features of that situation for the purpose at hand.
* Recognizing the need to revise a provisional model.
* Modifying the provisional model appropriately and efficiently.
* Justifying the revisions in terms of the inadequacies of the provisional model.
 |
| Additional Knowledge, Skills and Abilities | Other knowledge, skills, and abilities that may be required by this design pattern. KSAs make explicit the construct-irrelevant demands that the assessment situation places on the student which allows the assessment designer to acknowledge these sources of potentially irrelevant variance in a student’s score, i.e., variance caused by factors other than the student’s true knowledge skills and abilities that are the intended measurement targets of the assessment item. | * Ability to detect anomalies not explained by existing model (i.e., model evaluation)
* Familiarity with real-world situation
* Domain area knowledge (declarative, conceptual, and procedural)
* Familiarity with required modeling tool(s)
* Familiarity with required symbolic representations associated procedures (especially statistical methods)
* Familiarity with task type (e.g., materials, protocols, expectations)
* Ability to engage in model use
* Knowledge of model at issue
 |
| Potential Work Products | Ways in which students are able to demonstrate evidence of KSAs. These are things students say, do or make. | * Identification of reasons why the provisional model needs revision.
* Select new elements (via drag-and-drop) to add to provisional model to be consistent with logic of needed model revision
* Choice or production of revised model
* Explanation of reasoning for revised model
* Recordings or transcripts of what students said as they "thought aloud" while revising model
* Computer-kept records of inquiry steps in which model revision steps are embedded
 |
| Potential Observations | Particular aspects of a student’s work products that provide evidence of KSAs. These are qualities of the student’s work. | * Quality and appropriateness of model revisions in order to address inadequacies of provisional model.
* Degree of and appropriateness of general and/or domain-specific heuristics students use to revise their models.
* Quality of the basis on which students decide that a revised model is adequate
 |
| Characteristic Features of Tasks | Aspects of assessment situations that are likely to evoke the desired evidence. These are necessary features of any assessment informed by this design pattern. | * A situation to be modelled, a provisional model that is inadequate in some way, and the opportunity to revise the model in a way that improves the fit
 |
| Variable Features of Tasks | Aspects of assessment situations that can be varied in order to shift difficulty or focus. These are optional dimensions that the assessment writer might consider. | * Complexity of problem situation
* Complexity of the model; i.e., number of variables, complexity of variable relations, number of representations required, whether the model is runnable)
* Group or individual work?
* Is extraneous information provided?
 |

This design pattern can be applied to the revision of a model in many different domains and content areas, such as biology, earth sciences and economics. It can also be used to characterise general capabilities, both in domain-general and domain-specific formats.

# Assessment that supports a diverse range of learners

According to ECD, all assessment involves an inference about how an observed phenomenon relates to a construct of interest. Historically, the growth of standardised assessment has occurred in the context of *marginal inference* (a situation in which the all test conditions and items will favour some students and other aspects will favour others, but these are random differences that average out over the entire testing *population*). Another form of inference which has become of increasing importance is that of *conditional inference* in which the conditions (such as time allocation for test-taking) that change between testing situations are taken into account rather than ignored as an average measure ([R. J. Mislevy et al., 2013](#_ENREF_28)). This is relevant in cases of student accommodation provided in testing circumstances. The movement towards conditional inference can be understood within the context of student-centred learning, which is growing in its impact and adoption by schools.

## Universal Design for Learning

Universal Design for Learning provides a set of principles to guide the development of educational artifacts, including assessments:

**Principle I. Provide Multiple Means of Representation (the “what” of learning).** Students differ in the ways that they perceive and comprehend information that is presented to them. For example, those with sensory disabilities (e.g., blindness or deafness), learning disabilities (e.g., dyslexia), language or cultural differences, and so forth, may all require different ways of approaching content. Some may grasp information best when presented visually or through auditory means rather than the use of printed text alone. Other students may benefit from multiple representations of the content—a print passage presented with an illustrative photographs or line drawings and the use of an audio recording of the print passage.

**Principle II: Provide Multiple Means of Action and Expression (the “how” of learning).** Students differ in the ways that they can interact with materials and express what they know. For example, individuals with significant motor disabilities (e.g. cerebral palsy), those who struggle with strategic and organizational abilities (executive function disorders, ADHD), those who have language barriers, approach learning tasks very differently and will demonstrate their mastery very differently. Some may be able to express themselves well in text but not in speech, and vice versa.

**Principle III: Provide Multiple Means of Engagement (the “why” of learning).**Affect represents a crucial component to learning. Students differ markedly in the ways in which they can be engaged or motivated to learn. Some students enjoy spontaneity and novelty, while others do not, preferring strict routine. Some will persist with highly challenging tasks while others will give up quickly.

In reality, there is no one means of representation, expression, or engagement that will be optimal for all students in all assessment situations; providing multiple options for students is essential

The “infusion” of UDL principles into principled assessment design, such as ECD, described above, has been researched in the context of assessing students with a range of disabilities. This approach makes explicit the construct-irrelevant features of an assessment that may be modified without impacting on the validity of an item (that is, ensuring you are still assessing the knowledge skill or ability that you set out to assess). In a design pattern, they are expressed as Additional knowledge skills and abilities (additional KSAs).

## Assessing students with disabilities

ECD has been used to design for fair assessment in large scale assessment contexts in which students with high-incidence disabilities such as ADHD may attend general education classrooms and be supported with testing accommodations. In these contexts, it is argued that ECD can provide testing conditions in which the exam is made more equitable for all learners by reducing sources of construct-irrelevant variance which may limit students with disabilities ([Haertel et al., 2016](#_ENREF_17)).

The practice of ECD makes the assessment designer aware of the many kinds of additional KSAs that can contribute to faulty inferences about students’ assessment performances. Three broad types of additional KSAs may apply: (1) cognitive background (sometimes referred to as prerequisite knowledge), (2) student needs (perceptual, expressive, language and symbols, cognitive, executive processing, and affective) and (3) technology-related knowledge and skills. If these needs aren’t addressed in the testing situation, they can result in a student’s poor performance despite possessing the knowledge and skills of interest.

The process involves: identifying the focal KSAs that compose the construct being assessed; identifying the knowledge and skills required to successfully complete an item, *which* *are not the target of the assessment* (additional KSAs); and reduce the influence of these additional KSAs on a student’s assessment performance by identifying variable features that can be designed into the assessment and used to provide non-construct relevant supports (this may include additional representations, the use of text-to-speech, etc). This process of linking the additional KSAs to variable features that support performance without compromising the measurement of the construct of interest guards against inappropriate interpretations of a student’s test score. Next, identifying potential observations needed to provide evidence of whether a student has acquired the focal KSAs of interest, which forces the assessment designer to consider whether all students have an adequate opportunity to acquire the knowledge and skills required to perform the focal KSAs. These two processes—identification and mitigation of construct-irrelevant variance and “opportunity to learn”—are suggested to be a way to increase the fairness of the assessment for all students ([R. J. Mislevy et al., 2013](#_ENREF_28)). There are significant applications for using ECD to develop assessments for students with significant cognitive disabilities ([Flowers et al., 2015](#_ENREF_15)).

## Diagnosing students at risk of reading difficulty

The identification of students at the start of kindergarten who are likely to have reading difficulty is a condition for the prevention of future reading disabilities. The typical form of assessment is “static”, involving a single performance sample of letter identification and phonemic awareness skills, primarily associated with the ability to decode text. Limitations of these approaches are three-fold: measures of letter identification and phonemic awareness do not actually assess decoding; frequent floor effects; the single-measure that is used to make changes presumed to carry over into the future needs of a child is problematic.

The use of “dynamic” assessment in contrast uses a pre-test (typically resulting in floor effects) short instructional episode and post-test. The difference to static measures is that whereas the former is indicative of current ability, dynamic measures are considered to measure potential receptivity to instruction. It is therefore more appropriate for resource and support planning.

Prior studies have demonstrated that dynamic measures have yielded very high classification accuracies, whereas static measures tend to be susceptible to false positives. Children from a language background other than English are particularly at risk of such classification.

Petersen et al ([2016](#_ENREF_32)) compared the accurate classification ability of static and dynamic assessments by testing students at Kindergarten and then at the end of grade 1 to see whether the predictions were accurate. It was found that static measures had an accuracy of .72 and .70 (considered poor to fair) and dynamic measures had an accuracy of .83 and .85 (considered good) on a scale from 0.5 (chance) to 1 (perfect). The difference in accuracy was statistically significant.

Dynamic measures captured unique variance to static assessments and there was no benefit to combining the measures over the use of static assessment alone.

The teaching methods used in the process of dynamic assessment did not differ significantly in their classification ability, although the sound-by-sound teaching strategy was trending above that of the onset-rime strategy (see pp.203 – 204 for a short description of both strategies).

Dynamic assessment did not prove to be more accurate for Hispanic children, classifying them at a significantly higher false positive rate than the static assessments, suggesting that this measure may not be appropriate for children with language backgrounds other than English.

It is argued that overall, dynamic assessment is a more accurate tool in being able to predict future learning abilities in reading.

# Teacher judgement and teacher agency in student assessment

There is a fair consensus in the literature that social moderation of *standards based assessments* through teacher judgement is an effective alternative to statistical moderation by scaling tests. For example, social moderation, that involves teachers meeting together to discuss and negotiate judgments made on student work, is an alternative to statistical moderation and scaling ([Adie & Klenowski, 2016](#_ENREF_1)).

Moderation involves teachers in discussion and debate about their interpretations of the quality of student work. It is through discussion and debate that teachers negotiate a shared understanding of the qualities of the work in terms of the standards. Moderation is also understood as an important part of teacher professional learning that informs future teaching practice. This understanding informs teaching practice where standards are shared with students and are used formatively when feedback to the student can be directed at the standard achieved with further feedback provided about how the student might progress to the next level.

In education systems which include teacher-based assessment and teacher professional judgment, moderation is a form of quality assurance, increasing the dependability and comparability of the assessment results.

The purpose of moderation is to ensure that assessments align with established criteria and standards; are equitable, fair and valid; and that judgments are consistent, reliable, and based on evidence within the task response or assessed work. This process of moderation involves discussion of assessment tasks, evidence of learning, criteria, standards and judgment decisions to ensure the validity and reliability of assessments, and to improve the quality of the teaching and learning experience.

At moderation meetings teachers must:

* share the rationale for their judgments
* remain open and respectful to others’ perspectives
* consider only the evidence of student learning contained within the work sample in judgment decisions
* match the critical evidence against a standard descriptor.
* engage in conversations that critique and interrogate the qualities in student work
* record reasoning for the awarded grade is recorded based on the critical
* evidence that is indicative of the standard.

This is a process for making well-reasoned judgements about the quality of students’ work ([Adie & Klenowski, 2016](#_ENREF_1)).

Another activity that could be added to this list is the collaborative development of rubrics.

McMahon and Jones report a study on successful school based assessment involving five teachers and 145 secondary school students who completed a chemistry task. Test responses were de-identified and uploaded to a website where the teachers judged the student responses Teacher judgement offers promise for moderating teacher assessment, and improving student learning though the use of peer assessment ([McMahon & Jones, 2015](#_ENREF_25)).

Brookhart (2017) examined the use of teacher judgement for summative assessment in the USA. The research looked at studies of teacher classroom summative assessment, i.e. teacher grading practices and concerns about the validity and reliability of their judgements and studies of how teacher judgement accords with large-scale summative assessments. This article provided evidence that both kinds of studies have found mixed results and the paper concludes accountability polices in the USA have therefore generally preferred to use standardised tests instead of teacher judgement measures.

# Innovative practice trends in assessment

The Validation of Assessment for Learning and Individual Development (VALID) tests in New South Wales represent better practice in standardised tests. Three assessments are delivered: Science and Technology;6, Science 8 and Science 10.

All three VALID tests are interactive multimedia assessments that are completed online.

The tests contain both extended response tasks, which provide an opportunity to assess higher order thinking and deeper understanding of a scientific concept or big idea and short response and multiple choice items presented in item sets with a stimulus comprising videos, animations, graphics, audio and/or text. All items in a set are contextually linked to the stimulus.

VALID assesses stage content and skills as described in the Science Years K–10 Syllabus. Assessment items are framed in real-life situations related to self, the family and the community. The items allow students to demonstrate their understanding at both local and global levels. VALID also includes survey questions which identify student values and attitudes towards science (and technology).

The tests are all criterion-referenced (or standards based). VALID assessments are constructed against an assessment framework based SOLO.

Using SOLO, the mode of thinking required to demonstrate Stage 4 outcomes is the concrete symbolic mode. This mode becomes available and is developed by students in primary and junior secondary school years. Responses in the concrete symbolic mode include abstraction of concepts and operations using symbolic systems such as written language and numbers, and application of logic and order to describe and explain the experienced world.

Levels 1, 2 and 3 of VALID can be described as reflecting student responses that show a common sense understanding of science and science concepts using little technical language. Levels 4, 5 and 6 of VALID can be described as reflecting student responses that show understanding of science and science concepts using technical and/or scientific language as outlined in the syllabus.

As well as being a summative assessment, VALID also serves as a formative assessment. Thus VALID provides:

* High quality syllabus-based assessment frameworks that describe levels of achievement against standards. These can be used by teachers to inform school-based assessment practices.
* Online multimedia assessment tools that provide diagnostic data and time efficient reports.
* Data that can directly inform planning, programming and assessment across learning stages.
* Systems and processes to monitor student achievement of syllabus outcomes and content across stages.
* External student assessment data that can be used in combination with school-based assessments to inform whole school planning processes.

# Theoretical issues

Biggs and Collis see the SOLO framework as a framework for assessing a student’s work – performances or artifacts – and not the student (their knowledge, skills and abilities). Similarly, Adie and Klenowski see teachers moderating assessments as assessing the quality of student work. Neither see teachers as seeking to infer anything beyond the quality of the work – about the authors’ knowledge, skills, understandings.

By contrast, Evidence Centred Design does seek to infer from test performances knowledge, skills and abilities. “Special attention is given to the role of probability-based reasoning in accumulating evidence across task performances, in terms of belief about unobservable variables that characterize the knowledge, skills, and/or abilities of students.” ([Robert J. Mislevy, Russell G. Almond, & Lukas, 2003](#_ENREF_33)) Proponents of Evidence Centred Design refer to assessment as involving and evidentiary argument.- inferring cognitive states from task performances.

To put it more succinctly, Biggs and Collis and Adie and Klenowski see the practice of assessment as assessing student work – not the student. ECD see assessment practice as assessing the student, by way of inference from their work.

Unlike Biggs and Collis ([Biggs, 1995](#_ENREF_5); [Biggs & Collis, 1982](#_ENREF_6), [1991](#_ENREF_7)), or Adie and Klenowski ([2016](#_ENREF_1)), Evidence Centred Design appears to embrace cognitive science – that thinking can best be understood in terms of representational structures in the brain and computational procedures that operate on those structures. Cognitive Science has faced strong philosophical challenges:

**The world challenge**: Cognitive Science disregards the significant role of physical environments in human thinking, which is embedded in and extended into the world.

**The body challenge**: Cognitive Science neglects the contribution of embodiment to human thought and action.

**The dynamical systems challenge**: The mind is a dynamical or complex system, not a linear computational system.

**The social challenge**: Human thought is inherently social in ways that Cognitive Science ignores ([Thagard, 2014](#_ENREF_38)).

Evidence Centred Design is a valuable approach to developing assessments, particularly standardised assessments. Validity, reliability and efficiency are key characteristics of standardised tests. Can notions of validity, reliability and efficiency in assessment be understood without the theoretical trappings of cognitive science?

ECD seeks to generate assessments that in the words of Binkley “provide a window into students’ understandings and the conceptual strategies a student uses to solve a problem. Further, by making students’ thinking visible, assessments thus provide a model for quality practice.” ([Binkley, 2012, p. 25](#_ENREF_8)).

The SOLO framework also seeks to make student thinking visible, if we suppose that a student’s performance or artifact *is* their thinking. There is no need for any inference or argument to make thinking visible. But what would a concept like validity look like if we do not suppose that we are assessing the student and their states of mind?

A related issue is that Biggs and Collis and Adie and Klenowski see assessment as involving reasonableness. In the case of Biggs and Collis, assessment involves identifying the degree to which students bring a range of consideration (or reasons) to bear on a particular issue. In the case of Adie and Klenowski moderation assessments involves a reasoned discussion of criteria for evaluating the quality of students work. ECD, by contrast, uses more formal logic in making an inference, or rational argument, from the evidence of students work to their knowledge, skills and abilities.

Could ECD preserve its insights around notions of validity, reliability and efficiency by employing a notion of reasonableness closer to Biggs and Collis and Adie and Klenowski? These issues are not explored or even discussed in the literature. They are important questions for further research.

#  References

Adie, L., & Klenowski, V. (2016). Teacher Judgement and the challenges of moderation. In M. Peters (Ed.), *Encyclopedia of Educational Philosophy and Theory*. Singapore: Springer.

Amosa, W., Ladwig, J., Griffiths, T., & Gore, J. (2007). *Equity Effects of Quality Teaching: Closing the Gap*. Paper presented at the Australian Association of Research in Education, Fremantle

Anderson, L. W. a. K., D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives.* Boston, MA: Allyn & Bacon.

Barron, B., & Darling-Hammond, L. (2010). Prospects and challenges for inquiry-based approaches to learning. In H. Dumont, D. Istance & F. Benavides (Eds.), *The Nature of Learning: Using Research to Inspire Practice* (pp. 199-225). Paris: OECD, Centre for Educational Research and Innovation,.

Biggs, J. B. (1995). Assessing for learning: Some dimensions underlying new approaches to educational assessment. *The Alberta Journal of Educational Research, 41*(1), 1-17.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the Quality of Learning - the SOLO Taxonomy*. New York: Academic Press.

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. In H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement*. Hillsdale, NJ: Lawrence Erlbaum.

Binkley, M. e. a. (2012). Defining Twenty‐First Learning Skills. In P. Griffin, McGaw, B., & Care, E., (Ed.), *Assessment and Teaching of 21st Century Skills*. Dordrecht: Springer.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *The Phi Delta Kappan, 80*(2), 139-148.

Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: Longmans, Green.

Dinham, S. (2016). *Leading Learning and teaching*. Melbourne, Victoria: ACER Press.

Dunn, K., & Mulvenon, S. (2009). A Critical Review of Reserach on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education. *Practical Assessment, Research and Evaluation, 14*(7).

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: a review and analysis. *Studies in Science Education, 47*(2), 123-182. doi: 10.1080/03057267.2011.604476

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled Approaches to Assessment Design, Development, and Implementation. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 41-74). New York: Wiley-Blackwell.

Flowers, C., Turner, C., Herrera, B., Towles-Reeves, L., Thurlow, M. L., Davidson, A., & Hagge, S. (2015). *Developing a Large-scale Assessment Using Components of Evidence-centered Design: Did it Work?* Paper presented at the National Council on Measurement in Education, Chicago, IL.

Furtak, E. M., Morrison, D. E. B., & Kroog, H. (2014). Investigating the Link Between Learning Progressions and Classroom Assessment. *Science Education, 98*(4), 640-673. doi: 10.1002/sce.21122

Haertel, G., Vendlinski, T., Rutstein, D. W., DeBarger, A., Cheng, B. H., Zikar, C., . . . Ructtinger, L. (2016). Assessing the Life Sciences: Using Evidence-Centered Design for Accountability Purposes. In H. Braun (Ed.), *Meeting the Challenges to Measurement in an Era of Accountability*: Routledge.

Hanover, R. (2011). Effective Grading Practices in the Middle School and High School Environments. https://njctl-media.s3.amazonaws.com/uploads/Effective%20grading%20practices%20in%20the%20middle%20school%20and%20high%20school%20environments.pdf.

Hanover, R. (2014). *The Impact of Formative Assessment and Learning Intention on Student Achievement*. Washington DC: Hanover Research.

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. London: Routledge.

James, M. (2010). Educational assessment: overview. In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (3 ed., Vol. 3, pp. 161-171). Oxford: Elsevier.

Masters, G. (2013a). Reforming Educational Assessment: Imperatives, Principles and Challenges. Melbourne Victoria.

Masters, G. (2013b). Towards a growth mindset in assessment. Melbourne: Australian Council for Educational Research.

Masters, G. (2014). Why A to E grades don't tell the whole story. *The Conversation*.

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment *Assessment in Education: Principles, Policy and Practice, 22*(3), 368-389.

Meerbaum-Salant, O., Armoni, M., & Ben-Ari, M. (2013). Learning computer science concepts with Scratch. *Computer Science Education, 23*(3), 239-264. doi: 10.1080/08993408.2013.832022

Mislevy, R., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., . . . Vendlinski, T. (2013). A “conditional” sense of fairness in assessment. *Educational Research & Evaluation, 19*(2/3), 121-140.

NESA. (2016). Using A to E grades to report student achievement.

OECD. (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*: OECD.

Pegg, J. (2010). *Promoting the acquisition of higher-order skills and understandings in primary and secondary mathematics*. Paper presented at the Australian Council for Educational Research Research Conference, Crown Conference Centre, Melbourne.

Petersen, D. B., Allen, M. M., & Spencer, T. D. (2016). Predicting Reading Difficulty in First Grade Using Dynamic Assessment of Decoding in Early Kindergarten: A Large-Scale Longitudinal Study. *J Learn Disabil, 49*(2), 200-215. doi: 10.1177/0022219414538518

Robert J. Mislevy, Russell G. Almond, & Lukas, J. F. (2003). A Brief Introduction to Evidence Centred Design *ETS Research Report Series*.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the Classroom*. New York: Holt, Rinehart, and Winston.

Shavelson, R. J., & Kurpis, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science: Current Challenges and Future Directions* (pp. 13-26). Boston, MA: Sense Publishers.

Solem, M., Huynh, N. T., & Boehm, R. (Eds.). (2015). *GeoProgressions. Learning Progressions for Maps, Geospatial Technology, and Spatial Thinking: A Research Handbook*. Washington D.C.: National Center for Research in Geography Education.

Swan, G. M., Guskey, T. R., & Jung, L. A. (2014). Parents’ and teachers’ perceptions of standards-based and traditional report cards. *Educational Assessment, Evaluation and Accountability, 26*(3), 289-299.

Thagard, P. (2014). Cognitive Science. *Stanford Encyclopedia of Philosophy*.

Workman, E. (2012). Teacher Expectations of Students: A Self-Fulfilling Prophecy? The Progress of Education Reform. *ERIC, 13*(6).