

THE COMFORT OF COMPETENCE AND THE UNCERTAINTY OF ASSESSMENT

Patrick Griffin

Assessment Research Centre, The University of Melbourne, Australia

In 1977, Arieh produced his Handbook of Curriculum Evaluation (published by UNESCO and Longman) and in 1991 he produced 'National and school-based development' (published by UNESCO/IIEP). In both of these major publications, one major hurdle was to find ways of producing and using information from tests that could have a direct input to the formulation and revision of curricula. Patrick Griffin has shown how, by using modern psychometric techniques within a Vygotsky framework, this can be done.

Abstract

This article argues that a *probabilistic* interpretation of competence can provide the basis for a link between assessment, teaching and learning, curriculum resources and policy development. Competence is regarded as a way of interpreting the quality of performance in a coherent series of hierarchical tasks. The work of Glaser is combined with that of Rasch and Vygotsky. When assessment performance is reported in terms of competence levels, the score is simply a code for a level of development and helps to indicate Vygotsky's zone of proximal development where the student is ready to learn.

Measurement Theories

In the 1960s two developments provided insights into the interpretation of educational measurements; later, in the 1980s, when powerful, accessible computing technology became available, the value of the two developments became more apparent. Rasch (1960, 1980) and Glaser (1963, 1981) each opened new ways of thinking about

learning and measurement, and proposed the concept of underlying growth continua, or latent traits. They reasoned that the nature of these traits could be defined by the tasks that students performed; if the tasks were to be arranged in order of their increasing amounts of attribute required (student capability), then the nature of the trait could be defined by the nature and order of the tasks and the skills they demanded. Development of learning or competency could be traced by following progress along the trait or growth continuum.

Glaser (1963) proposed the concept of criterion-referenced interpretation of assessments. Like Rasch (1960, 1980), he described performance and development in terms of the nature and order of tasks performed. Initially, when criterion-referenced interpretations of assessment were used, observations were referred (or compared) directly to a single, fixed level of achievement or pre-specified criterion. If this level of performance was demonstrated, it was interpreted in terms of either mastery or non-mastery, by referencing to a single cut-off score. Only one threshold was used. Glaser (1963) originally used the term *criterion* to refer to a defined domain (area) of content or behaviour to which the test items were referenced. Despite its attraction in separating the individual's performance from those of his or her peers, the "can/cannot" or "mastery/non-mastery" interpretation could and did, for more than a decade, reduce assessment and curriculum to a level of trivia or to a checklist of potentially unrelated and non-cohesive sets of skills that were of little use in instructional terms. It led Glaser to expand on his original statement and clarify the purpose for criterion referencing. Glaser advised that the "mastery/non mastery" tasks should be ordered in coherent sets that lead to an overall interpretation of proficiency or competence. He argued that criterion referencing should "... encourage the development of procedures whereby assessments of proficiency could be referred to stages along progressions of increasing competence" (1981, p. 935)

For the most part, however, assessments were based on dichotomously scored "right/wrong" items mainly due to their ease of administration and the cost efficiency of scoring. Andrich (1978) and Master's (1982) work elaborating the Rasch model to allow for rating scales or partial credit, meant there was no longer a need to define tasks as having only one outcome, approach or solution, and no need to restrict the tasks to paper-and-pencil exercises scored in a predetermined way. Judgment could be used in interpreting performance on more complex tasks, and criteria of the performance quality. Like the ordering of correct responses to multiple choice test items, the quality criteria could be ordered according to the amount of competence required to demonstrate the level of performance quality described by the criteria. The performance could then be interpreted in terms, not only of the tasks completed, but also of the quality and order in which they were performed. When defined in this way, criterion referenced interpretation was based on a description of performance in terms of the order and sequence of tasks performed and quality of the behaviors displayed. Descriptions of these levels of quality then formed the descriptions of the *stages along progressions of increasing competence*. Interpretation in terms of a single task performance no longer sufficed. The idea of a criterion also had to change and be considered as a threshold separating levels on a developmental continuum rather than the domain of content (Glaser, 2005).

Once stage descriptions were established, a person's development could be interpreted by referring to the relative positions of stages in a continuum. At the time of Glaser's redefinition, there were few of these continua available but, increasingly, more

readily understood progressions of increasing competence became available, providing the framework for interpreting progress in terms of increasing competence in many domains of learning. The idea of developmental continua led to changes in curriculum at national levels in several countries such as Australia, UK, Hong Kong and Canada.

Two other developments made criterion referenced interpretation procedures easier to develop and use. The first was a generalization of the idea of testing. Instead of being seen as exercises that predominantly focused on multiple choice cognitive tasks, tests could be represented as tasks or procedures performed under specified conditions (Black, 1988). The elaboration of the Latent Trait Theory to encompass rating scales (Andrich, 1978) and Partial Credit (Masters, 1982) also eased the reliance on dichotomously scored items. The second development was the increasing availability of powerful microcomputers, which allowed analyses that could not be carried out some forty years ago when Glaser first formulated criterion referencing or when Rasch proposed his idea of the latent trait model.

Measurement

Thurstone (1925) and Thorndike (1927) and Rasch (1960, 1980) used the statistical characteristics of a group's responses to a set of items to determine the nature of measurement scales. Their ideas and later developments were grouped under the heading "latent trait models" to help interpret observations, but the trait definitions did not exist in any physical or physiological sense. This was also true of competency definitions: the descriptions of increasing competency helped us to observe and interpret behaviors by inferring a latent trait which we called competence. This helped in discriminating between persons based on their observed performance quality. Each developmental competency was a verbal description that aided observation and interpretation; it did not exist in its own right. When we modelled data on the observations, we took the construction one step further. We developed a statistical representation of a verbal description that, in turn, described an unobservable development of a trait.

Latent Traits

Latent trait theory was concerned with the relationship between the demands of tasks and the capacity of the person to perform them. Until recently the idea was restricted to the application of multiple choice tests and much of the literature still focuses on this. However, it was possible to generalise the idea of latent traits beyond those defined by multiple choice tests. There need be no restriction on the nature of the task and, in the most general of the Rasch models (Linacre, 1990), there were very few restrictions on the scoring procedures. The task could be a test question, a set of multiple choice items, an essay, a performance, a speech, a product, an artistic rendition, a folio, a driving test, the dismantling and reassembling of a motor car engine, building a brick wall, giving a haircut to a client, or whatever was related to some attribute of interest. The attribute could be an ability, an attitude, a physical performance, a procedure, an interest, a set of values or a generalised competence in an area of learning.

In standardised multiple choice tests, the tasks had a single correct outcome and candidates selecting the correct alternative were assumed to have demonstrated a specific

skill that underpinned the correct alternative. The items differentiated those who "can" from those who "cannot" and the probability of this happening was determined by the amount of the latent trait possessed by the candidate and the amount demanded by the item. The test item was defined by a single rubric or criterion that had to be demonstrated by the candidate. When test tasks did not have a single, correct outcome, a series of criteria that defined ordered thresholds of performance quality could be used. Each threshold could be considered as a single behaviour to be demonstrated. In this case the interpretation was whether a candidate demonstrated behaviour *at least as good as* the criterion behaviour. The criteria could be placed on a continuum according to the amount of latent trait required to demonstrate the behaviour described by the criterion. This was equivalent to the way the skills that defined the correct answers in a multiple choice test could be ordered according to their difficulty. Once the skills were ordered, a latent trait could be identified and described. This process was called a skills audit. Rasch also showed that, when the amount of latent trait possessed by the candidate was equal to the amount needed to demonstrate the criterion behaviour, the probability that the person could demonstrate the behaviour was 0.50. This was an important idea in defining a person's ability, but it was crucial to the assessment being used to improve learning, identify appropriate teaching resources and to development of curriculum policy.

Constructivists and Learning

The idea of ordering criteria and locating the criterion where the probability of success for each person is 0.50 can be linked to Vygotsky's research which was driven by questions about the development of human beings and the role that formal education plays in the process. The challenge for educators was to identify students' emerging skills and provide the right support at the right time at the right level. It was in this context that Vygotsky's construct of the Zone of Proximal Development (ZPD) - the zone in which an individual is able to achieve more with assistance than he or she can manage alone - was conceptualised. Burbules (1993) described the ZPD as a "state of readiness in which a student will be able to make certain kinds of conceptual connections, but not others; anything too simple for the student will quickly become boring; anything too difficult will quickly become demoralising" (p. 122).

The teacher therefore needed to be able to identify the ZPD or "the state of readiness" in the domain of learning being mastered. These ideas clearly have implications for teaching and learning practice particularly as evidenced in studies of "modelling" and "scaffolding".

The measurement theories of George Rasch were consistent with Vygotsky's ideas. The zone where success had odds of 50:50 pointed to the location on a continuum or trait where intervention had the best chance of assisting development. Glaser's words *stages along progressions of increasing competence* were also important in the assessment of competency development. Criterion referenced interpretation incorporated Vygotsky's ZPD and this, in turn, was formalised when ZPD and criterion referencing were linked to latent trait theory. This linked the relative positions of a person and an item on the developmental continuum to an interpretation of what a student could learn (with assistance) at the point

where the odds of success were 50:50. It shifted the emphasis away from focusing on a score indicating performance and yielded a substantive interpretation of the measurement that informed teaching and learning. It moved reporting away from the use of scores or grades to descriptions of levels of development.

The combination of these three theories gave meaning to measurements or test scores and enabled them to be interpreted in terms of Glaser's levels of increasing competence and, when interpreted in a Rasch-like manner, they provided an indicator of the point of intervention where learning could be "scaffolded". This was a radical change in the interpretation of a test score. It was not the summative information of the past. This interpretation of the score gave the starting point for instruction, not the end point indicating a level of attainment. In this way, testing was diagnostic and the score pointed to the ZPD where students were "ready to learn".

Teaching and Curriculum Implications

A serious challenge was how to make this explicit to classroom teachers. Psychometric methods were used by a small number of specialists in latent trait analysis and they, in turn, depended on sophisticated mathematical algorithms and computer programs. These methodologies were typically beyond the interest or training of classroom practitioners, but the theory provided a starting point for understanding the nature of the developmental progressions and providing a classroom approach to trait theory. The use of Rasch modelling software provided some easily interpretable graphics. These included graphics called variable maps, fit maps and Guttman scale distributions.

Variable maps provided a way of showing the relationship between tasks and persons as shown in Figure 1. These were produced by computer programs like Quest (Adams & Khoo, 1995), Conquest (Wu, Adams & Wilson, 1998), RUMM (Andrich, Lyne, Sheridan Luo, 2002) or Winsteps (Linacre, 1990). A schematic description of the procedure for interpreting competency progressions was provided by Griffin (2004) in a response to Russell's (2004) argument on validity of cross national testing and an example, based on the World Bank sample survey of Vietnam's Year 5 students (World Bank, 2004a, 2004b).

The analysis of a 60-item reading test is represented in Figure 1. The variable map shows students (represented by an "x") on the left of the figure and item numbers on the right side of the figure. The height on the scale of the x indicates the relative ability of the student x and the difficulty of the item is represented by the height of the placement of the item number. Where students x are at the same level as the item (number), the ability of the student is equal to the difficulty of the item and the odds of success are 50:50. This identified the ZPD for the students in terms of the skills required to answer the item correctly (or demonstrate the behaviour described in the item criterion). It was possible to define the stage of increasing competence by interpreting groups of item criteria where they appeared to cluster at similar levels on the latent trait.

Logits	Students	Reading Items
4.0	XX	
	XXXX	
	XXXXX	
3.0	XXXXXXXX	
	XXXXXXXX	
	XXXXXXXX	
	XXXXXXXX	54
2.0	XXXXXXXXXXXXXXXXXXXX	26 58
	XXXXXXXXXX	
	XXXXXXXXXXXX	21 48
	XXXXXXXXXXXXXXXXXXXX	44
	XXXXXXXXXXXX	33 34
	XXXXXXXXXXXXXXXXXXXX	3
1.0	XXXXXXXXXXXXXXXXXXXX	30 55 57 60
	XXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXX	22 43 45 51
	XXXXXXXXXXXXXXXXXXXX	9 14
	XXXXXXXXXXXXXXXXXXXX	1 13 42 50 53 59
0.0	XXXXXXXXXXXXXXXXXXXX	4 17 47 49
	XXXXXXXXXXXXXXXXXXXX	8 10
	XXXXXXXXXXXX	23 25 27
	XXXXXXXXXX	12 29
	XXXXXXXXXX	7 15 31
	XXXXXXXX	5 11 56
-1.0	XXXXXX	18 19 20 28
	XXXXXX	2 24 37 46
	XX	6 36 52
	XXX	
-2.0	X	16 32 35
	X	

Figure 1: Variable Map of 60-Item Reading Test

The variable map shows that items were grouped in clusters of similar difficulty levels. Because ability and difficulty were mapped on the same scale and in the same units, the students also could be grouped in clusters of approximately the same 'ability' range, adjacent to the items that had similar range in difficulty. The grouping of items (and students) identified a series of 'transition points' or thresholds in difficulty of the tasks and in the ability of the students and these could be situated between the clusters. They indicated where a discernable change in item difficulty was associated with a change in the kind of cognitive skill (or ability) required to provide correct answers. The item clusters are described in Table 1. Each item was analysed to identify the cognitive skill involved in responding correctly. The result of this is shown in Figure 2.

Item #	Difficulty	Level	Cognitive skill underpinning the correct response
41	4.2	6	Link a concept to a visual stimulus and bring outside knowledge to the
54	4.0	6	Combining several ideas and using outside knowledge (format-all of the
58	3.8	6	Combining several ideas, requiring interpretation beyond text level
26	3.8	6	Combining several ideas (format-all of the above)
21	3.5	5	Combining several ideas and using outside knowledge
48	3.4	5	Understanding figurative meaning of word (format-negative question)
44	3.3	5	Understanding author's main purpose on the basis of the title
34	3.1	5	Requiring interpretation beyond text level, unfamiliar topic
33	3.1	5	Deducing meaning from context
32	3.0	5	Combining several ideas (format-all of the above)
55	2.8	5	Understanding main idea, choosing a title
30	2.7	5	Inferring meaning from context (format-negative question)
60	2.7	5	Understanding figurative meaning
57	2.7	5	Locating specific information from text
40	2.7	5	Locating specific information from text
43	2.4	4	Locating specific information from text amid competing information
38	2.4	4	Locating specific information from text (too many details in long options)
22	2.4	4	Locating specific information from text amid competing information
45	2.3	4	Locating specific information from text amid competing information
51	2.3	4	Locating specific information from text amid competing information
9	2.3	4	Integrating reading and math skills
39	2.2	4	Locating information from text & illustration
14	2.2	4	Locating information from text (understanding signal words-'prediction' in
50	2.1	4	Inferring meaning from context (option d attracts some above average
53	2.1	4	Locating information from context amid competing information
59	2.1	4	Understanding author's main purpose
1	2.0	4	Match exact words and paraphrase from Chinese origin
13	2.0	4	Locating information from text (format - negative questions)
42	2.0	4	Locating information from text amid competing information
47	1.9	4	Understanding implications
4	1.8	4	Understanding implications
49	1.8	4	Locating information from text amid competing information
17	1.8	4	Understanding meaning of sentences
10	1.7	3	Understanding meaning of vocabulary
8	1.6	3	Integrating reading and math skills
27	1.5	3	Understanding meaning of sentences
25	1.4	3	Locating information from text
23	1.4	3	Locating information from text
12	1.4	3	Locating information from text (format - negative questions, using
29	1.3	3	Locating information from text
31	1.2	2	Locating information from text
15	1.1	2	Understanding meaning of words
7	1.0	2	Locating information from text
56	1.0	2	Locating information from text
11	1.0	2	Locating information from text
5	0.9	2	Understanding author's main purpose
18	0.8	2	Locating information from text
19	0.8	2	Understanding relationship between events in text
28	0.7	2	Locating information from text
20	0.7	2	Locating information from text
37	0.7	2	Locating information from text
2	0.6	2	Locating information from text
46	0.6	2	Understanding relationship between events in text
24	0.5	2	Understanding meaning of word
52	0.4	2	Locating information from text
6	0.4	2	Locating information from text
36	0.4	2	Matching word and visual stimulus
16	0.1	1	Exact match of text with adjacent text
32	0.0	1	Match exact words and paraphrase
35	0.0	1	Matching word and visual stimulus

Figure 2: Skills Audit for Each of the 60 -Item Reading Test

The students whose ability estimates were adjacent to the clusters of items can be shown to have odds of approximately 50:50 of being able to demonstrate the skills required by the items in the cluster. The odds of 50:50 at the transition points could be linked to a change in the type of cognitive skill and this could be translated into an implication for teaching. It was the zone of proximal development (ZPD) - a zone where the student was ready to learn. If the student were to improve a little, there would be a better than a 50:50 chance of succeeding on items in the adjacent group. It was not the level of development where success or failure was consistent. The main task of a teacher was to increase the odds of success of students in each of these competency levels to a point where the odds of success were greater than 50/50. To do this, the teacher needed to identify the kind of instruction needed to progress the student from one level on the variable to the next. This involved an understanding of the kinds of skills being developed by students at the relevant *stages along the progression of increasing competence*.

Table 1: Item Clusters

Group	Item in cluster
1	(from item 16 to item 35)
2	(from Item 7 to Item 52)
3	(from item 8 to item 29)
4	(from item 22 to item 49)
5	(from item 21 to item 60)
6	(from item 54 to item 58)

Grouping items on the variable map was a first step. The two pieces of information - item difficulty and underpinning skill - were then explored together to determine whether a common substantive skill interpretation could be found in clusters of items. There had to be an identifiable change in difficulty, and it was necessary for the skills audit to illustrate a change in the substantive skill on either side of the threshold between clusters and a common substantive interpretation of the items within a level between contiguous thresholds. The levels for the example reading test and their interpretation are presented in Figure 3.

Reading Skill Levels	
Level 1	Matches text at word or sentence level aided by pictures. Restricted to a limited range of vocabulary linked to pictures
Level 2	Locates text expressed in short repetitive sentences and can deal with text unaided by pictures. Type of text is limited to short sentences and phrases with repetitive patterns.
Level 3	Reads and understands longer passages. Can search backwards or forwards through text for information. Understands paraphrasing. Expanding vocabulary enables understanding of sentences with some complex structure.
Level 4	Links information from different parts of the text. Selects and connects text to derive and infer different possible meanings.
Level 5	Links inferences and identifies an author's intention from information stated in different ways, in different text types and in documents where the message is not explicit.
Level 6	Combines text with outside knowledge to infer various meanings, including hidden meanings. Identifies an author's purposes, attitudes, values, beliefs, motives, unstated assumptions and arguments.

Figure 3: Interpretation of the Reading Levels from the Analysis of Reading Test Item Sets

Intervention Strategies

This article suggests we need to focus on all students and identify the developmental level where their performance is inconsistent. This is the point of intervention for each student. The widely accepted interpretation of diagnostic assessment may need to be broadened (or abandoned) to enable targeted instruction at the ZPD level of development for each student, not just students at the lower levels of achievement.

Derived competency statement	Possible teaching activities and resources
<i>Level 1:</i> Matches text at word or sentence level aided by pictures. Restricted to a limited range of vocabulary linked to pictures	Reading conferences, logs kept by students, shared reading, retelling, drama activities based on reading. Books sorted into difficulty levels and students practice by reading aloud, reading to other students, parents, use 'take-home' books, reading by the teacher to students, reading simple repetitive language pattern picture-rich reading texts repeatedly alone and with other students,
<i>Level 2:</i> Locates text expressed in short repetitive sentences and can deal with text unaided by pictures. Type of text is limited to short sentences and phrases with repetitive patterns.	Reading to, with and by the students in the class reading centre. Shared reading with other students, parents, use 'take-home' books, reading by the teacher to students, guided reading and predicting stories with simple repetitive language pattern and picture-rich reading texts, repeatedly alone and with other students, Reading logs, sustained silent reading, retelling, running records readers' theatres and creative drama.
<i>Level 3:</i> Reads and understands longer passages. Can search backwards or forwards through text to for information. Understands paraphrasing. Expanding vocabulary enables understanding of sentences with some complex structure.	Comparing books and stories, identifying features, exploring common patterns using reading circles, sustained reading activities, discussions with other students and parents and recording reading logs and discussions; role plays, portfolios, individual reading conferences, guided reading programs at the individual student level; retelling; and links to writing instruction, collecting stories and other reading materials from the community. Small group activities and reading centres;
<i>Level 4:</i> Links information from different parts of the text. Selects and connects text to derive and infer different possible meanings.	Guided reading, small-group reading activities reading circles, reading logs reading materials from community and from non fiction, shared reading focussing on strategies for expository texts collected from a range of sources, and related to a range of curriculum learning areas; reading aloud, following directions, keeping reading diaries, writing letters; individual reading conferences, reading logs and response journals
<i>Level 5:</i> Links inferences and identifies an author's intention from information stated in different ways, in different text types and in documents where the message is not explicit.	Reading targets in terms of the number and range of texts and text types. Non fiction should be extensively used, unit and topic research activities as individual and group activities with work set for research at home and in the community. Critical analysis of text materials and evaluation of writers style and effect of style.
<i>Level 6:</i> Combines text with outside knowledge to infer various meanings, including hidden meanings. Identifies an author's purposes, attitudes, values, beliefs, motives, unstated assumptions and arguments.	Shared, guided and independent reading of a broad range of text types and from a range of sources. Retelling and reading circles predominate in teaching styles and activities, sharing insights, clarifying intentions analysing and evaluating texts using a range of criteria such as style, clarity, impact on the reader and so on. Book clubs, think and know charts based on reading, drawing conclusions from a range of texts, following directions, drama workshops, reading and writing discussion groups, literature response portfolios text cohesion analysis sessions.

Note: The examples provided in this article focuses on reading instruction and the examples are taken from the Literacy profiles by Griffin, Smith and Ridge (2001) and by Griffin, Smith and Martin (2003).

Figure 4: Level of Competence and Possible Intervention Strategy

If each level identifies the kinds of skills associated with a ZPD, then each level also needs a distinct teaching strategy. There is no point teaching the students at level six with

the same strategies or the same resources as the students at level one. This is self evident, but we continue to use a uniform approach to teaching as if the class were homogeneous. Accepting that the levels indicate the "readiness to learn" or the ZPD changes the idea of traditional diagnostic assessment which focuses on what some students cannot do and tends to ignore the better students who obtain high test scores.

Implementing this approach needs considerable planning and management at all levels. Adopting targeted teaching where the ZPD is identified requires differential approaches to the use of teaching strategies. These need different resources, which in turn mean that policy level decisions are needed to support such an approach at school, regional and system level. Professional development of teachers will be central to the implications of such an approach. Examples of suggested strategies for each level are shown in Figure 4 above.

For instance, early readers need activities like a listening centre, pocket charts, enlarged text, word walls, poem boxes and buddy reading. They need practice in visualising, retelling, and paraphrasing, all of which help students at the lower levels. Modelling of reading behaviour and group and guided reading activities are also useful. Higher level readers are encouraged by literature or book circles, directed reading, semantic webs, sketches and other innovative approaches. Teachers already know this, of course, but we still do not routinely organise classes to enable specifically targeted intervention.

There are five steps involved, as shown in Figure 5. The first step is the measurement using any form of assessment. From the measurement, identify the level, or, generalise to the ZPD in the domain of learning that the assessment represents. Different levels on the overall developmental progression suggest the use of different teaching strategies. If this is not done, teachers can fall into the trap of teaching to the test, where each item represents a specific skill. Practice sessions are used to help students answer similar test questions but this does not necessarily improve ability development. (For example, coaching for an intelligence test may improve the score, but not the intelligence). Different interventions need different resources. Most testing programs stop at the first step, measurement, and report in terms of scores. The zone of intervention is rarely identified, and subsequent teaching tends to focus on what the students cannot do; resources are not matched appropriately and critics of testing programs charge that testing does not improve learning. Why would we expect it to, when we ignore the information tests or other forms of assessments can provide?

It is argued in this article that the works of Rasch, Glaser and Vygotsky can be incorporated into teaching and assessment cycles. Developing profiles of learner development where stages of increasing competence are defined and used for intervention, resourcing and policy development shifts the emphasis in testing and assessment. A test score, for instance, signals where to start intervention for development, not the end point of instruction. An item or criterion level skills audit can provide the student with a detailed report on development. A developmental approach transcends the content of any one test and allows the student to be monitored on a general developmental construct. It encourages the second step - generalisation. Intervention is therefore based on a generalised development, not on a specific item-based interpretation of learning (or lack of learning). When this is done, intervention can be linked to appropriate provision of resources and this in turn leads to informed curriculum and learning policy.

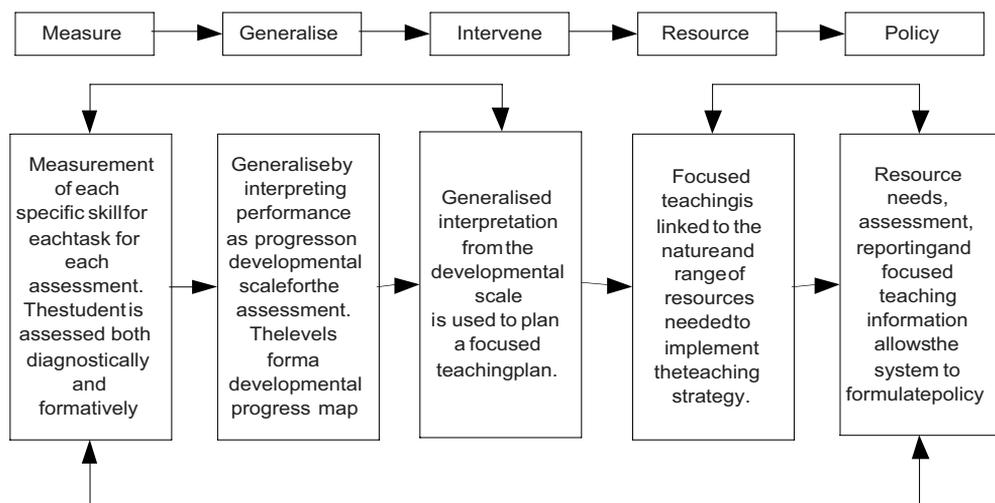


Figure 5: A Five-Step Approach to Developmental Assessment, Learning and Teaching

References

- Adams, R.J., & Khoo, S.T. (1995). *Quest: Interactive item analysis*. Melbourne: Australian Council for Educational Research.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Andrich, D., Lyne, A., Sheridan, B., Luo, G (2002). Rasch Unidimensional Measurement Model: RUMM (2010). Computer software. Perth Western Australia. RUMM Labs.
- Black, P. (1988). *The report of the Task Group on Assessment and Testing*. London, HMSO.
- Burbules, N.C. (1993) Savage inequalities: An interview with Jonathan Kozol. *Education Theory*, 43, (1).
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519-521.
- Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, 36, 923-936.
- Glaser, R. (2005). Personal email communication May 22, 2005.

Griffin, P. (2001). *Performance assessment and higher order thinking*. Paper presented at the Annual Meeting of the American Association for Research in Education. Seattle, April, 2001.

Griffin, P. (2004). Student test development and calibration. *Monitoring primary student achievement in Vietnam*, 2, 64-98. C.P. Shaw, Hanoi, World Bank.

Griffin, P., & Jones, C. (1987). *Assessing the development of reading behaviours: a report of profiles and reading band development*. Annual Conference of the Australian Association for Research in Education, University of New England, Armidale, NSW.

Griffin, P., Smith, P., & Ridge, N. (2001) *Literacy profiles in practice*. Portsmouth, NH: Heinemann

Griffin, P., Smith, P., & Martin, L. (2003). *Profiles in English as a second language*. Portsmouth, NH: Heinemann

Hill, P. (1997). *Professional support for teachers*. The Curriculum Agenda: 1997 and beyond: Fourth annual Curriculum Corporation Conference, Parkroyal Hotel, Sydney.

Linacre, J.M. (1990). *A user's guide to facets: Rasch measurement computer program*. MESA Press: Chicago.

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

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Neilson & Lydiche.

Rasch, G. (1980). *Some probabilistic models for the measurement of attainment and intelligence*. Chicago: MESA Press.

Russell, H. (2004) *Connections among factors in education*. *Curriculum Inquiry* 34:3, 300-307.

Thorndike, E. (1927). *The measurement of intelligence*. New York: Teachers College, Columbia University.

Thurstone, L. (1925). A method of scaling educational and psychological tests. *Journal of Educational Psychology*, 16, 433-49.

Vygotsky, L.S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Vygotsky, L.S. (1986). *Thought and language*. Boston: MIT Press.

World Bank (2004a) *Vietnam: Reading and Mathematics Assessment Study. Volume 2*. Hanoi: World Bank.

World Bank (2004b) *Vietnam: Reading and Mathematics Assessment Study. Volume 3*. Hanoi: World Bank

Wright, B., & Masters, G. (1983). *Rating scale analysis*. Chicago: MESA Press.

Wu, M.L., Adams, R.J. & Wilson, M. (1998) *Conquest: generalised Item Response Modelling Software*. Melbourne ACER Press.

The Author

PATRICK GRIFFIN is the Director of the Assessment Research Centre and the Deputy Dean of Education at the University of Melbourne. He specialises in item response modelling and criterion referenced assessment frameworks and their links to instructional strategies.

Correspondence: <p.griffin@unimelb.edu.au>