Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

Linking Cognitive Psychology and Item Response Models

Kelvin Lai and Patrick Griffin

Paper presented at the 2001 annual conference of the Australian Association for Research in Education, Perth, December 2-6.

Abstract

This paper describes some recent thinking in test development that illustrates the need for a measurement model that allows for appropriate interpretation and use of test data especially where the assessment of higher order thinking is involved. The paper proposed the application of measurement models that allow for dependent components in test items that address problem solving skills in Mathematics and yield evidence of higher order thinking.

Introduction

A central issue in test theory, as described by Gulliksen (1961), focuses on the relationship between the assessee's attribute that the test is measuring and the observed scores on the test. Items, designed specifically for the purpose of measuring a particular ability, are administered to a group and, based on their performance on these items, scores are derived as indicators of their abilities. Thus there are two aspects that test design needs to deal with, the design of items and the construction of scores. These are variously labelled as the*observation* design and the *measurement* design (Snow & Lohman, 1989; Lohman & Ippel, 1993). The former deals with the design of items to indicate the variable being measured. The latter concerns the assignment of a score to an individual.

Griffin and Nix (1991) described two more aspects related to tests as sequels to observation and measurement, they were assessment and evaluation. Assessment is defined as the purposeful observation, *interpretation and description* of evidence of achievement (Griffin and Nix, 1991). Although there are many ways to collect such evidence, tests remain one common and important means of collecting it. Evaluation is defined as the *judgment of value* or implication from results of assessment. Hence, there are, in all, four identifiable components of test development theory: observation, measurement, assessment and evaluation.

Modern views on these aspects

in the 1960s changes in Psychology and Psychometrics occurred that influenced the approaches generally used in observation and measurement. Behavioural psychology was found to be insufficient to account for the complex and varied behaviours in test performance and was in many cases supplemented or even replaced by reference to cognitive psychology. In psychometrics, classical test theory (CTT) was confronted by a new approach called "item response theory" (IRT) which offered a number of advantages over the classical approach (Choppin, 1987). They each had several implications for observation and measurement design in test development.

While there were changes in the emphases of observation and measurement designs, there were also changes in the expectation of the assessment and evaluation. Discontent with the widespread practices in assessment focussed on the routine counting of the number of correct answers or ranking students on a unidimensional. It was argued that assessment

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

should promote students' learning (Griffin and Nix, 1991) and should be in the service of instruction and learning (Snow, 1989). Collis and Romberg (1992) argued that the prime function of assessment and evaluation had changed from a norming or ranking device to become an aid for learning. Griffin and Nix (1991) showed that measurement, assessment and evaluation were hierarchical (p 3). It was not possible to change one of these aspects of test design without altering each of the others. Changes in expectations of assessment and evaluation, for instance, also had an effect on observation and measurement practices. It was clear that the manner in which tests helped teaching and learning depended on whether appropriate information could be derived from the tests and whether correct interpretations could be made and communicated to others based on this information. A careful design of items to elicit the appropriate responses from the students (observation) helps collect appropriate information. Measurement, assigning appropriate scores as indicators of the latent variables, helped in formulating an appropriate description and then influenced the nature of the interpretation.

In short, there were changes in the four aspects. Since the aspects of test theory: (observation, measurement, assessment and evaluation) are interrelated, changes in one aspect would be expected to induce changes in other aspects and combined changes in all these aspects may eventually build a new foundation of test theory.

A theoretical nicety - describing cognitive-information processing using item response models (IRMs)

In psychometrics, item response models (IRMs) founded on "item response theory" (IRT) overcame many of the shortcomings of the classical test theory (Choppin, 1987). In essence, an item response model postulated that a response to an item is determined by a latent, unobserved variable measured on a continuous scale. Each person j has a value, q_j, called the ability parameter on this scale. Each item i also has a value, d_i, called the difficulty parameter along the same scale. Both q_j and d_i are measured in the same units. In the simplest form, the model postulates the probability, P, for the person j to elicit a correct response to the item i as a function of the difference between the measure of the person's ability and the item's difficulty (q_{j-d_i}). Mathematically, it is expressed as

$P = f(q_{j-d_i})$;

where f is a monotonically increasing function of q_{j-d_i}. In the simple

Rasch model (Rasch, 1960), f is the logistic cumulative function

$$f(\theta_j - \delta_i) = \frac{\exp(\theta_j - \delta_i)}{1 + \exp(\theta_j - \delta_i)}.$$

In this setting, inferential procedures in statistics, such as those dealing with the estimation of parameters and testing hypotheses, can be applied to assess individual differences in ability.

However, while an item response model can be used to estimate an individual's position (ability, q_j) on the latent variable, based on the information of the observed test item responses, it does not define the meaning of the underlying trait, nor say anything about the construct validity of the trait (Hulin et al., 1983). It says nothing about what constitutes the difficulty of an item. Often there are situations in practice that one should "look beyond the simple universe of the IRT model - to the content of the item, the structure of the learning

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

area, the pedagogy of the discognitive psychology line, and the psychology of the problem-solving tasks the items demand" (Mislevy, 1993, p26). This is the process of defining and identifying the underlying variable. Griffin (2000) also shows that an analysis of the cognitive strategies needed to solve the items on the latent variable can lead to an interpretation of the nature of the construct. This is consistent with Stenner's (2001) call for the use of a construct theory as part of the test's observation design phase.

Consistent with this is the notion that cognitive information processing in cognitive psychology (CP) may provide a deeper understanding of the mental processes and the content knowledge that underpin the performance by breaking the task down into different, elementary, components. This approach can also lead to a better understanding of the observation design aspect of test development by explicating the various processes, strategies and knowledge that are involved in the item responses. This is what Embretson called the construct representation of a test (Embretson, 1985). In so doing, she also expounded the latent variable in an IRM. However, cognitive psychology does not offer any sufficiently developed methodology for assessing individual differences in the construct representation. It fails to explain why there is a difference in performance on an item if two persons follow the same strategy but yield a different measurement result.

Thus, if an item response model is used to describe the cognitive psychology of achievement, the weakness in each could be complemented by the strengths of the other. If an item response model is used to model each component of a cognitive process, there can be a quantitative model measuring individual person differences in ability and individual item differences in difficulty with respect to the component. The incorporation of a cognitive psychology model into an item response model can enhance the possibility of a deeper qualitative understanding of the latent variable underlying the individuals' performance. This may illuminate the implications from the analyses of the observed responses and may provide important feedback on both instruction and learning.

The integration of a cognitive psychology model and an item response model can lead to multidimensional indices with the dimensions representing different latent variables involved in processing information. Such a multidimensional indicator should be able to account better for the students' performance than a unidimensional indicator and has implications for fit and residuals. In this respect, integrating a cognitive psychology model with an item response model can help build a new foundation for more elaborate test theory. (Snow and Lohman, 1989, 1993; Embretson, 1985a, 1985b, 1993; Bejar, 1984; Messick, 1984). It may also provide a way of explaining more thoroughly, the development of the individual's achievement, using more than a single underlying trait. This has been the single most stringent criticism of IRMs generally used in studies of educational achievement. The latent trait(s) are considered to be multidimensional in most instances. The use of a single variable to account for all the non- randomness in students' responses to a set of tasks has been criticised to be "not a serious representation of cognition, but a caricature" (Mislevy, 1993, p25).

A practical demand - the necessity of looking at the process

Ideas linking cognitive psychology and item response models are usually applied to test items, which are normally scored dichotomously. According to Snow (1989), if assessment is in the service of instruction and learning, it has to be diagnostic or, as Lesh (1990) put it, it should provide profiles of the strengths and weaknesses of the individual students. Griffin (2001) has shown that the analysis and interpretation phase of assessment should also point to intervention and the identification of a student's readiness to learn. In this way, the teaching strategies and instructional material can be devised and organised so that improvements can be effected and strengths used for further learning activities. Thus, tests

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

should be designed in such a way that they enable the tester to read beyond a response or a series of responses.

Merely distinguishing responses to assessment tasks as correct or incorrect does not adequately serve these purposes. For diagnostic purposes, for example, it is necessary to know whether an incorrect response in a mathematics test is due to inability of understanding the information, or lack of the prerequisite mathematical skills or the use of an inadequate level of reasoning (Collis and Romberg, 1992). In other words, an item should be designed in such a way that the reasons for the incorrect response can be identified. For instance, according to Messick (1984), if a correct response to an item depends on an adequate subject knowledge and the possession of certain cognitive abilities, then the cognitive abilities should also be assessed, and assessed separately with achievement so that the source of failure in performance can be identified, whether it is due to inadequate knowledge or to deficiencies in the cognitive abilities. Griffin (2001) shows how these can lead to the identification of a learning hierarchy with recommended actions for teachers within a zone of proximal development (Vigotsky, 1978).

Even with correct responses to dichotomously scored items, assessments also should reveal the thinking that yields specific responses (Romberg et al., 1990) and the processes used (Lester & Knoll, 1993). Indeed a correct final answer can be the result of a whole sequence of correct operations or the results of two mistakes that cancel out one another. In situations where there are different strategies leading to the same answer, the strategies used may reflect different levels of understanding as exemplified by Carpenter's study on children's addition (Carpenter, 1985). The different cognitive strategies used by the children in performing addition actually reflect the different levels of sophistication reached by the children. Assigning the same score to the answers without considering the processes involved, may overlook important information and lead to an incorrect interpretation of the underlying variable. Griffin and Callingham (2000) approached this issue through the use of partial credit in performance tasks by allocating different scores to different quality of performance as defined by the cognitive strategies involved in solving the item or the item step. This enabled them to interpret the underlying trait in terms of cognitive strategies and then to link these to learning readiness, zones of proximal development and ultimately to intervention strategies appropriate for each level on the underlying latent trait.

Thus by examining the cognitive processes behind a student’s response, learning weaknesses and strengths can be identified as well as learning intervention strategies. Teachers can then use these intervention strategies to assist the students to progress. If students were aware of their developmental status and appropriate learning areas, they are in a better position to know which areas in which they should spend more time. Successfully identifying the students' developmental status helps the teacher make good use of those strengths.

For assessment to be of greater service to teaching and learning than is often the case, it is important to gather information on students’ learning using a range of item types. As described by Masters and Mislevy (1993), students learn in a continuous process of structuring and restructuring their framework of knowledge in order to accommodate new knowledge. At different stages, these structures represent different levels of understanding of a new concept. According to Masters and Mislevy, test items should take into consideration "the variety of types and levels of understanding that students have of the concepts" (p219).

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

Schema

Cognitive psychology contends that human memory comprises numerous schemas, each of which is a network of related pieces of knowledge, skills, algorithms or strategies (Marshall, 1990). How well these different pieces are interrelated reflects the different levels of students' understanding. According to Marshall, tests, in addition to testing the absence or presence of specific knowledge in a subject, should also be testing existing connections in the students' knowledge framework and also "the degree to which the student has developed a well-connected body of domain knowledge" (Marshall, 1990, p156).

The National Council of Teachers of Mathematics (NCTM) established a set of Curriculum and Evaluation Standards for School Mathematics, (1989). In them, the NCTM posited that "problem solving should be the central focus of the mathematics curriculum and it must also be the focus of assessment and assessment should determine students' ability to perform all aspects of problem solving". Closely connected to calls for assessment of problem solving is the call for assessing higher order thinking (Baker, 1990; Kulm, 1990a, 1990b; Romberg, Zarinnia and Collis, 1993), which involves non-algorithmic paths of solution, application of multiple criteria, self-regulation of thinking processes and search of structures in apparent disorder (Resnick, 1987).

This suggests that test items should be designed in such a way that both the process and the product are able to be examined. Understanding a concept or the connections between different pieces of knowledge in a schema can be seen from the process students' use, the concepts or related knowledge together to complete a task. Problem solving or higher order thinking skills are more often reflected in the process than in the final answer or product. The skills are revealed as the student interprets the problem, extracts the relevant information from the problem, integrates them to develop a strategy which, step by step, leads to the final solution to the problem.

There are other reasons why knowing the cognitive processes that lead to a response is important. One is the belief that tests should measure developmental and progressive learning (Griffin and Nix, 1991). Wilson (1989) argued that items should be constructed with the students' development and information processing in mind and should not attempt to tap abilities and performance beyond the students' cognitive range. On the other hand, there are important developmental phases in students' learning of a subject and the items should be able to locate such phases for individual students. As Messick (1984) put it, testing should be sensitive to developmental differences in subject-matter learning and performance and, when cast in developmental terms, considerations should not be given only to content but to structures and process as well. Baker (1990) who, with reference to Mathematics, argued that tests should specify the mathematical thinking processes shares his view and abilities expected of students at key points in their mathematical education and development. In order that the test items should meet this developmental criterion, item writers need to take into consideration the cognitive processes leading to each type of response, and the level of cognitive demand needed for each component and for each overall solution.

Tests need to be designed to meet these increasingly more demanding needs. The assessment and evaluative components of test theory need to be emphasised to a far greater degree than at present so that tests can become more informative and able to cope with the new demands for assessment and evaluation as well as observation and measurement. If they move in this direction much of the objection to tests may dissipate. To achieve these, items should be designed with the final answers in mind, and in such a way that the process yielding the answer can be observed. As summarised by Mislevy and Verhlest (1990), "In education, estimates of how students solve problems could be more

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

valuable than how many they solve, for the purposes of diagnosis, remediation, and curriculum revision" (p196).

The importance of dependent subtasks in an item

Lipson, Faletti, and Martinez (1990) argued that a testing system "that can present many different kinds of problems to a student, and that can track and interpret not only the answer, but the intermediate steps in a solution attempt" was needed (p126). They advocated the use of complex multistep problems in order to perform a kind of "cognitive cryptography" to decode the knowledge and the cognitive model that underlies and generates the student's performance (p128). Their view was echoed by Mislevy (1993) who stressed the importance of knowing the intermediate products in an answer to a task.."richer information can be accumulated if it is possible to track intermediate products of solution" and "inferences about skill profiles can be stronger if one can see which subtasks were attempted and their outcomes" (p32). In some measure Griffin and Callingham (2000) and Griffin (2001) have made considerable progress in this direction with the development of the multi step performance tasks yielding partial credit rubrics that reflect the different extent to which cognitive strategies are used in problem resolution.

For instance, in a problem-solving item, a student is provided with the initial state, the given conditions to begin with, and the goal state of the problem, involves completing what is asked to be done (Greeno, 1978). The task of solving the problem is equivalent to performing a sequence of subtasks to change the state of the problem from its initial state through a number of intermediate states eventually to the goal-state. Griffin and Callingham (2000) provided a problem solving framework based on a developmental learning sequence and this led to a sequence of subtasks to progressively lead the student to higher levels of cognitive operation through a number of intermediate steps eventually to the goal-state of hypothesis testing. Each approach used in assessing the problem-solving skill of a student, assumes that it is appropriate to divide the items into parts, each corresponding to a subtask in the sequence, in order to keep track of the overall process. In both approaches it is possible to locate the intermediate state that has been reached and which subtask is unsuccessful. In this way, the strengths and weaknesses of the student can be identified as well as the zone of proximal development together with its possible intervention strategy.

The test

In general, a test consists of a number of items. Each item asks the students to complete a task under some given conditions. To keep track of the process of the task for the purpose of assessment, the task has to be divided into a sequence of steps, each with meaningful subtask and information needs to be gathered on these subtasks. Accordingly, the item needs to be divided into parts corresponding to these subtasks as in Fig. 1.

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

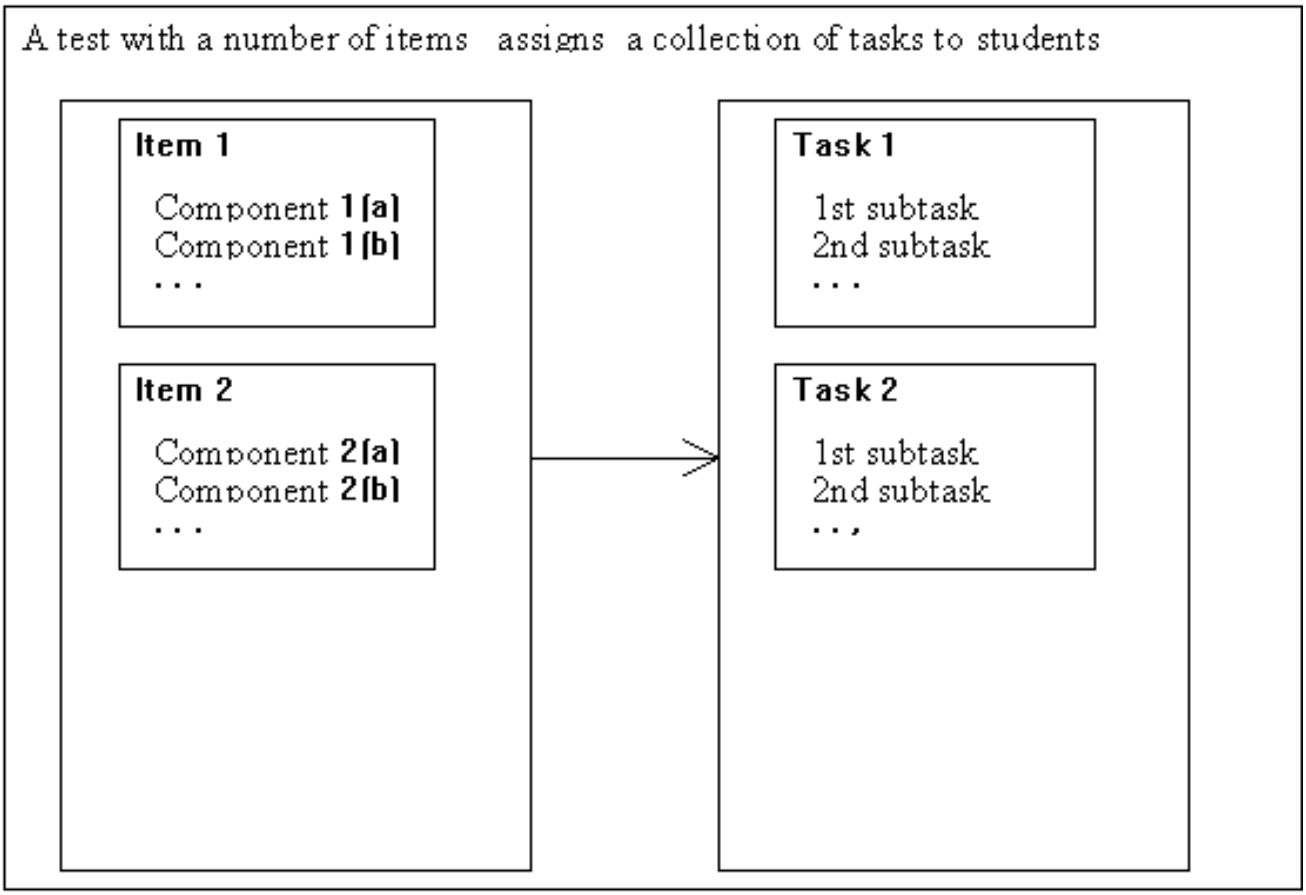

Figure 1 Item structure with tasks and subtasks.

These parts of an item are called components and such an item is said to be a multi-component item. Students are requested to respond to each of these components in turn as he cognitive processing become further embedded within the path towards the problem solution.

In order that the subtasks can provide valid and reliable information on the

students' performance throughout the solution process, items need to be designed with the subtasks related to each other as in the actual process of solution. According to Resnick & Resnick (1992), "Recent cognitive research into complicated skills and competencies shows that success in such areas depends not only on the number of components (subtasks) they engage but also on the interactions among the components (subtasks) ...effort to assess thinking and problem-solving abilities by identifying separate components of those abilities and testing them independently will interfere with effectively teaching such abilities" (p42-43).

There are two messages here. One is that students who are high achievers, confronted with tasks that require isolated skills, may not be successful problem-solvers. Suppose a problem-solving item is divided into various parts, corresponding to the sequence of subtasks necessary for solving the problem. If the item is designed with the subtasks independent of each other, then a student good at demonstrating the isolated skills may complete all the subtasks successfully. If the subtasks, however, are designed to be dependent upon each other as in the actual solution process, success in the item depends also on whether the students can relate the subtasks, and whether they can put together appropriate responses in the subtasks to think of a solution strategy. These are what

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

Resnick and Resnick (1992) refer to as interactions among the components. Thus, there is something missing if only isolated skills are tested with items consisting of independent subtasks. In fact, items with independent subtasks may be argued to be artificial but difficult to design. In order to free the subtasks within an item from dependence upon each other, additional information has to be supplied for each subtask and this may provide hints for the solution, making the response in the subtask fail to reflect the actual ability of the students to solve complex tasks.

The second message contained in Resnick & Resnick's work concerns the washback effect. If tests are to evaluate isolated skills, then in order to improve students' test scores, teachers may direct their instruction to these isolated skills instead of towards the higher-order cognitive processes. This "teaching to the test" effect or the "washback effect of the test" leads to a "curriculum shrinked to fit the test" (Lipson et al., 1990, p122) with some parts of the curriculum not often tested being neglected. This is the very point taken by Griffin and Callingham (2000) who deliberately used the teaching process to design the assessment tasks based on an assumption that positive washback could be helpful in promoting learning and the development of higher order thinking. Amid calls for promoting students' problem solving skills or emphasising higher order thinking in the curriculum, tests designed without this underlying approach would act in the opposite direction and would not be systemically valid in the sense described by Frederiksen & Collins (1989). As Elton and Laurillard (1979, p100) said "the quickest way to change student learning is to change the assessment system". Lefrancois (1991) also echoed this view by remarking that "instructional objectives are communicated very directly and very effectively to the students through the measurement device" (p373).

To achieve systemic validity, tests need to be direct reflections of what it is they are attempting to measure. They need to evaluate the cognitive skills as they are expressed and related in practical problem-solving situations. To achieve this purpose, tests need to consist of items with subtasks, which may be dependent upon each other, and allow the solution process to be monitored as the student progresses towards higher order thinking.

Combining a cognitive model and an item response model

The responses to components of an item serve as indicators of the students' performances on subtasks. Through the application of an item response model to the components, estimates on the difficulties of the components and the abilities of the students can be obtained. From this we can also begin to understand the cognitive processes that are used in solving the problem.

Hence, the components of a test item should be written with reference to the cognitive model for the task in the item. Figure 2 shows the roles of a cognitive model and an item response model in the design of a test. A cognitive model helps elucidate the processes involved in addressing various subtasks (the steps) a student goes through in order to complete a task; the abilities that are involved in completing the subtasks; the relationships among the subtasks and their significance in the overall solution process. It also helps to provide an indication of which subtasks need to be included in each of the item components and in what order the component steps should be arranged in the item. An item response model helps explain the variations among the students' performance in a component in terms of the differences in the estimates of student abilities and the component step difficulties. Exactly what these estimates refer to and how significant they are in the solution process can be explained through a cognitive processing model. The relationships between the models and the aspects of test theory are shown in Figure 2.

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

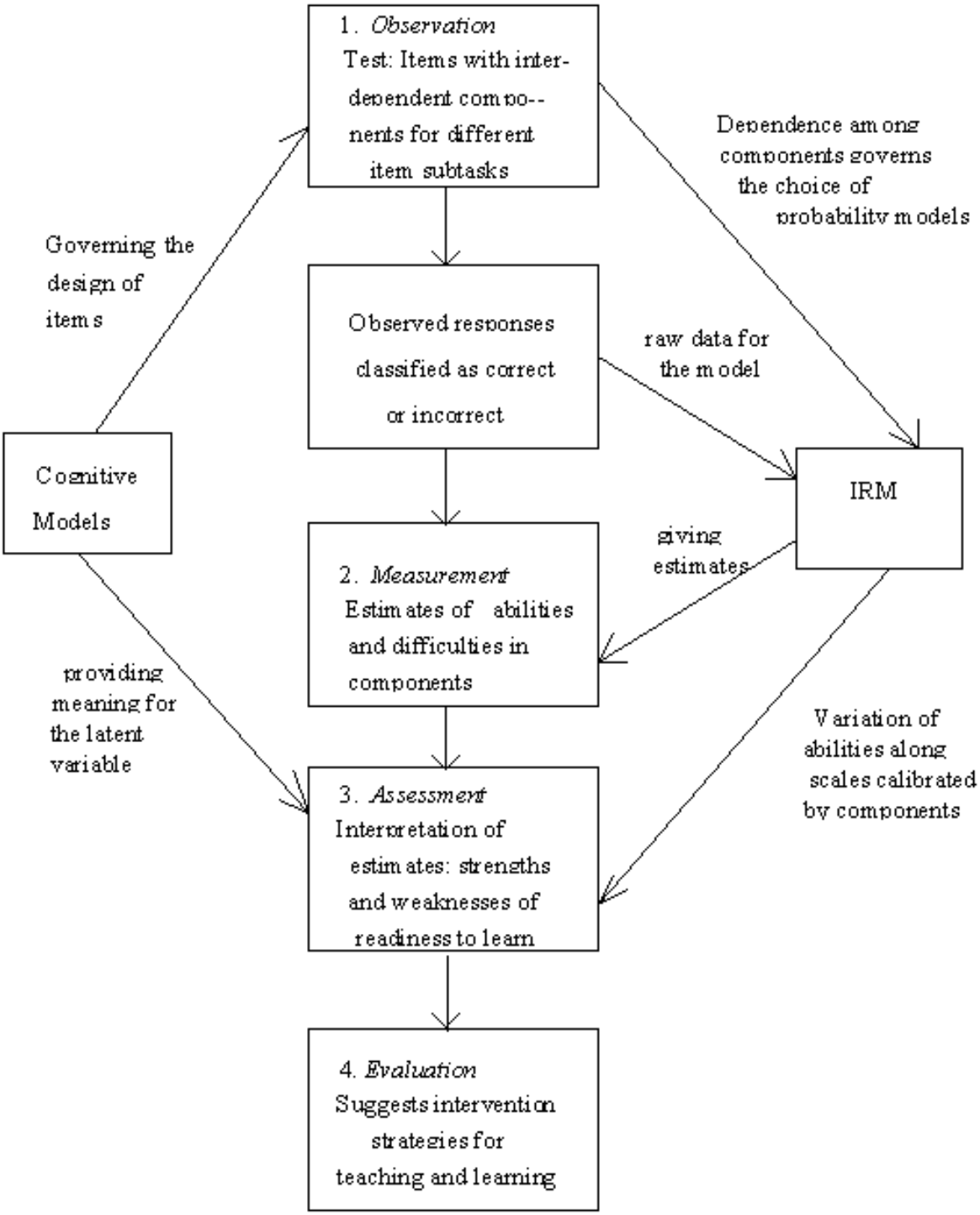

Figure 2. Cognitive and item response models in the four aspects of testing theory

The joint use of cognitive models and item response models presents a problem if the subtasks in the cognitive model are interdependent. In an item response model, the ability parameters and the difficulty parameters are estimated using maximum likelihood methods and this requires knowledge of the likelihood of the set of response data obtained by administering the test to the students. Often, the likelihood of each response in an item response model is calculated under the assumption of local independence, which means that a student's performance in one item is not affected by the performance in another item. But, under a cognitive model an item that has subtasks related to each other may lead to the situation where the components corresponding to these subtasks are dependent. A student, responding incorrectly on one component, is unlikely to respond correctly in a component

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

later in the item if the subtasks in these twocomponents are related and dependent. Hence, components within an item would violate the assumption of local independence. This problem needs to be overcome if item response models are used to model students' performance in multi-component items. These issues have been addressed by Lai (1998, 2001) and by Griffin, Lai, Wu and Mak (2001).

Conclusion

To incorporate the changes in the four aspects of test development, an application of an item response model for cognitive information processing is needed to meet the demands arising from examining students' performance in the process of completing a task. This is better done using interdependent components within an item.

Thus, a new item response model is needed to accommodate violation of local independence at the component level (Lai, 1998). Lai's work allows item response models and cognitive models to work together to provide useful feedback to teaching and learning. Models dealing with situations violating the condition of local independence exist. (Lai, 1998, Embretson, 1987need additional citations here). Many of these, however, approach the problem without reference to cognitive models and accordingly have shortcomings if used to model a cognitive process. Examples of integrating item response models with cognitive models making use of intermediate responses are scarce. Mislevy (1993) described the situation as "the application of the 20th century statistics to the 19th century psychology" (p19). Theoretical advances in modelling have not been fully used to model "students' internal representation of systems, problem-solving strategies, or reconfiguration of knowledge as they learn" (p19). This is an area in test theory and development that needs considerable attention.

Bibliography

Baker, E.L. (1990). Developing comprehensive assessments of higher order thinking. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science

Bejar, I.I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, *21*, 175-189.

Board of Studies, (1995). *Curriculum and standards framework.* Carlton, Victoria: Board of Studies.

Carpenter, T.P. (1985). Learning to add and subtract: An exercise in problem solving. In E.A.Silver (Ed.) *Teaching and learning mathematical problem solving: Multiple research perspectives*. Hillsdale, NJ:Lawrence Erlbaum.

Choppin, B.H. (1987). The Rasch model for item analysis. In McArthur, D. L. (Ed.) *Alternative approach to the assessment of achievement.* Boston, Dordrecht, Lancaster: Kluwer Academic Publishers.

Collis, K.F., & Romberg, T.A., (1992). *Collis-Romberg mathematical problem solving profiles [Kit].* Hawthorn, Victoria: Australian Council of Educational Research.

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

Elton, L.R.B., & Laurillard, D.M. (1979). Trends in research on student learning. *Studies in Higher Education*, 4, 87-102.

Embretson, S. E. (1984). A general latent trait for response processes. *Psychometrika*, *49*, 175-186.

Embretson, S.E. (1985a). Introduction to the problem of test design. In S.E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 3-17). New York: Academic Press.

Embretson, S.E. (1985b). Multicomponent latent trait models for test design. In S.E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 279-294). New York: Academic Press.

Embretson, S.E. (1991). A multidimensional item response model for learning processes. *Psychometrika*, *56*, 495-515.

Embretson, S.E. (1993). Cognitive processes. In Frederiksen, N., Mislevy, R. J.,& Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Fredericksen, J.R.,& Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18(9),* 27-32.

Greeno, J.G. (1978). A study of problem solving. In R. Glaser(Ed.), *Advances in instructional psychology* (Vol. 1, pp. 13-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Griffin, P. (2000). Compentency based assessment of hifher order competencies. Keynote address delivered at the NSW ACEA state Conference, Mudgee.

Griffin, P. (2001). Performance assessment of higher order thinking. Paper presented at the annual conference of the American Education Research Association, Seattle.

Griffin, P. & Callingham, R. (2000). Towards a framework for numeracy assessment. In J.Bana & A. Chapman (Eds.), *Mathematics Education Beyond 2000. Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia.*Freementle, WA: MERGA.

Griffin, Lai, Wu et al. (2001). Modelling atrategies in problem solving. Paper presented at the 2001 annual conference of the Australian Association for Research in Education, Perth.

Griffin, P. & Nix, P. (1991). *Educational assessment and reporting*, *A new approach.* New South Wales: Harcourt Brace Jovanovich.

Gulliksen, H. (1950) *Theory of Mental Tests*. New York: Wiley

.Hulin, C. L., Drasgow, F.,& Parsons, C.K. (1983). *Item response theory*, *application to psychological measurement.* Homewood, Illinois: Dow Jones-Irwin.

Kulm, G. (1990a). Assessing higher order mathematical thinking: what we need to know and be able to do. In G. Kulm (Ed.),*Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science.

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

Kulm, G. (1990b). New directions for mathematics assessment. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science.

Lai, C.P. (1999). Statistical modelling of students' performance in multicomponent tasks. Unpublished Phd thesis. University of Melbourne.

Lefrancois, G. R. (1991). *Psychology for teaching*. Belmont, California: Wadsworth Publishing Company.

Lesh, R. (1990). Computer-based assessment of higher order understandings and processes in elementary mathematics. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science.

Lester, F.K., & Kroll, D.L. (1990). Assessing student growth in mathematical problem solving. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science.

Lipson, J.I., Faletti, J., & Martinez, M. E. (1990) Advances in computer-based mathematics assessment. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science.

Lohman, D. F. & Ippel, M. J. (1993). Cognitive diagnosis from statistically based assessment toward theory based assessment. In Frederiksen, N., Mislevy, R. J.,& Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Marshall, S.E. (1990). The assessment of schema knowledge for arithmetic story problems: a cognitive science perspective. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp.1-4). Washington, DC: American Association for the Advancement of Science.

Masters, G.N., & Mislevy, R.J. (1993). New views of student learning: implications for educational measurements. In Frederiksen, N., Mislevy, R. J.,& Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*, 215-237.

Mislevy, R. J. (1993). Foundations of a new test theory. In Frederiksen, N., Mislevy, R. J.,& Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests.* Hillsdale, New Jersey:Lawrence: Erlbaum Associates, Publishers.

Mislevy, R.J., & Verhelst, N. (1990) Modelling item responses when different subjects employ different solution strategies.*Psychometrika*, *55*, 195-215.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Cophenhagen, Denmark: Danmarks Paedagogiske Institut.

Resnick, L.B. (1987). *Education and learning to think*. Washington, DC:National Academy Press.

Paper Presented at the AARE Annual Conference, Fremantle, 2001
Australian Association for Research in Education

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum:new tools for educational reform. In Gifford, B.R., & O'Connor, M.C. (Eds.) *Changing assessment.* Mass.:Kluwer Academic Press, 37-76.

Romberg, T. A., Zarinna, E.A., & Collis K. F. (1990). A new world view of assessment in Mathematics. In G. Kulm (Ed.), *Assessing higher order thinking in Mathematics* (pp. 1-4). Washington, DC: American Association for the Advancement of Science.

Snow, R.E. (1989). Towards assessment of cognitive and conative structures in learning. *Educational Researcher*, *18(9),* 8-14.

Snow, R.E.,& Lohman, D.F. (1989). Implication of cognitive psychology for education measurement. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-331). New York:Macmillan.

Snow, R.E., & Lohman, D.F. (1993). Cognitive psychology, new test design, and new test theory, an introduction. In Frederiksen, N., Mislevy, R. J.,& Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Stenner, J. (2001). The necessity of construct theory. Rasch Measurement Transactions 15:1 p804-5.

Vigotsky, (1978). *Mind in Society: The development of higher psychological processes.* Cambridge, MA:Havard University Press.

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 479-494.

Whitely, S.E. (1981). Measuring Aptitude process with multicomponent latent trait models. *Journal of Educational Measurement*, *18*, 67-84.

Whitely, S.E., & Schneider, L. (1980). *Process outcome models for verbal aptitude.* (Tech. Rep. NIE-80-1 for National Institute of Education). Lawrence: University of Kansas.

Wilson, V.L. (1989). Cognitive and developmental effects on item performance in intelligence and achievement tests for young children. *Journal of Educational Measurement*, *26*, 103-119.