DOCUMENT RESUME

ED 361 361 TM 020 427

AUTHOR Cheung, K. C.

TITLE On Meaningful Measurement: Issues of Reliability and Validity from a Humanistic Constructivist Information-Processing Perspective.

PUB DATE Aug 93

NOTE 31p.; Paper presented at the International Seminar on Misconceptions and Educational Strategies in Science and Mathematics (3rd, Ithaca, NY, August 1-4, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Affective Behavior; *Cognitive Processes; *Constructivism (Learning); Educational Assessment; Elementary Secondary Education; Error of Measurement; Evaluation Methods; Foreign Countries; *Humanistic Education; *Item Response Theory; Knowledge Level; *Measurement Techniques; Metacognition; Models; Student Evaluation; *Test Reliability; Test Validity

IDENTIFIERS *Meaningful Measurement

ABSTRACT

In the past decade, there have been ample interests in the assessment of cognitive and affective processes and products for the purposes of meaningful learning. Meaningful measurement (MM) has been proposed which is in accordance with a humanistic constructivist information-processing perspective. Students' responses to the assessment tasks are evaluated according to an item response measurement model, together with a hypothesized model detailing the progressive forms of knowing/competence under examination. There is a possibility of incorporating student errors and alternative frameworks into these evaluation procedures. Meaningful measurement drives us to examine the composite concepts of "ability" and "difficulty." Under the rubric of meaningful measurement, validity assessment (i.e. internal and external validities) is essentially the same as an inquiry into the meanings afforded by the measurements. Reliability, measured in terms of standard errors of measurement, is guaranteed within acceptable limits if testing validity is secured. Further evidences of validity may be provided by in-depth analyses of how "epistemic subjects" of different levels of competence and proficiency engage in different types of assessment tasks, where affective and metacognitive behaviors may be examined as well. These ways of understanding MM can be codified by proposing a three-level conceptualization of MM, where reliability and validity are central issues for an explication of this conceptualization. (Author)

On Meaningful Measurement: Issues of Reliability and Validity from a Humanistic Constructivist Information-Processing Perspective

K C Cheung
Faculty of Education
University of Macau
Macau

Paper presented at the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,
Cornell University, Ithaca, New York.
August 1-4 1993

On Meaningful Measurement: Issues of Reliability and Validity from a Humanistic Constructivist Information-Processing Perspective

K C Cheung
Faculty of Education
University of Macau
Macau

ABSTRACT

In the past decade, there have been ample interests in the assessment of cognitive and affective processes and products for the purposes of meaningful learning. Meaningful measurement has been proposed which is in accordance with a humanistic constructivist information-processing perspective. Students' responses to the assessment tasks are evaluated according to an item response measurement model, together with a hypothesized model detailing the progressive forms of knowing/competence under examination. There is a possibility of incorporating student errors and alternative frameworks into these evaluation procedures. Meaningful measurement drives us to examine the composite concepts of "ability" and "difficulty". Under the rubric of meaningful measurement, validity assessment (i.e. internal and external validities) is essentially the same as an inquiry into the meanings afforded by the measurements. Reliability, measured in terms of standard errors of measurement, is guaranteed within acceptable limits if testing validity is secured.

Further evidences of validity may be provided by in-depth analyses of how "epistemic subjects" of different levels of competence and proficiency engage in different types of assessment tasks, where affective and metacognitive behaviors may be examined as well. These ways of undertaking MM can be codified by proposing a three-level conceptualization of MM, where reliability and validity are central issues for an explication of this conceptualization.

INTRODUCTION

Towards the end of the 1980s, the establishment of a national examination system based on performance assessment for quality education reform has broad appeal (for a discussion of the USA scene, see Linn, 1993). The 3Ps - performance, portfolios, and products - were considered as more appropriate than the bubble tests, i.e. the multiple choice tests, in the high-stakes, authentic assessment of student outcomes of schooling. These are the visions being put forward by the educators and policy makers to raise educational standards equitably within the confines of a well-articulated "thinking curriculum" (e.g. California State Department of Education, 1990, p.17). The following two quotations summarize neatly some recent aspirations and concerns of the educational assessment communities regarding the thorny road ahead in the reform of a national examination system and the associated curriculum framework.

Performance-based, high-stakes assessments may well be preferable to the traditional tests in terms of their effects on the ways many teachers and students spend their time and the aspects of the curriculum to which they are forced to pay attention. However, such claims about authentic assessment are circular: They presume what is to be proven. And the witness of the high-stakes testing programs of the 80s regarding the corruption of instruction and of the tests themselves is not encouraging in this regard. Performance-based measures are as corruptible as any multiple-choice measure. In fact, advocates of authentic assessment as a policy lever offer the nation a variation on the old theme of measurement-driven instruction - the principal difference is the form of the measures. (Madaus, 1993, pp.20-21; citation omitted; emphases added).

I believe strongly that performance-based assessments in the hands of teachers, seamlessly integrated in normal classroom routines, should be more useful to them for formative and diagnostic purposes than traditional, standardized tests have ever been. But my analysis here does not deal with teacher use of the 3 Ps, which I strongly endorse, ... (Madaus, 1993, p.11; emphases added).

The question is really that if education is managed by results, will quality goes down. Black (1993)

remarked that assessment can be a powerful aid to the improvement of teaching and learning, or it can do great damage. The origin of this paradox is the tension between formative and summative assessment. He asserted further that the goal - that teaching to the test becomes the same as finding the best way to help students to learn - is very hard to achieve. This is because of the fact that while summative assessment can be improved by broadening the range of methods (e.g. the 3 Ps) and matching these to the learning aims of the curriculum, formative assessment, which is an underdeveloped art, is embedded in the planning and daily practice of teaching. In this regard, Madaus (1993, p.15) noted that much of the rhetoric about outcome-driven instruction and authentic assessment in USA tends to emphasize ends (e.g. standards, equity), not means (e.g. pedagogy, delivery and support system), thereby downplaying the contribution of formative and diagnostic assessment compared with summative assessment.

Consequently, there is a need to shift the focus from the ends of education back to the means for achieving these ends when designing and developing a curriculum and assessment framework. There is also a serious need for validation research to be an integral part of any new system of assessments, one main aim of which is to serve as motivators of student performance on valued activities in school (Linn, 1993, p.6-7). In

the past three years, research and development work with this focus in mind has begun under the rubric of Meaningful Measurement (MM). MM is still in the process of development and some of its concepts, technology, and exemplary studies have been documented in the literature (Cheung, et al, 1990; Mooi & Cheung, 1990-91; Cheung, et al, 1991; Cheung, 1992a; Cheung, 1992b; Cheung & Mooi, in press). What follows is a detailed description of what MM seeks to achieve, its research paradigm and issues of reliability and validity when undertaking MM in the assessment of cognitive and affective processes and products for the purposes of meaningful learning.

MEANINGFUL MEASUREMENT - A DEFINITION

A definition of MM was formally given in Cheung (1992a), where the concepts, technology and examples are discussed as well.

> *Meaningful measurement is quantitative measurement of conceptual and procedural knowledge with qualitative interpretations that should be firmly rooted in a theory of knowing, model of difficult learning, classroom realities, and educational objectives as intended in the programmes of study. (p.2)*

As suggested by this definition, the primary purpose of MM is to help students overcome learning difficulties so as to progress from a lower level of

cognitive functioning to the higher ones. In essence, MM is curriculum-specific, authentic measurement. As such, it entails a clear delineation of educational objectives and understanding of classroom realities and its processes so that students can develop progressively within their zone of proximal (potential) development.

MEANINGFUL MEASUREMENT - A 3-LEVEL CONCEPTUALIZATION

There is an intimate link between MM and test validity (or testing validity). Messick (1989, p. 56) had noted that "test validation in essence is scientific inquiry into score meaning - nothing more, but also nothing less". Thus, MM in essence is an inquiry into score meaning afforded by the measurement processes, as well as an evaluation of consequences of intended test use and its side effects. By linking meaningful measurement directly with test validity, it is hoped that Ebel's (1961) admonition that "it (test validity) is universally praised, but the good works done in its name are remarkably few (p.640)" can be rectified.

Since its inception in 1989, the MM inquiry process entails a three-level conceptualization to be concisely summarized below.

The Epistemological Level - this involves the use of a humanistic, constructivist, information-processing perspective of knowing to guide our conceptions on how

students get to know the curriculum materials presented to them, organize their everyday experiences and feelings into cognitive and affective structures, and engage in meaningful learning within a cultural, social, classroom, group context (for a detailed discussion of this perspective, see Cheung and Taylor, 1991; Cheung, 1993). This level of conceptualization of MM is important not only in the interpretation of score meaning, but also in suggesting humanistic and constructivist teaching approaches in overcoming learning difficulties.

The Conceptual Level - this involves the use of a multilevel, conceptual model of school learning explicating the process and context variables in order to explain how students can engage in their learning and assessment tasks, as well as to explain how students' learning progress can be monitored effectively and efficiently within a schooling context (for a discussion of some useful conceptual models of school learning, see Cheung, 1992a, p.3-7). This level of conceptualization of MM is vital in formulating a construct theory for the MM process, indicating variables and sources of variances needed for construct validation.

The Methodological Level - this level involves a revised rationale on the inquiry processes and psychomet-

ric methods, the rationale of which should be compatible with both the above-mentioned epistemological and conceptual levels of MM. As an example, notions of "ability" and "difficulty" so frequently and freely used in the psychometric literature have to be re-evaluated when "competence" and "affordance" become common basic terms in the cognitive science literature. In fact, "competence" and "affordance" are preferable for the purposes of MM because both of these terms convey the message that the primary purpose of learning and assessment tasks is to afford the students to deploy their conceptual and procedural knowledge so as to demonstrate their level of competence for task mastery. From the information processing perspective, students are aware of the information regarding the utility of the task environment before their competence are orchestrated.

Features discriminating between experts and novices are important considerations and they are very difficult to be measured quantitatively in MM. Hence, it should be reiterated that quantitative measurement with qualitative interpretation in MM is an ideal goal to be achieved by the combined efforts of psychometricians, curriculum experts and classroom teachers in the long-run.

Furthermore, the use of multitrait-multimethod (MTMM) matrices in the assessment of convergent and discriminant validities of psychological measures for

the purposes of construct validation, despite its popularity and continual development in the past three decades, is still a perennial issue yet to be resolved in the future (Campbell & Fiske, 1959). The many unsuccessful attempts of using MTMM raised the conceptual problem of **(1) how a variable and its measurement should be linked** and the methodological problem of **(2) how the nature of methods and their effects might be clarified** (Fiske & Campbell, 1993). The two problems, if adequately resolved, would render MTMM a powerful heuristic in protecting against construct underrepresentation and construct-irrelevant variance in the test (see Messick, 1989, p.34-36). Obviously, there is a need for a coherent three-level conceptualization of any measurement process. In this regard, the following section presents one inquiry paradigm exemplary to the undertaking of MM based on the afore-mentioned three-level conceptualization of the measurement process.

MEANINGFUL MEASUREMENT - AN ENQUIRY PARADIGM

An example of MM seeking to help students construct the "part-whole" concept of fractions for both conceptual and procedural understanding is presented here. In this example, the primary outcome of MM is that teachers can interpret the meaning of item and student score measures. They also identify cognitive processes and barriers so as to teach accordingly within the students'

zone of proximal (potential) development. This example has been documented in detail in Cheung et al. (1990) and Cheung (1992b). The various stages of the MM process are described concisely below:

Stage 1. Learning difficulties encountered by the target students in the acquisition (or "construction" from the constructivist, information-processing perspective) of the "part-whole" concept of fractions are analyzed. Based on an examination of the cognitive science and relevant literature, progressive levels of conceptual and procedural understanding of this fundamental concept are postulated (e.g. Progressive forms of knowing can be postulated to range from the lower level of recognizing names and symbolic notations of fractions, interpreting diagrammatic representations (continuous quantity versus discrete quantity models) of fractions and using diagrams to represent fractions, to the higher level of applying the "part-whole" concept of fractions to the more complex situations involving both the continuous and discrete quantity representation models). Key concepts, such as concepts of the parts and the whole and relationships of parts with the whole, together with the relevant knowledge schemes, such as the counting and partition schemes, are also studied as detail as possible.

Stage 2. The postulated levels of conceptual and procedural understanding provide a basis for the construc-

tion of an instrument measuring the "part-whole" concept of fractions (for details on the construction of the multiple choice test, see Cheung, 1992b). For the purpose of quantitative measurement, a continuum composed of items of varying difficulty to reflect primarily on the progressive forms of knowing the "part-whole" concept of fractions is needed. In this particular example, the item difficulties of the multiple choice questions are moderated by: (1) the different models of representation of a fraction (continuous versus discrete); (2) types of fractions (unit versus non-unit); (3) types of transition from one model of representation to another; (4) types of the perceptual cues or distractors. The context of the questions is also under control.

As far as item writing is concerned, the number of options required for each item is not uniform. From the constructivist perspective, this is the same as the number of predominant alternative conceptions, and therein maximum qualitative item discrimination is achieved by outcome. Also, since by definition alternative conceptions are conceptions viable to the students, guessing and perseverance are hopefully no longer key factors determining success on an item. Test performance would then depend less on luck and effort.

Stage 3 The test is calibrated against a sample of students who have had the opportunity to learn the "part-whole" concept before. Preferably, the students had not received instruction at a higher grade on other concepts of fractions, such as equivalent fractions, thereby preventing them from using the intended cognitive schemes built into the test. The responses are then tested for conformity with a measurement model - e.g. the Rasch Family Logistic Model. Misfitting students and items are treated separately because these would provide valuable information regarding the nature of the "part-whole" construct. The difficulty level (actually "affordance" in MM terminology) of the items, which are sample-free calibrations, are examined to see whether they are clearly segmented and ordered as designed. Simultaneously, each student can be measured quantitatively on both the logit linear and non-linear total score scales. Their performance can be ranked according to the linearized ability (actually "competence" in MM terminology) logit score, which are specifically objective as an endowment of the Rasch Family Logistic Models. Consequently, not only MM is criterion-referenced in accordance with the progressive forms of conceptual and procedural knowing, it is also norm-referenced measurement if a representative sample had been used in the calibration procedure. The Rasch Family Logistic Model, being a probabilistic model,

enables prediction of the students' most probable responses to the items when the levels of item affordances and student competence are known.

Stage 4 After the ordered, segmented progressive levels of the conceptual and procedural understanding of the "part-whole" concept of fractions has been established empirically, clinical interviews of selected students typical of each ordered segment of the construct hierarchy were then conducted to study the emergence and development of the cognitive schemes that undergird the progressive forms of knowing. In this way, inferences about the development of the cognitive schemes can be validated against the qualitative changes in forms of conceptual and procedural knowing of the "part-whole" concept of fractions. In particular, cognitive barriers preventing students from progressing from one level to the next may be studied as well.

Although MM described thus far is essentially conducted as a cross-sectional study, the students themselves are not viewed as the subject of enquiry in the clinical interviews. Rather, in line with Piaget's clinical research tradition, students are viewed as "epistemic subjects" exemplifying the qualitatively different mental structures of the "part-whole" concept of fractions. Hence, the subject of classification of responses is by no means in terms of the cognising stu-

dents, but in terms of the particular forms of knowing in the construct hierarchy which a representative "epistemic subject" possesses. In this way, MM is simulated to become a longitudinal study of "epistemic subjects".

Stage 5 MM is not limited to the modeling of narrowly focussed cognitive constructs. Analyses for problem-solver's actions in the form of problem-solving networks within a domain-specific knowledge base is underway (for a detailed discussion, see Cheung et al., 1991). The assessment of the emotional system using the MM technology has also proved to be very successful (Mooi and Cheung, 1990-91), although issues pertaining to the systematic and consistent response behaviors in the proper deployment of the Likert response scale deserves critical attention (Cheung and Mooi, in press). Moreover, the contribution of the emotional system such as motivational and attentional processes has not been ignored when episodes of problem-solving are analyzed (Cheung and Koh, 1992) and the notions of cognitive transfer based on the humanistic and constructivist perspective have also been included into the MM research agenda.

Of particular mention is that MM has ventured into statistical modeling of key types of problem solving errors committed by "epistemic subjects" of different levels of competence using the Dual Scaling method (see

Cheung, et al., 1991, pp.50-75). Rasch Family Logistic Models cannot be used in the scaling of errors hierarically along the problem solving proficiency continuum because of the unrealistic assumption that key types of errors can be aligned onto a unidimensional trait. Despite both Dual Scaling and the Rasch Family Logistic Models are based on radically different modeling assumptions and requirements, both methods are found to require consistent, systematic response behaviors from well-targeted students - otherwise there is nothing systematic to model and there is insufficient data where it is relevant (see Cheung and Mooi, in press). The use of Dual Scaling renders quantitative measurement possible when the underlying structure of the construct are uncertain regarding its conformity to a latent trait.

VALIDITY AND RELIABILITY IN MEANINGFUL MEASUREMENT

The three-level conceptualization of MM and the five stages of MM provide a concrete basis for a discussion of validity and reliability issues in MM. Validity theory has evolved continuously over time during the past century. In the past decade, there has been a rejection of the old trinitarian doctrine of test validity, i.e. content, criterion-related, and construct validity (for a discussion, see Shepard, 1993). Now, there is a full recognition of construct validity as the whole of validity theory. Messick's (1989)

unified theory of construct validity involves a progressive integration of test interpretation (i.e. score meaning) with test use (i.e. score meaning plus relevance of the test to the applied purpose and utility of the test in the applied setting), buttressed by a thorough examination of value implications of test interpretation, and social consequences and side effects of test use. Validity is formally defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p.13).

Since test validation is regarded as a never-ending process of collecting evidences for each particular interpretation and use of a test, there has been development on an **argument-based approach** to test validation, rendering test validation more manageable than that implicated by Messick's (1989) unified theory of construct validity (e.g. Cronbach, 1988 and 1989). This approach, as explicated and exemplified by Kane (1992), adopts the use of interpretative arguments (or practical arguments) as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. The four basic steps of an argument-based approach to test

validation are as follows:

> *One (a) decides on the statements and decisions to be based on the test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretative argument and refuting potential counterarguments. (Kane, 1992, p.527).*

In a nutshell, an argument-based approach of test validity, while accepting Messick's unified theory of test validity, can proceed to answer directly the question "Does the test do what it claims to do?" (Shepard, 1993, p.444). A distinctive feature of this approach is that interpretative arguments (or practical arguments) can be evaluated practically in terms of plausibility, rather than in terms of their truth values.

Since MM is firmly rooted in a theory of knowing and a conceptual model of difficult learning, a construct theory is readily available for the purposes of construct validation. Theory-based evaluations entails a process of theory testing involving an appraisal of hypotheses and claims, particularly the rival ones, as well as a process of evaluating implicit value assumptions, intended consequences and side-effects. Decisions, being outcomes of hypothesis testing and judgments, are based on rational argumentation and empiri-

cal verification of test interpretation and use. It is noteworthy that the availabilty of a construct theory allows the possibility of addressing both the internal structure of the test content (or the assessment procedure in general) and its responses, and external relations of the test content and its responses to other process and context variables as implicated by the construct theory. These internal and external components of construct validity have been termed **Internal/External Validities** in MM.

A. Internal Validity of MM

The internal validity of MM corresponds to Loevinger's (1957) "substantive" and "structural" components of test validity, or Embrestson's (1983) "construct representation" when she undertook statistical modeling of cognitive processes and products. Shepard (1993) has noted that the internal components of construct validity should "reflect all aspects of the theory that defines a construct, including its subdomains or subconstructs, the expected interrelationships among dimensions of the construct, and the processes believed to underlie test performance" (pp.417-418). She added that it "includes gathering data about all of the traditional psychometric questions regarding item intercorrelations and the like, but also includes questions about the appropriate weighting of different components and the influence of format on what is

tested" (p.418).

The probing of the cognitive processes productive of task performance based on constructivist, information-processing models of cognition and the inclusion of these cognitive processes and alternative cognitive frameworks in test construction are pertinent in informing the internal validity of MM (e.g. see Cheung, 1992b). Using item-response theories such as the Rasch Family Logistic Models whenever response data permit, the hypothesis-testing of the empirical responses in conformity with the progressive forms of knowing, which are resulted in part from expert judgment of content relevance and representativeness of the construct intended to be measured, is one hallmark of internal construct validation (i.e. on the "substantive" component of Loevinger's test validity) under the rubric of MM.

Cheung (1992a) noted that it would be useful to understand the simple Rasch logistic model, therein the same rationale would apply to the Rasch Family Logistic Models, by considering the responding process when a respondent is confronted with a problem task which affords certain sets of competence for its mastery.

The product term of the demonstrated level of competencies of a respondent and the designed level of affordances of a problem task can then be equated to the odds (ratio of success to failure) of accomplish-

ing the problem task. By taking logarithms on both measures of competencies and affordances in order to transform these scales of progression into mutually conformable measuring scale, the resulting probability model governing a respondent's mastery of the problem task becomes the famous Rasch item response model (Rasch, 1960). This revised understanding of the Rasch Model in terms of both terminology and conceptual framework acknowledges that not only the knowledge construction process is specifically objective, but also a similar situation occurs for the calibration and measurement of levels of understanding and proficiency of this acquired knowledge (Cheung, 1992a, p.11-12, emphases original).

This reconceptualization of the Rasch Family Logistic Models as applied to a progressive form of knowing from a constructivist, information-processing perspective is pertinent in evaluating the "structural" component of Loevinger's test validity, where scoring models should be rationally consistent with what is known about the structural relations inherent in behavior manifestations of the construct in question (Messick, 1989, p.43). In fact, under the rubric of MM, the outcome of measurement is that both the linearized logit score and the non-linearized total scale score can be meaningfully interpreted, rendering both norm-referenced and criterion-referenced measurement possible.

Lastly, conformity of responses to the measurement model is a matter of degree to be decided by the researcher. Reliability of both item and person parameters (i.e. the student competence and item affordances) are captured by their standard errors of measurement. Typically, calibration and measurement errors are least when the sample of students to be measured are well-targeted to the set of calibrated items, or vice versa.

External Validity of MM

The external validity of MM corresponds to Loevinger's (1957) "external" component of test validity, or Embrestson's (1983) "nomothetic span" in the modeling of a trait. The construct theory is likened metaphorically to a complex spatial network in which constructs are represented by knots interconnected by strands - the nomological network (c.f. Cronbach and Meehl, 1955). There is thus a distinction between "trait" and "nomological" validity - the former is concerned with the meaning of the measure as a reflection of the construct, whereas the latter with the meaning of the construct as reflected in the measure's relational properties and implications (Messick, 1989, p.46). As previously indicated, method and trait are intimately entangled together, rendering MTMM not so decisive in evaluating the convergent and discriminant validities of constructs, which is pertinent to a consideration of "trait" validity. Quantitative causal modeling of

construct relationships using a conceptual model of school learning such as those proposed under the rubric of MM would constitute a strong approach of construct validation which is pertinent to a consideration of "nomological" validity.

Inspired by Messick's unified theory involving both test interpretation and use, the construct theory should not only model the relations of the intended construct to other constructs, but also those relationships most centrally implicated by an intended test use (Shepard, 1993, p.419). Evidences of "criterion-related" validity is useful for connecting test scores to measures of applied criteria. Messick (1989, p.16) further added that we can "trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects".

Concerning examining the consequential basis of test interpretation and use, Messick (1989) has made the following valuable point :

Constructs are broader conceptual categories than the test behaviors, and they carry with them into score interpretation a variety of value connotations stemming from at least three major sources: the evaluative overtones of the construct labels themselves; the value connotations of the broader theories or nomological networks in which constructs are embed-

ded; and the value implications of still broader ideologies about the nature of humankind, society, and science that color our manner of perceiving and proceeding. (p.59)

Apparently, the humanistic perspective guiding the epistemology and philosophy of MM is extremely useful for an evaluation of consequences of test interpretation and use. Notions of fulfillment of intrinsic needs of students, realization of student inner potentials, acknowledgement of student self-concepts, utilization of concrete everyday experiences to assist student learning, and value inculcation for the development of an all-round personality, amongst others, are all important criteria of evaluating consequences of testing (for a discussion of humanism in mathematics and science education, see Cheung, 1993). Reliability, when assessed within the confines of external validity in MM, refers to the generalizability of decisions across occasions of test interpretation and contexts of test use. In this way, reliability is again a necessary condition for validity because generalization is a key inference in interpretative arguments (or practical arguments), but it is not a sufficient condition because generalization is not the only inference in the argument (Kane, 1992, p.529).

CONCLUSION

Meaningful measurement is a developing research agenda

researching into the statistical modeling of both cognitive and affective processes and products. Meaningful measurement, because of its firm roots in classroom realities and educational objectives as intended in the programmes of study, is in essence authentic measurement - the principal aim of which is to provide meaningful information (both formative and diagnostic) in order to monitor student progress and overcome their learning difficulties. Since the conceptualization of the three-level MM research paradigm is grounded on a humanistic, constructivist, information-processing perspective of knowing and a conceptual model of school learning, it is more context-specific and process-oriented than some of the ill-conceived practices of authentic measurement (for an extended treatment of this topic, see Mitchell, 1992).

The use of an item response model in testing the viability of postulated progressive forms of knowing, and the subsequent detailed examination of cognitive and affective processes and products mimicking a longitudinal study of "epistemic subjects" progressing from a lower level of conceptual and procedural understanding to the higher ones, are hallmarks of the MM research paradigm. This theory-based approach to validation provides a strong approach to collect and present evidences of both the internal and external validities of MM. On the other hand, an argument-based approach

to construct validation of test interpretation and use is also highly recommended since the required evaluation can be based readily on the epistemological and conceptual levels of the MM conceptualization. MM needs to strengthen its research programme, particularly on the evaluation of consequences of test interpretation and use, because this would contribute significantly to score meaning. Both theory- and argument-based approaches to construct validity allow the perfection of a construct theory, which is accompanied by an accumulation of practical wisdom of helping students not only to know more, but better than before within their zone of proximal (potential) development under the guidance of a humanistic, constructivist perspective of knowing and learning.

REFERENCES

Black, P.J. (1993). *Assessment and feedback in science education.* Paper presented at the International Conference "Science Education in Developing Countries: From Theory to Practice", Jerusalem, Israel, January 3-7, 1993.

California State Department of Education (1990). *Education Summit.* Sacramento, CA: Author.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cheung, K.C. (1990). To grow and glow: Towards a model of teacher education and professional development. In W.K. Ho & R.Y.L. Wong (Eds.), *Improving the quality of the teaching profession - An international perspective.* Institute of Education, Singapore. (Paper presented at the Meeting of the International Council on Education for Teaching, Singapore, July 27-31, 1990). (ERIC: ED338575)

Cheung, K.C. (1992a). *On meaningful measurement: Concepts, technology and examples.* Research Paper No.3, Centre for Applied Research in Education, National Institute of Education, Singapore. (Paper presented at the Regional Symposium on Educational Testing, Beijing, China, September 16-20, 1991). (ERIC: ED338649).

Cheung, K.C. (1992b). How teachers can construct multiple choice questions with meaningful test results. *Teaching and Learning - A Publication for Teachers,*

13(1), 71-77.

Cheung, K.C. (1993). Humanism and constructivism in mathematics and science education: Implications for curriculum reform for Chinese communities in Southeast Asia. In C.C. Lam, H.W. Wong & Y.W. Fung (Eds.), *Proceedings of the international symposium on curriculum changes for Chinese communities in Southeast Asia: Challenges of the 21st century.* Department of Curriculum and Instruction, Faculty of Education, The Chinese University of Hong Kong.

Cheung, K.C., Koh, W.K., Soh, K.C. & Mooi, L.C. (1990). *Meaningful measurement in the classroom using the Rasch Model: Some exemplars.* Institute of Education, Singapore. (ERIC: ED326544).

Cheung, K.C. & Koh, W.K. (1992). *Attributions, metacognition and mathematical problem-solving.* Paper presented at the international seminar on the state-of-the-art of research in science and mathematics education in Southeast Asia and the Pacific, SEAMEO-RECSAM, Penang, Malaysia, 17-19 February, 1992.

Cheung, K.C. & Mooi, L.C. (in press). Likert ordered response categories and related issues: A comparison between IRT Modelling and Dual Scaling. *Applied Psychological Measurement.*

Cheung, K.C., Mooi, L.C. & Loh, W.F. (1991). *Meaningful assessment of problem-solving activities in the classroom: Some exemplars.* Research Monograph No.2. Centre

for Applied Research in Education, National Institute of Education, Singapore. (ERIC: ED337488).

Cheung, K.C. & Taylor, R. (1991). Towards a humanistic constructivist model of science learning: Changing perspectives and research implications. *Journal of Curriculum Studies, 23*(1), 21-40.

Cronbach, L.J. (1988). Five perspectives on validation arguments. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.3-17). Hillsdale, HJ: Erlbaum.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Ebel, R.L. (1961). Must all tests be valid? *American Psychologists, 16*, 640-647.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Fiske, D.W. & Campbell, D.T. (1993). Citations do not solve problems. *Psychological Bulletin, 112*(3), 393-395.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and*

Policy Analysis, 15(1), 1-16.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*,635-694.

Madaus, G.F. (1993). A national testing system: Manna from above? An historical/technological perspective. *Educational Assessment, 1*(1), 9-26.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (Third Edition, pp. 13-103). New York: American Council on Education and Macmillan.

Mitchell, R. (1992). *Testing for Learning.* New York: The Free Press.

Mooi, L.C. & Cheung, K.C. (1990-91). On meaningful measurement: Junior college pupils' anxiety towards computer programming. *Journal of Educational Technology Systems, 19*(4), 327-343.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education, 13*, 405-450.