Issues in Attitudes and Personality Measurement in Educational Psychology: Validity of Measurement Scales and Causal Modeling*

Takashi MURAKAMI, Yumi SUZUKI, AND Yasuko NOGAMI**
(Nagoya University)

Abstract Constructing scales with sufficient reliability and validity and establishing causal relationships between psychological constructs are methodologically most important aspects to conduct correlational studies in educational psychology. Through reviewing articles recently published in the Japanese Journal of Educational Psychology and its international counterpart, the Journal of Educational Psychology, several problematic features are discussed. In order to facilitate validation of scales and appropriate use of causal modeling, the importance of theoretical elaboration of constructs and abundant information for validation is emphasized. A property of Japanese research style generating hypotheses rather than testing them is evaluated, since it is suitable to treat the complex character of psychological constructs and scales purported to measure them.

Key Words: reliability, construct validation, path analysis, covariance structure analysis

1. Introduction

In essence, the objective of educational psychology is to find ways of improvements of individual's intellectual, emotional, and motivational attributes empirically. One of the basic methodologies is to establish the causal relationships between controllable independent variables and resultant dependent variables. The most orthodox way to accomplish this is experimental research where subjects' mental attributes are observed under some distinct conditions set by researchers in the form of different levels of independent variables.

Unfortunately, we often face two difficulties when we wish to conduct such experimental research. The first problem is raised by the fact that the measurement of most mental attributes is not an easy task. Any scales are always largely contaminated by random errors of measurement. Moreover, their authenticities are often criticized due to the lack of evidences with regard to whether they really reflect the attribute to be purported to measure. This is a problem easy to solve because psychological attributes are conceptually more abstract ones than ordinal physical measures. In short, these are the problems of reliability and validity of measurement.

The second problem is about how to establish causal relationships, since many important psychological attributes in real life are traits, long-lasting properties, rather than states, temporally characteristics. Psychological traits are not easily changed by short-term experimental manipulations of independent variables by definition. As a result, independent variables have to be defined as long-term accumulations of everyday life events experienced by subjects. Therefore, we must rely on correlational research, where independent variables are not controlled by researchers, instead of experimental research. To obtain the objective of confirming the causal relationships between (conceptually) independent and dependent variables, several statistical methods of causal modeling such as path analysis, and its advanced form, covariance structure analysis, have been developed, and they are currently used in many studies. However, they seem to be a source of new methodological concerns (e.g., Breckler, 1990).

In this article, we will discuss these two issues, reliability and validity of measurement, and causal modeling. They are not only most important method-
ological aspects in many fields of psychology as well as educational psychology but also main research topics in psychological measurement and evaluation. To clarify existing problems in recent empirical studies, we reviewed several papers published in the Japanese Journal of Educational Psychology (JJEP), and its counterpart in the United States, the Journal of Educational Psychology (JEP). In the second section, 19 articles selected from two journals mentioned above, in which scales of self-efficacy are employed, are reviewed mainly from the viewpoint of how reliability and validity of scales used are guaranteed. In the third section, 35 articles in both journals, which use any method(s) of causal modeling, are reviewed with regard to how appropriately causal models are used.

Before reporting the results of review in each section, we will try to describe the current status of the theories as compact as possible, and to point out their potential problematic or questionable features in application works in order to demonstrate our conceptual framework of the review. We hope that these passages will also be useful as supplements to many textbooks, which ignore the advance of the theory of construct validation after Cronbach and Meehl (1955), and which do not refer to cautions for careless applications of causal modeling to psychological data.

Our main purpose is to look for any desirable forms of correlational research considering some specific features of psychological research, particularly the studies aiming at measuring individual differences of attitudes and personality by means of questionnaires. Reviewing papers published in Japan and the U.S., 'cross-cultural' comparisons were inevitable. In the preparatory stage of this project, our main objective was to point out shortcomings of Japanese articles in comparing with the Western ones. As the work proceeded, however, we slightly changed our own attitudes. Although we found several features discriminating Japanese articles from its counterparts, we do not consider now that all of them are necessarily shortcomings. We will return to this point in the last section.

2. Issues in reliability and validity of measurement; the case of self-efficacy scales

In this section, we focus on a construct of self-efficacy as one of the major psychological concepts in the field of educational psychology and review a number of recent articles that deal with self-efficacy. In order to discuss about the construct validation explained below, we have paid attention to several points; that is, how the constructs adopted in that research are defined, how the relationships of those constructs are illustrated; and accordingly what hypotheses are set up. Along with these points, we try to point out some shortcomings for the construct validation or hypothesis testing in terms of introducing the statistical methods employed in those studies.

2.1 General principles of the scale construction

Reliability and validity are two of the most important aspects of a scale within classical psychometrics (Rust & Golombok, 1989). We will explain the theoretical core of the notions briefly. Consider responses to $p$ items of a questionnaire with $m$-point rating scales by $n$ subjects. The responses will be treated as variables with ordered categories. A scale is usually defined as the simple or weighted sum of responses for each subject. (The term 'scale' in this sense must be differentiated from the $m$-point rating scale attached to each item. We will use the term scale in the former sense hereafter.)

The scale is expected to correspond to a quantitative attribute for each subject whenever it is more or less internally consistent; the items summed up for defining a scale correlate positively each other. On the other hand, if the scale is stable over time in the sense that the test-retest correlation, or its substitute (e.g., correlation between parallel forms), is rather high, the attribute to which the scale corresponds may be called a trait. Otherwise, it should be called a state. The concepts of the internal consistency and the stability cover the concept of reliability in classical test theory. They are two different aspects of a formerly unitary concept of the reliability. Namely, reliability has been defined as "the consistency of scores obtained by the same persons when reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions" (Anastasi, 1988, p. 109). While there are several techniques for estimat-
ing the reliability of a scale, they have been classified broadly into two categories mentioned above: stability over time and internal consistency. In general, reporting stability as well as internal consistency is desirable because it guarantees that the scale measures a trait rather than a state.

All types of reliability are expressed in terms of a correlation coefficient. It is difficult to say how high the reliability of a scale should be. Kline (1993) suggests the following criteria; in the case of reliability concerning stability over time, a correlation of .8 is required, taking at least a three-month gap between two testings and 100 subjects. The most straightforward is test-retest reliability. In a similar sense, coefficient alpha, a typical index of internal consistency, should not drop below .7. We consider that somewhat lower criteria will be acceptable in cases of attitudes and personality scales, especially when they are used for correlational research rather than assessment for individuals.

Validity of measurement has been defined as the extent to which the scale effectively measures what it is purported to measure (Anastasi, 1988; Rust & Golombok, 1989). However, the concept of validity evolved particularly after fifties, and it is now defined as “an integrated evaluative judgment of 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, 1993, p.13, italics in original).

When it is justified that a trait (or a state) is named any psychological concept, it may be called a construct. The justification, the (construct) validation, presupposes a psychological theory about the construct. Validation is not a procedure but a process consisting of a series of testing hypotheses that are derived from the theory through correlations with other scales, comparisons of means between groups, investigations of effects of interventions, and so forth. The empirical results that show substantial relationships predicted from the theory are called converging evidences, while weak or no relations with variables with which the theory implies irrelevance are discriminating evidences. Validation is a long and almost endless process since predictions derived from the theory may be usually tremendous.

As a result, one who wishes to propose new scales of some constructs has to report many evidences. For example, in presenting her 4 (sub-) scales of affiliation motivation, Hill (1987) gives correlations with 9 scales as convergent evidences, with 8 scales as discriminant ones, and with 4 role playing behaviors as well as coefficients of stability over time and of internal consistency. This may represent a standard style of this kind of articles.

One may find the great differences between the ways of verification of reliability and validity of a scale while the two concepts are often mentioned simultaneously. The former consists of almost the purely statistical computations once a set of items is given. In addition, reliability can be regarded as a property of the scale itself although it is sometimes inflated or deflated due to the range effects; a bias of the sampling of subjects. On the contrary, the psychological theory about a construct is indispensable for validation, and the statistical methods are only tools that inspect through the testing of hypotheses a possibility of legitimate use of the scale in the sense of the technical term of psychology. There is no single statistical method for validation except for the notorious simple package, multitrait-multimethod matrix (Campbell & Fiske, 1959). Moreover, validity is not a property of the scale per se but depends on its use. (A scale can be valid as the extraversion but not as the optimism.)

Even after the middle of fifties, when a monumental manifest about construct validity by Cronbach and Meehl (1955) was published, most textbooks includes the statement that the validation procedures of a scale involve various forms which are divided into three principal categories: content validity, criterion-related validity, and construct validity. Criterion-related validity is evaluated by a simple procedure that correlates a scale with a special variable, an external criterion. However, one cannot find any external criterion for most attitudes and personality variables. If no external variable is available, the scale is validated based on the content of items when one can define a universe of items about the construct he/she wants to measure. These two concepts, criterion-related validity and content validity, are not considered to be offering sufficient basis of validation cur-
rently, and they are not available in most studies constructing attitudes and personality scales.

For example, one who reads the paper entitled 'A personality scale of manifest anxiety' written by Taylor (1953) may be astonished by his easy ways of treating the scale which has been widely used so far. He chose its items from the MMPI based on the judgments of 5 clinicians who were informed the definition of manifest anxiety that follows the description by a psychiatrist. He selected the final 28 items mainly based on the internal consistency. Only an empirical result that he showed as an evidence of the relationships with any external attributes was a distribution of scores given by 103 psychiatric patients with the following comment; "The anxiety scale was developed for...a device for selecting experimental subjects, without regard to the relationship of the scores to more common clinical definitions. While defining degree of anxiety in terms of the anxiety-scale scores is a perfectly legitimate operational procedure, determining the relationship between this definition and clinical judgments might extend the applicability of both the scale and the experimental results found in the studies utilizing the scale" (Pp. 289-290). This example shows that the researchers in this era considered a scale to give an operational definition of a psychological concept, and gave the secondary importance on the relationships of the definition with the use of the term, e.g., anxiety, in wider contexts of the psychological research, and the ordinary usages by laypersons.

Recent psychologists seem to be more interested in the wider implications and generalizabilities of psychological concepts. In this sense, validation is a kind of semantics that clarifies the relationship between the psychological concepts (constructs) and operations substantiating them (scales). However, since many articles published recently continue to rely on the old fashioned framework, we will have to refer to the concepts of criterion-related and content validity in the sequel.

The relationship between reliability and validity is slightly complex. Many psychometricians (e.g., Nunnally, 1978) have argued that high reliability is a prerequisite of high validity. They imply the attenuation phenomenon. Let $\rho_{xy}$ be a correlation coefficient between true values of two variables, $x$ and $y$, $\rho_x$ and $\rho_y$ be coefficients of reliability of two scales, respectively, and $r_{xy}$ be an observed correlation coefficient. Then, it is shown that $r_{xy} = \sqrt{\rho_x} \sqrt{\rho_y} \rho_{xy}$, which means $r_{xy} < \rho_{xy}$ because $\sqrt{\rho_x}$ and $\sqrt{\rho_y}$ are generally less than unity. This reduction of the coefficient is called attenuation. If $x$ denotes a scale and $y$ an external criterion, validity of the scale evaluated by $r_{xy}$ is reduced when the reliability of the scale is low, and one may say that reliability is a necessary condition of validity.

On the contrary, there are some arguments (e.g., Cattell & Kline, 1977) that very high internal consistency might be antithetical to validity; high internal consistency is not a sufficient condition for validity. Any scales measure contents of some breadth, but if all items are highly consistent, then the scale is likely to be narrow and specific and thus not valid. In other words, one must select items having divergent meanings so as to give a sufficient breadth to the scale even if sacrificing the internal consistency to some extent in order to obtain a scale having substantial correlations with other variables. Cronbach (1990) calls this the bandwidth-fidelity dilemma. Since one does not have to consider the breadth of a scale in measurements of physical length, weights and so forth, balancing bandwidth and fidelity is a specific concern in psychological measurement.

2.2 Theoretical bases of the concept of self-efficacy

Since Bandura (1977) proposed the theory of self-efficacy, its conceptual importance and applicability in the field of educational psychology have been instantly recognized and incorporated into a variety of theoretical frameworks, and that is one of the reasons why this construct is treated in this review. Arguing that individual behaviors are not affected by immediate consequences but by the level of aggregate consequences through the influence of thought, Bandura introduced the concept of self-efficacy which refers to "the conviction that one can successfully execute the behavior required to produce the outcomes" (p.193).

The perceived self-efficacy influences the degree of which individuals make efforts to attain desired outcomes and to persist in aversive situations. Based on
his social learning theory, the perceived self-efficacy is purported to be formed and/or enhanced by some sources of efficacy information such as performance accomplishments (e.g., self-instructed performance), vicarious experience (e.g., live and/or symbolic modeling), verbal persuasion (e.g., suggestion), and emotional arousal (e.g., physiological states induced by attribution, relaxation or desensitization). Although the theory of self-efficacy was originally constructed to unify various modes of clinical treatment for behavioral change, Bandura and his colleagues have also applied this theory to educational settings (e.g., Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Bandura & Schunk, 1981; Zimmerman & Bandura, 1994; Zimmerman, Bandura, & Martinez-Pons, 1992).

2.3 General features of articles reviewed here

A result of computer search for the studies recorded in the PsycLit shows that there are 3741 journal articles that adopt the concept of self-efficacy as a research variable, in which the JJEP carries 22 articles whereas 49 articles are reported in the JEP. Focusing the last five years in Japan, 7 out of 13 journal articles that deal with self-efficacy have been published in the JJEP, implying the conceptual importance of self-efficacy in educational settings. Actually 40 articles focusing on self-efficacy have been reported for the last five years in Japan; however, the rest of them have been published as non-journal articles (i.e., 20 as university bulletin articles and 7 as annual reports issued by universities or research institutes).

Within 7 self-efficacy articles in the JJEP, one is a review article and accordingly excluded from the discussion in this section. The concept of self-efficacy in the rest of articles is specialized in a number of settings; two of them are self-efficacy for career decision-making (Tomiyasu, 1997; Urakami, 1996a); two are self-efficacy for social interactions (with disabled students, Kawauchi & Yokkaichi, 1998; with peers, Matsuo & Arai, 1998); self-efficacy for academic achievement (Ito, 1996); and a scale development of generalized self-efficacy (Narita, Shimonaka, Nakazato, Kawaai, Sato, & Osada, 1995).

In addition to them, 4 another journal articles in the JJEP deal with self-efficacy as a research variable, although self-efficacy is not the main construct in their studies. Consequently those studies will also be treated as well in this review. In their research the concept of self-efficacy is also specialized in career decision making self-efficacy (Shimomura, 1998; Urakami, 1996b), academic self-efficacy (Miura, Shimada, & Sakano, 1997), or general self-efficacy (Ohnogi & Miyakawa, 1996). We review these 10 journal articles in total, focusing on the descriptions of the definition(s), the relations of constructs if any, the hypothesis or hypotheses to be tested, and accordingly the ways of reliabilities and validities to be confirmed in their research.

As mentioned in the introduction section, these self-efficacy research conducted in Japan will be fittingly compared with the articles published in the JEP, although we will pay attention specifically to the studies between 1997 and 1998 in this review because there are 9 articles (that is approximately the same number as the Japanese articles to be reviewed in this section) adopting self-efficacy in their research just for two years. In contrast to the Japanese articles, self-efficacy treated as a research variable in the studies reported in the JEP is specialized especially in academic self-efficacy (Bong, 1997, 1998; Eaton & Dembo, 1997; Middleton & Midgley, 1997; Ryan, Gheen, & Midgley, 1998; Turner, Thorpe, & Meyer, 1998; together with general efficacy, Zohar, 1998) or learning self-efficacy (for schoolwork, Skaalvik, 1997; for dart-throwing skill, Zimmerman & Kitsantas, 1997). This specialization of self-efficacy for academic achievement or skill acquisition may indicate that researchers who contribute their articles to the JEP have been conscious of the importance of education as their research theme and settings.

2.4 Definition and theoretical descriptions of self-efficacy in reviewed articles

For conducting any piece of research the statement of clear definition is crucial, specifically in psychology, as many terms or constructs employed in psychological research are already used in everyday language and thus have vague colloquial meanings with wide variation (Roth, 1979). Even among psychologists in the same field, a total agreement would not be possible in respect of how a construct is treated as a
variable with a single scientific meaning if they relied upon vague understanding in possible common meaning. Of course the precise definitions inevitably abate richness of language connotations; nevertheless, it is very unlikely that a single definition covers the whole meaning of a particular construct.

The precise definition of constructs stipulates a possible valid method for measurement to be employed so that definitions must be clear even for the more easily measurable variables (Coolican, 1994). In addition, a construct is usually linked to other constructs in a theoretical framework, and further explanations and predictions could be possible if we concern the relationships between these constructs defined precisely. As for the statement of clear definition in the Japanese articles, however, it should note that a result of review is a little disappointing.

Out of 10 Japanese articles, only Matsuo and Arai (1998) define constructs in their article, although Urakami (1996b) cites one definition of the constructs treated in his article. The articles in which the word 'definition' has been used are just two of them. Matsuo and Arai deal with four psychological concepts and clearly define them. They specifically try to differentiate their main construct, social anxiety, from social anxiety. That is a noteworthy attempt, although its differentiation unfortunately seems to lack consistency in their introduction. On the other hand, Urakami (1996b) deals with three psychological concepts in which his main construct, self-growth motivation, is described in terms of borrowing a preceding study and the other two are mentioned ambiguously in his sentences. This retentive description is not a style peculiar to his article, rather a very common description of treating constructs. For example, the psychological concepts treated in Shimomura's (1998) and Tomiyasu's (1997) articles are not defined but described obscurely.

These descriptions are still a long way from the statements of clear definition; nevertheless, we unfortunately admit that something is better than nothing. Many of the articles (Kawauchi & Yokkaichi, 1998; Miura et al., 1997; Narita et al., 1995; Urakami, 1996a) describe just their main constructs in their introduction section. The description of one of four psychological concepts is missing in Ito's (1996) article. Ohnogi and Miyakawa (1996) introduce the subscales and item examples of the measures employed in their research, but it should not be equivalent to the definitions. Defining a construct in the form of what to be measured is not an authentic way, at least, in the context of traditional scientific methodology.

In contrast to the Japanese articles, the articles in the JEP define the constructs treated in their research explicitly in their introduction section (Bong, 1997; Turner et al., 1998), method section (Skaalvik, 1997; Zimmerman & Kitsantas, 1997), or both introduction and method sections (Middleton & Midgley, 1997; Ryan et al., 1998; Zohar, 1998). The other two articles (Bong, 1998; Eaton & Dembo, 1997) do not involve the statement of precise definitions of their constructs, although the operations necessary to measure their particular concepts are precisely worded. It should be noted here that any of these operational definitions are not the same as the description of constructs by introducing the subscales and item examples of the measures used. As pointed out above, any psychological concepts should not be defined by which the scales try to measure. All of the definitions in these Western articles take the form of statements which refer to their constructs, not simply enumerate the subscales of the measurements employed in the research.

Along with the precise definition of constructs, the relationships of those constructs must also be clearly illustrated if the research adopts more than one psychological concept. The stipulation of those relationships clarifies the theoretical framework of which the researcher fosters, and it is likely to lead to a meaningful hypothesis that is worthy to be considered and tested. All of the Western articles reviewed here thoroughly illustrate its theoretical framework and consequent hypotheses or research purposes. Six out of 9 articles expressly consider the expatiation on their treating concepts and these relationships by means of its respective subheadings.

In contrast, many of the Japanese articles reviewed here do not illustrate the relationships of their treating concepts. Although its space limit has been often pointed out, their attempts on the epitomization of theoretical frameworks seem to be insufficient. In other words, a number of studies appear to overreach

— 205 —
in respect to the construction of their theoretical frameworks. For example, Ito (1996) paid attention to some past findings indicating that causal attributions for failure in an academic achievement situation to one's effort enhanced his or her intrinsic motivation from a motivational standpoint, but at the same time that effort attributions did not influence one's self-efficacy from a cognitive perspective. Based on its inconsistency he proposes a working hypothesis that such children do not understand how to make an effort and consequently study without any effective learning strategies. Learning strategy is the main target variable in his research; however, the description of neither its definition nor the sub-constructs composing of its psychological concept is found. Accordingly, the applicability of this construct and then the justifiability of his research purpose cannot be discussed even if he has introduced a piece of research investigating its relations with motivation or with self-efficacy individually.

Such an overreaching tendency in theoretical frameworks has also been found in other articles. Shimomura (1998) investigated the learning effects of decision-making strategies on career decision by measuring the cognition of vocational decision-making training and career decision-making self-efficacy. For his assessment the undergraduates sample has been divided into two groups in terms of vocational readiness based on his assumption that the learning effects of vocational decision-making strategies (i.e., the cognition of its training and career decision-making self-efficacy) would be varied depending on vocational readiness which is regarded as a personality trait of vocational maturity. The relations of vocational readiness with the cognitive factors mentioned above, however, are not thoroughly illustrated in his introduction.

In a similar sense, Matsuo and Arai's (1998) study aimed at developing a social anxiousness scale for children and examining its validity in terms of the investigation of the relations of social anxiousness with public self-consciousness and social self-efficacy. They argue that the possession of serious social anxiousness must require not only high public self-consciousness but also some other factors. However, the concept of social self-efficacy, which is 'that fac-

cor', appears firstly in the statement of their hypothesis abruptly. Narita and his colleagues (1995) also developed a Japanese version of generalized self-efficacy scale and examined its reliability and validity. They adopt several personality factors such as depression and self-esteem in order to provide some supports of construct validity. However, these relations with generalized self-efficacy are not illustrated in their introduction.

In addition, no hypotheses are explicitly stated in a number of studies (e.g., Miura et al., 1997; Ohnogi & Miyakawa, 1996; Shimomura, 1998). The stipulation of constructs' relationships and that of consequent hypotheses are quite important for the justifiability of hypothesis testing or construct validation and that of statistical methods used. The authors thus should describe them clearly regardless of the fact that authors' intentions in the research are likely to become clear as readers are making an effort to read between lines.

2.5 Evidences of reliability and validity of scales

As noted above, both stability and internal consistency types of reliability should desirably be reported because both of them deal with the relationship of the sets of scores, but at the same time they are concerned with the very distinct aspects. If it is not possible, at least either of them must be reported. Unfortunately, any reliability coefficients are not reported in Miura and her colleague's (1997) research, regardless of six administrations. In addition, those administrations have the time interval of just one day, even ten days at the maximum. If a scale is unreliable, it is impossible for the scale to be valid. Details of reliability should have been reported because the original scales were modified and newly constructed. The U.S. articles reviewed here do not report test-retest reliability or others concerning stability over time. The sole exception is Zohar's (1998) research, where the reliability coefficient is not reported but stability of the sets of scores is discussed. In contrast, the half of Japanese articles reports test-retest reliability, although they involve two validation studies. Since validation studies usually consider such psychometric aspects of the scale extensively, it will be compared separately with their corresponding validation studies.
shortly.

Content validity of a scale examines whether its content reflects a representative sample of the domain (or item universe) to be measured. In this review Kawauchi and Yokkaichi (1998) report content validation of their social self-efficacy scale. However, it is somewhat questionable to regard their procedure as traditional content validation. Content validation procedure involves the systematic examination of the content relevance and the content representativeness, specifically the adequacy of item sampling (Messick, 1993). In other words, the application of content validity requires the domain of items to be clearly identifiable so that ability, achievement, and certain occupational tests are likely to be the case. Accordingly, for personality scales it might be often inappropriate and misleading because the content of personality scales must be verified empirically. Of course content validity is often evaluated more qualitatively than quantitatively (Rust & Golombok, 1989), but such evaluations must be systematic and should include a description of the test specification procedures, ensuring that the content areas to be covered and the objectives to be tested, together with the number of items in each category.

In the case of content validity it so happens that face validity and content validity overlap, especially in the content validation of personality scales (Kline, 1993). Content validity must not be confused with face validity. Face validity refers to the extent to which the scale appears to measure superficially, and it is not validity in the psychometric sense. Face validity is clearly different from content validity, although its effectiveness may be equally important; that is, the results of the scale must be meaningless if the participants consider its content as being irrelevant and accordingly provide poor cooperation.

On the other hand, criterion-related validity indicates the effectiveness of a scale in predicting an individual's performance by examining against a criterion measure. A result of review indicates that criterion-related validity, especially predictive validity, is one of the most popular forms of statistical validity. It might be due in part to the fact that prediction has been always regarded as a crucial part of the scientific method. Although criterion-related validation procedure is not described explicitly in most of the reviewed articles, 6 out of 9 articles in the JEP discuss either predictive or concurrent validity, indicating that the importance of criterion-related validity has been widely recognized at least in the U.S. Among the criteria employed frequently in examining validity reported in educational settings is some index of academic achievement, and all but one criterion-related validity reviewed here use achievement indexes such as academic grades. Criterion-related validity, however, is a useful form of scale validity only where good criterion indexes exist. In addition, correlations with the criterion test must be sufficiently high with some other advantage as well (Kline, 1993). It never is an easy task, and even academic achievement hardly becomes a clear criterion in many cases.

Many different kinds of validity have been proposed, but as implied above, there are still a wide variety of scales for which none of these methods seems appropriate. Most of the personality scales might be the case. For this difficulty, the notion of construct validity was introduced by Cronbach and Meehl (1955). They provide an important rationale for the examination of construct validity, making us focus on the role of definition and theory in scale construction and on the need of hypotheses to be tested in the validation procedure.

Construct validity is now the primary form of validation examining the extent to which the scale is purported to measure a theoretical construct or trait. Despite of the fact that theoretical constructs are frequently the object of investigation, few psychological concepts dealt with in the research can be measured directly so that its applicability is usually evaluated by examining the relations of the construct concerned with other concepts and/or relating phenomena which the theory predicts. Correlations between the target and other variables (constructs) involve two forms in order to demonstrate validity. One is the substantial correlations of the scale with others with which they should theoretically correlate (convergent evidences), and the other is the ignorable correlations with which they should not be related (discriminant evidences). We will again emphasize the importance of explicit definition of the constructs and theoretical considerations around them because
researchers cannot find any external variables giving convergent and discriminant evidences without appropriate guidance of such a clear statement.

Since construct validity embraces validity of every type (see Kline, 1993), and its importance is specifically emphasized in the studies whose main aim is the validation of scales newly constructed, Narita and his colleague's (1995) validation study of a generalized self-efficacy scale is discussed here in terms of introducing and comparing with their corresponding validation studies executed in the U.S. TABLE 1 presents reliability and validity examined and variables used as criterion, together with their correlation coefficients. In sum, the result of this review is consonant with the discussion above, indicating that Japanese validation studies consider both stability and internal consistency types of reliability, and that American validation studies pay more attention to relations with many external variables serving as convergent evidences. However, the U.S. studies also lack of discriminant evidences unfortunately, except for the correlation with the social desirability scale (see Sherer, Maddux, Mercandante, Prentice-Dunn, Jacobs, & Rogers, 1982).

### TABLE 1: Summary of validation studies

<table>
<thead>
<tr>
<th>No. Authors</th>
<th>Constructs</th>
<th>No. of groups</th>
<th>Factor analysis</th>
<th>Reliability</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Narita et al. (1995)</td>
<td>generalized self-efficacy</td>
<td>8(N =1524)</td>
<td>1 factor (principal axis factoring)</td>
<td>(internal consistency)</td>
<td>construct validity</td>
</tr>
<tr>
<td></td>
<td>depression</td>
<td>(13 - 59 years old)</td>
<td>(varimax rotation)</td>
<td>coefficient alpha from .86 to .89</td>
<td>(convergent evidences) p &lt; .001, p &lt; .05</td>
</tr>
<tr>
<td></td>
<td>self-esteem</td>
<td></td>
<td></td>
<td>(stability)</td>
<td>p &lt; .001, p &lt; .01</td>
</tr>
<tr>
<td></td>
<td>sex role (masculinity)</td>
<td></td>
<td></td>
<td>test-retest reliability</td>
<td>p &lt; .001, p &lt; .01, p &lt; .05, ns</td>
</tr>
<tr>
<td></td>
<td>perceived health</td>
<td></td>
<td></td>
<td></td>
<td>73 (2 weeks, different N = 67)</td>
</tr>
<tr>
<td>2 Sakano (1989)</td>
<td>generalized self-efficacy</td>
<td>7(N = 276)</td>
<td>N/A</td>
<td>(internal consistency)</td>
<td>construct validity ?</td>
</tr>
<tr>
<td></td>
<td>depression</td>
<td>(under 20 - over 71 years old)</td>
<td>(3 factors)</td>
<td>Spearman-Brown .66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>facial expression</td>
<td>(N = 2)</td>
<td></td>
<td>Kuder-Richardson .81</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(stability)</td>
<td>ps : N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>test-retest reliability .89</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(3 months, N = 85)</td>
</tr>
<tr>
<td>3 Sherer et al. (1982)</td>
<td>generalized self-efficacy</td>
<td>1(N = 376)</td>
<td>2 factors (general &amp; social)</td>
<td>(internal consistency)</td>
<td>content validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(N =288 for cross validation)</td>
<td>(extraction: N/A)</td>
<td>coefficient alpha .86 &amp; .71</td>
<td>(psychological meanings)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(proportion of the total variance)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>construct validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(convergent evidences) p &lt; .001 &amp; p &lt; .01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .001 &amp; p &lt; .001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .001 &amp; &amp; p &lt; .001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(criterion-related validity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(concurrent validity) p &lt; .01 &amp; ns</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .01 &amp; p &lt; .05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .05 &amp; p &lt; .01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .01 &amp; ns</td>
</tr>
<tr>
<td>4 Sherer &amp; Adams (1983)</td>
<td>generalized self-efficacy</td>
<td>1(N = 101)</td>
<td>N/A</td>
<td>N/A</td>
<td>construct validity</td>
</tr>
<tr>
<td></td>
<td>MMPI (13 scales)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>assertiveness</td>
<td></td>
<td>(general &amp; social)</td>
<td></td>
<td>(convergent evidences) ps &lt; .05, ns &amp; ps &lt; .05, ns</td>
</tr>
<tr>
<td></td>
<td>sex-role (masculinity)</td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .05 &amp; p &lt; .05</td>
</tr>
<tr>
<td></td>
<td>sex-role (femininity)</td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .05 &amp; p &lt; .05</td>
</tr>
<tr>
<td>5 Woodruff &amp; Cashman (1993)</td>
<td>generalized self-efficacy</td>
<td>1(N = 400)</td>
<td>2 factors (general &amp; social)</td>
<td>(internal consistency)</td>
<td>construct validity</td>
</tr>
<tr>
<td></td>
<td>mastery</td>
<td></td>
<td>(extraction: N/A)</td>
<td>coefficient alpha .84 &amp; .69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>self-esteem</td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .01</td>
</tr>
<tr>
<td></td>
<td>task-specific efficacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>generalized self-efficacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>grade expectation (A or B)</td>
<td>N/A</td>
<td>5 factors (3 general &amp; 2 social)</td>
<td>(internal consistency)</td>
<td>criterion-related validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2 factors)</td>
<td>(extraction: N/A)</td>
<td>coefficient alpha from .57 to .75</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(general &amp; social)</td>
<td></td>
<td>(concurrent validity) p &lt; .01 &amp; ns</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(t test)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(test-retest reliability) .89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(3 months, N = 85)</td>
</tr>
</tbody>
</table>
A result of review might clarify that many of the research conducted in Japan should take a clear description of definition and theoretical framework into consideration, whereas those conducted in the U.S. should regard both sides of reliability. As for the efficacy-scale, further efforts should be put to construct validation, especially to collecting more discriminant evidences in the U.S. as well as in Japan.

3. Several issues in applications of causal modeling to psychological measurement

Recently, the usage of path analysis and covariance structure analysis increases almost explosively in the field of educational psychology. Both are methods of causal modeling, and the latter is an integrated model having the features of confirmatory factor analysis. We can see many descriptions of causal relations as well as path diagrams, pictorial representations of causal relations in terms of boxes and arrows.

It is impossible to prove causal relations in the strict sense in correlational research where variations of independent variables are left for natural occurrences, while it is not the case in experimental research where they are controlled by the researcher through the randomization processes and/or similar techniques. However, there are general needs for causal inference even in the correlational research so that methods of causal modeling should at least partially meet the expectancy. In addition, it is true that problems for empirical research are formulated neatly when it is done in considering causal model. Moreover, a series of modification processes of causal models may be used for producing more persuasive mini theory, although their methodological bases are dubious as will be discussed shortly. In fact, modeling of causality is irresistibly attractive as one way of story telling. We should add that the methods of causal modeling relate to the validation of measurement because the confirmation of causal hypotheses can be a part of the validation procedure.

We will discuss some basic issues in the methods of causal modeling and a number of difficulties in their applications to psychological measurement, which will then lead to point out some problems in the case of careless use of causal models.

3.1 General difficulties in applications of causal modeling methods

Let us explain how causal inference becomes possible in a simple example of correlational research. Assume that a researcher, who has the hypothesis that the interest for mathematics raises achievement of mathematics, found .4 of the correlation coefficient between x, level of interests and y, achievement in a number of randomly sampled high school students. This result is not interpreted as the evidence that strong interest is the cause of high achievement. In fact, one may consider its reversed causal relationship that high achievement causes the interests for the subject is equally plausible. But we will not have any doubt on a direction of the cause and effect hereafter.

However, we can suspect the result \( r_{xy} = .4 \) as an evidence of the causal relationship because any third variable may be a cause of both the interest and the achievement. Let us consider that the researcher could develop an instrument to assess the aptitude \( z \) for mathematics in each student and obtained \( r_{xz} = .5 \) and \( r_{yz} = .8 \). Then, the numerator of the partial correlation coefficient between \( x \) and \( y \) obtained by controlling \( z \), \( r_{xy,z} \), is equal to \( r_{xy} - r_{xz}r_{yz} = .4 - .8 \times .5 = 0 \). This implies a possibility that the substantial value of correlation coefficient \( r_{xy} \) is so called pseudo correlation, which results from the third variable \( z \) (e.g., Bollen, 1989). Hence, one can deny the direct causal relationship between the interest and the achievement. This is a procedure of classical partial correlation technique of causal inference.

This example suggests that the causal inference in correlational research is very limited from a number of perspectives. First, it does not imply the direction of the cause and effect unless the occurrence of \( x \) precedes that of \( y \), which can be confirmed in the case of longitudinal studies. Second, even if non-zero partial correlation shows that \( r_{xy} \) is not pseudo one, the conclusion on the causal relationship between \( x \) and \( y \) remains tentative because one cannot rule out a possibility of the existence of (often unobserved) possible fourth and fifth variables that affect the value of \( r_{xy} \). In other words, a causal model would never be proved but only falsifiable. Third, the judgment whether the partial correlation is zero or not depends on a method of inferential statistics that is heavily
affected by the sample size, and the conclusion is thus also tentative because the larger sample is likely to reject the null hypothesis of the partial correlation. The first and second remarks imply that the causal inference in correlational studies is rather weaker than that in experimental research where the researcher can manipulate the cause directly, while the third remark suggests the common problem shared by both types of inferences.

Path analysis is a more general method that can verify the causal model depicted in the form of a path diagram by the use of multiple regression analysis and its resultant standardized partial regression coefficients. Path analysis was originally proposed in the early 20th century by Wright, a biologist (e.g., Bollen, 1989). In the example above, the hypothesis that the interest for mathematics has a causal effect on the achievement of mathematics will be rejected when the coefficient $b_x$ is regarded as zero statistically in the regression equation, $y = b_{xx} + b_{xz} + u$. Since the numerator of $b_x$ is $r_{xy} - r_{xz}r_{yz}$ that is exactly the same as the value of the partial correlation coefficient, $r_{xy,z}$, the conclusion of path analysis is identical with the intuitive method using the partial correlation in this simple example. By the way, the standardized partial regression coefficients in the equation above is called path coefficients, and interpreted as the strength of the causal effect, and $u$ is the residual term in the ordinal least squares regression.

Although path analysis makes it possible to confirm the more complex causal models including many variables, it does not resolve the three limitations already mentioned. As for the issue on the direction of causal effects, we must point out that there are many path diagrams that fit simultaneously to a set of correlations equally well. As a matter of fact, 4 models presented in Figure 1 are mutually equivalent because they fit the correlations between three variables in the example above. Therefore, one has to depend not on a statistical basis but on a theoretical supposition about the directions of the causalities. One may expect that such an indeterminacy would be reduced by increasing the number of variables. Unfortunately, it is not the case. For example, Mayekawa (1994) demonstrated that there could be many equivalent models in the case of more than three variables. Although much caution should be necessarily required for the problem of multicollinearity and the range effects on correlation coefficients, which could affect the results of path analysis, we will not discuss them here because they have been frequently pointed out.

![Figure 1](image)

**Figure 1** Path diagrams for four equivalent models

3.2 Specific difficulties in the applications to psychological data

We encounter some special difficulties beyond the general problems discussed above when applying path analysis to scales based on item responses in questionnaires (or item responses per se). They are mainly caused by low reliability and validity of scales (or item responses). There are a number of points that we should make much account of. First, we must consider the reduction of correlation coefficients, which is called attenuation. Results of path analysis are biased when reliabilities of analyzed scales are mutually different, because attenuation brings differential effects on correlations, on which path coefficients are computed. Hence, one must make reliabilities of scales as equal as possible before the application of path analysis. This caution has not been indicated in any textbooks to the authors' knowledge. Next, we have to consider the effect of response tendency, which is called response set. Social desirability and acquiescence exemplify the effect of this kind of constant error, although its evaluation is quite difficult in actuality.

Slightly subtler but potentially more serious problem is the overlapping of the domain of several con-
structs to be measured. It may be an unsolvable problem how to decide whether a correlation between two scales reflects some substantial laws holding between constructs, or it is a mere sign of the existence of some components (including the response tendencies) shared by two scales as long as we rely on the data sets obtained by the use of questionnaires. Considering these circumstances, it may be safe to restrict the application of causal modeling to the correlations between scales, obtained from different sources (e.g., peer- and self-ratings of personality), or from distinct levels of processing (e.g., responses to statements for some attitude objects and judgments to problems of arithmetic). One may not consider the non-zero correlations between these scales to be artifacts. However, this restriction limits the number of variables included in the analysis to be very small, and various interesting applications will become impossible.

Rather, more serious problem may be the following one; that is, a temptation to the post hoc modifications of the model including the changes of causal directions in order to improve the fitness of the model under a given correlation matrix becomes irresistibly strong once convenient tools of statistical packages are available for this procedure (Breckler, 1990). It is questionable in the light of the traditional principle of empirical sciences, which requires the collection of a new set of data to verify or falsify the modified model. Even if another sample is unavailable, at least one has to conduct cross validation in terms of dividing the data set into two parts, one for modifying the causal model and the other for its confirmation.

3.3 Covariance structure analysis: What was improved?

Covariance structure analysis (e.g., Bollen, 1989) is an advanced form of path analysis that implements the features of confirmatory factor analysis. The model consists of two sets of linear equations; one represents the assumed causal relations between several latent variables, which are not observed directly, and another (also assumed) relationships between latent and observed variables. These equations are fitted to the covariance matrix of observed variables, which assume multivariate normal distributions, in the sense of maximum likelihood method. One can execute the test of significance on the deviation of the model from the data using an approximate chi-square statistic. In addition, several indices for goodness of fit have been devised, and some information statistics like AIC (Akaike's information criterion) can also be used.

Although one can use spurious correlation coefficients to test fitness of the model when the number of variables to be analyzed exceeds that of estimated parameters, covariance structure analysis uses highly advanced theories of sampling from multivariate normal distributions in statistics in order to evaluate the model. Therefore, models in formulated following the method covariance structure analysis are said to be falsifiable, which is often regarded as a key concept discriminating science from non-science. It is not difficult to imagine that many social scientists accept covariance structure analysis enthusiastically. Moreover, it is a great improvement of classical methods of causal modeling that estimates of parameter in covariance structure analysis would not be affected by the attenuation as long as fitness of the model is good because the estimation is based on estimates of correlations through the rationale that is the same as the estimations of communality in classical factor analysis.

It should be noted here that covariance structure analysis is not very new technique. For example, Jöreskog (e.g., 1970) may be, at least, one of the founders. Although applications in Japan were very few in seventieth and eightieth, several good books have been published recently (e.g., Kano, 1997; Toyoda, 1998) and software packages have greatly facilitated its use.

3.4 Difficulties newly brought by covariance structure analysis

As pointed out above, in covariance structure analysis there happened two problems, the appearance of inappropriate solutions and the lack of identifiability, which are not observed in path analysis. These problems have been succeeded from confirmatory factor analysis and often discussed (e.g., Bollen, 1989), and consequently will not be treated here.

Probably more serious problem in psychological applications may be exemplified in the following case.
Suppose the situation where one wishes to analyze a data set consisting of responses for many items. If one formulates the simple model including the small number of latent variables and paths, then fitness of the model will be not good. On the contrary, if one modifies the model to have the sufficient number of latent variables and paths to get the satisfactory fit, then it is highly possible for him to have an inappropriate solution (with negative variances of error term). This problem is pointed out by Schönemann and Wang (1972) in the context of applications of classical factor analysis. To avoid inappropriate solution, it seems to be necessary to add some awkward assumptions, for example, correlated random errors. This is an attractive bypass from mathematical point of view, but it is possible to lead to another problem because selected items may cover narrow bandwidth.

We have to point out an implicit but important attitude change of the researcher. In the application of path analysis, large predictive power represented by the squared multiple correlation coefficient is highly evaluated, since the model is essentially multiple regression analysis. This is a good tradition because findings of interesting relationships between variables are most important in these studies. On the other hand, the degree of success of the analysis is assessed in terms of the goodness of fit to the total covariance matrix because the form of the model is essentially that of confirmatory factor analysis. As a result, even if the causal relationship of the variables, which should be predicted, is not sufficiently high, authors tends to report as if the study was successful because of the total satisfactory goodness of fit.

3.5 Comparisons of the usage of causal modeling between Japanese and international articles

In order to examine how methods of causal modeling are used in the research, we reviewed all the papers using path analysis or covariance structure analysis published in the JJEP between 1994 and 1998, and in the JEP between 1997 and 1998, which are the same as the range where papers adopting self-efficacy scales reviewed in Section 2. First, we excluded the articles that did not include attitude or personality scales. There remained 115 articles in the JJEP and 48 in the JEP, and 26 of the former and 9 of the latter employed causal modeling. The ratios of papers using causal methods are approximately equal in both journals. Eight out of 26 in the JJEP and 3 out of 9 in the JEP employ covariance structure analysis, and the rest of them use path analysis. Again, the ratios of use of covariance structure analysis in both journals were almost equal.

Only 2 within 26 articles by the same author (Urakami, 1996a, 1996b) adopt the longitudinal design with sufficiently long interval (about 6 months) to determine the directions of causality in the JJEP, as do 3 out of 9 articles in the JEP (Marsh & Yeung, 1997; Newman, 1998; Normandeau & Guay, 1998). Marsh and Yeung (1997) present a persuasive example where the longitudinal design is useful in the determination of causal direction. They measured academic achievement of students six times and their academic achievement three times in successive three years. Despite of the N-dependent conclusion, a set of longitudinal data collected thus makes it possible to decide whether self-concept facilitate the achievement or the reverse is the case by means of testing the significance of sets of two path coefficients lagged crossly, whereas it is difficult to deduce the direction from a purely theoretical basis.

Next, we will compare the causal models themselves in a quantitative way. Since some articles include several separate analyses of causal relations, the total number of analyses (path diagrams) is 40 in the JJEP, and 17 in the JEP. The mean values of the number of variables (observed variables in path analysis and latent variables in covariance structure analysis) in a path diagram are 7.8 (SD = 2.9) in the JJEP and 6.1 (SD = 2.2) in the JEP. The mean values of the number of paths in corresponding diagrams are 8.8 (SD = 4.7) in the JJEP and 7.4 (SD = 5.16) in the JEP. Therefore, together with the somewhat poor theoretical considerations pointed out in the preceding section, we may say that the Japanese researchers tend to do somewhat risky causal inferences by international standards even if the quantitative differences are small.

Variables included in path diagrams consist of scales based on scores of only one respondent in 25 out of 40 analyses in the JJEP while 9 out of 17 in the JEP. The other articles include any variables obtained from
other information sources; other ratings by others, achievement, and so forth, and it will serve to avoid the artifacts raised by response tendencies. Numerical difference between two journals is also small in this respect. Recently, some researchers have tried to fit a structural model to the covariance matrix between individual items directly. Reflecting the dilemma between the poor fitness and the inappropriate solution mentioned in 3.4 as well as the limitation of computational power, the number of items measuring a latent variable tends to be small. Some researchers have selected them from a larger set of the original items defining the scale. This procedure will lead to latent variables with narrow bandwidth that may make correlations between them lower. This type of application is found only in the JJEP (e.g., Shimoyama, 1995; Tanaka, Yamazaki, Yanai & Suzuki, 1998).

Finally, we must point out a disappointing tendency. Although more than half of articles using path analysis in the JJEP report squared multiple correlations, \( R^2 \), few in the JEP did so. No article employing covariance structure analysis reports \( R^2 \) in the JJEP as well as the JEP. This clarifies the ‘attitude change’ pointed out at the end of 3.4.

4. Concluding Remarks

Summarizing the discussions so far, we will list some points about the psychological measurement, which should be discriminating from a mere application of data analytical techniques based on mathematical statistics;

1. The clear, explicit statement of conceptual definition of constructs and these theoretical relationships with other constructs is crucial for conducting psychological research as long as its treating constructs are abstract, especially for validating the scale that measures the constructs and designing the empirical research including causal inferences.

2. For the validation of scales (and subscales), collecting the enough amount of information is very important, which includes the correlations with other measures, mean differences between groups, and the experimental results obtained by intervening the processes with which the constructs relates. It serves for the clarification of interpretations of an existing scale and suggests some way of refining it.

3. Two distinctive features in psychological measurement should be kept in mind in the application of several multivariate techniques, particularly methods of causal modeling because they have not been developed solely for psychological applications. One is unreliability of a scale; very large amount of random error causing the remarkable attenuation of the correlation. The other is the fact that any scale should be evaluated from the viewpoint of the extensiveness of its contents (bandwidth) as well as reliability (fidelity).

4. This multidimensional character of the psychological scale just mentioned suggests that the psychological constructs themselves inevitably must be more or less ambiguous. They may be conceptualized as a bundle of many mutually correlated but distinctive dimensions, and can be decomposed into several sub-constructs in many ways. The borders between constructs are vague like those between concepts in everyday life. Hence, the corresponding scale has the same property as being divided into several sub-scales in more or less arbitrary manners. Many psychological constructs are far more context sensitive than they are usually assumed.

5. Even if theoretical and empirical information are abundant, and the longitudinal research, which is critical for causal inferences in correlational research, has been conducted, the results of causal modeling should be accepted as tentative ones if we consider the limited liability of psychological measurement and the character of causal methods depending on the sample size.

We have to admit that neither the Western nor Japanese researcher in educational psychology has conducted their work fulfilling the conditions listed above. However, we noticed some stylistic features in Japanese research through the work of review. In short, on the one hand the articles published in the U.S. usually include the explicit description of definitions and hypotheses, confined analyses of data, and rather short discussion sections; on the other hand, many papers in Japan begin with obscure definitions of constructs and hypotheses without sound ratio-
nables, and they conclude with detailed discussions that are derived from an extensive analysis of data obtained by the use of insufficiently validated scales. Contribution of some Japanese articles could be regarded as propositions of new hypotheses rather than verifications of them. In other words, the former stress the first two characters of psychological measurement listed above while the latter the remaining three.

As for the scales used in the research, they are published and sold with thick manuals including information about their reliability and validity in the U.S. and many European countries, while they are revised repeatedly in individual studies in Japan. The most salient feature shared by Japanese studies is the amount of use of exploratory factor analytical techniques with rotations approximating the loading matrix to the simple structure. It is rather difficult to find papers without application of factor analysis in generic sense in Japan. Very few articles in the U.S. journal employ this procedure because the scales are likely to be fixed by preceding studies. Of course, since the range of our review is very limited, we should avoid oversimplifications. However, it will be fruitful to consider which approach will facilitate the future research because it is true that these two approaches do really exist.

From the viewpoint of traditional logical positivism, the approach often employed in the U.S. research should be promising. Ideally, one may expect that the falsifiable model building supported by covariance structure analysis will prompt the specialization of theoretical and empirical research, which is successfully done in physical sciences. However, this competitive cooperation will not work well in the current stage of educational psychology because any predictions based on most psychological theory will be easily falsified by large sample data. Otherwise, inevitably ambiguous characteristics of psychological construct mentioned above will lead to endless and unfruitful controversy, although we do not deny that it may serve for the unfolding of complicated conjectures. In a way, it is possible that the specific features found in Japanese studies, e.g., small-steps accumulations of empirical facts and constant modification of scales fit to these characters of educational psychology. In fact, while Angoff (1988) said that construct validation is a process, not a procedure, this property may fit the style of Japanese studies.

Unfortunately, most Japanese researchers seem to be unconscious about this good nature. In addition, they will have to learn more precise and explicit definition of constructs and theoretical elaboration of problems, and to put more effort to validate scales. This must be a fundamental condition for them to do collaborations and exchanges of information with the Western researchers. When Japanese researchers establish their own methodology combining their adaptable ways of thinking with rigorousness of the Western style, they will be able to produce original outcomes and to send them internationally.

References
Breckler, S.J. 1990 Applications of covariance structure modeling in psychology: Cause for


