CHAPTER TWO

LITERARURE REVIEW

As this research aims at acquiring a better understanding of the content validity of BCT on English Subject in Taiwan, the review of literature in this research lays emphasis on the following issues: (1) the development and content of BCT, (2) the content validity of a test, and (3) BCT as a CRT.

Basic Competence Test

Development of BCT

The Ministry of Education in Taiwan developed the Junior High School Curriculum Standards for English Education in 1994. New editions of the textbooks were issued in 1997, and a reformed entrance examination, called Basic Competence Test (BCT), was designed to meet the new curriculum objectives. As stated by Lin, S. H. (2000), Wang, Y. H. (2001), Wang S. Y. (2001), and Cheng (1998), the purpose of BCT is to evaluate junior high school graduates' learning achievement and potential English ability. The test focuses on the basic, central and important knowledge and competence of English, which could assist future learning (Wu, 2000). Therefore, the difficulty level of the test is based on the average ability of junior high school students. Students are not encouraged to increase scores by over-learning and rote-learning approaches, which are the traditional strategies for gaining high scores. The former Minister of Education indicates that BCT is intended to place students in an ideal and normal learning situations (Song, 1998).

Characteristics of BCT

According to the BCT Center, BCT is characterized as standardized, multi-functional, competence-oriented, and pedagogically beneficial to backwash (Lin, S. H., 2000).

The test, developed by a test research institution, complies with a standard procedure (Chiou, 2001). During the standard procedure, the two-way item specification for every subject field, namely, the sampling and writing of the test items, pilot testing, item analysis, and the check of reliability and validity, is meticulously controlled for better test quality. Along with test development, test administration and scoring procedures are all carefully conducted.

BCT is held twice a year. Different versions of BCT are equated so that candidates taking the test at different times have comparable scores (Tu, 2000; Chiou, 2001; Lin, H. J., 2001). Students could choose to participate in the test either once or twice. Test scores could be adopted to apply for senior high schools, vocational schools or even junior colleges if the score is high (Chang, W. J., 2001).

BCT is competence-oriented; that is, it evaluates students’ basic competence (Chian, 2001). The main focus is on the English language ability needed for daily life.

One goal of BCT is to create beneficial backwash on pedagogy, so that teachers could teach normally, and the students could learn happily (Lin, S. H., 2002).

Content of BCT

Every year the test content is announced prior to the test by the Minister of Education on the website and to every junior high school. English test content of 2002, including textbooks Volume 1 through Volume 5, complies with the principles of the Curriculum Standards and the Core Competence Indicators issued by the Ministry of Education.

Studies on BCT

Although BCT is an innovated test administered in 1989, surprisingly, little empirical research has been conducted on BCT. Some of the junior and senior high school principals such as Yu, L. (2000), Ho (2000), Chang, W. J. (2001), Chien (2001), and Shie (2001) were invited to offer their opinions about the effects BCT brought about and about the dates BCT being executed. Professors like Lin, S. H. (2000; 2002), who takes responsibility for BCT, and educational officials like the Ex-minister of Education (Song, 1998), made efforts to publicize the benefit of BCT. In terms of test score interpretation, Tu and Chang (2000) and Chiou (2001) explained the transformation from raw scores into standardized scores.

Given the widespread influence of BCT, researchers such as Cheng (1998), Wu (2000), Wang, S.Y. (2001), and Wang, Y. H. (2001) viewed BCT from different perspectives. Cheng (1998) referred to the indicators of *National Council of Teachers of Mathematics* from National Assessment of Educational Progress in the USA, and proposed the evaluation principles of BCT. Wu (2000) discussed the evaluation of course standards, taking Colorado Student Assessment Program as an example. Wang, S.Y. (2001) analyzed the development of Basic Competence Indicators and thus suggested that BCT should be precisely referred to as Basic Competence Test rather than other terms. About the execution and evaluation of BCT, Wang, Y. H. (2001) pointed out the ambiguity of the test function and the score explanation, and suggested that BCT should be considered a CRT. Although the researchers above have stressed relevant issues about BCT, concerns about content validity has been neglected. What the BCT content is and what the BCT items exactly measure are not mentioned in their studies.

Content Validity

Since the crucial question lies on whether BCT measures what it claims to measure, the researcher reviewed the definition, the factors that threat content validity, and the practice of content validity in order to better understand the nature of content validity.

Definition of Content Validity

Content validity is a form of validity based on the degree to which a test adequately and sufficiently measures the particular skills or behaviors it sets out to measure. A test is considered content valid if its content constitutes a representative sample of the language skills, structures, etc. which it purports to test (Anastasi, 1982; Henning, 1987; Bechman1990; Brown& Hudson, 2002; Hughes, 2003). Henning (1987) articulates that any test developer has to answer a fundamental question: Is the content of the test consistent with the stated goal for which the test is being administered? In other words, since congruence of the BCT and the curriculum refers to the coordination between the content of the test and the content of the curriculum, the test should adequately measure what it is supposed to measure, accurately reflecting the extent to which students have mastered the content of instruction, rather than containing material that is not encountered by students in the program of instruction.

Threats to Content Validity

From the above, it is made obvious that the higher content validity a test receives, the more likely it is to be an accurate measure of what it is supposed to measure. As a consequence, according to Henning (1987), the test developer has to avoid the threats to test validity such as misapplication of test, and inappropriate selection of content.

The most noticeable threat of validity arises from misapplication of test. Test developers must consider carefully the designated purpose of a test, including the exact content or objectives to be measured, and the type of examinee to whom the measurement is given (Henning, 1987).

Secondly, inappropriate selection of content commonly occurs when items do not match the objectives or the content of instruction. Also, this may happen when the test items are not comprehensive in the sense of reflecting all of the major points of the instructional program (Henning, 1987). The best safeguard against this is to write full test specifications and to ensure that the test content is a fair reflection of them (Hughes, 2003).

Studies on Content Validity

Various researches have been conducted concerning the investigation of content validity (e.g. Spool, 1975; Lawshe, 1975; Popham, 1978; Berk, 1980; Sireci, 1995).

Brown and Hudson (2002) also indicate that study of content validity requires a systematic investigation of the degree to which the items of a test, and the resulting scores, are representative of relevant samples of whatever content or abilities the test has been designed to measure (p.213). The method for studying content validity could generally be classified as subjective and empirical.

Subjective Method

Subjective methods refer to studies where subject matter experts (SME) evaluate test items and rate them according to the item relevance and representativeness to the tested content domain. Researchers such as Spool (1975) and Mehrens and Lehmann (1987), for instance, proposed some traditional steps of selecting the most suitable achievement test. First, define the domain of interest (often with a list of instructional objectives). Second, specify the tests selected for possible adoption. Third, identify a panel of content experts. Finally, have these experts review the test on an item-by-item basis and evaluate the degree to which items match the curriculum. For this procedure, they particularly advocated two chief steps, namely, writing item specifications and adopting expert judgment.

A. Writing Item Specifications

First, the development of elaborate specification is an approach to ensure that items in a variety of domains are representative of the objectives of instruction (Henning, 1987). In order to judge whether or not a test contains content validity, a specification of skills or structures, which is meant to be within test content coverage, is needed. Popham in *Criterion-Referenced Measurement* (1978) presented five components for writing item specifications: general description, sample item, stimulus attributes, response attributes, and specification supplement. Other researchers such as Brown and Hudson (2002) pointed out that content validity involves demonstrating a well-described course, which includes the objectives and item specifications, or a specification of content domain, which consists of a detailed description of the theoretical underpinnings and the item specifications related to those theoretical underpinnings. To help judgment, they also proposed a rating scale to evaluate the degree of content congruence and content applicability. In content congruence an item should measure what it is designed to measure, while in content applicability the rater judges whether or not the content is appropriate for a given language program.

B. Adopting Expert Judgment

Secondly, in order to guarantee content validity, it is usually necessary to seek the advice of content experts to compare test specification with test content (Hughes, 2003). It may be desirable to have a panel of experts thoroughly and intensively examine each part of the test and to determine its content and relatedness to test objectives at the appropriate level (Henning, 1987). Hughes (2003) advised that the panel of experts should be familiar with language teaching, testing and the test requirements, but should not be directly concerned with the production of the test. Spool (1975) also indicated that the use of experts is essential since content validity relies heavily upon expert judgment.

According to Brown and Hudson (2002), expert judgments approach to content validity first requires identifying experts. In an example of an ELI testing project, Brown and Hudson mentioned that it proves useful to get feedback from outside "expert" in order to have independent judgments of how well the items fit the objectives (p.223).

Empirical Method

While subjective methods rely heavily on expert judgments, empirical methods emphasize the procedures of analyzing data obtained from the test. Sireci (1998) used applications of multidimensional scaling, cluster analysis and factor analysis to evaluate subject matter experts' (SME) ratings of item similarity. Hughes (2003) compared students' performances to assess the impact of test on examinee performance (Hughes, 2003). To evaluate the influence of test on examinee performance, Hughes et al in 1996, in developing an English placement test for language schools, validated test content against the content of three popular course books used by language schools in Britain, compared students' performance on the test with their performance on the existing placement tests of a number of language schools, and then examined the success of the test in placing students in classes (Hughes, 2003).

To quantify the degree of fit between test items and curricula, Crocker, Miller and Franks (1989) in their *Quantitative Methods for Assessing the Fit between Tests and Curriculum* classified two types of content validation:

1. Percentage of Items, Klein and Kosecoff's Correlational Index, and Morris and Fitz-Gibbon's Index of Relevance were adopted to assess the overall fit between test and curriculum.
2. Hambleton's Index of Item-objective Congruence, Aiken's Validity Index, and Lawshe's Content Validity Ratio were used to measure the fit of individual items to a content domain.

Among these quantitative methods, to assess the overall fit between the test and curriculum, the researcher adopted the Percentage of Items due to the reason that it is easier to calculate and be understood. It is a simple index historically used for assessing the objectives of the local curriculum (p. 180). The importance of each objective was rated a 5-point scale ranging from “very unimportant” (1) to “very important” (5).

Aiken’s Validity Index also evaluates an item’s relevance to a particular content domain, using subject matter experts’ (SME) relevance judgments. His index takes into account the number of categories on the scale, which is used to rate the items, and the number of SME conducting the ratings. The index of content validity is calculated as:

$$V = \frac{\sum_{i=1}^{c-1} i n_i}{N\,(c-1)}$$

Where c is the number of categories on the scale to rate the item. The lower category is assigned a weight (or i value) of 0; the next lowest category receives a weight (or i value) of 1; and so forth. The number of judges who rate an item into the ith category (0, 1, 2…c-1) is designated as n_i , and N is the total number of judges. To illustrate, we are rating items with a simple three-category scale on which 0 represents not relevant, 1 uncertain, and 2 relevant. If the frequency distribution of six judges is

1, 3,and 2 over the three categories respectively, then we obtain the result:

$$V=\frac{1(3)+2(2)}{6(2)} = .58$$

The probability of obtaining by chance the particular frequency distribution that yields a given V value can be assessed by the formula:

$$\boldsymbol{P} = \frac{\boldsymbol{N! /c^{N}}}{\boldsymbol{n_{0}!\ n_{1}!\ldots n_{c-1}!}}$$

For the example just used,

$$P=\frac{6!/\boldsymbol{3^{6}}}{\boldsymbol{1!\ 3!2!}}= .01$$

For a fairly large number of raters, Aiken suggested using a normal approximation to ***p*** from

$$\boldsymbol{Z} = \frac{\boldsymbol{N(C-1)(2V-1)-1}}{\sqrt{\boldsymbol{N(C-1)(C+1)/3}}}$$

The computed value of ***Z*** would be significant if Z> 1.96 or Z< -1.96 at the alpha level of 0.05. by checking a standard normal Z table.

Aiken's Validity Index allows researchers to check statistical significance of research results, so that the researcher can double check the rating results more clearly than the results originated from merely adopting Percentage of Items approach. Besides, to make the judgment easier and more practical for junior high school teachers, the researcher adjusted the number of categories into two. Therefore, judges considered each item on the test and made a dichotomous decision about its match to content appropriateness.

Development of BCT as a CRT

Achievement tests have been widely applied to assess a student's performance at the end of a course of study. How much of the language material or skills outlined in the course objectives the students have learned has become the major issue of the content validity of criterion-referenced tests (CRTs). The importance of BCT as a CRT was discussed from two aspects: the nature of CRTs, and the reason why BCT should be considered a CRT.

Nature of Criterion-referenced Tests

Gronlund (1982) stated that criterion reference, strictly speaking, refers only to the method of interpreting achievement test results. Each student's score is meaningful without reference to the other students' scores. However, many language testing researchers such as Anastasi (1982), Bachman (1990) and Brown (1996) believed that criterion-referenced tests are specially designed to assess how much of the content in a course or program is being learned by the students. Brown and Hudson (2002) proposed that the interpretation of the CRT scores is intimately linked to assess well-defined criteria for what was being taught. CRTs thus provide information about an individual’s mastery of a given criterion domain or ability level.

The particular significance of CRTs lies on its measurement of the relationship between test content and instructional objectives. Lyman (1983) pointed out that careful specification and control of test development process is the best guarantee of content validity of a CRT. Based on the views given by Hambleton, Roid, Maladyna, and Popham(1983), Lyman (1983) outlined 12 steps to constitute content validity. Of the 12 steps, clarifying instructional intent is the first step. Then, test specifications operationalize skills and abilities specified in instructional objectives. The other steps detailed the second one. Lynch and Davison (1994) formed a workshop for teachers, attempting to translate curricular goals into tests. The approach linked curricula, teacher experience, and language tests in a CRT, hoping to produce positive washback effects. These studies all stress on the close relationship between CRTs and instruction.

BCT as a CRT

Good reasons to consider BCT a CRT rather than an NRT (Norm-referenced Test) are proposed. Language testers such as Anastasi (1988), Gronlund (1988), Brown and Hudson (2002), and Hughes (2003) made fundamental distinctions between NRTs and

CRTs. Brown and Hudson (2002) pointed out that the purpose of NRTs is to spread out the students along a continuum of general abilities of proficiency. On the contrary, CRTs are constructed to assess the amount of material being known or learned by each student. Like CRTs, BCT is designed to assess students' achievement by announcing the precise test content in advance. The relative scores (NRTs) transferred from the raw scores of the BCT only function as a tool for candidates to apply for secondary school. Besides, although the name of Basic Competence Test implies the general abilities and thus causes controversy (Wang S. Y., 2001;Wang, Y. H., 2001), BCT indeed measures a specific domain and objective-based language points.

For another, Brown and Hudson (2002) suggested that educators or policy makers choose to use CRTs when they wish to see how well students have learned the skills that they are expected to master. BCT items are planned according to the Curriculum Standards, which means, through BCT, the educators can know how well the students have achieved the objectives the educators have set. What's more, Brown and Hudson (2003) indicated that criterion-referenced testing is most useful to classroom teachers and curriculum developers because CRTs are specifically designed to assess how much of the content in a course or program is being learned by the students. With BCT, not only junior high school teachers but also teachers in senior high schools are aware of how much the students have learned from the three-year language program.

While Brown and Hudson proposed the kind of information educators and teachers can acquire from CRTs, Hughes (2003) further suggested that to develop beneficial achievement tests, it should be useful to make testing criterion-referenced. CRT scores are reported and interpreted with reference to a specific content domain or criterion level of performance. Brown and Hudson (2002) pointed out that a sound CRT could provide five types of information: students' needs, the goals and objectives, the tests themselves, the materials, and the teaching strategies. As long as the content of the test matches the content considered important to learn, CRTs give students, teachers, and parents more information about how much of the specific content has been learned. That is also the purpose of BCT. Wang. S. Y. (2001) thus strongly suggests that BCT should be regarded as a CRT.