

Chapter 2

Literature Review

2.1 Introduction

In *Statistical Analyses for Second Language Assessment* Bachman defines assessment as “the process of collecting information about a given object of interest according to procedures that are systematic and substantively grounded (Bachman, 2004, p. 7).” An assessment or assessments can also refer to the individual instruments, techniques and processes used to collect information and also the results of such procedures and instruments. For the purpose of this research a broad survey of literature was conducted regarding language assessment, standardized assessment and motivation.

2.2 A Brief History of Assessment

The first standardized examinations were the *Imperial Examinations* issued in China c. 2200 B.C., measuring a wide variety of skills including music, archery, arithmetic, writing and knowledge of rituals. The first written examinations were introduced between 202 B.C.–200 A.D. and have been used popularly to some degree ever since (Gregory, 2007). Assessments in educational settings have generally focused the measurement of “what learners know and can do” and have been used for a variety of purposes, such as determining eligibility, providing diagnostic feedback to learners, instructors and schools, and serving as a basis for graduation and credentialing (Reynolds, Livingston, and Willson, 2010).

2.3 Validity and Reliability in Assessment

The quality of an assessment is measured in terms of the validity and reliability with which its results can be interpreted (Bachman and Palmer, 1996). Assessments yield valid results when they successfully measure the qualities they intend to measure. Reliable assessments measure qualities with adequate consistency, both in terms of replicability and its application to the entire spectrum of participants it aims to measure. The validity of an instrument moreover depends to some extent upon its reliability (Bachman, 2004). Validity is often discussed in three varieties: construct validity, content validity, and criterion validity. Construct validity refers to “the degree to which a test measures what it claims, or purports, to be measuring (Brown J. D., 2000),” and encompasses the overarching aim and general sense of validity in language testing. High content validity implies the object of measurement is considered with sufficient depth, breadth and scope. High criterion validity implies the objects of measurement and the way they are measured maintain a good correspondence with external contexts (Pennington, 2003).

2.4 Language Assessment

Language assessments have generally corresponded with the going methodological trends popular at any given time. Grammar translation, the audio-lingual method and communicative approaches have all favored respective language assessments suitable to their individual aims, objectives, and theoretical underpinnings (Anderson, 1998; Nunan, 1988; Bailey, 1998). Discussion of language assessment often focuses on the major skills of language use and how each skill can be most efficaciously measured, sometimes in combination with other skills and activities. Productive skills require unique sets of procedures appropriate to their fundamental nature while receptive skills can often withstand objective measurements like multiple-choice tests (Read, 2000; Alderson, 2000; Douglas, 2000; Weigle, 2002; Buck, 2001; Purpura, 2004; Bachman, 2004). Others have pushed for language assessments that promote integrated views of language use, rather than repeatedly focusing on minute aspects of language use; the whole, they argue, is greater than the sum of its parts (Bachman and Palmer, 1996; Flowerdew and Miller, 2005).

2.4.1 Assessing Listening

Listening is the receptive auidial skill and must be assessed using means appropriate to the nature of the skill and its function in discourse (Bachman and Palmer, 1996). Assessing of listening may include the following: pre-listening activities; listening activities; post-listening activities; summary activities (Buck, 2001). Successful listening assessments make use of authentic tasks and maintain relevance to external contexts (Flowerdew and Miller, 2005). Popular formats include noise-tests, listening cloze tests, and dictation (Buck, 1988).

2.4.2 Assessing Speaking

Speaking is the productive vocal skill and successful speaking assessments target skills and microskills associated with speaking (Bachman and Palmer, 1996). Speaking examination tasks rely on one-to-one formats, paired formats, and group speaking tasks (Galaczi and Buggey, 2011). The level of interaction used in speaking examination formats can vary with some examination formats such as interviews involving a high degree of interaction and others, such as oral presentations, involving a low degree of interaction. Many speaking examinations are scored using rubrics that divide qualities of the speaking performance into sensible descriptors that provide greater depth and breadth to evaluation (Luoma, 2004). Like all language examinations, the degree to which speaking assessments pertain to external environments remains primary (Bachman and Palmer, 1996)

2.5 Multiple-Choice Assessments and Learning Development

Multiple-choice is among the most widely used assessment formats (Roediger III and Marsh, 2005). Multiple-choice assessments are valued for their internal reliability, cost-effectiveness and the ease with which they can be produced and administered. Developing good multiple-choice questions, however, can be difficult and, as a consequence, good questions are often recycled within learning institutions. Multiple-

choice has been criticized for the limiting what can be measured to low-level cognitive skills and an inherently poor range of expression. In a study of the advantages and limitations of multiple-choice testing Roediger III and Marsh (2005) found that multiple-choice assessments had, on the positive side, aided learner ability to recall information but also caused learners to apprehend misinformation contained within distractors.

In language learning, multiple-choice tests may have many limitations (Bailey, 1998). Neither speaking, writing nor interaction can be reliably measured using multiple-choice formats. Consequently, multiple-choice assessments most effectively measure receptive skills (listening and reading). Multiple-choice questions have also been associated with cheating. Lastly, distractor-stems may expose learners to high volumes of faulty input (Roediger III and Marsh, 2005).

Multiple-choice assessments, despite their acknowledged general limitations as means of measuring language use, serve many practical purposes in the field of second language learning assessment and are especially useful to specific given circumstances such as when large volumes of learners must be assessed or other resources are not available (Bailey, 1998; Roediger III and Marsh, 2005). Multiple-choice assessments formed the primary method of assessment for most of the language assessments used in this study. The multiple-choice format was chosen for use in this study primarily for its ease of administration and also because of their universality and the familiarity they could therefore provide to learners; other means of assessing listening and speaking, such as oral examinations or live conversation examinations, were not practical given size of classes and learning sections. Other formats such as essays and written responses were also avoided because they do not meet the criteria of the course design specifications, which prescribed a program of listening and speaking tasks only.

2.6 Assessment in the Era of Standardized Testing

Education as we know it emerged from the economic circumstances of the Industrial Revolution and the intellectual climate of The Enlightenment (Robinson, 2008; Gillard, 2011). The shift from farm labor to factory work meant that children were no longer needed to work on farms and needed a place to go while their parents worked in factories. Progressively minded individuals seized the opportunity to facilitate educations for large numbers of people. New theories underlying assessment took shape and technology with which they could be better produced came to be and so the modern era of standardized assessment was born (Gregory, 2007).

2.6.1 Theoretical and Empirical Perspectives

Educational philosophers since the time of Aristotle have wrestled with the question of how knowledge is known. "Knowing yourself," wrote Aristotle, "is the beginning of all wisdom." Whether knowledge can be measured using instruments and if doing so deepens individual understanding have additionally been considered. In *Democracy and Education* John Dewey wrote: "Were all instructors to realize that the quality of mental process, not the production of correct answers, is the measure of

educative growth something hardly less than a revolution in teaching would be worked (1916, p. 169).” Contemporary education, it would seem, often prefers the production of correct answers and knowledge in this paradigm, exits solely as isolated information bits supplied to learners then expected to reproduce them (Robinson, 2008). Freire argued that traditional education reduces learners to mere receptacles for knowledge into which the job of educators is to deposit knowledge. Regarding the teacher, “the more completely she fills the receptacles, the better a teachers she is.” Freire (1993, p. 52).

Many alternatives to traditional education and, correspondingly traditional assessment, have been proposed and used with varying degrees of influence and success. For other researchers like Alfie Kohn, while a proponent of nontraditional forms of education, alternative assessment does not diverge sufficiently from traditional assessment in its aim and scope and should be eliminated (Kohn, 1999; Kohn, 2007; Kohn, 2012; Kohn, 2000; Kohn, 2011).

The No Child Left Behind Act in the United States (NCLB) has been a source of great controversy. NCLB mandated states provide evidence of yearly progress relative to basic skills measured by standardized examinations and was sold in partly on the promise that so doing would lessen the gap in the achievement between high and low proficiency learners, specifically between minority and nonminority learners. Yet many argue this has not been the case, that, instead, the gap between well and poorly scoring learners has widened considerably, especially for minority learners (Altshuler and Schmutz, 2006; Diamond and Spillane, 2004; Sapon-Shevin, 2011; Frey, Fisher, and Nelson, 2013). Using high stakes standardized examinations to evaluate teacher performance is not a valid procedure, it has been argued, and has resulted in narrowed curriculums and higher dropout rates, in the meantime demoralizing rather than motivating teachers who, as a consequence of the narrow criteria for evaluation (student test scores) often avoid problem schools where their contributions might be poorly reflected and which often remain most in need of their service, and that, moreover, so doing has not been shown to induce learners to perform better (Baker, et al., 2010; Amrein and Berliner, 2002). In fact, between 2001, when NCLB was legislated into action and 2009 The United States dropped from 18th to 31st place worldwide in math and science after nine years of standardized testing (Darling-Hammond and McCloskey, 2011).

A loose association of consequences associated with standardized testing known as washback have been reported, with findings generally showing that testing influences “what and how English-language learners are taught” (Powers, 2010, p. 9). Consequently, were some skills emphasized on tests at the expense of others, it is probable the resulting diminished emphasis on the untested skills in classrooms could have serious negative consequences.

2.6.2 Standardized Testing and Language Learning

Standardized testing plays a pivotal role in the lives of many second language learners. Learners hoping to work or study in target language contexts depend upon minimum proficiency scores on standardized language examinations like TOEFL, TOEIC and IELTS (Educational Testing Service, 2014; Cambridge English, 2014).

The TOEFL and TOEIC have recently been amended to measure all four major language skills in order to assesses “the broader trait of communicative competence” rather than merely few skills in isolation (Powers, 2010, p. 10). Studies have found a poor correlation between TOEFL scores academic success (Ng, 2007; Xu, 1991) and that using TOEFL scores to place ESL learners university programs would not be effective (Kokhan, 2013).

The qualities of design used on the assessments included in this study mirrored some aspects of standardized language assessments. Language assessments constructed for this study were first piloted and then reconstructed based upon the findings of the pilot. The assessments used in this study favored multiple-choice questions and were issued to large numbers of learners many of whom, moreover, were preparing for standardized language examinations.

2.7 Alternative Assessment in Second Language Learning

As a result of the inherent limitations of using certain so-called traditional formats of assessments to measure certain language skills—for example, the inadequacy of using multiple-choice format to measure speaking and writing—alternative assessments in second language learning contexts have come to serve many useful purposes. Discussion here presents alternative assessments used in language instruction as a contrast to the traditional, predominantly multiple-choice assessments that were used to measure participants of the study.

2.7.1 The Distinction Between Traditional and Alternative Assessment

While some overlap may occur, alternative forms of assessment have been distinguished from traditional ones by similar criteria as that presented below, which summarizes Anderson (1998, p. 9):

Table 4 Traditional and Alternative Assessment

Traditional Assessment	Domain	Alternative Assessment
Single meanings	Knowledge	Multiple meanings
Passive	Learning	Participatory
Separation of process and product	Process	Emphasis on both process and product
Discrete, isolated bits of information	Focus	inquiry
To document learning	Purpose	To facilitate learning

Traditional Assessment	Domain	Alternative Assessment
Cognitive abilities are separate from affective and conative abilities	Abilities	Connects between cognitive, affective and conative abilities
Assessment is viewed as objective, value-free and neutral	Assessment	Assessment is viewed as subjective and value-laden
Dominated by hierarchical models	Power and Control	Power is shared
Learning is an individual process	Individual vs. Collaborative	Learning is a collaborative process

2.7.2 Varieties of Alternative Assessment

Some of the most frequently used alternative assessments include or have included:

Assessment for Learning/Assessment as Learning: Assessment platforms that emphasize using assessments as a means of learning rather than merely a measurement of learning; often such assessments use authentic tasks that aim to provide learning with their assessments (Sambell, McDowell, and Montgomery, 2013; Manitoba Education, Citizens and Youth, 2006).

Authentic Assessment: Authentic assessments see that learners engage in meaningful tasks that directly relate the skills and knowledge they must know (Wiggins, 1990; Wiggins, 2002; Wiggins, 2006).

Dynamic Assessment: An interactive approach to assessment based on the research of psychologist Lev Vygotsky (Haywood and Lidz, 2006).

Narrative Evaluation: Detailed, written descriptions of student performance providing in-depth analysis of student performance used at Brown University, Hampshire College, Oxford University, Yale Law School, among others (Wikipedia, 2013).

Portfolio Assessment: Collections of student work are assembled purposefully to reflect progress over the course of an instruction period (Mueller, 2012), with emphasis placed on the purpose of selection (Martin-Kniep, 1998). Portfolios vary in scope, with some purposefully arranged to “showcase” choice selections of student work and others arranged chronologically to include all assignments

Self-assessment: Any of a variety of procedures wherein learners provide a meaningful evaluation of themselves, their abilities, desires and performances.

2.8 Consequentiality of Learning

Assessment adds consequentiality to learning. Consequentiality here refers to the scope of impact the results of assessments yield. Assessments upon which important

decisions will be made are generally known as high stakes assessments. Because of the magnitude of their importance and their influence on the lives of learners and society, high stakes assessments must maintain the highest levels of validity and reliability. Assessments upon which less important decisions are made are called low stakes assessments and can be used as formative assessments and learning tools (Bachman, 1990). High stakes assessments are sometimes thought to create the occasion for motivation within learners; learners aware of the profound relevance of assessments perform especially well, both at the time of assessment and in terms of preparation (Wise and DeMars, 2003). High stakes assessments are also associated with two unwanted side effects: increased learner anxiety and cheating (Kohn, 2007; Wade, 2013). The scores of learners on high stakes standardized examinations are often used as the sole indicator of teacher performance. However, Baker, et al (2010) found that relying on learners' scores on high stakes standardized examinations to evaluate teachers does not provide a valid assessment of instructors. It moreover has been found that when consequentiality differs between the examinees and other stakeholders, the data supplied as a result might lack validity (Cole, S., and Osterlind, 2008; Guerriero, 2013). Amrein and Berliner (2002) found that learners, when independently measured, did not improve as a result of the stakes that had been imposed: "in all but one analysis, student learning is indeterminate, remains at the same level it was before the policy was implemented, or actually goes down when high-stakes testing policies are instituted."

In Finland and Norway high-stakes assessments are not administered to young learners, however their ultimate relative performance on international measures remains high (Moore, 2013; Hasselgren, 2000). On the other hand, in China, learners are conditioned using high-stakes assessments and Chinese learners also score in the top tier worldwide (McKinsey and Company, 2010). Stakes associated with student performance are notoriously low in Thailand—students cannot fail their classes, no matter how poorly they perform—and Thailand ranks low on educational international measures (Vongkiatkajorn, 2012; Fry, 2013; Halligan, 2011).

2.9 Communicative Language Teaching

Communicative Language Teaching (CLT) replaced grammar-based approaches to language learning in the 1970s and focused on communicative interaction typically using authentic materials. Instead of memorizing grammar forms and discrete language points, linguists began to stress communicative ability and correspondingly to assess communicative competence rather than knowledge of grammar rules (Richards, 2001). In communicative language teaching the emphasis is generally on fluency rather than accuracy, although the definition of fluency in language learning has sometimes been vague (Cucchiari, Strik, and Boves, 2000). The instruction provided for this study was rooted in CLT principles and made use mostly of authentic materials.

2.10 Second Language Acquisition and Learning: The Logical Problem of Language Acquisition

Primary in the treatment of second language acquisition and learning is the discussion of differences in first language learning (Lightbown and Spada, 1999). Our first language develops fluidly, naturally and above all successfully. Second language acquisition remains a choppy, unreliable process. In order to discover how most effectively to learn second languages, researchers look to the acquisition of native language, for which there exist what has been called the logical problem of language acquisition, which is, to wit that “the linguistic data to which children are exposed appear to be insufficient to determine, by themselves, the linguistic knowledge which children eventually attain (Bley-Vroman, 1990, p. 3).” In other words, we know and can use more language of a higher complexity than that to which we could have possibly been exposed. And yet many suggest that first and second language acquisition are not the same, that different rules apply and that Chomsky’s Universal Grammar, Language Acquisition Devices and the Critical Period Hypothesis active in children learning first languages, remain somewhat beside the point for adults attempting to learn a second or additional language. It is therefore often concluded that adult second language acquisition would mirror adult acquisition of other skills, such as musicality and athletics, but still would not predict an assessment-based structure to provide the greatest platform for learning (Bley-Vroman, 1990).

Much acquisition-based language instruction models favor the incidental acquisition of vocabulary, rather than, for example memorizing words to supply on tests (Ko, 2012; Dupuy and Krashen, 1993), and, many models that rely on acquisition-oriented frameworks, such as Krashen and Terrell’s Natural Approach, still emphasize the importance of evaluating learner performance (Krashen and Terrell, 1983). Communicative language learning approaches favor evaluation of communicative competence and generally view the evaluation of learners as necessary and facilitative of acquisition (Littlewood, 1981). This study measured the effect of frequent assessment on learner ability within a communicative language instruction paradigm. Whether the presence of assessments promoted or somehow interfered with acquisition, represented by the quality of ability, was at least in part measured.

2.11 Personal Characteristics

Personal characteristics are defined in Bachman and Palmer as “the individual attributes that are not part of test takers’ language ability which may still influence their performance on language tests (1996, p. 64)” and have been widely studied in second language research with respect to age (Snow and Hoefnagel-Höhle, 1978) and gender (Ehrman and Oxford, 1989). It is moreover generally acknowledged that different learners in general will respond differently to given instructions and assessments (Skehan, 1989). Such research bears primary relevance to the analysis of differences between assessed and non-assessed learners in the present study where differing performances by age and opinions toward the class were measured were analyzed based on treatment.

2.11.1 Age

The relationship between age and acquisition has been of fundamental concern both to first and second language research. Neurologically, it is posited that a child's brain is better suited to the acquisition of languages—childhood is the period when languages are most readily acquired. As the brain matures, lateralization occurs whereby brain functions are relegated to a given hemisphere of the brain making it difficult to acquire native-like proficiency in a second or additional language. While not disproven, the Critical Period Hypothesis is not universally accepted theory, except with respect to phonology where it has been quite well demonstrated that phonological forms are better acquired within a certain period (Brown D. H., 2000). Ultimately, however, it must be conceded that, where age is concerned, factors incidental to the sheer number play a significant role. Where learning second languages is concerned, it is generally thought earlier exposure to language yields superior long term results, nevertheless adult learners often apprehend individual components of material quicker than younger learners. There is some research that suggests that younger teenaged learners (ages 12–15) may acquire second languages faster and more completely than learners just a little younger or older (Bley-Vroman, 1990; Snow and Hoefnagel-Höhle, 1978). Of particular note has been the introduction of the Critical Period Hypothesis, which posits that there exists a certain phase of development beyond which language acquisition grows increasingly difficult (Brown D. H., 2000), beginning at the age of two and ending, according to some, at puberty (Lenneberg, 1967), and, according to other research, at the age of five (Krashen, 1973).

2.11.2 Gender

Gender roles in second language acquisition have been studied primarily relative to motivation and motivation to learn a given language and participate in language learning activities (Kissau, 2006; Ehrman and Oxford, 1989). Males and females are known to develop at different rates biologically, neurologically, and cognitively and are known to develop language and vocabulary at different rates, with females generally developing more quickly than boys overall and in terms of language and literacy (Klinger, Shulha, and Wade-Woolley, 2009). Nevertheless, gender has been shown to influence the way individuals use and develop languages (Shakouri and Saligheh, 2012). It has not, however, been determined that females perform better overall or vice versa or that gender plays a profound influence in the quality of development (Ellis, 1994). Ellis (1994) notes, for example, that Asian men in Britain attain higher levels of second language proficiency in English than Asian females largely because they interact with English-speakers more in business circumstances. Females have been shown in a number of studies to utilize strategies with better effect and to develop superior grammar and articulation skills sooner (Shakouri and Saligheh, 2012). Females have, moreover, been found to make greater use of resource strategies than men (Green and Oxford, 1995) and to generally possess a higher motivation to learn the target (Kissau, 2006). Societal influence has been proposed as one of the reasons motivations vary across genders (Kissau, 2006).

2.12 Motivation

Arbitration of the variance in learner success acquiring a second or additional language seems most often to rely upon the degree of learner motivation to learn a second or additional language (Bley-Vroman, 1990; Dörnyei, 2008; Gardner, 2007). Learners who are highly motivated to learn languages, or who have been taught or otherwise make use of learning strategies succeed where unmotivated learners or those who make no use of learning strategies, do not. In psychology, and sometimes in education literature and second language acquisition literature, it is acknowledged that circumstances authorizing autonomy predict higher levels of motivation and better performance than circumstances in which learner behavior is either manipulated or controlled (Pink, 2009). Whether the existence of these assessment structures actually has this effect has not been demonstrated, however numerous studies in psychology, economics and education show that individuals lose inherent interests in tasks when rewards are given (Lepper, Greene, and Nisbet, 1973; Greene, Sternberg, and Lepper, 1976; Deci, Koestner, and Ryan, 2001; Deci and Ryan, 2000; Enzle and Ross, 1978; Carlson, Miller Jr., Heth, Donahoe, and Martin, 2010; Deci and Ryan, 2000). It is possible the results of assessments serve as the carrots-on-a-stick and, even when effective in the short term, would have a shorter half-life than assessment borne of autonomy (Pink, 2009).

While autonomy is generally predicted to yield better motivation and results, learners are rarely given the opportunity to decide whether they will participate in language assessments or not—and neither were the learners in either treatment scenario where, on one hand learners were, without choice, administered several assessments or, on the other hand, though also without choice, administered no assessments. Assessment strategies likely influence learner motivation, as any aspect of a course be it instruction, materials, or even environment, and here the effect of the assessment strategy used was measured indirectly using the opinion survey questionnaire on which learners answered several questions pertaining to how likely they would be to apply what was learned in the future, sometimes referred to as willingness to communicate (WTC) or motivation to transfer (MT) (James, 2012) as well as more general questions regarding the effectiveness of the class. Poor responses to the class would reflect an impingement of learner motivation to perform well in the given scenario and possibly influence overall motivation to learn the target language (Gardner, 2007).

2.13 Related Studies

Several research studies have examined the relationships between intermediary assessments and final assessments and research speaks well of formative assessment as a means of preparing for summative assessments (Popham, 2008). Researchers in Taiwan found positive uses of practice tests towards the eventual success learners taking the TOEIC (Chen and Chiu, 2004). Learners who prepare for language examinations are known to perform better on average than those who “wing it.”

Age of acquisition research forms the bedrock of L2 research (Lightbown and Spada, 1999) and gender differences are widely studied (Ehrman and Oxford, 1989), largely

with respect to language learning strategies. Many independent studies have considered strategy use by gender relative to Thailand, the same context in which research for this investigation was conducted (Khamkhien, 2010; Phakiti, 2003; Prakongchati, 2007). Some studies have found scarcely any difference between male and female strategy use in Thailand (Khamkhien, 2010); one study found male metacognitive strategy use to be higher than female metacognitive strategy use, but no significant differences between results on a reading comprehension test and the corresponding use of cognitive strategies learners reported (Phakiti, 2003). A comprehensive study of L2 strategy use among university freshman at eight Thai public universities found that females consistently made better use of all measured strategies and that strategy use was a significant factor in determining learner success (Prakongchati, 2007).

2.14 Conceptual Framework

Figure 1 represents the conceptual framework for research utilized by this study.

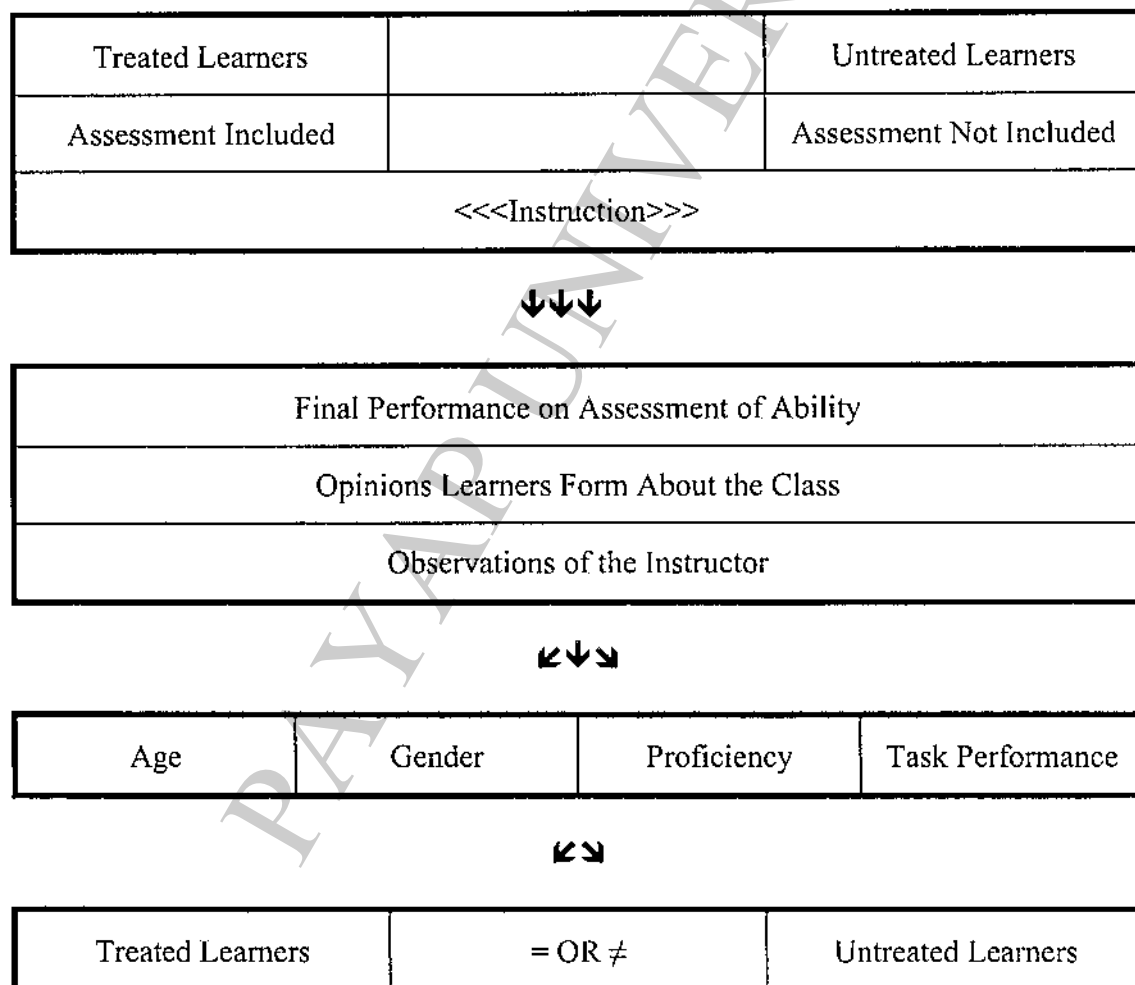


Figure 1 Conceptual Framework