CONSTRUCTING A NEW THEORETICAL FRAMEWORK FOR TEST WISENESS AND DEVELOPING THE KNOWLEDGE OF TEST-TAKING STRATEGIES (KOTTS) MEASURE

By

Hannah-Hanh Dung Nguyen

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Department of Psychology

2003

ABSTRACT

CONSTRUCTING A NEW THEORETICAL FRAMEWORK FOR TEST WISENESS AND DEVELOPING THE KNOWLEDGE OF TEST-TAKING STRATEGIES (KOTTS) MEASURE

By

Hannah-Hanh Dung Nguyen

One of the hurdles of taking a multiple-choice test is understanding the testing situation and knowing how to take the test effectively. A test taker's ability to clear this hurdle is commonly known as test wiseness (TW). In the present study, test wiseness was conceptualized as a test-taking construct consisting of three integrated psychological components: cognitive, behavioral, and metacognitive. Because test takers' acquisition of general strategies in taking multiple-choice tests was considered fundamental in this TW framework, the Knowledge of Test-Taking Strategies (KOTTS) instrument was developed and evaluated mainly to assess individual differences in the declarative knowledge of test-taking strategies (TTS). This measure was also modified to assess self-reported use of strategies on a given cognitive ability test in the present study. The result of a series of reliability analyses, confirmatory and exploratory analyses was a reliable and valid measure of 11 dimensions. These dimensions were differentially correlated with test takers' performance on a cognitive ability test, test-takers' characteristics (GPA, ACT or SAT scores, age, gender, previous training in TTS and 2 personality traits), and with other test-taking skills and attitudes (test-taking metacognitive strategies, test-taking motivation, test-taking self-efficacy, and multiple-choice test-taking self-efficacy). The practical use of the KOTTS measure in the employment selection context was discussed.

This thesis is dedicated to my parents, Mrs. Kim-Dung and Mr. Hien Huy Nguyen.

ACKNOWLEDGEMENTS

This thesis grew out of a series of conversations with my advisor, Professor Ann Marie Ryan. She saw the need for this line of research and inspired me to engage in the endeavor. Ann Marie's comments on my drafts were themselves a course of critical thoughts upon which I constantly drew to finish my thesis. She has the capacity of combining constructive critique with an immediate empathy for her students' struggles and a commitment towards their progress. I am indebted to her support and guidance.

Special thanks to my committee faculty, Professor Neal Schmitt and Assistant Professor Fred Oswald, for their invaluable input. I thank Neal for his genuine caring for students' progress and his technical advice regarding my use of LISREL. I thank Fred for giving me several unofficial occasions where I practiced defending this topic before my official thesis defense.

I might not have had the chance to go to graduate school in the first place if my parents had not persevered in their 2-decade efforts to secure our migration to the United States. They made sacrifices, leaving their home, pets and financial security behind, becoming underemployed in America, so that their children could have a chance at a good education and intellectual freedom. For that, they have my heartfelt thanks. I am also grateful to my kid sisters, Hoa, Hop, Hien and their families, for their unconditional love and the special bond among us.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Overview

Cognitive ability tests (CA-Ts) have been used for selection purposes in both academic and employment settings. However, there are consistently observed differences in mean CA-T scores between Caucasians and minorities (African Americans and Latino Americans; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). For decades, researchers have tried to analyze the problem of adverse impact associated with CA-T scores in order to remedy it. The psychometric properties of CA-Ts were thought to be the problem until empirical evidence has verified that these properties do not explain the observed adverse impact (e.g., Bartlett, Bobko, Mosier, & Hannan, 1978; Hunter, Schmidt, & Hunter, 1979).

Research attention has since turned to other possible determinants of a CA-T score besides the true score. The question is whether there are additional ethnicity-related constructs being measured in CA-Ts besides general mental ability. The determinants being investigated range from minority psychological identity (Ogbu, 1991), stereotype threat activation (Steele & Aronson, 1995), to the difference in the use of test-taking strategies between Blacks and Whites (Ellis & Ryan, in press). Among these determinants, the concept of test wiseness (TW), as well as the related construct of test-taking strategies (TTS), has long been recognized as a source of variance in CA-T scores in the measurement literature (Thorndike, 1949, 1951). Because some studies found differences in TW or the use of TTS between Whites and African Americans or Latinos

(Anastasi & Cordova, 1953; Ebel, 1965; Ellis & Ryan; Kalachstein, Kalachstein, & Docter, 1981), research targeting the relationships between CA-T scores and TW or TTS may shed light on the group mean differences in CA-T scores in both academic and employment testing settings. For example, Guion (1998) believed that research on TTS should be given "a prominent place on the research agenda" (p. 478). His rationale was that TTS likely constituted "the third variable" problem—the "unknown and unmeasured variables influencing predictor/criterion measures or moderating predictor-criterion relationships"—and that there might be strategy differences associated with group membership.

Prior to engaging in such research endeavors, it is necessary that we first have a clear understanding of the concept of TW and a valid, reliable measure of this construct or related constructs. At first glance, TW is presented as a consistently defined and measured construct in research literature. However, a closer review of the literature reveals the falsity of this presumption. TW has been interpreted as a specific, unidimensional or carefully stipulated concept (e.g., Woodley, 1975), or as a broad collection of cognitive skills and abilities (Flippo, Becker, & Wark, 2000). Even the term "test wiseness," that is familiar to many test constructors, is not always favored by researchers in the field. For example, this phenomenon was also denoted as "test insight" (Thorndike, 1951), "test wisdom" (Diamond & Evans, 1972; Preston, 1964), "testmanship" (Huff, 1961), or "test sophistication" (Anastasi, 1976; Erickson, 1972). Occasionally, the phrase "test-taking skills" or "test-taking strategies" are used interchangeably with TW. Unfortunately, the conceptualization and operational definitions of the construct are as messy as the labels used to identify them. The variety

of denotations of the concept either reflects what some researchers call a "conceptual confusion" (p. 41; Green & Stewart, 1984), or a disagreement among TW researchers with respect to how this concept should be theorized, interpreted and, subsequently, measured with an appropriate instrument or a battery of instruments. This study aims at clarifying the construct and developing a measure that assesses an aspect of the construct of TW.

The Structure of the Study

First, background literature is reviewed to provide a broad picture of issues related to the evaluation and interpretation of CA-T performance, including (a) the minority-majority cognitive ability score difference and (b) continuing research efforts to identify determinants of measurement error in CA-T scores. Existing definitions of the construct of TW in the research literature will be introduced, and some effects of TW training will be evidenced. This section will set the stage to call for a new TW theory.

Second, a new conceptual model for TW is proposed, integrating three psychological components: a cognitive factor, test-taking behaviors, and test-taking metacognition (level of awareness of testing process). This theory of TW explains not only the manifestation of the construct but also the process of one's becoming strategically proficient in taking multiple-choice tests.

Third, TTS as a related yet distinguishable from and more narrow construct than TW will be reviewed, along with a review of the measures of TTS in the literature. A review of the empirical effectiveness of individual TTS will also be presented.

Fourth, steps toward developing and validating a measure of knowledge of TTS will be proposed, along with a set of hypotheses about the relationships between TTS as

measured with the newly-developed instrument and test takers' characteristics and attitudes.

Last, the results of the measure development and evaluation study will be reported and discussed.

Cognitive Ability Tests and Test Wiseness as a Determinant of Variance

CA-Ts are a class of individual tests intended to measure general mental ability (*g*) that is empirically related to job proficiency for most jobs (Hunter & Hunter, 1984) and to training success in all job families (Hunter, 1980; Hunter & Hunter, 1984; Pearlman, Schmidt, & Hunter, 1980). CA-Ts are also good predictors of employees' job success and economic gains for organizations (e.g., Hunter, 1981; Schmidt & Hunter, 1998). However, there is the problem of adverse impact associated with the use of CA-T scores in academic admission or employment selection: a mean score difference of approximately one standard deviation tends to exist between African American and Caucasian subgroups with the difference favoring Caucasians; the Latino/Chicano American group mean test score is between .5 and .8 standard deviations below that of Caucasians (Berryman, 1983; Dearman & Plisko, 1981; Hennessy & Merrifield, 1978; Hunter & Hunter; Jensen, 1980; Loehlin, Lindzey, & Spuhler, 1975; Sackett & Wilk, 1994). Only the East Asian group mean is typically higher (d = .20) than the Caucasian group mean (i.e., Hernstein & Murray, 1994).

Conducting a meta-analysis, Roth et al. (2001) found that the ethnic group differences in cognitive ability in academic and employment settings are close to the previously reported values (the overall mean standardized differences d's for g are 1.10

for the Black-White difference and .72 for the Latino-White difference). However, the sizes of group differences, which Roth et al. termed the "generally accepted effect size," might not be accurate unless people take several moderator effects into account. The identified moderators were job complexity (i.e., Black-White *d* for high-complexity jobs was .63, lower than that for low-complexity jobs, $d = .86$), the use of within-job versus across-job research designs (i.e., for incumbents, *d*s were .38 for within-job design and .92 for across-job studies) and the use of applicant versus incumbent samples (i.e., for applicants, within-job *d* was .83 and across-job *d*s were from 1.0 to 1.23).

Possible causes of the observed mean score differences have been extensively studied for decades. A substantial body of research has effectively refuted the hypotheses that the psychometric properties of CA-Ts (e.g., differential validity, differential prediction, criterion bias, test item bias) are the causes of adverse effects in employment settings (e.g., Gottfredson, 1986; Schmidt, 1988) and in educational settings (e.g., Cleary, Humphreys, Kendrick & Weisman, 1975; Jensen, 1980; Linn, 1973; Stanley, 1971). For example, studies with adequate statistical power have concluded that CA-Ts are equally valid for Whites, Blacks, and English-speaking Latinos (Bartlett et al., 1978; Hunter, Schmidt, & Hunter, 1979; Linn, 1978; Schmidt, Pearlman, & Hunter, 1980). Studies using the fairness model proposed by Cleary and Hilton (1968) have shown that lower ability test scores are consistently correlated with lower levels of performance across racial subgroups (e.g., Bartlett et al., 1978; Campbell, Crooks, Mahoney, & Rock, 1973; Gael & Grant, 1972; Wigdor & Garner, 1982). Even when differential prediction is found, it is often slightly over-predicting rather than under-predicting job performance of Blacks and Latinos. Regarding racial and gender bias in test items, Hunter and Schmidt

(2000) asserted that studies that had detected test item bias (differential item functioning) either failed to control for measurement error in ability estimates, violated the assumption of unidimensionality required by bias detection methods, or relied on the statistical significance of tiny effect sizes of differential item functioning detected in very large samples. These artifacts taken into account, there was no evidence that ability test items functioned differently in different racial and gender groups.

Because the psychometric properties of CA-Ts do not explain the aforementioned group mean differences, investigators have tried to identify other determinants of variance in CA-T performance besides a true score. According to the Classical Test Theory (Nunnally, 1978), a test score is an imperfect observed measure of a latent variable; in other words, a test score is the combination of a true score and nonsystematic measurement error. Johnson (1984) later elaborated on this definition. She posited that a test score was affected by three potential factors and their interactions: (1) the test itself, (2) the characteristics of those taking the test, and (3) the characteristics of the setting and testing procedures. Examples of the test characteristics included the item scaling model and methods, the format of test questions, item content and features, and the test development samples. Examples of test takers' characteristics included individual differences in ability, school experiences, motivation, ways of thinking, and interests and preferences, as well as demographic components such as race, gender, and socioeconomic status. Examples of the testing setting and procedures included expectancies and beliefs of test administrators, physical surroundings, and the mode of test preparation or presentation.

Of particular interest for educational researchers and psychologists is the influence of test-takers' characteristics on CA-T performance. Some investigated characteristics that potentially contribute to a CA-T score beyond individual cognitive ability include test wiseness (e.g., Millman, Bishop, & Ebel, 1965; Sarnacki, 1979; Thorndike, 1951), test takers' dispositions (e.g., Arvey, Strickland, Drauden, & Martin, 1990; Schmit & Ryan, 1992; Barbera, 1994), test takers' motivation (e.g., Chan, Schmitt, DeShon, Clause, & Delbridge, 1997), minority psychological identity in opposition to the imposing society (Ogbu, 1991), and the feeling of being stereotypically threatened (Steele & Aronson, 1995). Among these variance sources, test wiseness (TW) is probably the longest-standing construct that has received speculation and scrutiny for years. Industrial-organizational psychologists have recently paid more attention to TW and/or the related construct of TTS (i.e., Doverspike & Miguel-Feruito, 2001; Ellis & Ryan, in press; Guion, 1998; Nguyen, O'Neal, & Ryan, 2003). In the next section, I will review the theoretical background of TW.

Test Wiseness

In this section, various viewpoints of TW as a theoretical construct will be presented along with the relationship between TW and general mental ability. The conceptualization section will be followed by a review of the empirical evidence of TW existence based on results in the TW training research literature. I will then present my own theoretical framework of TW and a discussion of the psychological elements of this framework.

Definition of Test Wiseness

The construct of TW has a long history. Cronbach (1946) first conceptualized TW as a persistent characteristic of test takers, positing that some examinees could acquire a systematic way of test taking and, thus, increase their score beyond their measured knowledge of the content areas. Thorndike (1951) confirmed that individuals' general ability to comprehend instructions and TW "are likely to enter into any test score, whether we want them to or not" (p. 569). Sarnacki (1979) writes, "Test wiseness is widely recognized as a source of additional variance in test scores and is a possible depressor of test validity" (p. 253).

According to Sarnacki (1979), there are three theoretical approaches to the analysis of TW. The first approach in TW research views TW as mostly the product of faulty test construction, "...being quite specific to the cues in poorly written items" (Sarnacki, 1979, p. 266). This interpretation is commonly credited to Millman et al. (1965). The second, less prevalent approach holds that TW is a cognitive trait that some individuals develop or possess, while some others do not, thereby partly explaining individual differences in test taking abilities. This interpretation was first proposed by Thorndike (1951) and has its share of believers. Sarnacki claimed that both approaches were insufficient to fully explain the construct of TW, and he himself proposed the third approach that synthesized both the characteristics of test takers and those of the test format and test-taking situation. The following sections are the discussion of these approaches.

Test wiseness as a result of test idiosyncrasies. Table 1 presents the definitions reflecting the first approach in TW conceptualization. Following the footsteps of Ebel

and Damrin (1960), who defined test-wiseness as the condition in which one's knowledge of the test design and format determines how one correctly responds to a test item, Millman et al. (1965) stressed that (a) TW was a test taker's ability to take advantage of multiple-choice test characteristics, (b) TW helped improve test performance beyond cognitive ability, and (c) TW did not logically relate to test takers' knowledge of the content area being tested. According to the authors, TW is mostly an individual characteristic resulting from the properties of the format of *objective* achievement and aptitude tests, excluding other types of tests (e.g., essay writing).

Note that this perspective may lead to the belief that test constructors are able to control for or at least mitigate the effects of TW on the validity of a predictor by developing a "good" test. For example, Thorndike, Cunningham, Thorndike, and Hagan (1991) give 11 recommendations for writing multiple-choice test items. Most of the recommendations specifically deal with the problem of test idiosyncrasies, such as using plausible distracters (incorrect alternatives), avoiding the use of "none of the above" or "all of the above" as an alternative, making each alternative of equal length and similar grammatical structure, paying attention to stem-alternative grammatical agreement, and randomizing the position of the correct alternative. Following these recommendations seems to effectively remove "TW" as it is defined as a test-dependent phenomenon (i.e., test takers could not take advantage of test idiosyncrasies to gain points any more).

Table 1

Definitions – TW as a Result of Test Idiosyncrasies

Author(s)	Definition
Ebel & Damrin (1960)	Test-wiseness is defined as the condition where "the subject's response is based upon his knowledge of the design and format of objective tests" (p. 1511).
Millman, Bishop & Ebel (1965)	"A subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score. Test-wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures" (p. 707).
Gibb (1964)	"The ability of a testee to react to the presence of secondary cues in ways advantageous to himself on a multiple-choice test of knowledge of factual information" (p. 5).
Erickson (1972)	"The ability to use test characteristics as an aid" (p. 142).
Oakland (1972)	"The ability to manifest test-taking skills which utilize the characteristics and format of a test and/or test-taking situation in order to receive a score commensurate with the abilities being measured" (p. 355).
Diamond & Evans (1972)	"The ability to respond advantageously to multiple choice items containing extraneous clues and to obtain credit on these items without knowledge of the subject matter" (p. 145).

However, Millman et al.'s (1965) perspective is narrow. The researchers consider TW more a situation-specific construct that relies on the low validity of a test (e.g., because of faulty test construction) than as an individual cognitive factor (e.g., a person's "capacity"). This perspective partially reflects a cynical, extreme viewpoint that high test-scorers who perform well on objective tests can be test wise but their knowledge is second rate and superficial (Hoffman, 1962).

Despite these weaknesses, the contribution of Millman et al.'s (1965) work is their important taxonomy of TW. They logically outlined what they called five test-taking principles grouped into two broad dimensions (Appendix A). Note that some of the principles if executed would involve some knowledge of the subject matter, which is contrary to the researchers' definition that TW is "independent of the examinee's knowledge of the subject matter." For example, in the category of deductive reasoning principle, the researchers list individual strategies such as elimination of "options which are *known to be incorrect* and choose from among the remaining options" (p. 716), or the restriction of "choice to those options which encompass all of two or more given statements *known to be correct*" (p. 717). In other words, a test-wise examinee needs to have some prior knowledge of the subject content or related information of the field to use the deductive reasoning strategies.

Gibb's classic study (1964) is the first empirical and most frequently cited evidence of the effect of TW, which was defined as the consequence of test idiosyncrasies. Gibb operationalized TW as the ability to detect secondary cues imbedded in multiple-choice test items. Some other investigators in the early 70's also defined TW

as the result of reactive behaviors of examinees with respect to test format and/or situation (e.g., Diamond & Evans, 1972; Erickson, 1972; Oakland, 1972; see Table 1).

Interpreting the construct of TW as the product of test idiosyncrasies has resulted in the development of instruments designed to measure how well test takers are able to detect cues embedded in a test (i.e., faulty test items), or superficial test-taking preparation programs which train examinees only on a few skills whose effectiveness is questionable as far as well constructed or standardized objective tests are concerned. Test takers' ability (or lack of it) to recognize cues and take advantage of them to obtain higher scores than one's ability permits is equated to being "test-wise" or "test-naïve." This approach in conceptualizing TW also creates the belief that improving test validity by constructing better multiple-choice tests would control for TW as the third variable explaining test score variance. In reality, although taking advantage of test idiosyncrasies to optimize one's test scores may be commonly considered a demonstration of TW, such behavior may not represent the full scope of the construct of TW.

Test wiseness as an individual trait. Thorndike (1951) brought up the concept of TW as a persistent and general characteristic of individuals who developed a systematic test-taking method to over-perform on a test (see Table 2). Although Thorndike distinguished general test-taking skills and techniques from general TW, later researchers interpret his comments more broadly, explaining TW in terms of cognitive abilities and dispositional traits of individual examinees (Pike, 1978; Spielberger & Vagg, 1995; Woodley, 1975). A few researchers also draw attention to the experience component in TW (e.g., English & English, 1970; Woodley, 1975). The contribution of this approach is the acknowledgement of individual differences in test-taking ability, which can be

proactively developed over time. In other words, the focus has been shifted from conceptualizing TW as a reaction to poorly constructed tests to defining it as a cognitive ability that may operate beyond one's taking advantage of test idiosyncrasies. This approach of TW conceptualization also opens the door to question the link between TW and general intelligence, which I will further discuss later.

Table 2

Definitions – TW as an Individual Trait

Author(s)	Definition
Thorndike (1951)	"Two rather special types of persisting general factors deserve some particular mention. These are the general ability to comprehend instructions and what we may speak of as 'test-wiseness.' These factors are mentioned because they are likely to enter into any test score, whether we want them to or not. That is, performance on many tests is likely to be in some measure a function of the individual's ability to understand what he is supposed to do on the test. Particularly as the test situation is novel or the instructions complex, this factor is likely to enter in. At the same time, test score is likely to be in some measure a function of the extent to which the individual is at home with tests and has a certain amount of sagacity with regards to tricks of taking them" (p. 568-569).
Woodley (1975)	TW is considered "as a cognitive factor... that is measurable and subject to change through either specific test experience or training in a test-taking strategy" (p. 7).
Pike (1978)	TW refers to "that set of skills and knowledge about test-taking itself that allows individuals to display their abilities... to their best advantage" (p. 6).
Spielberger & Vagg (1995)	TW is defined as "cognitive activities that affect the organization, processing, and retrieval of information [in test situations]" (p. 201).
English & English (1970)	TW refers to a student's being "experienced in taking tests; knowing how to increase one's scores by evasion of some of the standard requirements" (p. 550).

Test wiseness as a "synthesis." Between the aforementioned extreme viewpoints of TW, there are researchers who prefer an integrative theory of TW, referring to the construct as a mixture of learned and inherent abilities.

Green and Stewart (1984) take the mid-range position, considering TW as an "artifact" of one's general cognitive ability, which is the natural and inevitable result of the interaction of one's developed mental ability with one's experience via training (general training as well as training with test-taking skills; see Table 3). This stance implies that TW can be trained and improved, and the more intelligent a test taker is, the greater TW that person can acquire, controlling for training intensity. The researchers suggest that test performance is influenced by intellectual or cognitive abilities, and the comprehension and appropriate response to the tasks required by the test (i.e., test-taking skills). Recent TW investigators supported this view point (Flippo et al., 2000; Parham, 1996). In other words, TW is considered a composition of cognitive skills, general test-taking skills, as well as personal attributes. Therefore, TW may be an inherent trait but it can also be learned.

Table 3

Definitions – TW as a Synthesis

Author(s)	Definition
Green & Stewart (1984)	"Test-wiseness seems to be most accurately considered as an artifact of one's general cognitive ability... Test-wiseness is the natural and inevitable consequence of a highly developed cognitive, mental, and reasoning ability intertwined with one's general experience and training, and specific experience and training with tests and test-taking skills" (p. 422).
Parham (1996)	"Test-wiseness will be viewed in its broadest sense as a cognitive factor that may be employed on a variety of tests and that encompasses both the method of measurement and the characteristics of the test taker" (p. 6).
Flippo, Becker & Wark (2000)	TW is a broad collection of skills and traits that "in combination with content knowledge promotes optimal test performance" (p. 224).

Summary. The topic of how TW should be defined has been discussed and debated for decades. One school of thought views TW as the joint product of less valid tests (i.e., with faulty test items) and test takers' cleverness in detecting and taking advantage of such idiosyncrasies. Those who support this perspective suggest that developing better, more valid tests should rectify the problem and control for the extraneous influence of TW on test scores. Another school of thought considers TW a lasting individual characteristic that will always contribute to the variance in CA-T scores regardless of test quality. The contemporary perspective changes foci, defining TW in the context of general mental ability.

The Relationship Between Test Wiseness and General Mental Ability

The different viewpoints of TW, either as a situational influence (i.e., the characteristics of a test or a testing situation) or an individual cognitive trait, have led to a related question about the potential links between TW and general intelligence: Is TW related to or independent of cognitive ability or general mental ability?

Some researchers believe that TW is independent of general cognitive ability or at least independent of the factual knowledge of the subject matter being tested (Diamond & Evans, 1972; Millman et al., 1965). For example, some of the test-taking principles cited in the TW outline of Millman et al. seem to be straightforward and unrelated to any knowledge of mathematics, verbal ability or analytic ability whose test results are used to infer one's general mental ability (i.e., time-using strategies, guessing strategies, or cue-using strategies).

Other researchers (e.g., Pike, 1978; Scruggs & Lifson, 1985; Woodley, 1975) posit that there is a relationship between TW and general mental ability (*g*), given the fact that test-wise individuals performed consistently at a high level regardless of test formats (e.g., standardized tests, classroom essays). Furthermore, some researchers suggest that TW is a highly specific trait of general mental ability (Jacobs, 1975), having its own unique characteristics (Diamond & Evans, 1972). Bridgeman and Morgan (1996) supported this view of TW as a specific type of mental ability when they examined test scores of high-achieving high school students in an advanced placement (AP) program. Despite their equal general mental ability, students who generally scored higher on essay exams but lower on multiple-choice exams also scored higher on the essay portion than on the multiple-choice portion of an AP test, while those who generally scored higher on

multiple-choice tests and lower on essay tests performed significantly better on the multiple-choice portion of the AP test. The researchers concluded that the two test formats (essay tests and multiple-choice tests) tapped different constructs of intelligence; therefore, TW, defined as being multiple-choice test-related, tended to be a specific type of cognitive ability.

This fact leads to other questions about the covariance between an individual's TW level and cognitive ability as measured by a CA-T. Is there a strong, positive relationship between TW and cognitive ability? Are those who are highly intelligent highly test-wise? In other words, is it true that there are individual differences in TW because of observed individual differences in cognitive ability? If the answers to these questions are affirmative, an implication is that TW is part of the true variance of CA-T scores and thus it is not necessary to try control for TW as an extraneous determinant to CA-T scores. Another implication is that those at a lower level of general intelligence might never develop a sophisticated test-taking ability regardless of any TW training that they receive. Furthermore, controlling for cognitive ability in studies investigating the effects of TW on test taking would underestimate such effects.

In my opinion, TW, as a cognitive activity during test taking, should be linked to general mental ability, but there may be aspects of TW that are less *g*-loaded than other TW aspects. Furthermore, even though they may be related to one another, the magnitude of the link between the cognitive ability that a CA-T is purported to measure and the cognitive ability implied by TW manifestation is not necessarily large. For example, an intelligent person may be somewhat test naïve whereas another individual with lower

mental capacity may be more sophisticated in test taking. These points will be elaborated later when I propose a more integrated conceptual framework for the construct of TW.

Evidence of Test Wiseness Effects

Regarding what is the proportion of variance in CA-T scores that TW may explain, some researchers insist that it is an important, large source of variance (e.g., Kalechstein et al., 1981; Sarnacki, 1979; Wahlstrom & Boersma, 1968). A few authors argued that, though the existence of TW as a source of variance in test scores has been empirically supported, its effect on test scores is relatively small, not a substantial source of variance in test scores as some might assume. For example, Scruggs (1985) claimed that the notion of TW as a *substantial* source of measurement error was a myth stemming from "a confusion (among researchers on TW) between the terms 'statistically significant' and 'practically important'" (p. 343) because the portion of variance in test scores that TW actually accounts for was statistically significant but not large.

The empirical evidence for the effects of TW comes mostly from the findings in TW training literature. Past studies have provided some evidence that statistically significant improvement in test performance is associated with TW or training in TW (e.g., Dolly & Vick, 1986; Fagley, 1987; Gross, 1975; Rowley, 1974; Wahlstrom & Boersma, 1968). For example, Shuller (1979) assessed the effectiveness of Mini Tests, a comprehensive TW skills instructional program aimed at teaching elementary students strategies in test taking (e.g., time use, guessing strategies) to improve their performance on reading achievement tests. Data analysis regarding the use of Mini Tests in the New York City Public Schools in 1974-1976 indicated that the schools that used the program made more significant gains in the numbers of their students on grade level than the

schools not using the program. Rowley (1974) administered vocabulary and mathematics test items in both free response and multiple-choice items and measures of TW and risk-taking, partialing out free response scores. Controlling for content knowledge, the researcher found significant partial correlations between test taker vocabulary scores and TW (r = .27) and risk-taking (r = .14) measures. In addition, TW accounted for 7% of the variance in vocabulary test performance. However, TW and risk-taking were not correlated with mathematics multiple-choice items (partial r's approaching zero). In another investigation, Fagley (1987) examined the importance of TW, positional response bias (e.g., the tendency of test takers' choosing a response based on its physical or logical position among other responses such as in the middle), and guessing strategy, as predictors of performance on multiple-choice tests of learning. He found that these three factors combined explained a respectable 18% of the variance in test scores. These findings lend some empirical support for the existence of TW effects, at least with younger test takers in educational settings.

The question is whether TW training programs similarly influence adult test takers in employment settings. Unfortunately, there has not been much investigation done in this area. A few studies that did investigate (e.g., Holden, 1996; Quinn, 1993) failed to support the effectiveness of TW training. For example, Holden designed three training programs to improve test performance for applicants to entry-level clerical positions in a utility company. These intervention programs included either two hours of training of test-taking skills (i.e., cue-using strategies, deductive reasoning), two hours of basic mathematics training, or four hours of test-taking skills and basic mathematics (the "both" condition); there was also a control condition where no training was offered. The

cognitive ability test used in this study was a subset of the Basic Skills Test (Psychological Services, 1985), which participants took before and after their respective training programs. In terms of participants' characteristics, those in the treatment conditions (the training programs) had all failed a pre-employment test; the control group consisted of volunteers who had successfully passed the pre-employment selection test. The researcher found that the post-training test scores of those in all experimental conditions did not differ significantly from one another. Compared with the control group, the researcher found that taking a refresher course of math benefited participants the most (treatment participants' post-training math score was almost twice as much as that of those in the control group, 48% versus 25%, due to an improvement in the performance on the arithmetic subtest). Holden suspected that the result regarding the lack of TW effects was attributable to the fact that participants in the test-taking skill condition had used the strategies incorrectly or been too distracted by the application of these skills to perform effectively on the test (i.e., participants not having enough time to practice). Although these explanations were reasonable, the researcher did not offer any evidence (i.e., follow-up interviews) to support her beliefs. In my opinion, there were two other explanations that were related to the characteristics of the TW program offered in this study. First, participants were trained on test-taking skills that might not be particularly useful for or applicable to the test they took; in other words, the TW training program was too superficial to produce any observable effects on test scores. Second, the TW training program duration might be too short for participants to acquire the skills.

Pike (1978) distinguished three different approaches to increasing TW in the research literature: (a) short-term instruction (providing TW instructions in a short period

of time); (b) coaching (providing very brief instruction on general TW and practicing answering questions similar to those appearing on a test), and (c) intermediate-term instruction. We can argue that short- and intermediate-term instructions imply time constraints in efforts to improve test scores by providing special instructions. Meanwhile, coaching has more to do with the nature of the instruction than with the duration of instructions. According to Anderson and Sauser (1995), experimental evidence indicates that both short- and intermediate-term approaches can be effective in increasing students' TW levels. However, Samson (1985) conducted a meta-analytic review of 24 studies investigating the effects of training programs designed to improve test-taking skills on elementary or secondary academic achievement. The researcher found that longer training programs tended to produce greater score gains than crash courses. Most studies in this sample provided basic instructions in general test-taking skills (e.g., following directions, proper use of time, instructions in the use of answer sheets and in checking answers). Several studies included more sophisticated skills such as deductive reasoning strategies and secondary-cue strategies. Samson found an overall mean effect size of .33, indicating that an average student in the TTS training group scored at the 63rd percentile on an achievement measure relative to the 50th percentile for an average control-group student. This effect size suggested that training in TTS had a moderate and significant effect on academic achievement test performance, supporting the need for TTS training programs for young students. To support the advantage of longer periods of training time, the researcher found that training programs lasting from five to seven weeks yielded a much greater mean effect size (.56) than those of one week or less (.22). Programs lasting three to six weeks yielded a significant mean effect size of .37. Note that the training

programs studied were specifically designed for children. It is unclear whether similar results would be observed in training programs targeting adult test takers in the employment testing setting. However, it is reasonable to hypothesize that a well-designed TW training program would improve adult test takers' test performance, particularly for those who were diagnosed as having a low level of test-taking sophistication.

Summary

The construct of TW is considered a possible determinant of CA-T scores in addition to true content knowledge being measured. There is also a speculation that TW might partially explain the observed mean score differences by ethnicity in employment selection tests. Therefore, the construct is worth being investigated. A review of the definitions of the construct showed some discrepancies in how TW should be conceptualized: from mainly depending on the characteristics of tests or testing situation, to mainly depending on test takers' cognitive characteristics. Possible links between TW and general mental ability were also reviewed and discussed. Regarding the effects of TW, empirical evidence from TW training literature seems to support the positive influence of TW on test scores, at least among young students. The evidence of TW training effects for adult test takers in selection situations is still scarce and inconclusive.

In the coming sections, I will propose a new framework for the construct of TW, building on the existing literature of TW but expanding to include the psychological mechanisms underlying TW as a process. Please note that, although individual TW may transcend various test formats (i.e., multiple-choice tests, essays tests or oral tests) for different assessment goals (i.e., ability, achievement, non-cognitive assessment), the investigation in this study is restricted to TW aspects that are related to the paper-and-

pencil multiple-choice test format, which is the most common format of standardized CA-Ts.

Reconceptualization of Test Wiseness

The Need for a New Theory

The most recent definition of TW involves the conceptualization of the construct as a test taker's lasting and general characteristic. For example, in their review of test preparation and test taking skills, Flippo et al. (2000) describe TW as a broad collection of individual skills and traits that "in combination with content knowledge promotes optimal test performance" (p. 224). Although the strength of this concept is the focus on individual ability to handle multiple-choice tests, it is insufficient to help understand the construct because the concept is descriptive, providing no further information about the psychological mechanisms underlying the construct.

To better understand TW, one should shift perspectives and examine TW as a dynamic mental process in test taking. From this perspective, TW is not only about knowing a set of test-taking tips and tricks and (mechanically) applying them to multiple-choice tests or test items; TW is also about how the knowledge of test-taking techniques is developed, and about the insight which guides a test taker in using such knowledge to his or her advantage when and where it is necessary to do so.

In the following sections, I will first present the process of how a hypothetical test taker develops her TW and uses her test-taking skills in a testing situation. Second, I will propose a new definition and a theoretical model of TW construct. Third, I will discuss the role of test-taking metacognition as an integrated component of TW, which has been examined separately or in conjunction with TW in the literature but never incorporated

with the construct of TW. Fourth, the interpretations and implications of the proposed TW theory will be discussed.

Identifying a Test-Wise Individual

Traditional views of TW have focused more on what the construct is, but less on how it is developed and how it operates. From analyzing how an individual goes through the process of becoming "test-wise," utilizing both content knowledge and transcending knowledge of strategic behaviors to their best advantage in a test-taking situation, we can help clarify the psychological mechanisms underlying the construct of TW.

Let us consider the example of Song Le, an international student who used to be unfamiliar with objective, multiple-choice tests. When preparing for a college-placement objective test, Song knew that she would be tested on verbal, quantitative and analytical skills. Therefore, she began to review related course materials on the subject matters to be tested, learning the right terminology for each subtest, and reinforcing her content knowledge.

Concurrently, Song learned what a sample multiple-choice test item looked like (e.g., including a stem and alternatives), how to respond to it (e.g., darkening a letter corresponding to the correct choice on an answer sheet), or how to correct a clerical mistake (e.g., erasing her mark completely). She learned to pay attention to sample test instructions, and commits the information to her memory. In other words, Song had gradually acquired general test-taking techniques that could transcend all multiple-choice tests.

More important, Song made an effort to learn "tips" that pertained to the test she was going to take. She found these tips or tactics particularly useful when she could not

directly respond to a "difficult" test item based on her content knowledge alone. For example, while pondering a question about which she was uncertain, Song evaluated her acquired test-taking skills in light of the test item characteristics and made a decision about what strategy or combination of strategies she would need to solve the problem at hand (e.g., saving time by marking a more difficult item and skipping it to easier ones, with the intention of working on it later), and then executed the strategy(ies). Moreover, during the test, Song might occasionally take a mental pause to evaluate her own performance thus far or monitor her test-taking process (e.g., asking herself task-related questions such as, "Am I doing well on this test?" "I am confused about the stem; should I need to reread the alternatives to see if I can understand the problem better?" or "Should I eliminate some answers first before taking a guess at this question?").

Gradually, Song became more at ease with a multiple-choice test or test items. For example, she was aware that some test-taking techniques, such as time allocation and changing answers, appear so general that they could be broadly and uniformly applied to most multiple-choice tests. Some other strategies were, however, very specific to a subtest item (e.g., spotting the keyword in a reading comprehension question and scanning the passage for that keyword or its synonym; comparing, not calculating, quantitative comparison questions). She also noticed that several cue-recognition strategies she picked up in some test-preparation materials should not be applied blindly to every test (e.g., she could find some grammatical cues on a classroom exam but not on a standardized test), some strategies required more thinking than other strategies, or some strategies could even be detrimental to your test performance and should be avoided. In

short, Song has become an insightful person as far as taking a multiple-choice test is concerned.

Definition of Test Wiseness

Taking an objective test is considered a multiple-hurdle process. Knowing how to take such a test is the first hurdle that an examinee must clear before she can tackle the subject content being tested. A test taker is at a distinct disadvantage compared with other examinees if she lacks a general knowledge of the characteristics and format of a multiple-choice test and test items to begin with, as well as lacking an understanding of the testing situation, the mental process involved in test taking, or the importance of strategizing in taking an objective test. The reason is that the "test naïve" person may not be capable of displaying her abilities effectively, nor optimizing her test performance by incorporating learned strategies with her knowledge of the subjects being tested. Therefore, a basic level of TW is conceptualized as an essential prerequisite to successful test taking, a construct that involves three integrated psychological components: cognitive, behavioral, and metacognitive elements (see Figure 1 for a diagram of the TW construct).

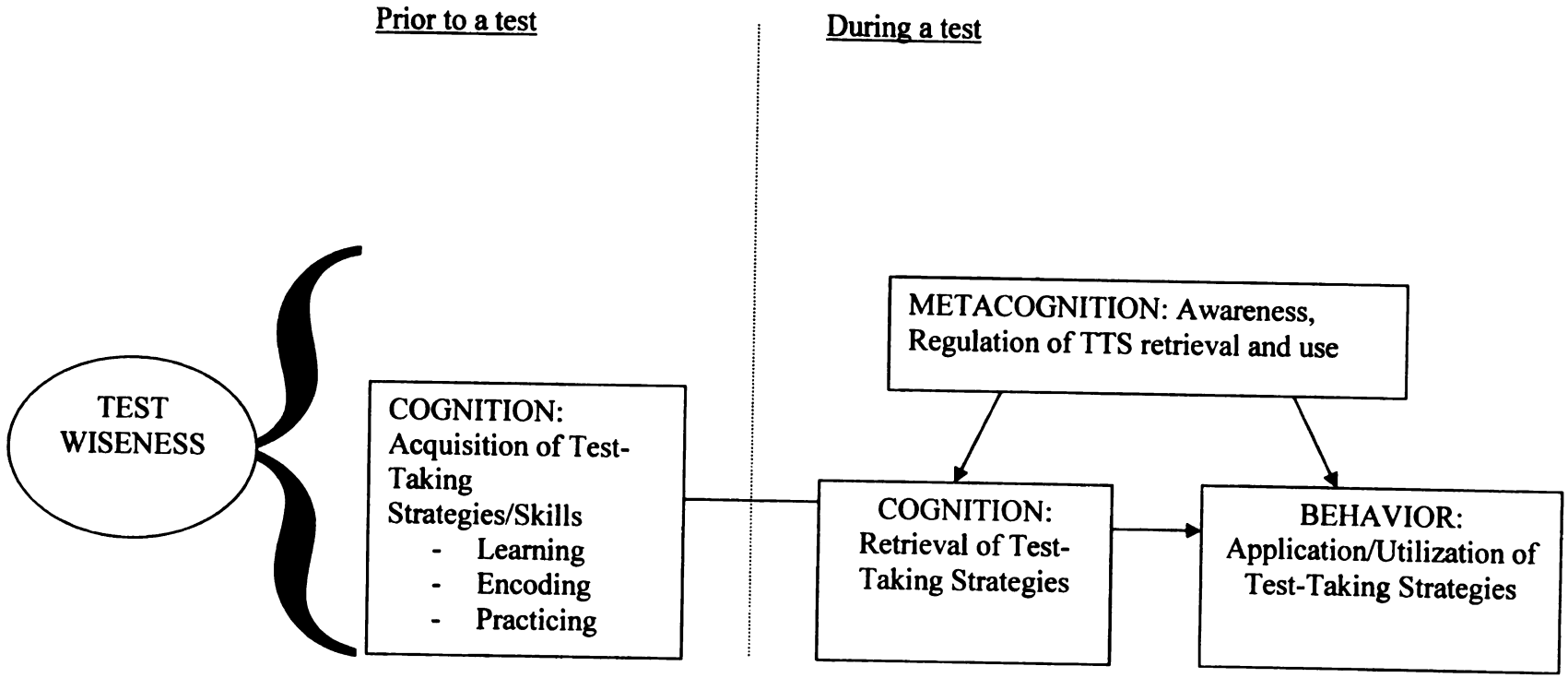

Figure 1. A conceptual diagram of test wiseness

The cognition element of TW refers to the process in which individuals *acquire* the knowledge of strategies and principles in taking multiple-choice tests, *encoding* them in one's memory, and *retrieving* them for use with a test or a test item in the absence or inadequacy of direct content knowledge. The cognitive process of TW begins long before one takes a test and continues to operate throughout the test. A detailed discussion of the strategy acquisition stage will be further presented in a later section.

The behavioral element of TW refers to the actual *application* or *utilization* of test-taking skills learned where the need for such skills arises in a particular test-taking situation. Note that TTS retrieval is necessary for TTS application but the two processes are distinct from each other (i.e., a test taker may be able to retrieve a strategy from her recollection of TTS but for some reason has decided not to use it on the test at hand). Again, I will discuss more about the utilization of TTS when I present the research plan for developing a measure of TTS.

The application of certain TTS can be automatic and subconscious (e.g., with sufficient practices or experience, or when the strategy is simple). Applying TTS can also be a more structured and conscious process, first requiring a test taker to generalize strategies they have acquired and then make decisions about what type of strategy is suitable for what type of test or test items. This decision-making phase is under the influence of a high-order construct called *metacognition*, which refers to an individual's overall *awareness*, *regulation* and *evaluation* of the proper utilization of TTS in a particular test-taking situation.

In other words, TW is a multidimensional psychological construct whose components represent mental processes that occur both prior to and concurrently with

test-taking mental activities. Logically, the cognitive process precedes the behavioral component, and the metacognition emerges during the test-taking process as an umbrella construct, regulating the processes of retrieval and applying TTS where necessary. The particular role of a metacognitive component in this proposed model is further discussed in the next section.

The Role of Test-Taking Metacognition in Test Wiseness

In the above example, Song Le possesses a test-taking capability that, for some reason, has not received its deserved attention in the TW research literature: her heightened awareness and monitoring of the test-taking situation, of the test itself, and of her cognitive ability and test-taking ability. In other words, besides her subject-matter knowledge and knowledge of test-taking skills, our fictitious test taker seems capable of monitoring her test-taking cognitive processes and behaviors throughout the test by taking occasional mental pauses and evaluating her progress as well as the application of strategies.

Indirect evidence for the existence of such a state of awareness and its relationship with positive test performance can be found in several studies related to TW testing. For example, Huck (1978) hypothesized that the enhanced awareness about the level of difficulty of certain test items would cause test takers to pay more attention to those items (e.g., reading them more carefully). The findings supported the researcher's hypothesis: knowing in advance the difficulty of an item significantly increased test scores. However, Huck did not discuss how this awareness functioned in bringing about the observed effect.

The importance of self-reflection or general test-taking awareness is emphasized in Kirkland and Hollandsworth's (1980) study. When investigating different designs of test anxiety-reduction programs, the researchers found that the most effective program is the one that focused on deficits in three areas: effective TTS, adaptive self-instructional statements, and attention-control skills, which are self-instructional statements such as "I will think about that later; now back to the test." Besides improving test takers' performance on anagram problems under stress-inducing testing conditions, the test takers also reported fewer off-task thoughts (e.g., thinking less frequently about their level of ability, less often about how much time was left), which were associated with test anxiety, and reported more concentration on each test item.

Studying effective, high performing test takers, Hollandsworth, Glazeski, Kirkland, Jones, and van Norman (1979) identify two basic kinds of self-instructions that these test takers often used: on-task statements (e.g., "I have plenty of time; read the questions carefully") and positive self-evaluations (e.g., "I will perform well on this test because I am well prepared").

Schraw (1997) calls this phenomenon of giving oneself instructions during a test "test-taking metacognition," and suggests that there is a positive relation between high performance and high level of test-taking monitoring. He writes, "Effective test taking depends on two important skills: selecting correct responses to test questions and monitoring one's performance accurately" (p. 135). Selecting correct answers depends on a test taker's access to relevant content knowledge (i.e., knowledge of mathematical formulas or vocabulary) and access to test-taking skills, as previously discussed. Accurately monitoring one's performance is the function of general metacognitive

knowledge (Pressley, Borkowski, & Schneider, 1987; Stock, Kulhavy, Pridemore, & Krug, 1992; Webb, Stock, & McCarthy, 1994). According to test-taking metacognition theories, test takers evaluate their performance using domain-independent metacognitive processes (e.g., checking their comprehension of test questions, comparing given test information to information in memory, allocating their resources effectively, and rereading when they fail to fully understand a question). This ability is assumed to transfer across knowledge domains, and thus, qualifying as a lasting individual characteristic that may contribute to CA-T scores.

Is it possible that test takers with metacognitive skills are able not only to monitor how well they perform on the test but also to evaluate how appropriately certain test-taking skills are applied to supplement or substitute for subject-matter knowledge? Logically, the answer is "yes." Rogers and Bateson (1991) proposed a cognitive model of test-taking behavior of test-wise test takers in which they identified what they called the "cognitive monitor for testing" or an "executive governor" for selecting the correct answer for a multiple-choice test item. According to this model, the process of responding to a test item follows a defined path. First, a test taker uses her content knowledge to attempt to recognize the correct alternative. If the content knowledge is insufficient to find the answer, an unskilled test taker will either randomly guess or omit the question, while a skilled test taker—who has developed a cognitive monitoring system for testing, as well as having partial knowledge about the content being measured—will apply her TW principles to go through the elements of the set of stem and choices in a cyclical manner for a test-wiseness element-item cue match. When a match is made, a test-wise response (as opposed to pure knowledge) is recorded. If not, a

test-wise test taker makes a random response and moves on to the next item. In short, the characteristic that distinguishes an unskilled (test-naïve) test taker and a skilled (test-wise) one is the possession of a cognitive monitor for testing, according to Rogers and Bateson's model.

In order to test their model, the researchers developed a 2-section Test of Test-Wiseness (TTW) based on the outline of Millman et al. (1965) and the most frequently occurring item faults in British Columbia's provincial high school examinations administered in a previous year. The first section of the measure was composed of 24 items selected from earlier TW tests (e.g., Gibb, 1964; Millman, 1966; Slakter et al., 1970), tapping three deductive reasoning strategies (eliminating options known to be incorrect; choose neither or both of two options which imply the correctness of each other; choosing neither or one of two options, one of which, if correct, would imply the incorrectness of the other) and one cue-recognition strategy (recognizing and using similarities between the stem and the options). The second section of the TTW involved the use of guessing whenever the chance of score gain was positive (e.g., making an educated guess after eliminating more than one option). The scoring system for the TTW was designed so that a high score reflected frequent use of TW reasoning and appropriate use of guessing strategy. From a pretest with a sample of 936 high school students, Rogers and Bateson identified test-wise students and test-naïve ones, a sub-sample of who were later interviewed to ascertain their response strategies (the cut-off scores were one standard deviation above and below the mean on Section 1 of the TTW).

The researchers found only two strategies (eliminating incorrect options; choosing neither or both of two options) that significantly distinguished between test-wise and test-

naïve students in terms of frequency of use. Test-wise participants also tended to use other strategies more frequently, though not significantly, than did the test-naïve. The researchers suggested that (1) there was a relationship between test takers' cognitive ability and the use of TW strategies (i.e., the classified TW group was rated by their high-school principals as more academically talented, enrolling in a greater number of academic courses and being more likely to enter university than the test-naïve sample), (2) partial knowledge of test items combined with knowledge of the TW principles would determine the level of performance (e.g., TW students with low content knowledge and test-naïve students with low TW knowledge would perform less well than students high on both accounts) and (3) the observed difference in test-taking approach appeared to be explained by students' differences in cognitive monitoring (e.g., TW students were better able to take deliberate follow-up actions in the presence of partial knowledge). Therefore, Rogers and Bateson concluded that TW seems to be the function of "partial knowledge about what an item measures and ability to take advantage of item cues" (p. 210).

This study has several major limitations in terms of research design. While they integrated a metacognitive factor (which they called cognitive monitor of test taking) in a theoretical model of test takers' response to multiple-choice test items, Rogers and Bateson made the same mistake of building their TW instrument and test-taking model around a few TTS as had some earlier researchers, therefore limiting the explanatory power of TW in their model. For example, they measured only four strategies that can be directly applied to test item responding, excluding other test-taking skills, such as time allocation skills, which can be applied regardless of the presence or absence of test knowledge. Further, they did not quantitatively assess the metacognitive factor in their

model (e.g., their conclusion regarding the role of a cognitive monitor in TW was based on the qualitative description of a small number of post-hoc interviews). Nevertheless, the researchers were insightful in making their theoretical assumptions.

Harmon (1997) was the first to empirically investigate the direct relationship of test-taking metacognition with TW. She hypothesized that test-taking metacognition (which she termed "metastrategic knowledge") would play an important role in the construct of TW. The metastrategic knowledge was defined as test takers' perceptions and knowledge of their actual strategies of TW, and was measured by two constructed instruments based on the taxonomy of TW by Millman et al. First, Harmon assessed students' perceptions of how they perform in test situations with the constructed 70-item Report Strategy Instrument (RSI). This 5-point Likert-scale measure was designed to assess metastrategic knowledge in terms of identifying the participant's preferred strategy use under particular test situations, based on the assumption that a participant's beliefs about effective test strategies are represented in their report of their preferred strategy use (e.g., what they reported they would do in a test-taking situation). The RSI consisted of Millman et al.'s six categories with low internal reliability: (1) time-use (coefficient alpha = .34), (2) error avoidance (.53), (3) guessing, (.26), (4) deductive reasoning, .63, (5) intent of the test constructor or purpose of the test (.60), and (6) cue use (.58).

Secondly, test takers responded to another designed metastrategic measure, the Best Strategy Instrument (BSI), assessing test takers' awareness regarding test-taking process both in general monitoring strategies and the awareness and use of efficient strategies for arriving at correct solutions under particular test-taking conditions. The BSI consisted of two parts: a set of 15 multiple-choice questions which set up test conditions

under which participants had to select the best strategy option based on their knowledge of TTS; and another set of 23 multiple-choice questions for which the correct response must be determined by using TW strategies. In addition, test takers were required to provide a written explanation as to how they arrived at the solution for the given condition. Similar to the RSI, the BSI had low sub-scale coefficient alphas (ranging from .22 to .64). Participants' TW level was assessed by the Gibb Experimental Test of Testwiseness (GETTW; Gibb, 1964). Participants in this study were 220 college undergraduate students.

Controlling for participants' general intelligence (as assessed by the intelligence scores from the Shipley Institute of Living Scale; Shipley, 1986), Harmon found that metastrategic knowledge regarding students' perceptions of their test behaviors/strategies (as assessed by the RSI) explained a significant 6% of the variance in the test-wiseness score. Also, metastrategic knowledge in terms of students' understanding and actual use of the six test-taking skills (as assessed by the BSI) significantly explained 13% of the variance in TW score. The researcher concluded that metacognition of test-taking process was an integral part of TW.

The use of GETTW, which focuses on cue-use strategies, limited the scope of TW in this study. General intelligence of test takers as assessed by a cognitive ability test (the SILS) was controlled, but it is possible that the SILS scores were partly influenced by test takers' TW because its vocabulary subtest was in a multiple-choice format. That means TW was underestimated in this study. More importantly, reliability levels of both metastrategic instruments were below the desirable level and their validity levels have yet to be established. Nevertheless, Harmon's theoretical framework is logically acceptable

and her results provide some initial support for a link between test-taking metacognition and selected TTS.

Interpretations and Implications of the Proposed Test Wiseness Theory

Several interpretations and implications of the aforementioned definition of TW are of particular interest.

First, TW is a test taker's acquired ability about knowing how to take a multiple-choice test in general. An individual who is handicapped in this area would be systematically penalized for his or her deficiency, increasing the level of measurement error. A proactive solution for this problem is to train test takers actively in terms of test-taking strategic proficiency, thus making the test more valid.

Second, TW is a source of additional variance in multiple-choice test scores that test constructors are only able to control to a certain extent: for example, test constructors can improve the validity of a test by eliminating poorly-constructed test items to render cue-using strategies futile. However, test-wise individuals may be able to apply other appropriate strategies to a multiple-choice test. In other words, a test-wise examinee may perform beyond her cognitive ability being tested in the presence of test idiosyncrasies as well as in the absence of faulty test items.

Third, knowledge of TTS is necessary but insufficient to define TW operationally. One should take into account the high-order mental activities or metacognition during a test that governs the use of certain skills or strategies to complement or substitute for pure content knowledge.

Fourth, the proposed theory of TW goes beyond the traditional view of the concept in that it emphasizes the active role of test takers in the process of learning and strategizing prior to and during a test.

Last, it appears that whether or not the construct of TW as newly conceptualized is measurable depends on whether we have valid measures of the dimensions of this construct (cognition, behavior, and metacognition).

In order to test this theory of TW, it is essential to develop reliable and valid instruments assessing the 3 psychological elements of the TW construct. Because the acquisition of TTS is a fundamental cognitive step in the model, there is a need for a good measure of test takers' knowledge of TTS, which, unfortunately, is still lacking. Therefore, the second goal of this study is to develop and evaluate a diagnostic test of knowledge of TTS for use in research and applied settings. Before doing so, I will review what we have learned about TTS from the literature, the relationship with TW, a few common TTS measures, and some empirical evidence of the effectiveness of such strategies in test taking.

Test-Taking Strategies

Definition of Test-Taking Strategies

The most common definition of test-taking strategies or skills (TTS; the terms "strategies" and "skills" will be used interchangeably in this study) is that these skills are components of TW that are more reasonably studied than the construct itself (e.g., Green & Stewart, 1984). Pike (1978) refers to TTS as "that set of skills and knowledge about test-taking itself that allows individuals to display their abilities... to their best advantage" (p. 6). Sarnacki (1979) agrees that TTS are the operational definition of TW,

including a set of skills that a test taker can use to improve his or her test score regardless of the content area of a test.

Some researchers specifically define a test-taking strategy as any tactic, rule, or procedure used for interpretation and solution of multiple-choice test questions only (e.g., Bruch, 1981a, 1981b). On the other hand, other researchers argue that some TTS are general enough to enable a test-wise person to perform well on forms of assessment other than multiple-choice tests, such as essay writing, and true-false tests (e.g., Pike, 1978; Scruggs & Mastropieri, 1992; Woodley, 1975). My preliminary review of TTS in recent commercial test preparation materials indicates that there are specific skills associated with taking a specific subtest, such as verbal reasoning (i.e., "Break an unfamiliar word down into components (prefixes, suffixes, roots) to find a clue to its meaning;" Brownstein, Weiner, Green, & Hilbert, 1999), quantitative reasoning (i.e., " Whenever possible, don't calculate: just compare;" Brownstein et al.) or analytical reasoning ("Classify and connect things and events to analyze the arguments and draw conclusions;" Jaffe & Hilbert, 1998). There are also general test-taking techniques that are applicable to most multiple-choice tests (i.e., "When marking (filling in) an answer on the answer sheet, do it completely;" Bobrow, 1999).

For the scope of this study, I will limit my investigation of TTS to studying the strategies that are essential for taking a paper-and-pencil, multiple-choice cognitive ability test. The strategies can either be general, applicable to all subsections of a test, or specific to a type of subtest. Furthermore, some strategies may be more suitable to one type of test than to another (e.g., speeded versus power tests; classroom examinations versus standardized tests).

Taxonomy of Test-Taking Principles or Strategies

In my theory, the application of TTS to taking an objective test is considered the behavioral component of TW. In research reality, the boundary between TW and TTS has often been blurred. Many researchers have continuously equated TTS with TW (e.g., Gibb, 1964; Parrish, 1982; Dolly & Vick, 1986) and proceeded to investigate a set or subset of TTS as the operational level of TW, inferring the results as the manifestation (or lack) of TW.

This tendency is reflected in Millman et al.'s (1965) "An Outline of Principles of Test-wiseness" (see Appendix A). The researchers compiled and categorized a list of general test-taking principles from studies examining the rationalization of students' choices in taking tests, and the literature of principles of test construction and taking examinations. The taxonomy of TW that they developed, which is a profound contribution to the TW research body and serves as a theoretical framework for numerous subsequent studies, is in fact *a categorization of test-taking skills* that successful examinees seem to embrace. Millman et al. grouped these skills or strategies into two major groups in terms of their relation with test characteristics. The group of categories which is "independent of test constructor or test purpose" (p. 711) includes (1) Time-using strategy, (2) Error-avoidance strategy, (3) Guessing strategy, and (4) Deductive-reasoning strategy. The TTS that stem from a test taker's capability to detect idiosyncrasies in test items or test purpose are (5) Intent consideration strategy and (6) Cue-using strategy.

Millman et al. (1965) also provided some empirical evidence and examples for a few strategies on their list. For example, for the skill of "Set up a schedule to progress

through the test," they cited Cook's findings (1957) that showed a periodic check on progress rate helps test takers maintain proper speed in taking timed tests. Subsequent researchers have added or reduced the number of Millman et al.'s categories. For example, Parham (1996) expanded the TTS and grouped them into 15 factor categories based on perceived effectiveness of the strategies; Doverspike and Miguel-Feruito (2001) reduced Millman et al.'s categories into 4 groups, (1) Time use, (2) Guessing, (3) Error avoidance, and (4) Elimination alternatives.

The taxonomy of TTS that Millman et al. (1965) came up with is logically sound. Several of its principles have been used by subsequent researchers to develop instruments to measure test-taking strategies. For example, Ferrell's Form Z, assessing deductive reasoning strategies and cue-using strategies, had an internal consistency coefficient of .84 (Borrello & Thompson, 1985); a test-taking strategy measure adapted from Millman et al.'s framework yielded an overall reliability coefficient of .81 (Nguyen et al., 2003). Some of the most popular measures of TTS based on the work of Millman et al.'s will be discussed in the next section.

Measuring Test Wiseness or Test-Taking Strategies

A common practice in TW measurement is developing a fiction-based multiple-choice test whose problems examinees theoretically cannot solve based on prior knowledge (i.e., all the information is made up by the test developers) unless they know how to use a few strategies effectively. These measures usually provide only an assessment of a set of TTS, being generalized as the assessment of TW. In other words, although their authors would call them measures of TW, they are actually measures of a specific set of TTS.

Gibb (1964) is the first investigator who examined a subset of TTS that coincided with the principle of "Cue-using strategy" in Millman et al.'s taxonomy (1965). Gibb trained his participants to detect seven types of secondary cues (e.g., longer correct alternatives, more precise language, grammatical clues). He then tested test takers' ability to identify and respond to the presence of these secondary cues in a constructed 70-item multiple-choice test called the Experiment Test of Testwiseness (ETT). The test items appear to be difficult history questions but can be answered only by using the non-content cues imbedded in the questions and answers. Gibb did not report what cutoff scores should be used to represent the high and low levels of TW, although Gibb found a mean TW score of 29.67 for 92 untrained college students (i.e., having not received a training on his TW skills). Later, using the same instrument, Harmon, Morse, and Morse (1996) obtained a mean score of 24.08 from a sample of 173 untrained students. Nevertheless, Gibb interpreted a high score on the ETT as a manifestation of TW in light of poorly constructed items.

Such an interpretation seems theoretically problematic to contemporary researchers in two ways. First, because he theorized that TW was dependent on faulty test characteristics, the strategies Gibb chose to train test takers and the measure were incomprehensive and non-representative as far as TTS are concerned. The best inference we can draw from such findings is that Gibb's trained participants were skillful in taking advantage of faulty test items, but they did not necessarily master other TTS (at least those on Millman et al.'s taxonomy, 1965). Secondly, it was considered a waste of effort, or even an unethical practice by some researchers (e.g., Scruggs & Mastropieri, 1992) to

extensively train test takers in strategies whose effectiveness might not extend beyond classroom examinations or would cease to exist once a test is rid of faulty test items.

Note that Gibb's measure, which is founded on a rather limited conceptualization of TW, has received considerable research attention. There is empirical support for the reliability (Fagley, Miller, & Downing, 1990; Miller, Fagley, & Lane, 1988) and validity of the Experimental Test of Testwiseness (Fagley et al.; Harmon et al., 1996; Miller, Fagley, Dowing, Jones, Campbell, & Knox-Harbour, 1989).

Nguyen et al. (2003) adapted Millman et al.'s framework into a 29-item measure, the Test-Taking Strategies Survey (TTSS). These strategies were classified into 6 original categories as suggested by Millman et al. However, subsequent exploratory factor analyses and item analyses revealed that a revised 6-factor, 25-item measure fit the data better. The revised 6 categories of TTS include "Using logical or physical cues," "Using grammatical cues," "Treating test items with caution," "Deductive reasoning," "Time management" and "Guessing." The category of "Using logical or physical cues" that had a satisfactory reliability coefficient (Cronbach's alpha was .87); other subscales had acceptable internal consistency levels (between .66 and .78). An overall scale reliability coefficient of .81 was obtained. The limitation of this measure is that the sample size used for item analysis did not meet the desirable standard of 1:10 ratio. In addition, the subscale of Guessing contained only one item. Analyzing the data collected, I found that the "Guessing" category was positively and significantly correlated with cognitive ability performance on both the verbal subtest ($r = .204, p < .01$) and the quantitative subtest ($r = .195, p < .05$); the category of "Using physical cues" was

negatively and significantly correlated with performance on a verbal subtest ($r = -.231$, $p < .01$) and a quantitative subtest ($r = -.16$, $p < .05$).

Other measures were developed to assess TW for use in training (e.g., Bajtelsmit, 1975; Dolly & Vick, 1986; Ferrell, 1972; Crehan, Gross, Koehler, & Slakter, 1978; Parrish, 1982). For example, a popular measure of TW is Ferrell's (1972) Test of Test-Wiseness (Form Z). This 46-item instrument includes multiple-choice and true-false questions taken from Gibb's and other sources. It is intended to assess only two particular categories of TTS: deductive reasoning and cue-use strategies. Its internal consistency coefficient ranged from .66 to .84 in subsequent studies. Using five samples of high school students to test this instrument, Ferrell found that the measured construct which he believed was TW accounted for between 16% and 28% of the variance in students' classroom examination scores after controlling for general mental ability. One weakness of this study lies in the fact that the researcher controlled for general mental ability, which was also measured with a multiple-choice test; this fact might result in the underestimation of TW effect size. The other limitation of this study is that there were no standardized operational definitions of participant achievement and general mental ability in this study (i.e., using *different* classroom tests and *different* standardized general mental ability tests across samples).

Dolly and Vick (1986) aimed at identifying individuals in need of some TW training. They developed two 25-item multiple-choice tests that would be administered before and after a one-hour training workshop on TW. Each test consisted of a set of 10 items that the researchers believed to be susceptible to four types of TTS (which the authors called "cognitive strategy category") on various subject matters, and another set

of 15 well-constructed items that was assumed to be non-susceptible. The skills taught in the workshop were limited, including three cue-recognition skills (the correct choice is one of the two middle options; the correct choice is the longest option; the correct choice is one of two items which imply the oppositeness of each other) and a deductive reasoning skill (the correct choice contains a repetition of words or ideas appearing in the item stem). The investigators found a significant improvement in test performance due to training, although higher pre-training test scorers did not gain as much on the post-training tests as those who had the lower pre-training test scores. Grade point average also significantly related to the amount of gain: low-GPA test takers gained more points on the post-test than those reporting a high GPA. The reliability and validity of the TW pretest and posttest are not reported in this study. Another limitation of this measure is that it focuses on a few test-taking skills that exploit faulty test items and, therefore, has little validity in terms of measuring TW.

Some TW measures have been developed for the restricted use of a population of test-takers. For example, the Measure of Obscure Medical Information was developed to test health sciences students' TW levels (Harvill, 1984a). Using obscure, factual medical information, the researcher wrote two versions of a 24 multiple-choice test items to test several TW strategies (similar options, umbrella term, item give-away, convergence principle, length of correct option, and stem-option cue). There was some evidence of internal consistency and alternate form reliability; however, the norm sample size was very small (54 first-year medical students).

The general limitation of these measures is that they assess selected TTS, not TW per se. The implication is that, where some TTS were assumed to be the operational level

of TW, the presence (or absence) of TW in several aforementioned reports of TW investigations may be alternatively explained by the incomprehensiveness or lack of representativeness of the measure used. In other words, while this type of instrument may be used to assess examinees' mastery level of selective TTS suitable for use in classroom objective examinations, by no means should they be utilized to evaluate whether a test taker is test-wise or test-naïve in its broad (and correct) sense. This fact, in turn, calls for a more valid and comprehensive measure of TW.

Parham's (1996) study is an exception in terms of the comprehensive inclusion of TTS. Compared with that of earlier investigators', her research design is markedly improved in several aspects. First, she specifically focused on developing a valid and reliable measure of TTS used for multiple-choice tests, *not* a measure of TW, and subsequently drawing proper inferences from her results. Secondly, recognizing the narrow range of TTS that had been examined in the past, Parham searched for and compiled a fairly comprehensive list of 225 strategies for taking tests from published TW and coaching research literatures (from 1951 to mid-1990's), as well as from five focus group sessions with 28 undergraduate participants. Working from this list, the investigator developed a checklist-type measure of 78 TTS, which was supposed to differentiate test takers who engage in effective test-taking behaviors more frequently than those who do not. Initially, Parham categorized these 78 TTS into six categories, five of which were based on Millman et al.'s (1965) research: Time-use, Error avoidance, Deductive Reasoning, Guessing, and Cue-use strategies. She omitted Millman et al.'s category of Intent consideration strategies but added a 1-item category of Cheating.

One hundred twenty nine college students were asked to attend two experimental sessions (with an interval of 2 weeks). (Parham's sample size for Session 1 was larger but I only quoted those who attended both sessions.) In both sessions, besides responding to a battery of measures, participants worked on a standardized cognitive ability test (the Wonderlic Scholastic Level Exam) and then were asked to rate 78 TTS on two separate scales pertaining to the test that they had completed: (a) frequency of strategy use (i.e., 1 = I never use this strategy; 4 = I always use this strategy), and (b) perceived effectiveness of strategies (i.e., 1 = Not at all effective, 2 = Somewhat effective, and 3 = Very effective). The frequency of use ratings was later recoded dichotomously: the rating of 1 was recoded into 0, indicating *not used*; other values were recoded into 1, indicating *somewhat used*. Parham argued that by recoding the ratings into two discreet categories, she would provide "a simple way of examining how many of the strategies within each category each student endorsed, which was the original rationale for using a checklist" (p. 27). The mean effectiveness for each category of strategies was obtained by taking the average of effectiveness ratings of the strategies in a category.

To compute the reliability levels of each strategy category on both scales (use frequency and effectiveness), Parham (1996) used a series of statistical analyses of internal consistency reliability, test-retest reliability, and within participant agreement (mean intraclass correlation coefficient). Because most of the subscale reliabilities were below the investigator's cutoff point of .70 for both frequency of use and effectiveness, the investigator went back to the original list of strategies, asking a group of judges to resort the 78 strategies into the 5 aforementioned categories of Millman et al. Based on the proportion of judges' agreement and item-total correlations (for frequency of use

ratings only), Parham reduced her measure to a multidimensional instrument of 54 items; then she reconducted the reliability analyses for this new measure on both the frequency of use and effectiveness ratings.

However, because the obtained reliability coefficients for the 5 categories in this new measure were not clearly improved compared with the previous ones, Parham became doubtful of the wisdom of her initial theoretical approach of measure development. She writes, "it may be that the six strategy categories are not sufficient to describe the strategy use of test-takers," (p. 46) and "a confirmatory factor analysis was not appropriate for this study both because the six strategy categories were not empirically derived and because there was little theory to suggest that there should only be six categories of test-taking strategies" (p. 47). Therefore, Parham decided to switch to an empirical approach of measure construction. Specifically, she conducted an exploratory factor analysis on the frequency of use ratings (not effectiveness ratings) of the original 78 strategies to explore the factor structure of her measure. The end product was a 78-item, 15-factor Test-Taking Strategy Checklist.

Although I agreed with Parham about the need for developing a TTS measure that was more representative of the construct of TTS as well as reflecting the dimensionality of this phenomenon, her research design and findings were unfortunately wanting on several aspects. Among the weaknesses of this study, the most important limitation is the atheoretical and haphazard fashion from which the final measure was derived. Beginning with Millman et al.'s TTS taxonomy, Parham later rejected this framework as being atheoretical after she tried but failed to improve the scale reliability of the categories of strategies based on this framework.

Exploring the factor analytical structure of her measure, Parham arrived at 15 factors some of which were not clearly interpretable. For example, in Factor 4, "Check and recheck answers," there were two items whose content did not fit the category ("When reading the questions, cover-up the answer choices and look only at the question;" "Do not hesitate to change an answer if you feel you should"). In Factor 13, "Always guess if right answers only are scored," the item "Use the first answer that looks correct and go on to the next question" inferred something different than the meaning of the category (i.e., a careless manner of guessing). Some factors simply do not make sense; for example, Factor 15, "Use scrap paper," consisted of two items that did not relate to each other in terms of content: "Use scrap paper to figure answers before looking at answer choices," and "Don't read too much into a question."

Another limitation is Parham's use of post-test assessment of TTS. Considering test takers' knowledge of TTS as a stable characteristic, or a "trait," it is conceptually problematic to measure test takers' specific TTS use on a particular test—or a "state"—and try to infer the trait. Her turning the Likert-type ratings of frequency of TTS use into discreet categories might result in a loss of information.

Nevertheless, Parham's (1996) effort to construct the Test-Taking Strategies Checklist was important in that it reflected a shift in the operational definition of TW: from measuring a subset of TTS with limited use to assessing multiple facets of TTS to infer TW.

Summary. Considering the role of knowledge or acquisition of TTS in understanding TW conceptually, it is apparent that in order to assess TW, we need a valid and reliable measure of TTS knowledge. However, the current measures of TTS (some

are commonly known as the measures of TW) have certain limitations at the conceptual level and/or psychometric level. Because knowledge of TTS contributes to the understanding of TW conceptually, the first apparent step toward developing an integral TW measure should be the development of a diagnostic test of TTS knowledge.

While most researchers put their effort in studying general sets of TTS, some other investigators paid more attention to verifying the effects of a specific strategy on test performance.

Effectiveness of Test-Taking Strategies

TW coaching and training materials or programs often provide their consumers with a list of TTS that are presumably helpful with regard to optimizing a test taker's performance. The problem is that not all TTS are supported by empirical evidence in terms of effectiveness. Some test-taking advice found in existing training guides actually is harmful to test takers' performance if applied because it tends to decrease test takers' chance to gain extra points (e.g., the myth of not changing initial answers where research has shown that changing answers is likely to improve one's test scores; Geiger, 1991; Mueller & Wasser, 1977).

Researchers have assessed test-taking skill effectiveness with two main methods: judges' ratings and performance-based measures.

Judges' Ratings of Effectiveness of Test-Taking Strategies

In her study, Parham (1996) asked undergraduate students to rate the frequency of use and general effectiveness of each test-taking strategy that they had used on a cognitive ability test. Perceived effectiveness of 78 TTS was measured on a 3-point scale

(1 = not at all effective, 2 = somewhat effective, 3 = very effective). Effective strategies were defined as items whose mean rating was greater than the average score of 1.5. Among these 78 strategies, 12 items were perceived as ineffective (mean ratings were smaller than 1.5). A few examples are "Choose answers that will make a design on your answer sheet," "Read the answer choices from the bottom up (i.e., D, C, B, A)," and "Look at your neighbor's answers if you are not sure of the answer."

The strategies were factor analyzed based on *frequency of use* scores, yielding 15 factors. Examining the mean effectiveness ratings for these factors, the researcher concluded that participants perceived some factors of TTS as effective (e.g., Watch for "Give-aways"; Be careful; Use scrap paper). The factor that was perceived as having the lowest effectiveness was "Guess a consistent choice." The researcher also found that several strategies perceived as effective by participants were not related to participants' test performance.

We should be cautious in interpreting Parham's findings for several reasons. First, undergraduate participants in Parham's study were not pre-screened for their expertise in taking multiple-choice tests; thereby, using their judgment on the (in)effectiveness of strategies is not the most appropriate way to construct a measure of TTS. Secondly, a closer examination of item percent ratings versus item mean ratings reveals that there are a couple of strategies whose ineffectiveness was endorsed by a majority of test takers (over 50% gave it a score of 1 = not at all effective), yet they were still categorized as "effective" based on their higher mean rating (i.e., "when guessing, pick an answer choice at random," $M = 1.52$; "use the first answer that looks correct and go on to the

next question," $M = 1.56$). This is a result of the use of a biased scale favoring effectiveness ratings (e.g., no "neutral" score in the scale).

Ellis and Ryan (in press) used a panel of five experts in test construction (e.g., graduate students and faculty) to re-categorize Parham's (1996) Test-Taking Strategy Checklist into effective and ineffective strategies. Ineffective skills were defined as items having the potential of reducing an individual's test score. The opposite was true for effective skills. The researchers came up with a list of 15 ineffective strategies and 42 effective ones. A few examples of the ineffective items are "Answer all the questions in order without skipping any," "Use the first answer that looks correct and go on to the next question," and "Always choose the most complicated answer." The implication drawn from these findings is that a comprehensive training program on TW or TTS should teach test takers to be aware of ineffective skills besides teaching good strategies.

Empirical Evidence for Effectiveness of Test-Taking Strategies

Investigators have empirically measured the level of effectiveness of TTS based on the relationship of TTS with test performance. For example, McClain (1983) explored what strategies a successful student might use, compared with an average student. She instructed a group of "A" students and a group of "C" and "F" students who took a multiple-choice exam in an introductory psychology course to verbalize their test-taking procedures while taking the exam. McClain found that the scores of the two groups significantly differed from one another. Further, on average, "A" students were significantly more likely than "C" and "F" students to apply certain TTS. These strategies included (1) reading all alternative answers, (2) reading the answers in the order in which they are presented in the test, (3) anticipating an answer to a test item before reading the

alternatives, (4) analyzing and eliminating incorrect alternatives to help determine the correct answer, and (5) skipping questions about which they are unsure and coming back to them later. She concluded that there was a relationship between strategies a test taker used and their grades.

In the following sections, I will review several individual TTS (i.e., changing answers, informed guess, elimination marks) that have been empirically examined for their effectiveness in improving test scores. Note that the research attention individual strategies receive is not equivalent: some skills or strategies have been investigated more often than others.

Changing answers. Changing original answers on a multiple-choice examination is one of the most researched test-taking behaviors. Yet, its effectiveness is often underestimated and underutilized by test takers.

In his early exploration of TW, Millman et al. (1965) mentioned that the tendency to rethink and change one's responses is the basic aspect of TW. Consistent research evidence shows that changing answers produces higher test scores. Past studies (i.e., Best, 1979; Johnston, 1975, 1978) found that test takers made wrong-to-right changes more frequently than right-to-wrong changes. Mueller and Wasser (1977) and later Geiger (1991) administered classroom multiple-choice tests and examined students' answer sheets under high illumination for erasure marks associated with answer changing. They found that, on average, for every point lost due to changing answers on a multiple-choice exam, two to three points were gained. In addition, Geiger found that students' maturation affected the mean percentage of net point gain, with upper-level business students tending to gain more points due to changing answers than introductory

students. Overall, 63.5% introductory and 69.4% upper-level students had point gains from engaging in answer changing. Frederickson (1999) obtained a more conservative result in his study. Overall, students who changed answers improved 58% of the answers changed. The conclusion is that answer changing is a test-taking strategy moderately beneficial to a student's grade.

Despite the evidence, the common belief among students is that changing answers will lower their scores. This belief probably originates from recommendations of early educational psychologists. Frederick (1938) says, "Your first thought is generally best" (p. 345). Dressel and Jensen (1955) advise, "Don't change any of your answers unless you find you have made an obvious error" (p. 33). This myth is still endorsed by some commercial test preparation materials (e.g., "Do not second guess;" Passbook, 1998; "It's usually best not to change based on a hunch or a whim;" Brownstein et al., 1999).

Benjamin, Cavell, and Shallenberger (1984) summarized 33 studies that examined the skill of answer changing. They found that between 70% and 100% of students completing multiple-choice tests endorsed the strategy of avoiding changing their initial answers because of the myth. Geiger (1991) found the majority of test takers (65%) underestimated the benefit of answer-changing strategy; only 9% overestimated the outcome of their behavior, whereas 26% correctly perceived the outcome of their behavior. Nevertheless, Benjamin et al. still estimated a median of 84% of students as changing at least one answer when taking tests. No significant gender differences are associated with answer changing or with successful outcomes of answer changes (Frederickson, 1999; Skinner, 1983), although more men than women perceived answer changing as beneficial (Geiger).

In research on reasons for changing initial answers, Schwarz, McMorris, and DeMers (1991) found that over 90% of the answers are changed for reasons other than clerical accuracy (e.g., marking an unintended answer in the first place). Further, Lynch and Smith (1975) found a significant positive correlation between the difficulty of an item and the number of students who changed the answer to that item, meaning that the more difficult a question was, the more students tended to rethink and change their initial responses.

Other investigators examine whether an answer-changing strategy works for everybody. The results show a significant relationship between multiple-choice test performance and individual characteristics, such as cognitive ability, and anxiety level of test takers. Investigators found that in general, the higher test scorers made more changes than the lower test scorers (Lynch & Smith, 1975; Mueller & Schwedel, 1975; Penfield & Mercer, 1980). Higher test scorers gained more points by changing answers than did lower test scorers (Geiger, 1991; Mueller & Schwedel; Penfield & Mercer). Regarding attitudinal influences, low-anxiety students tended to change more answers and did so more successfully (e.g., gaining more points from those changes) than did high-anxiety students, though both groups did gain points (McMorris & Leonard, 1976).

Apparently, answer changing after some reflection and contemplation is overall an effective strategy that test takers should be encouraged to adopt.

Anticipating test items' difficulty. Huck (1978) advised test takers about how difficult previous examinees had found certain items in a multiple-choice examination to be. The researcher hypothesized that students might read these items more carefully if they were made aware of how difficult those items had been for previous test takers. He

found that knowing the difficulty level of an item had a significant and positive effect on test takers' test scores. In other words, judging levels of item difficulty before responding to the items may help examinees prioritize efforts and subsequently apply other strategies to solve the problem.

One implication of this finding is that effectively taking a test may involve not only simple, straightforward skills but also a certain level of awareness and/or overall judgment of test items and the test itself.

Marking strategy. A related strategy to the above "Anticipating test difficulty" strategy is the use of answer option elimination marks. From analyzing Nguyen et al.'s (2003) data, we found test takers who eliminated incorrect options prior to selecting the correct answers tended to score better on their CA-T than those who did not ($r = .21$, $p<.01$). Kim and Goetz (1993) allowed test takers to write or mark on a multiple-choice test booklet and later examined how frequently test takers made elimination marks (e.g., crossing out an assumingly incorrect alternative). They found that elimination marks significantly related to students' test scores. In other words, increased test scores were associated with greater frequency of marking of eliminated options.

Also, the frequency of marking on a test increased as the difficulty level of test items increased. The researchers explained this positive relationship in terms of cognitive efficiency. It is probable that test markings in general and elimination marks in particular may serve as an aid in facilitating retrieval of information from long-term memory, assisting test takers in focusing on important information, and decreasing information load.

Strategy of planning test time. Cook (1957) found that a periodic check on progress rate helps test takers maintain proper speed in taking timed tests. On the other hand, based on Nguyen et al.'s (2003) data, I found that the individual strategy of "setting up a schedule of progress" was negatively correlated with cognitive ability test performance ($r = -.25, p < .01$). The more test takers planned their test progress, the less likely they improved their score. This counter-intuitive finding may be explained by the fact that keeping track of time may be an interfering task with the test-taking process or causing test takers to slow down on a speeded test. Carr, Bell, Ryan and Kilanowski (2002), on the other hand, did not find a significant relationship between test performance and the subscale of "Time use" strategies (defined as "the extent to which test-takers effectively used their time, e.g., balancing speed and accuracy," p. 21).

Note that time-management strategies in general may not be particularly helpful when one takes a power (not timed) test.

Test-Taking speed. Paul and Rosenkoetter (1980) studied the speed with which test takers finish multiple-choice exams. They reported no significant relationship between the order in which test takers finished an exam and the scores they earned. Johnston (1977) found that the mean performance of "early" and "late" finishers did not significantly differ. However, the researcher noted that the variability of test performance in these two groups was greater than the performance variability of test takers who finished neither early nor late. This finding suggested that both good and poor test takers tended to be among the early and late finishers.

On a different but somewhat related note, analyzing Nguyen et al. (2003) data, I found a relationship between cognitive ability performance and the strategy of working

as rapidly as possible on the test ($r = .18, p < .05$). Test takers who worked as quickly as possible throughout a test tended to perform better than those who did not.

Skipping items strategy. There are mixed results regarding the effect of skipping questions and returning to them later as a strategy in test taking. Rindler (1980) reported that "middle ability" students (GPA was between 2.20 and 2.79) skipped questions on a verbal aptitude test more frequently than did "high ability" (GPA > 2.79) and than "low ability" (GPA < 2.20) students. However, test takers' cognitive ability interacted with the effect of this strategy. Among high ability test takers, those who skipped questions earned higher final scores than those who did not. Among low and middle ability test takers, those who did *not* skip questions earned higher test scores than those who did.

In another study, McClain (1983) found that "A" students skipped significantly more questions on a classroom multiple-choice examination than "C" and "F" students. However, the exam scores of those who skipped questions did not significantly differ from the scores of those who did not skip questions, controlling for level of ability.

Guessing strategy. Most if not all training materials in test-taking skills strongly endorse guessing as an effective strategy, educated guessing (e.g., after eliminating incorrect choices, or based on prior knowledge) more so than blind or random guessing. Basic statistics show that by chance alone, a blind guess on a 5-alternative test item on average gives test takers 20% of the probability of answering the question correctly.

A correlational analysis on the data collected by Nguyen et al. (2003) revealed that guessing whenever right answers were scored was significantly related to cognitive ability performance ($r = .23, p < .01$). Carr et al (2002) found a significant relationship ($r = .10, p < .01$) between guessing strategies and test scores among a sample of job

applicants for the position of firefighters. However, random guessing with some regularity is verified as an ineffective strategy in the literature. For example, a blind guess yielded a negative correlation with performance on a passage-independence reading test ($r = -.30, p < .05$; Powers & Leung, 1995). Random guessing or omitting a difficult question inversely related to test performance ($r = -.19, p < .01$; data from Nguyen et al., 2003).

Personality traits have been found to moderate the effects of guessing strategy in general. For example, Brenk and Bucik (1994) examined the relationship between personality differences and differences in guessing tendencies or preferences (high versus low) on the performance of a difficult test about whose content the test takers had little prior knowledge (i.e., a test of knowledge of foreign words). The personality measure was the Cattell's Sixteen Personality Factor Questionnaire (Form C). The sample included 276 participants (average age was 26.4 years). Brenk and Bucik found that, for those who scored high on the test, those who were more likely to guess tended to be more radical, critical, dominant, aggressive and determined, those who were less prone to guessing were more spontaneous, uninhibited and unconventional. Participants with lower test scores were described as careless, cooperative and enthusiastic persons but also sophisticated, shrewd and impatient.

Use of reasoning strategies to answer reading comprehension test items. In 1950, Bloom and Broder reported that undergraduate students who were trained in general problem solving techniques (including the ability to reason logically, comprehension of test directions, and understanding of the nature of specific test questions), but *not*

additionally trained in subject-matter knowledge, made significant gains in subsequent achievement test scores.

A test-taking phenomenon that has long fascinated researchers is the "passage independence," or the ability of test takers to correctly answer test questions related to a reading comprehensive passage without seeing the passage itself, using prior knowledge and reasoning ability to "fill-in-the-blank" (e.g., Chang, 1979; Powers & Leung, 1995; Doverspike & Miguel-Feruito, 2001). For example, Powers and Leung examined the extent to which various test-taking skills were being utilized to answer reading comprehension questions on a new version of the Scholastic Aptitude Test. The group of TTS was composed of (1) Reasoning strategies (e.g., trying to determine the meaning of a word, or phrase, or the way in which it was used, from the other questions in the set); (2) Personal knowledge or experience (e.g., I recognized a passage or knew where it came from); (3) Strategies for Vocabulary, and (4) Guessing. Test takers were asked to answer sets of reading questions without the reading passages and then their TTS were assessed. The investigators found that students were able to attain a level of performance that exceeded chance level on the SAT reading questions. The researchers also reached the conclusion that the strategies for answering questions without reading passages reflected participants' use of verbal reasoning (using the consistencies within the question sets to reconstruct the theme of the missing passage; attempting to reconstruct the theme of a missing passage from all the available questions and answer choices) more than their relying on the characteristics of questions or answer choices (e.g., choosing the typical – and incorrect – meaning of a word that appeared as one of the alternatives).

In their review of TTS, Flippo et al. (2000) drew readers' attention to a little-known but sophisticated strategy called the use of convergence cues. Similar to test-taking secondary cues, the convergence cues tap on a characteristic of multiple-choice test items. However, unlike secondary cues, which fail with well-constructed test items, convergence cues can be drawn from the principle of constructing good objective test items. Basically, this strategy requires a test taker to have the ability of abstract reasoning in order to examine and detect the logical relationships between the multiple-choice stem and alternatives. In constructing good objective multiple-choice items, the principle is that all distracters must be plausibly related to the stem in order to distinguish test takers who really know the test domain from those who do not. That means the alternatives might include some but not all dimensions underlying the stem, but only the correct answer will include all of the stem-related dimensions.

The use of convergence cues was first brought up and empirically investigated by Smith (1982). The researcher trained a group of high school students to look for dimensions that underlie the alternatives for multiple-choice questions and consider how these dimensions relate to the dimensions in the stem. Because all the dimensions in the stem converged only in the correct answer, participants were taught how to find the convergent point and identify the correct choice. The control group received general test-taking instruction. Smith compared the pre- and post-training test scores of both groups (scores on the Preliminary Scholastic Aptitude Test (PSAT) and the Scholastic Aptitude Test). The researcher found a significant mean point gain of 39 on the verbal subscale for the experiment group, adjusted for the PSAT covariate, due to the training in

convergence cues. However, this strategy did not affect performance on the mathematics subscale.

Three deductive reasoning strategies proposed in Millman et al.'s framework ("I chose neither of two options which implied the correctness of each other;" I chose one of two statements, which, if correct, would imply the incorrectness of the other," and "I restricted choice to those options which encompassed all of two or more given statements known to be correct") did not significantly relate to score gains in Nguyen et al.'s (2003) study either. Carr et al (2002) did not find a significant relationship between deductive reasoning strategies and test scores among job applicants.

Note that in the research literature, the strategy of deductive reasoning is considered as going beyond a test-taker's subject-relevant reasoning ability (e.g., reading comprehension) that a test is designed to measure, as verified by the above "passage independence" findings, even though reasoning ability in general is part of an individual's general mental ability.

Use of secondary cues. Using cues to correctly respond to items when having little knowledge of the test content is a reliable yet controversial strategy in taking poorly constructed teacher-made tests. Note that this class of test-taking strategies is very controversial because some researchers view this group of strategies as the sole operational level of the TW construct (e.g., citing these cue-use strategies whenever they refer to TW; Gibb, 1964). Some researchers even oppose teaching students and test takers to take advantage of these test idiosyncrasies because it is neither ethical nor practical to do so (e.g., Scruggs & Mastropieri, 1992).

When students taking a teacher-made multiple-choice examination have to make a blind guess (e.g., having no clue about how to answer a stem and its alternative choices), they may significantly increase their chances of picking out a correct response if they choose (1) the longest alternative answer, and/or (2) the answer in the second or third position (B or C). Jones and Kaufman (1975) observed that inexperienced test writers have a tendency to (1) provide the most complete information to make an alternative correct and thus make it the longest choice, and (2) hide the correct alternative in the "B" or "C" position in a multiple-choice alternative set, probably thinking that the correct choice will stand out in the "A" or "D" position and be too obvious.

How prevalent are secondary cues in classroom examinations? To answer this question, Brozo, Schmelzer, and Spires (1984) investigated to what extent college teacher-made multiple-choice tests contained unintended cues that could be used to identify correct answers. They reviewed 43 undergraduate examinations (at a total of 1220 multiple-choice questions) from 5 colleges and universities, written by faculty members at various tenure levels. The researchers found that almost half of the test items (44%) contained a secondary cue, and that 70% of these items could be answered correctly by applying the cue. According to Brozo et al.'s findings, the most often discovered cue was "direct opposites" (e.g., an alternative was directly opposite to the correct answer). "Key word association" was the cue that test takers could apply the most successfully, resulting in a correct choice.

The implication of these findings is that the frequency of these secondary cues might depend on the expertise and experience of the test constructor. Theoretically speaking, secondary, unintended cues are more likely to be imbedded in classroom

examinations with greater frequency than in standardized tests whose constructors are more experienced with measurement metrics. Does that mean that the cue-use strategies would be rendered useless when being applied to standardized CA-Ts?

Flynn and Anderson's (1977) study provided a partial answer for the above question. Four types of secondary cues were investigated for their effects on undergraduate students' scores on standardized tests measuring mental ability and achievement. The cues taught were (1) options that were opposites, so that if one were correct, the other would be incorrect (e.g., "the war started in 1812" versus "the war ended in 1812"), (2) length of correct response (longer correct options), (3) use of specific determiners (e.g., always, never), and (4) resemblance between the correct option and an aspect of the stem (e.g., grammatical cues). Participants were classified as either test-wise or test-naïve based on a pre-test TW measure, and later they were trained to recognize the four cues. The researchers found that teaching these particular cues did not result in gains on several classroom ability and achievement tests. One explanation that the researchers offered was that the four target cues might not be present in the ability and achievement tests because these tests had been validated with other groups of students over the years. (The researchers did not mention whether they had crosschecked the administered tests for the target cues and test validity.) Moreover, identified test-wise students tended to score higher than test-naïve students in this study. Therefore, Flynn and Anderson suggested that their test-wise examinees must have used other TTS than cue-using strategies.

Regarding the effectiveness of this category of TTS, some studies found no relationship between using imbedded cues and test score improvement (e.g., Flynn &

Anderson, 1977; Jones & Kaufman, 1975). My analysis of my data in Nguyen et al. (2003) data set showed that test takers' taking advantage of grammatical cues in stems and/or alternatives did not significantly relate to their test performance. Furthermore, I found an inverse relationship ($r = -.23, p < .01$) between analytical performance and the use of logical or physical cues (e.g., choosing an answer based on its length and/or its position). One possible reason for the observed negative correlation is that the analytical subtest used in this study was overall easy, so that the only people who would need to resort to the cue-using strategies would be less able students. This explanation might not be plausible, however, because a peruse of the subtest score distribution showed that participants' analytical test performance was somewhat positively skewed (i.e., a larger proportion of test takers got low scores than the proportion of those scoring highly on the test), indicating that this test was relatively difficult for this sample. Therefore, such a negative relation with test performance cast doubts about the effectiveness of the strategies of cue-use on CA-T.

Summary. The investigation of the effectiveness of individual TTS has yielded mixed results. Some strategies when applied tend to improve a test taker's performance over and above his or her knowledge of the subject matter(s) being tested. Other strategies appeared to have no influence on test takers' test scores or even hinder their performance.

A summary of the above research findings is presented in Table 4.

Table 4

Summary of Empirical Evidence for TTS Effectiveness

Strategy	Effective		Moderator(s)	Source
	Yes	No		
- Changing answers	+		- Cognitive ability levels; Test anxiety levels	- Mueller & Wasser (1977); Benjamin et al. (1984); Geiger (1991); Frederickson (1999)
- Anticipating difficulty level of a test item	+		- Gender	- Huck (1978); Bielinski (1999),
- Marking strategy (elimination of incorrect options)	+			- Kim & Goetz (1993); Nguyen, O'Neal & Ryan (2003)
- Periodic check on test progress rate	+	+	- Test format	- Cook (1957); Nguyen et al. (2003)
- Skipping item and returning to it later	+	+		- Rindler (1980); McClain (1983)
- Test-Taking speed	+	+		- Paul & Rosenkoetter (1980); Johnston (1977); Nguyen et al. (2003)
- Guessing:				
- Informed guessing	+		- Cognitive ability; Personality traits	- Brenk & Bucik (1994); Nguyen et al. (2003); Carr et al. (2002)
- Random guessing	+	+	- Frequency of use	- Powers & Leung (1995); Bloom & Broder (1950)
- Reasoning skills (general problem solving techniques; use of convergence cue)	+	+	- Training	- Chang (1979); Powers & Leung (1995); Nguyen et al. (2003); Carr et al. (2002)
- Use of secondary cues	+	+	- Faulty test items	- Gibb (1964); Flynn & Anderson (1977); Nguyen et al. (2003)

Moderators of Effectiveness of Test-Taking Strategies

Based on the above review, levels of effectiveness or ineffectiveness of TTS have been uniformly defined so far. That means a strategy, once being verified "effective" is expected to apply as well for one level of item difficulty as for another level, and from one testing situation or one test format to another situation or format.

However, there has been research conducted on the moderating effects of individual characteristics on the level of effectiveness of TTS, such as gender, personality traits, and cognitive ability. A few studies explored test characteristics, such as type and format, as possible TTS effectiveness moderators. For objective multiple-choice test formats, whether an ability or achievement test is well constructed and standardized influences the effective use of secondary-cue strategies. Gibb's (1964) test takers, well trained in recognizing cues in teacher-made examinations, would have performed as well (or as poorly) as Flynn and Anderson's (1977) participants, whose previous secondary-cue training effect was tested against standardized ability and achievement tests and failed to improve performance. High frequency of use for a statistically sound strategy such as random guessing (where there is nothing else to lose) also affects its effectiveness (e.g., being related to low test scores; Powers & Leung, 1995). Some of Millman et al.'s (1965) time-using strategies (e.g., begin to work as rapidly as possible with reasonable assurance of accuracy) are not applicable—thereby becoming non-effective —in power testing (e.g., tests are not timed).

Bridgeman and Morgan (1996) compared the relationship between scores on the essay and multiple-choice portions of advanced placement (AP) tests and scores on similarly formatted examinations. They found that the group of high multiple-choice/low

essay scorers performed significantly better on other multiple-choice tests than did the group of low multiple-choice/high essay scorers, and vice versa. Combined with the fact that participants in this study were high-achieving high school students, the findings provided evidence that there must be different test-taking skills that are effective for different test formats (essay versus multiple-choice), and that test-taking skills must exist independent of general cognitive and intellectual abilities.

In addition, while it is reasonable to assume that certain general strategies transcend most multiple-choice testing situations (e.g., Scruggs & Mastropieri, 1992), there must be domain-specific strategies that work for mathematical ability reasoning or analytical reasoning items, for instance, but that are generally ineffective with verbal reasoning items, and vice versa. An examination of different test-taking tips for different CA-T sections offered by selected, higher-quality test preparation materials supports this fact (e.g., Brownstein et al., 1999). Moreover, Rowley (1974) found that level of TW (e.g., time use, guessing) accounted for 7% of the variance in elementary students' vocabulary test performance, but these skills were not correlated with performance in mathematics. In other words, the strategies that were effective for verbal tests might be different than the effective strategies used for quantitative tests. For reading comprehension tests, different reasoning strategies were used when the passage was omitted versus when it was not (e.g., Doverspike & Miguel-Feruito, 2001; Powers & Leung, 1995).

Although past studies have supported the effectiveness of several common strategies, they also show us that such effects may vary due to different reasons. This fact calls for future investigations of what may mitigate the effects of a test-taking skill.

Further, empirical research shows that some popular "strategies" considered effective by many students turn out to be a myth (e.g., sticking to your first choice), implying that we cannot fully rely on students' subjective perceptions in order to determine whether a strategy is effective or not in a test-taking situation.

Summary

Some of the common strategies in test taking have received research attention and verified as either effective or ineffective in helping test takers improve standardized test scores beyond their content knowledge depending on the circumstances. Also, the effectiveness of other TTS seems to be moderated by test takers' individual characteristics and attitudes, such as gender, motivation and state test anxiety.

As mentioned earlier, this study aims at two purposes. The first goal was to propose a new theoretical framework for the construct of TW in which test takers' knowledge and actual utilization of TTS were important elements. The second goal is to develop a measure of general TTS knowledge, which can be also used to assess specific TTS utilization during test taking.

Developing a Measure of Knowledge of Test-Taking Strategies

As aforementioned, the majority of current TW or TTS measures have been developed to test a few sets of skills or tactics in test taking; these skills or tactics are mostly related to or based on the logical framework of TW principles compiled by Millman et al. (1965). Most of these measures are similar in format (i.e., consisting of multiple-choice questions); the items in some measures are content-free (e.g., including fictitious facts) in order to control for test takers' prior knowledge of subject matter(s)

being tested. The underlying theoretical assumption of these measures is identical and simplistic: people who score higher on these tests over chance probability are inferred to be "test wiser" than those who score lower. In addition to the limitations and validity problems previously discussed, these measures seem to assess what cognitive psychologists call "procedural knowledge" of TTS, or an individual's knowing how to perform a skill.

An exception is the TTS inventory checklist that Parham (1996) developed. This instrument aimed at measuring the "declarative knowledge" aspect of TTS, or factual knowledge of principles that an individual can report or describe. According to Anderson (1993), declarative knowledge is operationally defined as the factual knowledge that one is able to report or describe, whereas procedural knowledge is knowledge people can only manifest in their performance, or the knowledge that can only be inferred from one's behaviors. An example of declarative knowledge is the knowledge that Washington, DC, is the capital of the United States. Our ability to speak English is an example of procedural knowledge. Simply speaking, an individual develops a skill by first acquiring declarative knowledge via examples or principles of a domain area. Declarative knowledge is then used to solve additional problems, and this experience develops procedural knowledge in an individual.

Applying to TTS, a test taker possesses a declarative knowledge of the strategies she has learned, or the knowledge of individual strategies and their applicability. Being able to use these acquired strategies during test taking is the evidence of test takers' procedural knowledge of TTS. In other words, TTS declarative knowledge is the prerequisite for developing procedural knowledge of TTS.

When discussing the theoretical framework for the construct of TW, I have mentioned the roles of the cognitive and behavioral components (see Figure 1), which correspond to a test taker's declarative knowledge of TTS, and TTS actual use (procedural knowledge). In the next sections, the acquisition and utilization of TTS will be described to develop a conceptual foundation for the development of a measure of TTS.

Test-Taking Strategy Acquisition

Figure 2 presents a diagram of the path along which our fictitious test taker, Song Le, acquires her test-taking strategies.

First, Song gathers information about TTS by reading test-guide materials or attending training courses in taking tests. The information she may acquire includes principles of taking multiple-choice tests, ranging from very general strategies that can apply for any test-taking situation to very specific ones applicable to a subtest only (e.g., mathematics or reading comprehension). She also studies examples of the use of these strategies. The characteristic of this phase is that the test-taking strategies and examples that Song gathers may be incomprehensive and/or misleading, depending on the quality of the information source, and she may or may not be aware of the need to correct these flaws (incorrect and missing information).

Next, Song encodes the strategies and examples in her memory to be retrieved later, either in practice or in an actual test-taking situation. At this point, Song is not only able to describe or declare what strategies she knows, but also whether or not she intends to apply these TTS when taking a multiple-choice test.

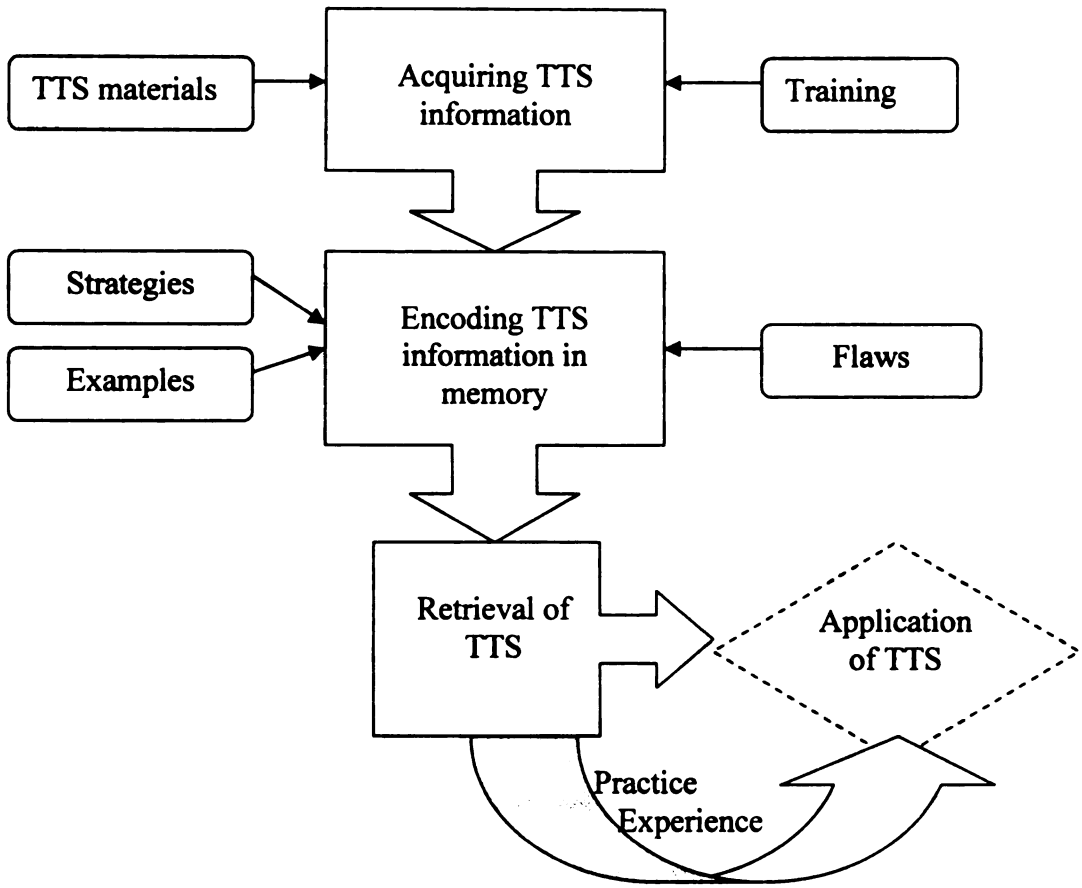

Figure 2. Simplified path diagram of test-taking strategy acquisition

Note that in this model, the skill acquisition process occurs some time prior to the application of these skills, and there should be some amount of practice involved before the cognitive skills can be smoothly transferred to test-taking behaviors. The implication is that it is necessary to have a time lag between TTS knowledge acquisition and TTS utilization on an actual test. The necessity of building up some experience by using the strategies in practice may help explain why the evaluation of the effectiveness of a TTS training program (i.e., Holden, 1996) yielded negative results: assessing test takers' proficiency of some TTS immediately after a training program might be unable to detect any manifestation of TTS effects. Such an evaluation might take place too early, before the participants had the opportunity to practice and master the newly acquired skills.

Another implication is that a deficiency in the early stage of TTS acquisition (e.g., learning too few strategies or learning ineffective ones) may impair the effectiveness of the application of TTS. Therefore, it is necessary to take into account how long the acquisition and storage phases have been, how the acquisition occurs, and how many adaptive strategies are acquired, in order to determine how sufficient the knowledge of TTS of a test taker is. In other words, in addition to inferring the existence of TW from test-taking behaviors as far as the use of TTS is concerned, it makes sense to develop a self-report measure assessing declarative knowledge of TTS based on a theory of TW.

Utilizing Test-Taking Strategies

The following description of how TTS are used in a test-taking situation is based on the assumption that a test taker is motivated enough to perform well on a test. At the onset of (and throughout) a multiple-choice CA-T and before actually responding to individual test items, a test taker with some knowledge of TTS will retrieve from his or

her memory a few basic, general strategies, which are uniformly applicable to the multiple-choice test format. For example, a test taker would plan to use some time-allocation strategies for a speeded test, or not to use them at all where a test is not timed. Making sure she knows how to properly mark a response (e.g., blackening or circling correct answers) and paying attention to test directions are also some examples of the use of test-taking strategies.

Responding to individual test items, the test taker's content knowledge will predict whether or not she would need to retrieve one strategy or a set of strategies, and whether the strategies would be domain general or domain specific (e.g., applicable to most test items or just to a specific kind of questions) in order to solve the problem. Specifically, if she has full knowledge of the item content being measured, she may not need to engage in the use of additional TTS beyond some general skills to select a correct alternative. On the other hand, where her content knowledge is absent or less than adequate (e.g., she has some doubt about the choices), her utilization of TTS to make decisions beyond her subject matter ability will be in full force.

Figure 3 outlines the route along which a test taker's mental journey proceeds when responding to a multiple-choice test item.

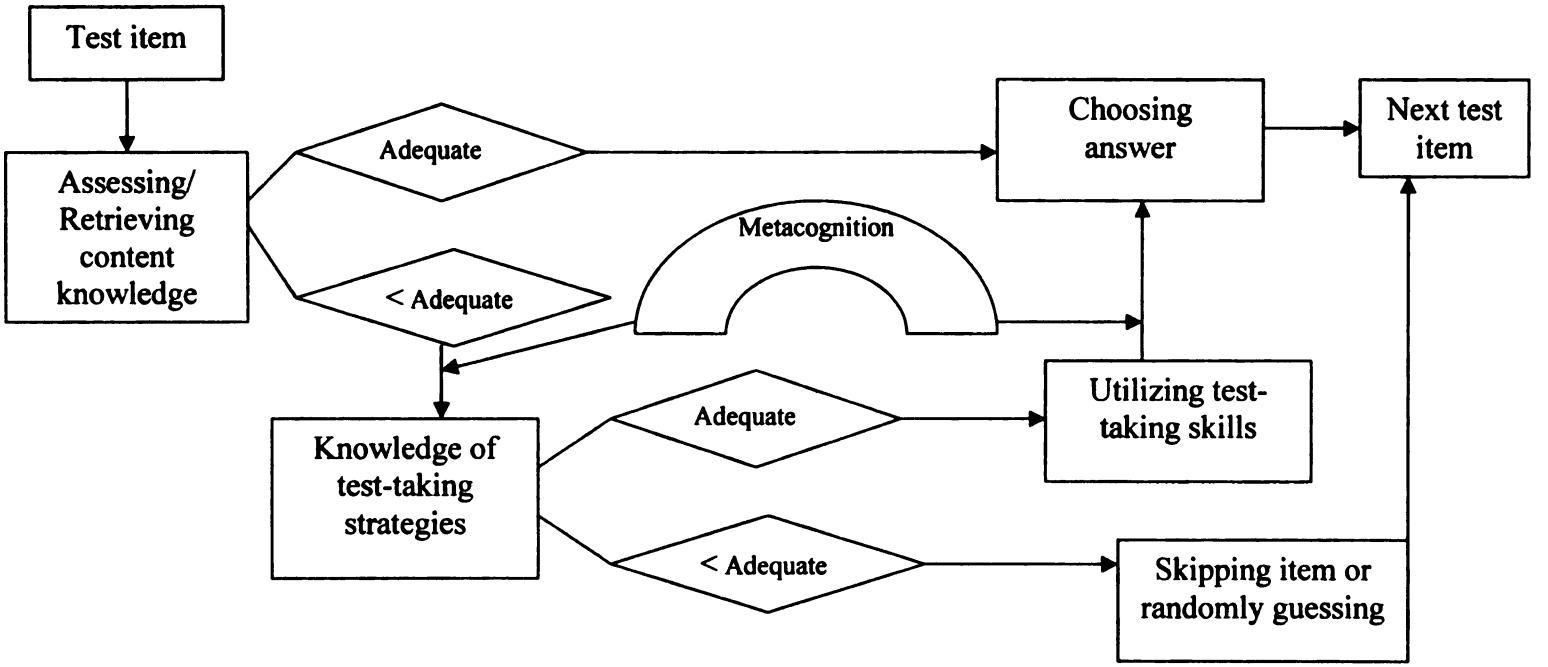

Figure 3. Simplified path diagram of how a test taker responds to a multiple-choice test item

It appears to be a simple matter when one can answer the question with certainty. However, if one has less than sufficient knowledge (e.g., she has some doubts or uncertainty in her mind about the correct answer of a question), the status of her knowledge of TTS will be assessed. In the case that the test taker is test naïve (e.g., having little knowledge of TTS) and absolutely knows nothing about the answer, she may choose to skip the item (with no intention to return to it later) or take a random guess, then proceed to the next test item. On the other hand, having acquired a set of supposedly effective TTS, the test taker may use one strategy or a combination of strategies (e.g., eliminating incorrect choices, then guessing based on her background knowledge) to complement or substitute for her less-than-adequate knowledge of subject matter. Test-Taking metacognition may be present throughout the test-taking process, not only as an "executive governor" (Rogers and Bateson, 1991), monitoring the use of acquired TTS, but also as a buffer against competing cognitive and affective constructs (e.g., test anxiety, distracting thoughts) that may interfere with the mainstream test-taking task. Therefore, one can speculate that the lack of good cognitive monitoring skills may negatively affect the whole test-taking effort in general and the application of TTS in particular.

Two implications can be drawn from this model of test-taking process. First, holding everything else constant, it is apparent that pure content knowledge leads to true CA-T score and mitigates the use of TTS (i.e., reducing the frequency of strategies used). Secondly, it is reasonable to hypothesize there would be a curvilinear relationship between content knowledge and the frequency and number of TTS used: (a) adequate knowledge secures a quick, correct response and thus the use of TTS throughout the test

would be minimal; (b) partial content knowledge or even some related, background knowledge would increase the frequency of TTS used, and (c) the total absence of content knowledge or prior background knowledge may restrict the application of TTS to the last resort: the strategy of random guessing. When the test is about to end, a test taker may retreat to the use of general strategies (e.g., completely erasing marks on the answer sheet; checking for clerical errors).

Developing a Measure of Knowledge of Test-Taking Strategies

As mentioned in a previous section, it is essential to assess how well a test taker knows her TTS to establish the foundation for a subsequent measure of TW. Therefore, I will focus on the construction of an instrument measuring declarative knowledge of TTS, the Knowledge of Test-Taking Strategies (KOTTS) measure, expanding the taxonomy of TW of Millman et al. (1965), and adopting a similar philosophical approach as that of Parham (1996). This is a self-report instrument assessing a test taker's factual knowledge of various types of TTS that are associated with paper-and-pencil multiple-choice tests.

Operationalizing knowledge of individual strategy. The first issue is how to operationalize the concept of declarative knowledge of TTS at the item level. Earlier, I have posited that conceptually, the declarative knowledge of a single strategy consists of knowing the strategy of interest as well as knowing about its applicability in test taking. Therefore, the declarative knowledge of an individual test-taking strategy is operationalized as a composition of 2 elements: (a) the extent to which a test taker is *familiar* with a strategy (being aware of its existence from having learned it or been told about it), and (b) the extent to which the test taker is inclined to use this strategy (reflecting her knowledge of the applicability of the strategy). The rationale is that it is

not sufficient to measure one's familiarity with a test-taking strategy without assessing her stand on the strategy. In other words, the rating of an individual strategy will be a composite of these two ratings. A scale score will then be the aggregate of the composite scores of individual strategies in the scale.

One might argue that a test taker rates a strategy as something he or she is less inclined to use because he or she thinks that it is not useful for most tests, not because he or she is not familiar about it. In other words, there might be an inverse relation between familiarity and use inclination ratings. In this study, a subsequent examination of the correlations between these two types of ratings (familiarity and inclination of using) at the item level showed that most correlation coefficients were positive. The magnitudes were mostly in the range of .60's or .70's with the exception of the Guessing subscale (half of the correlations were in the .30's). This fact supported the notion that the more likely a person reports knowing of a certain strategy, the more likely he or she rates highly on the inclination of use scale. Whether or not such a declared knowledge about a strategy, inferred by the composite item score, is accurate or functional may depend on the actual effectiveness of the strategy of interest. For example, a low composite score on an "ineffective" strategy (i.e., related negatively to a criterion) would indicate that a test taker's declarative knowledge of this strategy is functional, no matter how this low score is derived (i.e., either because he or she knows of the strategy but does not think it is particularly useful, or because he or she is not very familiar with this knowledge and has a hunch about not using it).

Multidimensionality. At the scale level, a conceptual issue is whether TTS knowledge is a unidimensional or multidimensional construct. The literature suggests that

there may be multiple factors or dimensions in the construct of TTS. Millman et al.'s (1965) TW taxonomy includes 6 categories of strategies (or TW principles as the authors called them): (1) Time-using, (2) Error-avoidance, (3) Guessing, (4) Deductive reasoning, (5) Intent consideration, and (6) Cue-using. Doverspike & Miguel-Feruito (2001) reduced Millman et al.'s categories into 4 factors, (1) Time use, (2) Guessing, (3) Error avoidance, (4) Elimination alternatives. Testing Millman et al.'s taxonomy on the data of reported use of TTS using an exploratory factor analysis, Nguyen et al. (2003) found a 6-factor construct of TTS after discarding or rearranging some original strategies based on the factor analytic results. These factors included (1) Using logical/physical cues, (2) Using grammatical cues, (3) Treating test items with caution, (4) Deductive reasoning, (5) Time management and (6) Guessing. In short, Millman et al.'s taxonomy was not perfectly confirmed in subsequent studies but some dimensions in their taxonomy tended to be stable concepts.

I will use the method of logically grouping existing strategies and skills into multiple categories or TTS dimensions based on strategy content. The issue of how many a priori dimensions of TTS knowledge should be included in the KOTTS measure would depend on the variety of strategies gathered for this study. However, I anticipated that the scope of the a priori dimensions of TTS knowledge in this study would be broader than Millman et al.'s (1965) taxonomy because I might find some individual TTS whose content was not mentioned in their outline.

There is an implication for conceptualizing the KOTTS measure as being multidimensional: the various dimensions of TTS might differentially relate to other TW components (utilization or application of TTS and test-taking metacognition), other test-

taking constructs (i.e., test-taking motivation, test-taking self-efficacy) and test-takers' personal characteristics (i.e., demographic information, general intelligence). The formulation of individual strategies and creation of TTS dimensions is detailed in the method section.

TTS could also be grouped according to their level of domain specificity. Domain specificity refers to the fact that some strategies seem to be less applicable to general multiple-choice tests and test items than other strategies. It is speculated that TTS may lie on a continuum between high (specifically applying to certain types of test items) and low level of specificity (most general strategies). Therefore, I employed both fashions of strategy classification (based on item content and based on item domain specificity level) in developing this measure.

In addition, if appropriately modified (i.e., changing verbs or verb tenses from the present to the past), the same measure of general TTS knowledge can also be used to assess test takers' reported utilization of the strategies on a specific multiple-choice test that they have taken. For example, the strategy "I know that the test constructor may place the correct answer in a certain physical position among the options (e.g., in the middle)" on the general knowledge version will become "The test constructor made the correct answer grammatically consistent with the stem" on the version of TTS utilization. In this study, both versions of this TTS measure will be administered: the general TTS knowledge assessment (TTS knowledge) being administered to test takers prior to a cognitive ability test, and the reported actual use of TTS (TTS use) after test takers complete the test.

Major Relationships between TTS Knowledge, TTS Use and Other Test-Taking Constructs

In the diagrams depicting the process of test-taking strategy acquisition and the mental process of a test taker's responding to a multiple-choice test item (Figure 2 and Figure 3), a test taker's knowledge of TTS is related to her utilization of these strategies, assuming that she wants to apply both her content knowledge and her knowledge of TTS to ensure optimal performance on a cognitive ability test. In other words, there will be a positive relationship between a dimension of TTS knowledge (the combined familiarity and inclination) and the corresponding dimension of TTS use on a specific test. The magnitude of such relationship will, however, vary depending on the applicability of the strategy dimension of interest on the test.

Hypothesis 1: Self-reported TTS knowledge will positively correlate with self-reported TTS use.

On the one hand, given the assumption that effective TTS are an antecedent of test takers' effectively handling multiple-choice test items as indicated in my TW model, the knowledge of certain appropriate TTS dimensions might relate positively to better test scores in this study. On the other hand, given that the test used in this study was a well-constructed test (i.e., no test idiosyncrasies embedded), the dimensions of knowledge of TTS that help test takers recognize and take advantage of these idiosyncrasies might bear no relationships with test takers' test performance because such knowledge was not applicable to the test at hand. In fact, past research findings showed that ineffective TTS such as random guessing, if used, might be even detrimental to TTS (Nguyen et al., 2003; Powers & Leung, 1995).

Hypothesis 2: Knowledge of effective TTS that are not linked to test idiosyncrasies (e.g., strategies that help test takers avoid clerical or careless errors; guessing; working carefully and thoroughly) will be positively correlated with test scores.

While participants' general TTS knowledge may relate to their test performance, their self-reported use of certain strategies should be correlated with their test scores because past research showed that the use of certain TTS in taking tests was correlated with test performance. For example, for every point a student might lose for changing answers on a multiple-choice test, he or she might gain 2 to 3 points (Geiger, 1991; Mueller & Wasser, 1977). The strategy of eliminating incorrect options by marking them on a test significantly and positively related to examinees' test scores (Kim & Goetz, 1993; Nguyen et al., 2003). However, the use of secondary-cue strategies did not link to one's performance on standardized tests (Flynn & Anderson, 1977) or even inversely related to CA-T scores (Nguyen et al, 2003). Therefore, I hypothesized that the use of certain strategies will relate to test takers' performance on the CA-T in this study, but not the strategies dealing with test idiosyncrasies.

Hypothesis 3: Utilization of effective TTS that are not linked to test idiosyncrasies (e.g., strategies that help test takers avoid clerical or careless errors; guessing; working carefully and thoroughly) will be positively correlated with test scores.

While cognitive ability tests such as the one used in this study measure one's general ability levels, students' grade point average (GPA) is often used as an indicator of one's achievement level or one's knowledge of specific subject contents. In the literature, the use of TW (as measured with cue-using strategies and deductive reasoning strategies) was found to correlate positively with students' GPA (i.e., $r = .34$, $p < .01$; Gentry &

Perry, 1993). I hypothesized that those who possessed the knowledge of these strategies would also tend to have a higher GPA because their knowledge might facilitate the use of these strategic dimensions. Little is known about the relationship between GPA and other TTS dimensions as specified in this study; these relationships, therefore, will be explored.

Hypothesis 4: (a) Test takers' GPA will be positively correlated with the cue-using and deductive reasoning dimensions of TTS knowledge; (b) Test takers' GPA will be positively correlated with the cue-using and deductive reasoning dimensions of TTS use.

Past findings showed that cue-using strategies and deductive reasoning strategies could be effectively applied to a multiple-choice test (i.e., gaining points on classroom examinations; Chang, 1979; Flynn & Anderson, 1977; Gibb, 1964), but these strategies might have little or no observed effects on a well-constructed test (i.e., a standardized test; data from Nguyen et al., 2003). Therefore, it was reasonable to speculate that participants' standardized test score (i.e., self-reported ACT or SAT scores) would not relate to how well one knew these dimensions of strategies. However, knowledge of other types of strategies might benefit test takers to a certain extent (i.e., positively related to better scores). Similar directions of relationships between TTS use and standardized test scores might also be detected.

Note that this hypothesis is different from Hypothesis 2 in that (a) Hypothesis 2 pertained to the particular CA-T being used in this study whereas this hypothesis was about standardized tests in general (as indicated by test takers' ACT or SAT scores), and (b) Hypothesis 2 explored the possible effectiveness of some TTS dimensions on specific test performance whereas this hypothesis particularly focused on verifying the debatably

(in)effectiveness of a set of strategies on standardized tests in general. Nevertheless, the findings from the two hypotheses could be parallel in that they might complementarily shed light on the nature of the relationship between TTS and general mental ability.

Hypothesis 5: (a) Knowledge and/or use of cue-using strategies and deductive reasoning will not relate to participants' standardized test score (ACT or SAT scores); (b) Other dimensions of TTS knowledge and/or use will be positively correlated with participants' standardized test score.

Conceptually, the ability to monitor one's test-taking process and regulating strategy use, or test-taking metacognition, is an important component of the construct of TW (see Figure 1). Test takers' metacognition should relate to the retrieval and application of TTS knowledge during a test. Empirically, Harmon (1997) found that examinees' awareness of the utilization of TTS and test-taking process explained a small but statistically significant amount of variance in their use of cue-using strategies and deductive reasoning (as measured by Gibb's Experimental Test of Testwiseness). Nguyen et al. (2003) found that the general use of TTS was correlated with Shraw's (1998) test-taking metacognitive strategies of Cognitive Knowledge ($r = .47, p < .01$) and Cognitive Regulation ($r = .54, p < .01$). Therefore, I hypothesize that the more adept one is in metacognitive strategies, the more likely one reportedly uses TTS while taking a test.

Hypothesis 6: Test-Taking metacognition will positively relate to one's usage of TTS.

Research evidence has long shown that individuals put forth more effort to perform on a test or get engaged more in a learning task when they are highly motivated. For example, test takers worked faster or attempted to solve more problems when they

For example, test takers worked faster or attempted to solve more problems when they were offered a monetary incentive than when they were not (Maller & Zubin, 1933; O'Neil, Sugrue, & Baker, 1995/1996). Highly motivated job applicants engaged more in metacognition and learning strategies, which were in turn associated with higher test performance (Clause, Delbridge, Schmitt, Chan, & Jennings, 2001). However, regardless of his or her cognitive ability, a test taker might choose not to put forward as much effort as he or she should to perform well on a certain test because of different reasons: a test taker might have neither desire nor will to perform well on the test (e.g., adopting the attitude of "I don't care"); he or she might immerse in his or her own feelings of individual powerlessness (e.g., low on test-taking self-efficacy), or he or she might have a withdrawal attitude due to a previous failure with a similar test (Engelhardt, 1979). A test taker might also consider the testing process itself as purposeless, or dislike the examiner's negative attitude; therefore, his or her test-taking motivation might be reduced. To ensure that participants were motivated to perform well on the cognitive ability test in this study, and thus trying to use TTS as much as possible, I introduced a monetary incentive into the research design of this study to make the task of test taking more meaningful to participants (i.e., having one of three chances to win a cash prize if scoring high on the test). However, the incentive might not guarantee that *all* participants would feel highly motivated to do well on the test at hand. Therefore, there might still be individual differences in the pretest test-taking motivation level, which might positively link to individual differences in participants' report of TTS use. Note that the relationship between test-taking motivation and knowledge of TTS is not hypothesized here because

Hypothesis 7: The more motivated one is in taking the cognitive ability test, the more likely one reports using TTS.

Similar to the possible effect of test-taking motivation, a stronger belief in one's ability to perform well on an upcoming test might be related to one's tendency to use more TTS in taking the test. Nguyen et al. (2003) found a moderate positive link between test-taking self-efficacy and the use of general TTS. Dolly and Vick (1986) found test-taking self-efficacy to be a significant predictor of some TTS. Therefore, it was expected that the use of TTS would be positively related to a test taker's beliefs in her own ability to perform well on the CA-T that he or she was about to take. In other words, the more self-efficacious one was, the more one reportedly used TTS. Similar to test-taking motivation, test-taking self-efficacy was a state construct in this study (i.e., pertaining to the upcoming CA-T). Therefore, there was no hypothesis about the relationship between TTS knowledge and test-taking self-efficacy.

Hypothesis 8: Participants' TTS use will positively correlate with their test-taking self-efficacy.

Multiple-choice test-taking self-efficacy is a concept first proposed by Miguel-Feruito (1997). Although the concept is defined as test takers' belief in their ability to perform well on a multiple-choice test, the measure of the construct consists of items assessing one's *general* confidence and positive attitude toward the multiple-choice test format (i.e., "I do NOT worry before taking multiple-choice tests;" "I usually perform well on multiple-choice tests;" "I would rather take an essay test than a multiple-choice test (reverse scored)"). Note that the multiple-choice test self-efficacy is conceptually different from the test-specific self-efficacy mentioned in Hypothesis 8 in that the former

test (reverse scored)"). Note that the multiple-choice test self-efficacy is conceptually different from the test-specific self-efficacy mentioned in Hypothesis 8 in that the former construct pertains to test takers' general self-confidence about all multiple-choice tests, whereas the latter construct specifically assesses how efficacious a test taker may feel about the test that he or she is about to take. Nevertheless, it is reasonable to speculate that these two conceptually distinct constructs are also related to each other because they both may tap into a general construct of test-related self-efficacy. (In fact, a correlation analysis later verified that test-specific test-taking self-efficacy and general multiple-choice test self-efficacy were moderately and positively correlated with each other, $r = .33, p < .01$.)

Examining the relationship between this construct and "TW" (cue-using and deductive reasoning strategies), Miguel-Feruito (1997) did not find any empirical evidence to support this link. However, I would explore the possibility that an adept person in TTS knowledge would feel more efficacious and positive about the prospect of taking a multiple-choice test. This positive feeling would be, in turn, related to one's tendency to report more use of TTS on the test.

Hypothesis 9: (a) Participants' TTS knowledge will positively correlate with their general self-efficacy about taking a multiple-choice test; (b) participants' TTS use will positively correlate with their general self-efficacy about taking a multiple-choice test.

Past research linked personality as measured by the MBTI with the use of some strategies of test-taking (e.g., Borrello & Thompson, 1985; Frederickson, 2000). However, the MBTI is an ipsative instrument, not designed to provide normative information about inter-individual differences; therefore, it is inappropriate to use the

MBTI to assess personality for such a research purpose. Would we be able to detect the potential link between individual personality and TTS use, using a more appropriate instrument such as a Five-Factor personality test? Further, what personality traits would relate to TTS use? Logically speaking, the characteristics that Conscientiousness encompasses (dependable, careful, thorough, responsible, organized, hard-working, achievement-oriented, and persevering; Digman, 1990) were important attributes to a test taker's accomplishment and might be linked to how likely a test taker used TTS in a test-taking situation. The characteristics of Emotional Stability (e.g., being low on anxiety, worry, nervousness) might also increase the utilization of TTS to respond to a test item.

Hypothesis 10: (a) Test takers' conscientiousness and (b) emotional stability will be positively related to TTS use.

Note that the above hypotheses do not represent all possible relationships among the measured predictors (i.e., the links among motivational factors and dispositional factors). However, these relationships are very likely to exist and will be reported.

Chapter 2

METHOD

Measure Development

The development of the Knowledge of Test-Taking Strategies measure (KOTTS) consisted of a sequence of 4 phases as follows:

Phase 1: Gathering test-taking strategies and screening item content

I searched for existing TTS in the research literature and in major commercial test-preparation materials. An important inclusion criterion for a strategy was that the strategy should be applicable to paper-and-pencil multiple-choice tests, thus excluding several strategies specific to computerized tests. Next, I screened the content of these acquired items to avoid redundancy, rephrasing several items for clarification.

Specifically, I started with the strategies listed on Millman et al.'s (1965) taxonomy of TW (Appendix A) and Parham's (1996) TTS measure (Appendix B). For example, Parham reportedly searched for strategies and skills in test wiseness literature up to 1995. She also ran focus group meetings with undergraduate students to explore commonly used TTS. The result was a list of 225 individual TTS, which she later reduced to a 78-item measure.

Other main sources of strategies in this study included journal articles and books on TW or TTS (in addition to those that had been cited in Parham's unpublished dissertation), as well as from the section of test-taking strategy tips or advice in selective self-taught test-preparation materials for standardized tests published between 1965 and 2002. The basic criteria for my selection of test-preparation materials were the credibility

of a test-preparation publisher (i.e., inferred from the popularity of their publications in standardized test preparation) and/or the variety of these materials (i.e., preparing test takers for either college admission or professional certification). My search resulted in a compilation of approximately 450 individual strategies from various sources (see Appendix C for source references).

After gathering these strategies, I systematically screened the strategies for content relevance and repetitiveness. Strategies that seemed not relevant to strategies of taking a specific test (i.e., pertaining to how to prepare oneself physically or mentally prior to taking a test) or redundant (i.e., similar strategies worded differently) were discarded. Because of the huge amount of available *general* TTS (those being applicable to most multiple-choice tests or multiple-choice test items), I also made the decision to exclude strategies specific to solving a certain type of test problem (i.e., strategies for solving quantitative reasoning problems, strategies for dealing with reading comprehension questions). For the remaining strategies, those whose wording was judged unclear were rephrased. Upon the completion of this screening and modifying process, I came up with a list of 85 original strategies (see Appendix D).

Phase 2: Preliminarily classifying strategies

Partly based on Millman et al.'s (1965) framework, the research literature on TW and/or TTS, and partly based on my analysis of the content of the acquired strategies, I classified the 85 strategies into 11 preliminary categories. Each category was defined based on its assumed function or utility in taking multiple-choice tests. Table 5 shows the definitions of these preliminary dimensions.

Table 5

Definitions of 11 Preliminary Dimensions of Test-Taking Strategies

Dimension	Definition
Knowing how multiple-choice tests work[a]	Test-takers' basic knowledge of how multiple-choice tests work (i.e., test characteristics and features) and how to respond to a multiple-choice test item.
Optimizing time efficiency and effectiveness	Strategies that enable test takers to work thoroughly on multiple-choice test items with no waste of time while producing satisfactory results.
Avoiding clerical or careless errors	Strategies that help test takers avoid clerical or careless mistakes.
Using physical cues[b]	Strategies that help test takers take advantage of faulty test items in terms of their physical characteristics.
Using grammatical or contextual cues[b]	Strategies that help test takers take advantage of faulty test items in terms of grammar and/or hints from other test items.
Deductive reasoning	Deductive reasoning strategies based on the "exclusiveness" characteristic of multiple options, and other reasoning strategies.
Guessing	Strategies based on the "chance" nature of choosing a correct multiple-choice option.
Changing answers	(self-explanatory)
Working carefully and thoroughly[a]	Strategies that allow test takers to exercise caution and thoroughness in responding to test items.
Staying in control[a]	Strategies that allow test takers to maintain their focus on the test they are taking.
Trouble-shooting and using aids in recall[a]	Strategies that allow test takers to effectively deal with questions that they are uncertain about.

Note. [a]These dimensions were neither included in Millman et al.'s (1965) framework of TW principles nor mentioned at length in the research literature; however, they were logically derived from the variety of strategies I had found. [b]The strategies in these 2 categories were traditionally treated as elements in a general construct of "cue-using strategy" (Millman et al., 1965). However, I decided to categorize them separately because of their content dissimilarities.

Phase 3: Expert categorization

To establish content validity for this measure, I invited 2 Industrial-Organizational psychology professors and 4 senior graduate students to serve as judges or expert raters in this study. The judges had extensive experience with test construction or taking multiple-choice tests. Their task was to classify the 85 strategies (on a mixed-item list) into 11 a priori categories (see Appendix E for instructions for this task). As a rule of inclusion, any strategy that received two-thirds of expert votes (4 out of 6) as belonging to a certain category was retained in this category. In addition, based on two judges' suggestions, some strategies were slightly modified (i.e., rewording or being shortened). Three new items were written for one dimension that had only a couple of items left after the item rearrangement. The result was a 78-item measure categorized into 11 a priori dimensions. This measure would be administered to a development sample to be evaluated for its reliability and dimensionality.

Furthermore, the judges were asked to rate the domain specificity of each test-taking strategy on a 7-point Likert scale (1 = very general, uniformly applicable to any multiple-choice test; 7 = very specific, only applicable to certain type of test items). Initially, I proposed to use the ratings of domain specificity to classify TTS in addition to the content dimensions. However, one of the judges wrote me a comment that the instruction regarding this rating task was unclear: "I found this [the task of judging the domain specificity of strategies] very difficult because I didn't know what assumptions we should make about the test (i.e., is there a *scantron answer sheet*? Is the test *standardized*?)" Note that a certain level of ambiguity was unavoidable because there was no sufficiently clear distinction between strategies used for standardized tests and

strategies used for classroom examinations in the literature. However, I acknowledged the judge's concern and, therefore, decided not to use the ratings on domain specificity of strategies in further data analysis in this study.

Phase 4: Evaluating the measure

To evaluate the soundness of my measure of TTS knowledge, I employed the method of reliability analysis, confirmatory factor analysis and exploratory factor analysis. Specifically, I administered the measure to a sample of undergraduate students whose ratings on the items were used to assess the psychometric properties of this measure. The procedure of this phase, the characteristics of the development sample, and additional instruments used in this study to address other related questions are detailed in the following sections.

Measure Evaluation

Procedure

Two questionnaires and a cognitive ability test were administered to undergraduate students in large-group settings (i.e., approximately 20-70 participants per session). The purpose of the study (developing a measure of TTS knowledge) was revealed to participants at the onset (see Appendix G for the Informed Consent Form). Participants were asked to respond to a battery of pre-test measures, to take a timed cognitive ability test, and to respond to another set of post-test measures. Note that participants were required to turn in their completed pre-test survey prior to taking the cognitive ability test and the post-test questionnaire. The purpose of this specific procedure was to minimize the possibility that participants' self-report of TTS use was

not a function of their actual use of the surveyed strategies but a function of their ratings on the pre-test measure of TTS knowledge (i.e., participants' copying the same ratings from the pre-test to the post-test TTS measure). It took all participants approximately 1 3/4 hours to finish the study, in exchange for 4 experiment credits to fulfill a course requirement. They also received a debriefing form that explained the rationale for this study (Appendix F).

In addition, to motivate participants to exert efforts when taking the cognitive ability test, I announced that the top 20% scorers on the test would be entered into a drawing for one of 3 cash prizes ($50 each). I also emphasized that the cash prizes only applied to the cognitive ability test to prevent the monetary incentive from creating response distortion on other measures. The three winners for the prizes were later identified and contacted, each receiving a check for the specified amount.

Development Sample

Three hundred and seventy eight undergraduate students at Michigan State University (MSU) participated in this study in exchange for 4 experimental credits to fulfill their course requirement. Participants were mainly recruited from the MSU Department of Psychology's Subject Pool; some participants took part in the study in exchange for extra course credits in an upper-level psychology course. The data of 9 participants were dropped from subsequent analyses because these individuals failed to respond to a major portion of the key measures in this study. Therefore, the final sample size was $N = 369$.

Age, gender, and ethnicity. In this sample, participants were typically young (M = 19.61 years, $SD = 2.23$; min = 17, max = 45). A large proportion of the sample was

female (N_{female} = 266 or 72.1%; N_{male} = 99 or 26.8%; missing = 4). In terms of ethnic composition, the sample was predominantly Caucasians (N = 278 or 75.3%). African Americans were 7.9% of the sample (N = 29); 6.0% were Asians or Asian Americans (N = 23); 2.4% were Latino Americans (N = 9); 0.5% were Native Americans (N = 2), and 7% of participants categorized themselves as "other ethnicity" (N = 26). Three respondents did not declare their ethnicity.

Class standing. The largest groups in this sample were freshmen (N = 119 or 32.2%) and sophomores (N = 123 or 33.3%), followed by juniors (N = 74, 20.1%) and seniors (4 years and more; N = 50 or 13.6%). Three respondents chose not to declare their class standing.

Cumulative grade point average (GPA). Because the data collection occurred in the fall semester, a proportion of freshman participants had not learned of their cumulative college GPA. In this case, participants were asked to report their high school GPA instead. On the average, those who reported college GPA (N = 274) did fairly well as far as academic achievement was concerned (M = 3.12, SD = .53; min = 1.5, max = 4.0). Likewise, on the average, participants who reported high school GPA (N = 86) had performed well prior to attending college (M = 3.58, SD = .27; min = 2.9, max = 4.0). There was less variance in the high school group than in the college group because of the university's admission policy on high school GPA (i.e., admitting those with high GPA). Given the fact that both forms of GPA positively and significantly related to participants' corresponding standardized test score (r = .32, p < .01 for college GPA; r = .27, p < .05 for high school GPA), I made the decision to use either GPA form as an indicator of participants' academic achievement.

Standardized test scores. Most participants reported their ACT score ($N = 325$) and some participants reported their SAT score. For those who reported both test scores, their ACT score was used. Thirty-one participants did not report their standardized test scores, probably either because they did not have it, did not want to report it, or they forgot their test score (i.e., indicated by participants' putting a question mark next to the question, explicitly saying that they did not remember, or leaving the question blank). Overall, participants' mean ACT test score was 24.35 ($SD = 3.45$); mean SAT was 1161.33 ($SD = 133.44$). Compared with the respective national means standardized test score (ACT = 21, $SD = 4.7$; SAT = 1016, $SD = 226$), this sample's means standardized test score were above national average. Participants' standardized test scores were then standardized for use as an indicator of participants' cognitive ability in subsequent analyses. Note that I did not try to combine participants' GPA and standardized test score to yield a general indicator of ability although the relationship between these two variables was positive and moderate ($r = .34, p < .05$). The reason is that participants' GPA and standardized test score might have different relationships with TTS as indicated in the literature (i.e., certain dimensions of TTS knowledge might correlate more highly with one's GPA than with one's standardized test scores).

College major. Participants were asked to report either their college major or the fact that they had not declared a major ($N = 367$). Based on the classification of MSU academic programs, participants' majors were grouped into 15 categories. Overall, social science and business were the two largest groups of majors (28.5% and 17.7%, respectively). Nearly one-fifth of participants had not declared a major.

Test-Taking preparation. Participants (N = 370) reported their background in formal and informal test-taking preparation. Only a small proportion of the sample (16.6%) claimed that they had previously had some formal training in TTS (i.e., taking at least a course in test-taking preparation). Among those who reported about their level of informal training in TTS (N = 322), only 30.1% of participants endorsed the fact that they had studied self-taught materials involving test-taking strategies (i.e., by choosing "strongly agree" or "somewhat agree"), compared with 40.7% who did not endorse this fact (i.e., by choosing "strongly disagree" or "somewhat disagree"); 16.5% chose "neither agree nor disagree."

Measures

The pre-test questionnaire (Appendix H) consisted of five measures: a test-taking motivation scale, a test-taking self-efficacy scale, a multiple-choice test self-efficacy scale, an assessment of two personality dimensions (Emotionality and Conscientiousness), and a measure of TTS knowledge using two rating scales (Familiarity and Inclination). The pre-test questionnaire was followed by a short cognitive ability test (Appendix I). The post-test survey packet (Appendix J) consisted of a measure of participants' reported usage of TTS while taking the test, a general test-taking metacognition measure, and a demographic survey. These measures and the cognitive ability test are further described in the following sections.

Test-Taking motivation. Participants' test-taking motivation was measured with a 5-point Likert scale with 7 items adapted from the Test-Taking Motivation (TTM) subscale of the Test Attitude Survey (Arvey et al., 1990). Sample items included "I am extremely motivated to do well on this test," and "I just don't care how I do on this test."

A composite score was obtained for each participant by averaging one's ratings on these items. A higher score means being more motivated to do well on the upcoming test. In this study, the scale internal consistency in the data was desirable, $\alpha = .87$, corrected item-total correlations ranged from .55 to .71.

Test-Taking self-efficacy. Participants' test-taking self-efficacy was assessed with a 5-point Likert instrument with 5 items adapted from the self-efficacy subscale in Pintrich and DeGroot's (1990) Motivational Beliefs Scale. Sample items were "Compared with other applicants taking this test, I expect to do well," and "I am confident that I will receive a high score on this test." A person's test-taking self-efficacy score was computed by averaging his or her ratings; a higher score means being more efficacious about one's ability to successfully perform on the upcoming test. In this study, the scale internal consistency in the data was desirable, $\alpha = .87$, corrected item-total correlations ranged from .60 to .82.

Multiple-choice test-taking self-efficacy. Miguel-Feruito (1997) developed a 9-item Test-Taking Self-efficacy Measure (TTSEM) to assess the extent to which test-takers believe in their ability to successfully perform on multiple-choice tests in general. The scale composite score was computed by averaging individual responses; a higher score indicated higher self-efficacy. A peruse of the content of this measure, however, revealed that there were two types of items included in the TTSEM: six items assessing test takers' confidence with the multiple-choice test format, (i.e., "I usually perform well on multiple-choice tests;" "I am good at taking multiple-choice tests"), and 3 items assessing test-takers' attitude toward the test format ("Multiple-choice tests are usually tricky," "Multiple-choice tests are usually difficult," and "I would rather take an essay test

than a multiple-choice test"). However, this measure had a desirable level of internal consistency, $\alpha = .86$; corrected item-total correlations ranged from .44 to .76, indicating the items in this scale were likely measuring the same construct.

Personality. Two personality traits, Conscientiousness and Emotional Stability, were assessed with two 5-point Likert scales (10 items each) adapted from the International Personality Item Pool, short version (Goldberg, 1999). An average score for each scale was obtained; a higher score on each scale meant more conscientious or more emotionally stable. The internal consistency coefficients for the personality dimensions used in this study (10 items each domain) yielded satisfactory reliability coefficients: Conscientiousness, $\alpha = .84$, item-total correlations ranging from .35 to .64; Emotional Stability, $\alpha = .83$, item-total correlations ranging from .50 to .68.

TTS knowledge. The 78-item measure of Knowledge of Test-Taking Strategies (KOTTS) was administered to participants. Participants rated each strategy on two 5-point Likert scales. Specifically, participants read, "Objective (multiple-choice) tests are part of students' academic life. Below is a list of 78 tactics or strategies when taking multiple-choice tests. Please indicate, using the scales below, (1) to what extent you are familiar with each test-taking strategy, and (2) how inclined you would be to use each test-taking strategy." As aforementioned, an item score was the average of both ratings. A sub-scale score for each category of strategies was the average of all item composite scores in that scale. Means and standard deviations of items and scales were presented in Appendix H.

Cognitive ability test. Participants took a timed, 20-item cognitive ability test (8 quantitative reasoning problems; 12 verbal reasoning problems; see Appendix I). This

test was a shortened version of the "Air Force Officer Qualifying Test" (AFOQT), a selection test developed by Skinner and Ree (1987). The selection test originally includes six subtests (3 verbal, 3 mathematical) with a total of 150 test items. Note that I randomly selected various groups of test items whose content was most likely representative of the AFOQT original quantitative and verbal problems. Because of the variety in test items selected (i.e., from three different sections of the AFOQT verbal subtest) and because of the relatively short test length, internal consistency coefficients for both the verbal subtest and the mathematical subtest were not very high: Verbal $\alpha = .60$, Mathematical $\alpha = .50$. An important implication is that any observed bivariate correlation between test scores and other variables would be attenuated because of the multidimensionality in the criterion, meaning that such relations would be underestimated.

Participants had 20 minutes to work on the test: 5 minutes to solve 12 verbal problems and 10 minutes to solve 8 quantitative problems. The time limit was approximated from the average time that test takers could spend on each AFOQT problem, and pilot tested with a group of 3 undergraduate students. Means and standard deviations were presented in Appendix I.

TTS usage. The same 78-item KOTTS with modification (i.e., changing verb tenses or rephrasing statements; Appendix J) was administered to participants post-test to gauge the extent to which participants had reportedly used certain strategies or groups of strategies during the test. The directions for this measure read, "Now that you have finished taking the cognitive ability test, please indicate the extent to which you agree (or disagree) with each of the following statements, pertaining to any test-taking strategies

that you might have used when taking the test." Depending on how the KOTTS was refined, the subscale means and standard deviations of TTS usage would be analyzed.

Test-Taking metacognition. Schraw's (1997) 10-item General Monitoring Strategies Checklist (GMSC) with slight modification (e.g., changing verb tenses from present to past) was used to measure participants' levels of awareness and regulation of the test-taking process. This 10-item instrument was developed to include skills that were assumed to be related to good test-taking monitoring skills or test-taking metacognition (Pintrich & DeGroot, 1990; Presley et al., 1987; Zimmerman & Martinez-Pons, 1990). Sample items included "I asked myself periodically if I was doing well," " I found myself pausing regularly to check my comprehension," and " I changed strategies when I failed to understand a problem." The scale score is computed by taking the sum of all item scores. Cronbach's alpha of the measure was satisfactory, $\alpha = .78$; corrected item-total correlations were from .33 to .53 though, indicating that this measure might not assess a uniform construct.

Demographic questionnaire. Participants' personal data such as age, gender, race, class standing, grade point average (GPA) and standardized test score (i.e., ACT or SAT score) were obtained in a demographic survey. Note that using participants' self-reported GPA and/or standardized test scores in lieu of participants' official records is a common practice in educational-psychological research because these types of self-reported statistics are acceptably accurate. For example, Cassidy (2001) conducted a study on the accuracy of self-reported GPA and SAT scores, and found a significant correlation between self-reported and actual cumulative GPA, $r = .97, p < .01$, and between self-reports and actual total SAT score ($r = .88, p < .01$). Although the self-report of SAT

verbal subscale was less reliable ($r = .73, p < .01$) than that of the math subscale ($r = .89$, $p < .01$) in Cassidy's study, the average accuracy was still within reasonable guidelines (as suggested by Nunnally & Bernstein, 1994).

In addition, participants were asked for a history of their learning experience of TTS. The descriptive statistics obtained have been reported above.

Data Analysis

Data preparation and reverse coding of variables. I first checked the frequencies of all variables measured in this study to detect and fix any detected clerical errors in data entry. Second, I reverse-coded specified items in the established measures (i.e., motivation, metacognition) as instructed by their test constructors.

Reverse coding TTS items. As aforementioned, knowledge of TTS is defined as consisting of 2 components: strategy familiarity (i.e., participants recognizing a strategy or skill as familiar) and intention to use a strategy (i.e., participants' inclination to use a strategy or skill in a test-taking situation). In other words, it was unnecessary to reverse code items on the "strategy familiarity" scale because the target of interest was to determine to what extent participants were familiar with a certain strategy. However, participants' inclination for using a strategy theoretically depends not only on one's familiarity level with such a strategy but also on one's implicit evaluation of the effectiveness of this strategy. For example, those who indicated their willingness to use a known ineffective strategy would be expected not to be as test wise as those who are less inclined to use such a strategy. In other words, items in the "strategy inclination" scale that are deemed as ineffective (as indicated in the literature) should be reverse coded based on their content prior to any analyses.

The problem with this approach of reverse coding lies in the inconclusiveness of the extent to which a strategy is considered effective or ineffective in a certain test-taking situation. Consider Table 4, which summarizes the strategies empirically verified as either effective or ineffective in past studies. Some strategies, such as random guessing, could be both harmful and helpful depending on the situation (i.e., frequently guessing randomly vs. random guessing as the last resort). Given the ambiguity of strategy effectiveness in the literature, I made the decision to refrain from reverse coding TTS items based on item content, but proceeding to explore the psychometric properties of the knowledge scales instead. Based on inter-item correlation matrices, I would determine whether an item needed reverse coding. Subsequent scale analyses later verified the fact that almost all items were positively correlated with one another in their respective subscale; a few negative correlations were slightly below 0 (i.e., r = -.025 between Item 43 and Item 48 in the Deductive Reasoning dimension), suggesting that no TTS items should be reverse coded.

Missing data. Next, I searched through each item for any missing values in order to detect whether such data missing (between-person or within-person) was random or non-random, especially for the scales of KOTTS administered pretest and posttest. Based on this process, I dropped the data from 9 cases where participants appeared to fail to respond to a major portion of a scale(s) on purpose. Therefore, the final sample size was N = 369.

Note that I retained two cases where participants did not respond to the last few items of one TTS scale, likely because these participants accidentally missed them (i.e., skipping a page on which there were the last four items of the TTS scale). The rest of the

missing values at the item level in all measures appeared random (i.e., rarely approaching 2% of total scale items in the case of TTS scales), not revealing any noticeable pattern. Therefore, it was not necessary to employ an imputation technique to replace the remaining missing data in this study.

Data analysis. To develop and evaluate the KOTTS measure, I used reliability analyses, confirmatory and exploratory factor analyses. I used correlation analyses to test the relations of test takers' characteristics, attitudes, and their strategic knowledge and/or usage (Hypotheses 1-10).

Chapter 3

RESULTS

In this chapter, I will present the analytic results of the measure development and evaluation process, and the findings about the relationships between the KOTTS measure (knowledge and use) and test takers' attitudes, skills and dispositions, as well as with other external correlates.

KOTTS Measure Development

Refining KOTTS Measure With Reliability and Item Analyses

Overview

I used Classical Test Theory (CTT) procedures to develop the KOTTS measure. The main index of discrimination was item-total test score correlations (r_{it}, corrected for overlap) that were used to increase scale internal consistency (coefficient alpha) as suggested by Nunnally and Bernstein (1998). Arbitrarily, discriminating items are those whose item-total correlation is $r_{it} \geq .3$. However, items with $.2 > r_{it} < .3$ (moderately discriminating) might be kept in their respective scale if they improved the scale reliability value. Because the KOTTS is conceptually a multi-dimensional measure, the appropriate index of items in a subscale is the correlations of items with the subscale score instead of the total test score.

Further, item-scale test score correlations (corrected for inflation using scale reliability coefficients) were used as another item discrimination index. The minimum criterion was that an item was correlated more strongly with its own scale than with other

scales. However, if the difference between the correlation of an item in Scale A with its own scale and that with Scale B is $r_{iA} - r_{iB} \leq .06$ (an arbitrary value), this item might be dropped from Scale A, after taking into account the item content. In case that an item was correlated more strongly to another scale than with its own scale, the item might be switched to the other scale if its content fit the content of the latter scale better. The item analysis process was repeated and coefficient alphas were recomputed until I reached the point where the reliability coefficient of a scale leveled off or began to decrease because I had selected the best possible items to be used on the final version of the KOTTS measure.

Reliability Analysis (Wave 1)

With the data that I had collected from the measure development sample, I first conducted the reliability analyses for 11 a priori dimensions of TTS knowledge, then refining each subscale based on higher item-total correlations corrected for overlap to produce relatively homogeneous scales (higher scale reliability). Table 6 describes the psychometric properties of each scale (original and refined) and scale items, followed by a detailed description of decisions made concerning the first-wave scale refinement process.

Table 6

Internal Consistencies of 11 TTS Dimensions: Original and Refined Scales (Wave 1)

Dimension	Item			Scale Cronbach alpha (unstandardized)	
	Item	Corrected item-total r_{it}			
	No.	Original	Refined	Original	Refined
(1) Knowing how multiple-choice tests work	**8**	.37	.47	.55	.57
	9	.33	.31		
- Originally 9 items	**7**	.31	.43		(Items 1-6 were
- Refined scale: 3 items	2	.29	-		dropped)
	3	.25	-		
	4	.25	-		
	1	.24	-		
	5	.19	-		
	6	.17			
(2) Optimizing time efficiency & effectiveness	**16**	.50	.51	.76	.75
	18	.47	.50		
- Originally 13 items	**19**	.46	.45		(Items 10 & 20
- Refined scale: 11 items	**11**	.43	.41		were dropped)
	15	.41	.39		
	17	.38	.36		
	14	.36	.37		
	12	.36	.38		
	22	.34	.34		
	21	.33	.32		
	13	.32	.30		
	10	.29	-		
	20	.27	-		
(3) Avoiding clerical errors	**23**	.53	.61	.66	.75
	24	.49	.54		
- Originally 6 items	**27**	.48	.53		(Item 28 was
- Refined scale: 5 items	**26**	.45	.42		dropped)
	25	.44	.51		
	28	.17	-		
(4) Using physical cues	**29**	.56	Same	.67	.67
- Originally 4 items	**30**	.53			
- Refined scale: 4 items	**32**	.43			
	31	.33			
(5) Using grammatical or contextual cues	**38**	.53	Same	.76	.76
	35	.50			
- Originally 7 items	**39**	.49			
- Refined scale: 7 items	**37**	.49			
	34	.46			
	36	.45			
	33	.40			

(6) Deductive reasoning	**45**	.51	.54	.73	.71
- Originally 10	**47**	.47	.42		
items	**42**	.42	.46		(Items 40 & 48
- Refined scale: 8	**46**	.42	.39		were dropped)
items	**43**	.41	.46		
	49	.38	.34		
	41	.34	.30		
	48	.33	-		
	44	.32	.34		
	40	.28	-		
(7) Guessing	**50**	.34	.40 (7a)	.58	(7a) .58
- Originally 5 items	**51**	.37	.47 (7a)		(7b) .69
- 2 Refined scales:	**52**	.42	.53 (7b)		
(7a) 3 items &	**53**	.30	.53 (7b)		
(7b) 2 items	**54**	.28	.33 (7a)		
(8) Changing answers	57	.40	N/a	.36	N/a
(3 items)	55	.18			(Dropped all)
	56	.08			
(9) Working carefully &	**61**	.54	Same	.76	.76
thoroughly	**64**	.54			
- Originally 7 items	**59**	.54			
- Refined scale: 7	**62**	.54			
items	**63**	.44			
	60	.39			
	58	.38			
(10) Staying in control	**72**	.58	Same	.77	.77
- Originally 8 items	**71**	.57			
- Refined scale: 8	**69**	.52			
items	**65**	.51			
	68	.45			
	70	.42			
	66	.40			
	67	.38			
(11) Trouble-shooting and	**75**	.57	Same	.74	.74
using recall aids	**77**	.55			
- Originally 6 items	**76**	.52			
- Refined scale: 6	**74**	.48			
items	**78**	.43			
	73	.36			

Note. Bolded fonts indicate that the items are retained in the refined subscales.

Dimension 1: Knowing how multiple-choice tests work. The internal consistency of this 9-item scale was low ($\alpha = .55$); the corrected item-total correlations (r_{it}) were between .17 and .37. Six items (Items 1 to 6) were dropped from this scale because they had lower item-total correlations than the cutoff value of .3 and because their content did not clearly fit in with the scale content. I then reconducted a reliability analysis for the revised 3-item scale: the Cronbach alpha was .57, and the r_{it}'s ranged from .31 to .47.

Dimension 2: Optimizing time efficiency and effectiveness. The internal consistency of this 13-item scale was satisfactory ($\alpha = .76$). However, a couple of items (Items 10 and 20) had lower corrected item-total correlations than the cutoff point ($r_{it} = .29$, and $r_{it} = .27$ respectively). Considering item content, Item 10, "If a question is taking too long, I guess and go on to the next question, then come back to it at the end if I have time," was wordy, and Item 20, "In order to save time, I put slash marks through clearly wrong answers on the test booklet (if allowed) so that I would not waste time rereading them," read more like a strategy of elimination incorrect answers than that of time management. Therefore, these 2 items were removed from the scale. The internal consistency for the refined 11-item scale was $\alpha = .75$.

Dimension 3: Avoiding clerical errors. The internal consistency was $\alpha = .66$ for this 6-item scale; the r_{it}'s ranged from .44 to .53, except for Item 28 whose item-total correlation ($r_{it} = .17$) was lower than the cutoff value. The item content ("To reduce clerical errors, I mark my answers on the test booklet (if allowed) as I work with the questions, and periodically transfer a block of answers to the answer sheet") was more closely related to a marking strategy than the strategy of avoiding errors. Therefore, Item

28 was dropped from the scale. The Cronbach alpha for the refined 5-item scale was .75, and the r_{it}'s ranged from .42 to .61.

Dimension 4: Using physical cues. The internal consistency of the 4-item scale was $\alpha = .67$; the r_{it}'s ranged from .33 to .56. Two items, 29 and 30, were more highly correlated with each other than with the other 2 items, 31 and 32; in fact, removing Items 31 and 32 from the scale would increase scale reliability (i.e., $\alpha = .80$). However, the content of Items 29 and 30 was similar ("Among the options, I choose the answer that is longer than other options" and "I choose the answer that is shorter than the rest of the options"), whereas the content of Items 31 and 32 added meaningfully to the scale. Therefore, I made the decision to keep this scale intact.

Dimension 5: Using grammatical or contextual cues. The internal consistency of the 7-item scale was satisfactory ($\alpha = .76$); the r_{it}'s were moderately strong (from .40 to .53), indicating that this scale might not need further refinement.

Dimension 6: Deductive reasoning. The internal consistency of this 10-item scale was satisfactory ($\alpha = .73$). Item 40 had a lower r_{it} (.28) than the cutoff point; its content ("When responding to a multiple-choice test item, I should first eliminate apparent incorrect answer choices after careful scrutiny") was not as clear as another item with similar content (Item 43, "I eliminate answer choices which have some similarities"). Therefore, I dropped this item from the scale. A subsequent reliability analysis with the 9-item scale indicated that Item 48 should also be dropped because of its low r_{it} (.274) and its content was redundant ("I eliminate options which are known to be incorrect and choose from among the remaining options"), considering the similar content of Item 43.

The refined scale then consisted of 8 items with r_{it}'s ranging from .30 to .54 and an internal consistency value of .71.

Dimension 7: Guessing. The internal consistency of this 5-item scale was low (α = .58); the r_{it}'s were from .28 to .42. Two items, 52 and 53, were correlated much higher with each other (r = .53) than with other items in the scale. A peruse of item content revealed that both items 52 and 53 measured a similar concept ("I tend to guess a particular choice (e.g., A or C) as soon as something looks unfamiliar or difficult," and "I guess as soon as something looks unfamiliar or difficult"), whereas the other 3 items assessed a different approach toward guessing (i.e., "I guess only after eliminating as many wrong answers as I can"). Therefore, I made the decision to split this dimension into two scales, (7a) and (7b). Scale 7a was renamed *Guessing*, consisting of 3 items 50, 51 and 54. Its alpha level was relatively low (.59) but the r_{it}'s were moderate (from .34 to .42). Scale 7b was renamed *Careless guessing* (inferred from the item content), including Items 52 and 53; the scale reliability was α = .69.

Dimension 8: Changing answers. The internal consistency of this 3-item scale was very poor (α = .36); the r_{it}'s were very low except for that of Item 57. Item 55, "I go with my first instinct and don't change my answers," and Item 56, "I do not hesitate to change my answers if I feel I should," did not relate to each other (r = -.09), which made sense considering the opposite meanings of these 2 items, although they were both positively and significantly correlated with Item 57 (.34 and .20, respectively). This fact suggested that either new, better items should be added to this scale, or this scale should be dropped altogether. Given the content of the dimension that makes writing new items unlikely, I decided to drop this scale from subsequent analyses.

Dimension 9: Working carefully and thoroughly. The internal consistency of this 7-item scale was satisfactory ($\alpha = .76$); the r_{it}'s were moderate to moderately high (from .38 to .54), indicating that this scale might not need further refinement.

Dimension 10: Staying in control. The internal consistency of this 8-item was satisfactory ($\alpha = .77$); the r_{it}'s were moderate to moderately high (from .38 to .57), indicating that this scale might not need further refinement.

Dimension 11: Trouble-shooting and using recall aids. The internal consistency of this 8-item scale was satisfactory ($\alpha = .74$); the r_{it}'s were moderate to moderately high (from .36 to .57), indicating that this scale might not need further refinement.

Obtaining Item-Scale Correlations and Refining Scales (Wave 1)

I continued to use another item discrimination index, the item-scale correlations, to detect whether scale items were correlated with their own scale mean more than with that of other scales. Examining the zero-order correlation matrices (Tables 7 to 17) in the following sections, I identified items that were either misplaced (i.e., correlating more strongly with another scale mean instead of their own scale mean) or non-discriminating (i.e., relating as highly to other scales as to their own).

Dimension 1: Knowing how multiple-choice tests work. Table 7 shows that Item 9, "I am familiar with the columns and ovals on a multiple-choice answer sheet and how to completely fill in the ovals," should be switched to Dimension 3, *Avoiding clerical errors,* because this item was correlated higher with this scale ($r = .43$) than with its own scale ($r = .31$), and its meaning logically fits Dimension 3. Therefore, there were 2 items remaining in Dimension 1 (7 and 8). The new scale internal consistency was $\alpha = .59$; item correlation was $r = .42$.

Table 7

Item-scale Correlation Matrix: Dimension 1

Item	D1[a]	D2	D3	D4	D5	D6	D7a	D7b	D9	D10	D11
7	*.43*	.24	.20	.04	.16	.10	.20	.04	.20	.13	.13
8	*.47*	.36	.32	.14	.27	.25	.32	.07	.28	.21	.24
9	*.31*	.24	.43	-.07	.22	.09	.26	-.07	.30	.17	.27

Note. [a] Italic fonts indicate that *r*'s were corrected for overlapping.

Dimension 2: Optimizing time efficiency and effectiveness. Table 8 shows that Item 15 ("I pay attention to how much time is left so I can finish (e.g. a test section) in the allotted time") and Item 21 ("I know that time element is a factor in taking tests, but accuracy should not be sacrificed for speed") should be dropped from this dimension because they were correlated more strongly with at least another scale (i.e., Dimension 3; Dimension 11) than with their own scale, but the relationships were not strong enough nor discriminating enough to consider rearranging them. Item 13 ("To get all of the easier questions answered before time runs out, I mark questions that can use further consideration and those that are very unfamiliar or difficult, skip them and go back to work on them later"), and Item 22 ("I take the same time I might spend on the single hardest question to answer three easier questions") were correlated non-discriminatingly with other scales besides their own scale according to the aforementioned rule of thumb (*r* difference < .06). Their content was either wordy or did not clearly fit this dimension; therefore, they were dropped from the scale. The new scale then consisted of 7 items (11, 12, 14, 16, 17, 18, and 19); its internal consistency was satisfactory, $\alpha = .71$; r_{it}'s ranged from .31 to .53.

Table 8

Item-scale Correlation Matrix: Dimension 2

Item	D1	D2[a]	D3	D4	D5	D6	D7a	D7b	D9	D10	D11
11	.33	*.41*	.23	.08	.26	.19	.16	.17	.32	.27	.23
12	.05	*.38*	.09	.18	.09	.15	.03	.20	.09	.12	.08
13	.08	*.30*	.27	-.05	.15	.12	.16	.05	.30	.18	.25
14	.11	*.37*	.12	.17	.30	.29	.12	.10	.07	.13	.19
15	.31	*.39*	.40	.06	.33	.25	.32	-.01	.33	.26	.34
16	.21	*.51*	.19	.28	.31	.36	.13	.15	.28	.26	.24
17	.23	*.36*	.22	.08	.14	.21	20	.20	.17	.19	.21
18	.19	*.50*	.18	.21	.22	.17	.12	.16	.20	.20	.21
19	.19	*.45*	.26	.13	.34	.28	.22	.06	.17	.13	.30
21	.28	*.32*	.33	.01	.30	.24	.21	.01	.30	.28	.32
22	.28	*.34*	.27	.07	.25	.17	.21	.04	.17	.25	.27

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 3: Avoiding clerical errors. Examining the item-scale correlations in Table 9, I found that all items in this scale were correlated more strongly with their own scale than with other scales. However, Item 26 ("I check the general accuracy of my work at the end of a test section") related non-discriminatingly with 2 other dimensions (Dimensions 9 and 11) according the rule of thumb (*r* difference < .06) and thus should be dropped from the scale. Including Item 9 (being switched from Dimension 1), this refined 5-item scale had a satisfactory level of internal consistency ($\alpha = .76$) and moderately strong r_{it}'s (from .46 to .58).

Table 9

Item-scale Correlation Matrix: Dimension 3

Item	D1	D2	D3 [a]	D4	D5	D6	D7a	D7b	D9	D10	D11
23	.32	.33	*.61*	-.03	.28	.15	.25	-.05	.39	.31	.35
24	.28	.29	*.54*	.02	.25	.08	.24	.01	.38	.33	.30
25	.27	.31	*.51*	-.11	.26	.13	.23	-.05	.34	.26	.29
26	.21	.32	*.42*	.07	.26	.20	.14	.04	.41	.34	.40
27	.31	.25	*.53*	-.05	.21	.12	.38	-.08	.35	.23	.30

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 4: Using physical cues. All items in this scale (Items 29-32) were correlated more strongly with their own scale than with other scales (Table 10). Therefore, no change was made to this scale.

Table 10

Item-scale Correlation Matrix: Dimension 4

Item	D1	D2	D3	D4 [a]	D5	D6	D7a	D7b	D9	D10	D11
29	.04	.17	-.01	*.56*	.17	.28	.01	.21	-.04	.08	.05
30	.08	.20	-.08	*.53*	.10	.19	-.06	.22	-.08	.10	.00
31	.03	.06	-.01	*.33*	.17	.16	.04	.18	-.01	.01	.04
32	.10	.21	.04	*.43*	.29	.20	.13	.20	.03	.13	.17

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 5: Using grammatical or contextual cues. Table 11 shows that Item 33 ("I use relevant content information in other test items and options to answer an item (e.g., one item "giving away" the answer to another question") was correlated more strongly and non-discriminatingly with 2 other scales (Dimension 3 & Dimension 11) than with its own scale. Therefore, it was dropped from the scale. Including Item 49

(being switched from Dimension 6), the refined 7-item scale (Items 34-39 and 49) had a satisfactory internal consistency, $\alpha = .76$; r_{it}'s ranged from .42 to .53.

Table 11

Item-scale Correlation Matrix: Dimension 5

Item	D1	D2	D3	D4	D5[a]	D6	D7a	D7b	D9	D10	D11
33	.331	.322	.498	.043	*.402*	.224	.315	-.033	.360	.275	.497
34	.170	.353	.134	.285	*.463*	.329	.104	.080	.206	.218	.257
35	.222	.354	.213	.232	*.502*	.291	.219	.111	.222	.268	.356
36	.098	.254	.172	.136	*.450*	.314	.087	.092	.229	.212	.265
37	.110	.265	.124	.206	*.489*	.327	.149	.028	.176	.276	.230
38	.269	.311	.378	.098	*.532*	.307	.247	.067	.368	.272	.455
39	.110	.199	.158	.177	*.494*	.294	.158	.077	.194	.200	.241

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 6: Deductive reasoning. Table 12 shows that Item 49 ("Among the choices, I look for the answer that converges ALL the dimensions in the stem") was correlated more strongly with another scale (Dimension 5, *Using grammatical or contextual scale, r* = .41) than with its own scale (*r* = .34; *r* difference > .06). Its content also fit Dimension 5 better; therefore, it was switched to Dimension 5. Item 41 ("I restrict my choice to those options which encompass all of two or more given statements known to be correct") and Item 44 ("I try to work backwards from the answers to the stem, especially with math questions") were non-discriminatingly correlated with other scales (i.e., Dimension 11; Dimension 5) according to the *r* difference criterion. Therefore, they

were dropped from this scale. The refined Dimension 6 then consisted of 5 items (Items 42, 43, 45, 46 and 47) with internal consistency $\alpha = .68$ and r_{it}'s ranging from .41 to .50.

Table 12

Item-scale Correlation Matrix: Dimension 6

Item	D1	D2	D3	D4	D5	D6[a]	D7a	D7b	D9	D10	D11
41	0.133	0.241	0.215	0.106	0.234	*0.300*	0.162	0.129	0.208	0.186	0.247
42	0.102	0.242	0.033	0.185	0.171	*0.462*	0.041	0.178	0.101	0.161	0.124
43	-0.067	0.162	-0.079	0.289	0.178	*0.458*	0.035	0.185	0.082	0.113	0.070
44	0.126	0.232	0.087	0.233	0.286	*0.336*	0.084	0.099	0.132	0.152	0.179
45	0.118	0.225	0.031	0.263	0.266	*0.538*	0.116	0.059	0.145	0.169	0.173
46	0.150	0.257	0.212	0.097	0.279	*0.394*	0.274	0.038	0.273	0.204	0.284
47	0.184	0.275	0.231	0.059	0.343	*0.417*	0.251	-0.062	0.268	0.229	0.350
49	0.180	0.296	0.176	0.113	0.411	*0.343*	0.154	0.059	0.266	0.218	0.286

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 7a: Guessing. Table 13 shows that Item 54 ("I guess only after eliminating as many wrong answers as I can") should be switched to Dimension 9, *Working carefully and thoroughly,* because it was correlated more strongly with this scale ($r = .43$) than with its own scale ($r = .33$) and because the item content also inferred carefulness. The refined 2-item scale (50 and 51) had a relatively low internal consistency, $\alpha = .58$. The item correlation was .41.

Table 13

Item-scale Correlation Matrix: Dimension 7a

Item	D1	D2	D3	D4	D5	D6	D7a[a]	D7b	D9	D10	D11
50	0.269	0.256	0.161	0.152	0.179	0.189	*0.402*	0.152	0.149	0.102	0.215
51	0.202	0.180	0.214	0.003	0.162	0.105	*0.474*	0.139	0.158	0.121	0.166
54	0.265	0.247	0.373	-0.046	0.266	0.250	*0.330*	0.108	0.431	0.198	0.310

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 7b: Careless guessing. Table 14 shows that both items in this scale were correlated more strongly with their own scale than with any other scale.

Table 14

Item-scale Correlation Matrix: Dimension 7b

Item	D1	D2	D3	D4	D5	D6	D7a	D7b[a]	D9	D10	D11
52	0.063	0.174	0.022	0.215	0.116	0.112	0.226	*0.526*	0.142	0.114	0.066
53	0.000	0.169	-0.078	0.295	0.053	0.154	0.081	*0.526*	0.044	0.156	0.017

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Dimension 9: Working carefully and thoroughly. All items in this scale were correlated more strongly with their own scale than with other scales (Table 15). However, Item 58 ("I will reconsider my answers if I finish the test before time is called") and Item 60 ("I give thought to what the answer should include before reading the answer choices") were also non-discriminatingly correlated with other dimensions (i.e., Dimension 3; Dimension 10 and 11; r difference < .06); their content also did not fit their scale content well. Therefore, these 2 items were removed from this scale. Combined with Item 54 (being switched from Dimension 7a), the refined 6-item scale (Items 59, 61, 62, 63, 64, and 54) had satisfactory internal consistency, $\alpha = .76$; r_{it}'s ranged from .41 to .55.

Table 15

Item-scale Correlation Matrix: Dimension 9

Item	D1	D2	D3	D4	D5	D6	D7a	D7b	D9[a]	D10	D11
58	0.152	0.226	0.342	0.035	0.216	0.100	0.151	0.075	*0.378*	0.246	0.248
59	0.227	0.275	0.418	-0.033	0.203	0.135	0.252	-0.006	*0.535*	0.235	0.320
60	0.235	0.295	0.211	0.070	0.298	0.284	0.182	0.172	*0.386*	0.349	0.332
61	0.228	0.281	0.309	0.020	0.372	0.353	0.257	0.079	*0.539*	0.329	0.359
62	0.225	0.225	0.366	-0.070	0.203	0.179	0.159	0.002	*0.534*	0.339	0.285
63	0.238	0.231	0.298	-0.091	0.182	0.183	0.263	0.142	*0.444*	0.360	0.323
64	0.161	0.240	0.424	-0.081	0.223	0.196	0.154	0.000	*0.536*	0.359	0.352

Note. [a] Italic fonts indicate that r's were corrected for inflation.

Dimension 10: Staying in control. Table 16 shows that Item 66 ("If I get annoyed for some reason (e.g., an "off-the-wall" question) and my concentration lapses, I will take a quick break, relaxing and regrouping for a minute") should be dropped because it was correlated more strongly with Dimension 11 (r = .40) than that with its own scale (r = .40). After conducting a reliability analysis, I found that removing 2 items, Item 67 ("During breaks (if any), I do not talk to other test takers: I keep my energies focused on the test") and Item 70 ("If I get annoyed for any reason during the test, I try looking over the material I have successfully completed to recover and remotivate myself"), would increase the scale reliability. Also, the content of these items did not fit well in this scale. The refined 5-item scale (Items 65, 68, 69, 71 and 72) had a good internal consistency, α = .78; r_{it}'s ranged from .51 to .67.

Table 16

Item-scale Correlation Matrix: Dimension 10

Item	D1	D2	D3	D4	D5	D6	D7a	D7b	D9	D10[a]	D11
65	0.175	0.271	0.314	-0.016	0.290	0.278	0.181	0.028	0.412	*0.513*	0.372
66	0.150	0.286	0.224	0.101	0.184	0.158	0.109	0.174	0.261	*0.398*	0.404
67	0.014	0.160	0.180	0.146	0.182	0.115	-0.046	0.036	0.151	*0.382*	0.209
68	0.228	0.220	0.325	-0.002	0.217	0.178	0.302	0.050	0.345	*0.446*	0.275
69	0.218	0.239	0.346	0.133	0.252	0.195	0.228	0.139	0.323	*0.517*	0.274
70	0.118	0.242	0.239	0.172	0.280	0.199	0.019	0.261	0.309	*0.420*	0.336
71	0.096	0.213	0.214	0.042	0.239	0.216	0.065	0.057	0.338	*0.568*	0.278
72	0.134	0.251	0.263	0.003	0.237	0.227	0.150	0.013	0.375	*0.576*	0.216

Note. [a] Italic fonts indicate that r's were corrected for inflation.

Dimension 11: Trouble-shooting and using recall aids. Table 17 shows that Item 73 ("If I draw a temporary blank, I recite the question to myself, or even write it out to help recall the material") should be dropped for correlating more strongly with Dimension 10 (r = .37) than with its own scale (r = .36). Item 75 ("If I am uncertain of what a word or phrase means, I try to resolve the vagueness or ambiguity by defining the word or phrase according to what I think it might be, then use the definition to solve the problem") should be dropped because it was correlated non-discriminatingly with Dimension 5 besides its own scale (r difference < .06). Likewise, Item 78 ("If an item looks unfamiliar, I brainstorm about what I know about the problem topic, even if it seems only tangentially related, to get enough information to answer the question correctly") should be dropped for relating non-discriminatingly to Dimension 10 (r difference < .06) and because the item was wordy. The refined scale then consisted of 3 items (Items 74 and 76-77) with the internal consistency α = .648; r_{it}'s ranged from .44 to .50.

Table 17

Item-scale Correlation Matrix: Dimension 11

Item	D1	D2	D3	D4	D5	D6	D7a	D7b	D9	D10	D11[a]
73	.204	.295	.300	.129	.239	.227	.130	.037	.285	.371	*.359*
74	.117	.350	.310	.053	.263	.239	.171	.012	.281	.264	*.481*
75	.212	.358	.356	.108	.525	.336	.311	.057	.411	.404	*.566*
76	.123	.181	296	.049	.258	.165	.161	.001	.290	.210	*.521*
77	.175	.318	.315	.036	.327	.260	.252	.051	.365	.304	*.549*
78	.206	.233	.271	.045	.349	.256	.160	.045	.362	.372	*.431*

Note. [a] Italic fonts indicate that *r*'s were corrected for inflation.

Item-scale Analyses (Wave 2)

After each step of refining a dimension based on the discrimination index of item-scale correlations, I had recomputed each scale reliability coefficient, which were reported in the sections above, and obtained its corrected item-total correlations. Because the values of these r_{it}'s were greater than the cutoff point of .3, I conducted another item-scale correlational analysis (Table 18) to see whether the scales could be further refined to improve their discriminant validity while maintaining an acceptable level of reliability. Note that the correlations between a set of scale items and their own scale were corrected for overlap.

In Dimension 1, I found that Item 8, "I know that I may work only on the section the test administrator designates and only for the time allowed (e.g., I may not go back to an earlier test section)," was correlated non-discriminatingly with Dimension 2 (*r* difference = .06). However, consider the item content was more related to its own scale than with Dimension 2, I decided to keep Dimension 1 unchanged.

Table 18

Item-Scale Correlations After Scale Refinement

Item	D1	D2	D3	D4	D5	D6	D7A	D7B	D9	D10	D11
7	**0.42**	0.24	0.20	0.04	0.14	0.10	0.20	0.04	0.20	0.13	0.09
8	**0.42**	0.36	0.32	0.14	0.27	0.17	0.32	0.07	0.28	0.21	0.17
11	0.33	**0.379**	0.23	0.08	0.27	0.12	0.16	0.17	0.32	0.27	0.14
12	0.05	**0.383**	0.09	0.18	0.09	0.13	0.02	0.20	0.09	0.12	0.06
14	0.11	**0.383**	0.12	0.17	0.30	0.26	0.12	0.10	0.07	0.13	0.20
16	0.21	**0.513**	0.19	0.28	0.33	0.30	0.13	0.15	0.28	0.26	0.20
17	0.23	**0.311**	0.22	0.08	0.15	0.19	0.20	0.20	0.17	0.19	0.20
18	0.18	**0.529**	0.18	0.21	0.23	0.11	0.12	0.16	0.20	0.20	0.16
19	0.18	**0.420**	0.26	0.12	0.35	0.24	0.22	0.06	0.17	0.13	0.24
23	0.32	0.33	**0.583**	-0.02	0.23	0.13	0.25	-0.05	0.39	0.31	0.32
24	0.28	0.29	**0.484**	0.02	0.20	0.03	0.24	0.00	0.38	0.33	0.24
25	0.27	0.31	**0.582**	-0.11	0.20	0.10	0.23	-0.05	0.34	0.26	0.27
27	0.31	0.25	**0.568**	-0.05	0.16	0.06	0.38	-0.08	0.34	0.23	0.25
9	0.53	0.24	**0.464**	-0.07	0.17	0.05	0.26	-0.07	0.30	0.17	0.19
29	0.04	0.17	-0.01	**0.556**	0.19	0.25	0.01	0.21	-0.04	0.08	0.03
30	0.08	0.20	-0.08	**0.528**	0.13	0.17	-0.06	0.22	-0.08	0.10	-0.01
31	0.03	0.06	-0.01	**0.326**	0.17	0.14	0.04	0.18	-0.01	0.01	0.01
32	0.09	0.21	0.04	**0.434**	0.28	0.16	0.13	0.20	0.03	0.13	0.11
34	0.17	0.35	0.13	0.28	**0.46**	0.27	0.10	0.08	0.21	0.22	0.20
35	0.22	0.35	0.21	0.23	**0.48**	0.24	0.22	0.11	0.22	0.27	0.27
36	0.10	0.25	0.17	0.14	**0.47**	0.27	0.09	0.09	0.23	0.21	0.21
37	0.11	0.26	0.12	0.21	**0.53**	0.21	0.15	0.03	0.18	0.28	0.14
38	0.27	0.31	0.38	0.10	**0.50**	0.28	0.25	0.07	0.37	0.27	0.37
39	0.11	0.20	0.16	0.18	**0.50**	0.20	0.16	0.08	0.19	0.20	0.17

49	0.18	0.30	0.18	0.11	**0.42**	0.30	0.15	0.06	0.27	0.22	0.22
42	0.10	0.24	0.03	0.19	0.21	**0.44**	0.04	0.18	0.10	0.16	0.10
43	-0.07	0.16	-0.08	0.29	0.19	**0.44**	0.03	0.19	0.08	0.11	0.04
45	0.12	0.22	0.03	0.26	0.30	**0.50**	0.12	0.06	0.15	0.17	0.11
46	0.15	0.26	0.21	0.10	0.27	**0.41**	0.27	0.04	0.27	0.20	0.25
47	0.18	0.27	0.23	0.06	0.34	**0.41**	0.25	-0.06	0.27	0.23	0.33
50	0.27	0.26	0.16	0.15	0.17	0.19	**0.407**	0.15	0.15	0.10	0.19
51	0.20	0.18	0.21	0.00	0.13	0.11	**0.407**	0.14	0.16	0.12	0.13
52	0.06	0.17	0.02	0.22	0.13	0.08	0.23	**0.526**	0.14	0.11	0.04
53	0.00	0.17	-0.08	0.29	0.07	0.14	0.08	**0.526**	0.04	0.16	0.01
59	0.23	0.27	0.42	-0.03	0.17	0.13	0.25	-0.01	**0.46**	0.23	0.28
61	0.23	0.28	0.31	0.02	0.38	0.30	0.26	0.08	**0.55**	0.33	0.27
62	0.23	0.23	0.37	-0.07	0.20	0.14	0.16	0.00	**0.54**	0.34	0.19
63	0.24	0.23	0.30	-0.09	0.18	0.15	0.26	0.14	**0.49**	0.36	0.29
64	0.16	0.24	0.42	-0.08	0.22	0.16	0.15	0.00	**0.54**	0.36	0.28
54	0.27	0.25	0.37	-0.05	0.26	0.21	0.65	0.11	**0.41**	0.20	0.26
65	0.17	0.27	0.31	-0.02	0.29	0.28	0.18	0.03	0.41	**0.51**	0.27
68	0.23	0.22	0.32	0.00	0.19	0.20	0.30	0.05	0.34	**0.52**	0.21
69	0.22	0.24	0.35	0.13	0.22	0.17	0.23	0.14	0.32	**0.53**	0.17
71	0.10	0.21	0.21	0.04	0.24	0.18	0.07	0.06	0.34	**0.52**	0.17
72	0.13	0.25	0.26	0.00	0.24	0.19	0.15	0.01	0.38	**0.67**	0.15
74	0.12	0.35	0.31	0.05	0.25	0.20	0.17	0.01	0.28	0.26	**0.45**
76	0.12	0.18	0.30	0.05	0.23	0.13	0.16	0.00	0.29	0.21	**0.44**
77	0.17	0.32	0.32	0.04	0.29	0.24	0.25	0.05	0.36	0.30	**0.50**

Note. Bolded numbers indicate that the correlation between an item and its own scale was corrected for inflation.

In Dimension 2, Item 11, "I vary my test-taking speed from test section to section, depending on how much time I have for a particular section," was non-discriminatingly correlated with 2 other scales besides its own scale. Although removing this item reduced the scale reliability, I decided to drop Item 11 from the scale and conducted another reliability analysis. This new dimension has an acceptable internal consistency, $\alpha = .68$; r_{it}'s ranged from .30 to .51.

In Dimension 3, Item 9, which had been switched from Dimension 1 based on a strong item-scale correlation with Dimension 3, now related more strongly with Dimension 1 than with Dimension 3. Given the unstable nature of this item, I decided to drop it instead of rearranging it; then I recomputed the scale reliability, yielding a coefficient $\alpha = .75$; r_{it}'s ranged from .49 to .59.

In Dimension 9, Item 59, "At the end of a test section, I go back to the test questions that gave me difficulty and verify my work on them," was non-discriminatingly related to Dimension 3. Item 54, which had been switched from Dimension 7a based on a stronger item-scale correlation with Dimension 9, now related more strongly with Dimension 7a than with Dimension 9. Both items were dropped from this scale. The internal consistency of this 4-item scale was $\alpha = .72$; r_{it}'s ranged from .49 to .54.

There were no noticeable changes associated with Dimensions 4, 5, 6, 7a, 7b, 10 and 11. Note that in subsequent sections, Dimension 7a (Guessing) was renumbered Dimension 7, and Dimension 7b (Careless guessing) became Dimension 8.

Table 19 presents the KOTTS measure with its 11 refined dimensions and the respective strategies.

Table 19

The Knowledge of Test-Taking Strategies (KOTTS) Refined Measure, k = 44

Old No.	New No.	Dimension & Item formulation
		D1. KNOWING HOW MULTIPLE-CHOICE TESTS WORK
7	1	- I am aware that some exams may penalize you for wrong answers (e.g., deducting points from your total score).
8	2	- I know that I may work only on the section the test administrator designates and only for the time allowed (e.g., I may not go back to an earlier test section).
		D2. OPTIMIZING TIME EFFICIENCY AND EFFECTIVENESS
12	1	- To save time, I scan the test for more time-consuming/difficult questions and leave them for last.
14	2	- To save time, I memorize the (unchanged) directions for each type of questions in advance and only skimming through them when taking tests.
16	3	- I figure out how many minutes per question on average and spend the same amount of time on each question.
17	4	- I work as rapidly as possible with reasonable assurance of accuracy.
18	5	- I first scan the test for question types (e.g., certain types of questions require more thoughts and processing) and plan strategy accordingly (e.g., budgeting my time).
19	6	- I look for "short-cuts" to save time (e.g., doing a quick estimate to quickly eliminate some answer possibilities, or dividing an unfamiliar word into its prefix, suffix and root).
		D3. AVOIDING CLERICAL ERRORS
23	1	- I am careful not to make or leave any stray marks on my (machine-scored) answer sheet.
24	2	- I periodically check my answers to catch careless or clerical mistakes.
25	3	- When I skip a question, I remember to skip the corresponding row of answer choices on my answer sheet.
27	4	- I erase the initial answer completely (when I change my mind about an answer).
		D4. USING PHYSICAL CUES
29	1	- Among the options, I choose the answer that is longer than other options.
30	2	- Among the options, I choose the answer that is shorter than the rest of the options.
31	3	- I do not answer a series of questions with the same letter choice (e.g., all *A*s).
32	4	- I know that the test constructor may place the correct answer in a certain physical position among the options (e.g., in the middle).
		D5. USING GRAMMATICAL OR CONTEXTUAL CUES
34	1	- I consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.
35	2	- I notice that the test constructor may qualify the correct answer more carefully (e.g., more precise and specific in meaning), or make it represent a higher degree of generalization.
36	3	- I pay particular attention to negatives.
37	4	- I recognize and make use of resemblance between the options and an aspect of the stem.
38	5	- I recognize and make use of specific determiners (e.g., always, never), disclaimers (e.g., best, all, or none), and/or "hedging" words (e.g., probably, most likely).
39	6	- I know that the test constructor may make the correct answer grammatically consistent with the stem.
49	7	- Among the choices, I look for the answer that converges ALL the dimensions in the stem.

		D6. DEDUCTIVE REASONING
42	1	- I choose neither of two options which imply the correctness of each other.
43	2	- I eliminate answer choices which have some similarities.
45	3	- I choose one of two statements, which, if correct, would imply the incorrectness of the other.
46	4	- I eliminate the choice "*all of the above*" when there are opposite choices in the answers.
47	5	- I rule out choices that contradict the question.
		D7. GUESSING
50	1	- I know if I randomly choose an answer out of four options, I'll have 25% chance of getting it correctly.
51	2	- If there are only a few minutes left for a test/test section and there is no penalty for wrong answers, I will fill in the remaining problems with guesses (randomly or picking a particular answer choice such as B or C) before time is called.
		D8. CARELESS GUESSING
52	1	- I tend to guess a particular choice (e.g., A or C) as soon as something looks unfamiliar or difficult.
53	2	- I guess as soon as something looks unfamiliar or difficult.
		D9. WORKING CAREFULLY AND THOROUGHLY
61	1	- I read the test items carefully, determining clearly the nature of the question.
62	2	- I read all instructions/directions carefully to make sure I understand them, determining clearly the nature of the task and the intended basis for response.
63	3	- When I do not understand something about directions, I ask the examiner for clarification.
64	4	- I read all information provided, even when I see an immediate answer.
		D10. STAYING IN CONTROL
65	1	- I avoid internal distractions by directing attention away from negative self-evaluative thoughts.
68	2	- I don't panic if I cannot answer a question: I keep calm and move on.
69	3	- I do not become impatient or discouraged if I cannot start at once effectively.
71	4	- During the test, I divorce myself from everything (e.g. the proctor, my surrounding), and let no disturbance influence my concentration.
72	5	- I don't waste time worrying during the test but work efficiently.
		D11. USING RECALL AIDS
74	1	- Before answering test questions, I note in the margins of my test booklet (if allowed) certain short-term memory items.
76	2	- If I draw a temporary blank, I try to visualize the place in the book where this material appeared—this might lead me to relate facts and increase chances of recall.
77	3	- I sometimes make a quick drawing or a formula to clarify the questions and initiate and aid in recall.

Obtaining Inter-Scale Correlations for the Refined Scales

The last step was to conduct inter-scale correlational analyses to find further evidence of discriminant validity of the KOTTS subscales. Table 20 shows that, although a majority of the scale means were positively correlated with each other (after being corrected for inflation), the strengths of these relationships ranged from very weak (i.e., r = .04) to moderate. For example, the strongest corrected relation was between Dimension 9 and Dimension 10, r = .66, a value that was smaller than the cutoff point of r > .80. This fact indicated that the measure of KOTTS seemed to be multi-dimensional, unlikely to be assessing one general factor of TTS knowledge.

Furthermore, some dimensions tended to relate better to certain dimensions than to other ones. For example, Dimension 3 (Avoiding clerical errors) related positively to Dimension 9 (Working carefully and thoroughly), corrected r = .59; whereas Dimension 9 was correlated positively with Dimension 10 (Staying in control) and Dimension 11 (Using recall aids), corrected r = .66 and r = .51, respectively.

Another example is that Dimension 4 (Using physical cues) was correlated positively and moderately with Dimension 2 (Optimizing time efficiency and effectiveness), corrected r = .42, Dimension 5 (Using grammatical or contextual cues), corrected r = .40, Dimension 6 (Deductive reasoning), corrected r = .39, and Dimension 8 (Careless guessing), corrected r = .43. On the other hand, Dimension 4 did not relate to Dimension 10 (Staying in control), corrected r = .06, and it was negatively correlated with Dimension 9 (Working carefully and thoroughly), corrected r = -.13. The implication is that the KOTTS subscales, although measuring different aspects of the phenomenon, were not truly independent constructs.

Table 20

Observed and Corrected Inter-scale Correlations for the 11 Refined Scales (Wave 2)

Scale	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
D1	(.59)	.27	.30	.11	.24	.16	.26	.06	.25	.20	.15
D2	.42	(.68)	.25	.28	.41	.33	.19	.24	.18	.23	.29
D3	.45	.35	(.75)	-.05	.27	.10	.26	-.05	.44	.37	.36
D4	.17	.42	-.07	(.67)	.29	.26	.09	.29	-.09	.05	.06
D5	.35	.57	.36	.40	(.76)	.39	.19	.11	.31	.33	.36
D6	.25	.49	.15	.39	.54	(.68)	.18	.12	.24	.28	.25
D7	.44	.30	.40	.15	.29	.29	(.58)	.17	.15	.19	.19
D8	.10	.35	-.07	.43	.15	.18	.27	(.69)	.07	.08	.03
D9	.39	.26	.59	-.13	.42	.34	.23	.09	(.72)	.49	.35
D10	.30	.32	.49	.06	.43	.39	.29	.11	.66	(.78)	.27
D11	.24	.44	.51	.10	.51	.38	.31	.04	.51	.38	(.65)

Note. Numbers in the upper triangle of the matrix are *observed* inter-scale correlations; numbers in the lower triangle are *corrected* inter-scale correlations. Numbers in parentheses are scale reliability coefficients. *D1*: Knowing how multiple-choice tests work. *D2*: Optimizing time efficiency and effectiveness. *D3*: Avoiding clerical errors. *D4*: Using physical cues. *D5*: Using grammatical or contextual cues. *D6*: Deductive reasoning. *D7*: Guessing. *D8*: Careless guessing. *D9*: Working carefully and thoroughly. *D10*: Staying in control. *D11*: Using recall aids.

Summary

After a series of item analyses and scale reliability analyses, I came up with a final product of the 11-dimension measure of KOTTS, consisting of 44 individual strategies. Compared with the original measure, there were a couple of changes in this final version of the measure: the original dimension of *Changing answers* was discarded due to its poor psychometric properties; the dimension of *Guessing* was split into two 2-item scales of *Guessing* and *Careless guessing* (being renamed based on the item content

of each scale). In terms of scale reliability, 4 scales had internal consistency values in the range of .70's, 5 scales in the range of .60's, and 2 scales in the upper .50's.

Item-scale correlations and inter-scale correlations showed evidence for the discriminant validity of KOTTS dimensions. The next step was to analyze the factor analytic solutions for this measure.

Factor Analyses: Confirmatory and Exploratory

Confirmatory Factor Analyses

My hypothesized model of TTS knowledge was a 11-factor model based on the concept that one's knowledge of TTS might include multiple facets of knowledge of a test-taking process. Some facets may be general, involving a broad understanding of the general characteristics of a multiple-choice test; other types of strategies involve test takers' specific knowledge of a strategy or a set of technique in handling a multiple-choice test problem. Although the different facets of TTS knowledge are likely to relate to one another because they are all the strategies that test takers may use to handle a test in the multiple-choice format, they should not be conceptually similar to the point where measuring a single aspect would be sufficient to infer about an individual's knowledge of TTS. Another alternative model was that of a zero-correlation 11-factor one; this model, if confirmed, would provide evidence for the independence of the subscales.

These multi-dimensional models were tested against a competing 2-factor model that explained the covariance in the data with two general factors. The theoretical assumption for a two-factor model was derived from Millman et al.'s (1965) taxonomy. In their work, Millman et al. categorized their strategic test-taking behaviors into two broad categories: "Elements independent of test constructor or test purpose" and

"Elements dependent upon the test constructor or purpose." This categorization could be interpreted as distinguishing strategies related to test takers' characteristics from strategies that were used to take advantage of test idiosyncrasies. In the context of this study, the covariance of all strategies, except those in the 2 cue-using dimensions, the deductive reasoning dimension, and the careless guessing dimension; might be explained by the first general factor, whereas the second general factor might explain the variance in the rest.

To test these hypotheses, 3 confirmatory factor analyses were conducted using LISREL: the first CFA tested the measurement model of 44-item, correlated 11 dimensions; the second one testing the model of a 44-item, zero-correlation 11 dimensions, and the last one testing a 44-item, correlated 2-factor model.

The correlated 11-factor model ($df = 847$) seemed to fit the data the best compared with the zero-correlation model and the 2-factor one. Although Normal Theory Weighted Least Squares χ^2 was huge because of the large sample size in this study (1580.18, $p < .01$), other major fit indices were reasonably good considering the number of indicators and latent variables in this model, which tended to reduce fit indices in general. The Root Mean Square Error of Approximation (RMSEA) = 0.05, indicating a good fit; the Standardized Root Mean Square Residual (SRMSR) = 0.06, indicating a moderate fit; the Non-Normed Fit Index (NNFI) = 0.80, and the Comparative Fit Index (CFI) = 0.82, being lower than the good-fit criterion of .90. Overall, the 11-factor model fit the data well.

Note that in the matrix of modification indices of λ_X (see Appendix K for Modification Indices of λ_X and Completely Standardized Solution matrices), there were 2

items in Dimension 6 (Item 4, "I eliminate the choice 'all of the above' when there are opposite choices in the answers," and Item 5, " I rule out choices that contradict the question") that covaried highly with several other dimensions. Therefore, removing them from this scale would reduce the inter-scale correlations. In fact, when Items 4 and 5 were dropped, the internal consistency of Dimension 6 was $\alpha = .697$ for a 3-item scale (compared with $\alpha = .683$ for a 5-item scale). After dropping Items 4 and 5, the KOTTS measure was consisted of 42 strategies arranged in 11 categories.

For the zero-correlation 11-factor model ($df = 902$), overall, the model fit indices were not good: RMSEA was 0.08, indicating a moderate fit. However, the Normal Theory Weighted Least Squares was $\chi^2 = 3305.93$ ($p < .01$); SRMSR = 0.14; NNFI = 0.63, and CFI = 0.65; these fit indices were below the desirable level for these indices, respectively (see Appendix L for the Completely Standardized Solution matrix). Taken together, the fit indices suggested that the 11 subscales of KOTTS measure were not independent constructs.

Regarding the correlated 2-factor model ($df = 901$), the fit indices were not impressive: $\chi^2 = 3466.12$, $p < .01$, RMSEA = 0.088, indicating a moderate fit; SRMSR = 0.09, indicating a bad fit; NNFI = 0.49, and CFI = 0.52, far below the good-fit criterion of .90 (see Appendix M for the Completely Standardized Solution matrix).

Overall, the 2-factor model provided a poor fit to the data. In other words, compared with the 2-factor model derived from Millman et al. (1965)'s taxonomy of strategies in test taking and the zero-correlation 11-factor model, the correlated 11-factor model of TTS knowledge, which was proposed in this study, seemed to provide the most respectable fit for the data.

Exploratory Factor Analyses

Item factor analysis. Because some of the fit indices of the correlated 11-factor model indicated that the model only moderately fit the data, an exploratory factor analysis was conducted with a principal axis factoring of the 42 item ratings, and a Varimax rotation. Eleven factors were extracted and rotated (Table 21).

At first glance, all factors were interpretable; most variables in a specific scale loaded more highly on their corresponding factor, unsurprisingly so considering the series of refinement steps that had been done to the measure. However, when examining the factor loadings more closely, I found 3 problematic items: Items 4 and 6 in Dimension 2 (Factor 5), and Item 3 in Dimension 9 (Factor 8). These variables loaded non-discriminatingly on factors other than their respective one. To better interpret these factors, I made the decision to drop these items from their scales, although this fact might result in a lowered scale reliability coefficient for these dimensions. As a result, the KOTTS measure was reduced to 39 items.

Table 21

Item Factor Loadings (Principle Axis Factoring, Varimax Rotation)

Item	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
D5_4	**.65**	.12	-.05	.07	.09	.06	-.04	.13	-.07	.09	-.04
D5_6	**.58**	.07	.07	.08	.02	.02	.02	.05	.04	.07	-.03
D5_5	**.54**	.08	.33	-.02	-.09	.06	.27	.09	.11	-.02	.16
D5_3	**.49**	.09	.10	.01	.04	.11	.13	.08	.10	-.14	.04
D5_2	**.49**	.11	.07	.12	.12	.03	.19	-.02	.04	.11	.11
D5_1	**.45**	.07	.03	.18	.18	.15	.09	-.02	-.03	.06	.05
D5_7	**.43**	.08	.02	-.02	.14	.13	.10	.16	.00	.01	.10
D10_5	.09	**.80**	.09	-.02	.11	.06	.03	.07	-.04	.00	.00
D10_4	.16	**.58**	-.03	.02	.11	.04	.05	.23	-.04	-.02	-.06
D10_2	.07	**.55**	.27	-.03	-.09	.09	.12	-.02	.07	.15	.17
D10_3	.10	**.54**	.28	.13	.00	.03	-.01	.09	.14	.10	.06
D10_1	.18	**.52**	.08	-.08	.04	.08	.21	.20	.00	.02	.06
D3_3	.10	.12	**.70**	-.13	.09	-.01	.12	.04	-.01	.04	.03
D3_4	.03	.07	**.63**	-.02	.03	-.05	.11	.18	-.07	.19	.07
D3_1	.10	.16	**.57**	-.01	.06	-.02	.22	.14	-.07	.06	.13
D3_2	.08	.21	**.45**	.07	.15	-.16	.11	.18	-.03	.07	.09
D4_1	.08	.00	-.04	**.78**	.13	.20	.03	-.03	.07	-.04	.04
D4_2	.04	.04	-.15	**.70**	.22	.13	.00	-.10	.08	-.09	.13
D4_4	.28	.00	.00	**.42**	.12	.05	.00	-.01	.11	.15	-.04
D4_3	.21	-.12	.08	**.30**	-.05	.05	-.05	-.06	.19	.13	-.04
D2_5	.13	.04	.02	.12	**.68**	-.03	.04	.11	.05	.02	.09
D2_1	.00	.00	.06	.13	**.52**	.09	-.03	.05	.15	-.05	-.02
D2_3	.22	.11	.07	.15	**.47**	.18	.09	.10	.09	-.01	.17
D2_2	.26	.06	.03	.05	**.37**	.18	.14	-.14	.00	.08	.03
D2_6	*.31*	.01	.18	-.04	**.36**	.09	.15	-.11	-.03	.19	.08
D2_4	.04	.10	*.19*	-.06	**.24**	.18	.12	-.07	*.19*	.12	.17
D6_2	.11	.03	-.09	.20	.03	**.69**	.00	.09	.10	.01	-.14
D6_1	.09	.10	-.06	.06	.13	**.66**	.06	.00	.09	-.03	.11
D6_3	.25	.06	.00	.12	.11	**.52**	.00	.04	-.02	.05	.05
D11_2	.12	.03	.19	.03	-.05	-.03	**.59**	.12	.03	.00	.01
D11_3	.16	.12	.15	.00	.07	.08	**.57**	.07	.01	.10	.05

D11_1	.14	.11	.09	-.01	.24	.02	**.53**	.05	-.05	.10	-.01
D9_2	.08	.23	.19	-.04	.07	.01	.05	**.70**	-.04	.01	.07
D9_1	.28	.16	.18	-.06	.04	.14	.13	**.54**	.06	.06	.06
D9_4	.09	.23	.28	-.10	.00	.03	.23	**.39**	.03	-.08	.06
D9_3	.02	*.33*	.16	-.11	.00	.01	.25	**.36**	.17	.06	.14
D8_2	.01	.05	-.11	.17	.15	.09	.00	-.03	**.76**	-.02	.02
D8_1	.06	.03	-.04	.10	.10	.05	.00	.08	**.62**	.20	.00
D7_2	.07	.10	.23	-.08	-.05	.03	.03	-.04	.12	**.68**	.04
D7_1	.08	.05	.05	.11	.10	-.01	.14	.07	.08	**.53**	.16
D1_1	.04	.05	.09	.04	.09	.02	.03	.07	.00	.05	**.68**
D1_2	.18	.09	.23	.07	.12	.03	.02	.12	.01	.24	**.50**

Note. Factor 1 consisted of items in Dimension 5 (Using grammatical or contextual cues). *Factor 2*: those in Dimension 10 (Staying in control). *Factor 3*: those in Dimension 3 (Avoiding clerical errors). *Factor 4*: Dimension 4 (Using physical cues). *Factor 5*: Dimension 2 (Optimizing time efficiency & effectiveness). *Factor 6*: Dimension 6 (Deductive reasoning). *Factor 7*: Dimension 11 (Using recall aids). *Factor 8*: Dimension 9 (Working carefully and thoroughly). *Factor 9*: Dimension 8 (Careless guessing). *Factor 10:* Dimension 7 (Guessing). *Factor 11:* Dimension 1 (Knowing how multiple-choice tests work). Bolded numbers indicate variables loading on a particular dimension.

Scale factor analysis. Regarding the dimensions in the KOTTS measure, there might be the question whether the 11 subscales covaried with one another. To address this question, I conducted an exploratory factor analysis for the 11 scale mean scores using principal axis factoring and Varimax rotation. There were 3 factors extracted as the result. The rotated factor matrix is presented in Table 22.

Table 22

Scale Factor Loadings (Principle Axis Factoring, Varimax Rotation)

Scale	Factor 1	Factor 2	Factor 3
D9 - Working carefully and thoroughly	**.71**	-.01	.05
D3 - Avoiding clerical errors	**.58**	-.14	.43
D10 - Staying in control	**.57**	.12	.14
D5 - Using grammatical or contextual cues	**.48**	.43	.12
D11 - Using recall aids	**.45**	.12	.20
D4 - Using physical cues	-.06	**.64**	.10
D6 - Deductive reasoning	.18	**.54**	-.10
D2 - Optimizing time efficiency and effectiveness	.24	**.48**	.15
D8 - Careless guessing	-.04	**.40**	.16
D7 - Guessing	.13	.12	**.51**
D1 - Knowing how multiple-choice tests work	.26	.14	**.36**

Note. Bolded numbers indicate higher loadings on a factor.

The 5 dimensions that loaded highly on the first factor were Dimension 9 (Working carefully and thoroughly), Dimension 3 (Avoiding clerical errors), Dimension 10 (Staying in control), Dimension 5 (Using grammatical or contextual cues), and Dimension 11 (Using recall aids). This factor, therefore, seemed to reflect the positive

characteristics that a test taker should have in order to handle a multiple-choice test in general in a calm, thorough and resourceful manner.

The 4 scales that loaded highly on the second factor included Dimension 4 (Using physical cues), Dimension 6 (Deductive reasoning), Dimension 2 (Optimizing time efficiency and effectiveness), and Dimension 8 (Careless guessing). This factor revealed a relation among the categories of TTS that were identified in the literature as test-dependent (i.e., relying on either test-construction characteristics or the characteristics of the multiple-choice test format). Note that some of the categories loading on this factor were verified as negatively relating to performance on a standardized cognitive ability test in the present study (i.e., Careless guessing; Using physical cues). Also note that Dimension 5 (Using grammatical or contextual cues) also loaded almost as highly on this factor as on its own factor, implying that the strategies in this scale might tap not only test takers' knowledge of English grammar or the ability to recognize contextual cues but also the availability of such cues in a test.

The third factor included two remaining scales: Dimension 7 (Guessing) and Dimension 1 (Knowing how multiple-choice tests work). This factor was less interpretable than the other two factors though it might partly reflect test takers' general ability to handle multiple-choice tests.

Summary

Based on the ratings on the original 78-item measure of KOTTS, I conducted a series of item analyses and factor analyses (both confirmatory and exploratory) to refine the measure and find the evidence to support the hypothesized multidimensional model of TTS knowledge. After a series of reliability analyses, the number of items in the measure

was reduced to 44. Through the confirmatory factor analyses, 2 more items were dropped from their respective subscales. The item exploratory factor analysis resulted in another 3 items being dropped. The end product was an instrument of 39 variables loading on 11 factors, which were corresponding with the proposed dimensions.

Table 23 presents these dimensions, item formulation, scale statistics and item statistics of the final version of the KOTTS measure. Table 24 presents the observed inter-scale correlations and the correlations adjusted for inflation of this version. Generally, compared with the correlations in Table 20, the correlations among the subscales of the final KOTTS were slightly smaller.

Table 23

The Knowledge of Test-Taking Strategies (KOTTS) Measure, Final Version, k = 39

Item No.	Dimension & Item formulation	*M* (*SD*)	R_{it}
	D1. KNOWING HOW MULTIPLE-CHOICE TESTS WORK (α = .59)	**4.12 (.81)**	
1	- I am aware that some exams may penalize you for wrong answers (e.g., deducting points from your total score).	4.08 (1.00)	.42
2	- I know that I may work only on the section the test administrator designates and only for the time allowed (e.g., I may not go back to an earlier test section).	4.16 (.93)	.42
	D2. OPTIMIZING TIME EFFICIENCY & EFFECTIVENESS (α = .64)	**2.90 (.80)**	
1	- To save time, I scan the test for more time-consuming/difficult questions and leave them for last.	2.94 (1.16)	.41
2	- To save time, I memorize the (unchanged) directions for each type of questions in advance and only skimming through them when taking tests.	2.93 (1.19)	.32
3	- I figure out how many minutes per question on average and spend the same amount of time on each question.	2.75 (1.13)	.45
4	- I first scan the test for question types (e.g., certain types of questions require more thoughts and processing) and plan strategy accordingly (e.g., budgeting my time).	3.00 (1.10)	.51
	D3. AVOIDING CLERICAL ERRORS (α = .75)	**4.51 (.55)**	
1	- I am careful not to make or leave any stray marks on my (machine-scored) answer sheet.	4.51 (.74)	.59
2	- I periodically check my answers to catch careless or clerical mistakes.	4.29 (.81)	.49
3	- When I skip a question, I remember to skip the corresponding row of answer choices on my answer sheet.	4.57 (.74)	.56
4	- I erase the initial answer completely (when I change my mind about an answer).	4.67 (.61)	.55
	D4. USING PHYSICAL CUES (α = .67)	**2.51 (.74)**	
1	- Among the options, I choose the answer that is longer than other options.	2.29 (1.04)	.57
2	- Among the options, I choose the answer that is shorter than the rest of the options.	2.06 (.92)	.53

3	- I do not answer a series of questions with the same letter choice (e.g., all *A*s).	2.83 (1.10)	.33
4	- I know that the test constructor may place the correct answer in a certain physical position among the options (e.g., in the middle).	2.86 (1.12)	.43
	D5. USING GRAMMATICAL OR CONTEXTUAL CUES (α = .76)	**3.57 (.64)**	
1	- I consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.	3.44 (1.04)	.45
2	- I notice that the test constructor may qualify the correct answer more carefully (e.g., more precise and specific in meaning), or make it represent a higher degree of generalization.	3.65 (.94)	.48
3	- I pay particular attention to negatives.	3.44 (1.05)	.47
4	- I recognize and make use of resemblance between the options and an aspect of the stem.	3.24 (1.03)	.53
5	- I recognize and make use of specific determiners (e.g., always, never), disclaimers (e.g., best, all, or none), and/or "hedging" words (e.g., probably, most likely).	4.06 (.91)	.50
6	- I know that the test constructor may make the correct answer grammatically consistent with the stem.	3.46 (1.06)	.50
7	- Among the choices, I look for the answer that converges ALL the dimensions in the stem.	3.71 (1.0)	.42
	D6. DEDUCTIVE REASONING (α = .68)	**3.05 (.78)**	
1	- I choose neither of two options which imply the correctness of each other.	3.06 (.95)	.53
2	- I eliminate answer choices which have some similarities.	2.83 (1.03)	.55
3	- I choose one of two statements, which, if correct, would imply the incorrectness of the other.	3.26 (.98)	.46
	D7. GUESSING (α = .58)	**4.21 (.77)**	
1	- I know if I randomly choose an answer out of four options, I'll have 25% chance of getting it correctly.	4.17 (.92)	-
2	- If there are only a few minutes left for a test/test section and there is no penalty for wrong answers, I will fill in the remaining problems with guesses (randomly or picking a particular answer choice such as B or C) before time is called.	4.24 (.92)	-
	D8. CARELESS GUESSING (α = .69)	**2.51 (.85)**	
1	- I tend to guess a particular choice (e.g., A or C) as soon as	2.75	-

	something looks unfamiliar or difficult.	(1.04)	
2	- I guess as soon as something looks unfamiliar or difficult.	2.28 (.90)	-
	D9. WORKING CAREFULLY AND THOROUGHLY (α = .68)	**4.10 (.66)**	
1	- I read the test items carefully, determining clearly the nature of the question.	4.13 (.79)	.52
2	- I read all instructions/directions carefully to make sure I understand them, determining clearly the nature of the task and the intended basis for response.	4.13 (.82)	.56
3	- I read all information provided, even when I see an immediate answer.	4.04 (.92)	.42
	D10. STAYING IN CONTROL (α = .78)	**3.63 (.69)**	
1	- I avoid internal distractions by directing attention away from negative self-evaluative thoughts.	3.70 (.90)	.52
2	- I don't panic if I cannot answer a question: I keep calm and move on.	3.90 (.88)	.52
3	- I do not become impatient or discouraged if I cannot start at once effectively.	3.43 (.97)	.53
4	- During the test, I divorce myself from everything (e.g. the proctor, my surrounding), and let no disturbance influence my concentration.	3.41 (1.03)	.52
5	- I don't waste time worrying during the test but work efficiently.	3.72 (.99)	.67
	D11. USING RECALL AIDS (α = .65)	**4.01 (.73)**	
1	- Before answering test questions, I note in the margins of my test booklet (if allowed) certain short-term memory items.	3.87 (1.04)	.45
2	- If I draw a temporary blank, I try to visualize the place in the book where this material appeared—this might lead me to relate facts and increase chances of recall.	4.14 (.88)	.45
3	- I sometimes make a quick drawing or a formula to clarify the questions and initiate and aid in recall.	4.03 (.91)	.50

Note. R_{it}: Corrected item-total correlation.

Table 24

Observed and Corrected Inter-scale Correlations of the Final KOTTS (k = 39)

Scale	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
D1	(.59)	.22	.30	.11	.24	.09	.26	.06	.22	.20	.15
D2	.36	(.64)	.17	.30	.35	.29	.11	.23	.15	.19	.23
D3	.45	.25	(.75)	-.05	.26	-.04	.26	-.05	.44	.37	.36
D4	.17	.46	-.07	(.67)	.28	.32	.09	.29	-.07	.05	.06
D5	.36	.50	.35	.39	(.76)	.30	.17	.11	.33	.32	.34
D6	.15	.43	-.06	.46	.41	(.70)	.06	.18	.12	.18	.11
D7	.44	.19	.40	.15	.26	.09	(.58)	.17	.11	.19	.19
D8	.10	.34	-.07	.43	.16	.26	.27	(.68)	.03	.08	.03
D9	.35	.23	.61	-.10	.46	.18	.17	.04	(.68)	.45	.32
D10	.30	.27	.49	.06	.42	.25	.29	.11	.62	(.78)	.27
D11	.24	.35	.51	.10	.49	.16	.31	.04	.48	.38	(.65)

Note. Numbers in the upper triangle of the matrix are *observed* inter-scale correlations; numbers in the lower triangle are *corrected* inter-scale correlations. Numbers in parentheses are scale reliability coefficients. *D1*: Knowing how multiple-choice tests work. *D2*: Optimizing time efficiency and effectiveness. *D3*: Avoiding clerical errors. *D4*: Using physical cues. *D5*: Using grammatical or contextual cues. *D6*: Deductive reasoning. *D7*: Guessing. *D8*: Careless guessing. *D9*: Working carefully and thoroughly. *D10*: Staying in control. *D11*: Using recall aids. In parentheses are scale reliability coefficients.

Correlation Analyses

In the following sections, the findings related to the hypotheses were presented along with the discussion related to each hypothesis.

Tables 24 and 25 present the correlational matrices reflecting the relationships between test takers' TTS knowledge and their self-reported TTS use during a test-taking situation, between TTS knowledge and CA-T performance, and between TTS use and CA-T performance.

Relations Between TTS Knowledge and TTS Use

Hypothesis 1 stated that the relationships between TTS knowledge and TTS reported use would generally be in the positive direction. The rationale was that the more aware test takers were of the general strategies used in the testing situation, the more likely they reported using these strategies. The zero-order correlation matrix (Table 25 diagonal) shows evidence to support this hypothesized relationship, although the magnitudes of knowledge-use correlations varied across dimensions.

Specifically, significant and moderately strong correlations (i.e., in the range of .30's to .50's) were detected between knowledge and use in most dimensions (Dimensions 2, 3, 4, 5, 6, 7, 9, and 10). The magnitudes of knowledge-use relations of Dimensions 8 and 11 were significant and moderate. However, there was no knowledge-use relationship for Dimension 1 (Knowing how multiple-choice tests work; $r = .03$, $p > .05$). Generally, for most of the KOTTS dimensions, the more participants had endorsed certain strategies, the more likely they reported using them while taking the test. Therefore, Hypothesis 1 was mostly supported.

Table 25

Correlations of Self-reported TTS Knowledge and TTS Usage

Var.	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
D1U	.03	.06	.13*	.09	.18**	.07	.15**	.03	.08	.06	.09
D2U	.12*	.44**	.07	.26**	.33**	.25**	.08	.14*	.44	.20**	.10
D3U	.20**	.20**	.52**	.09	.27**	.08	.19**	-.01	.34**	.29**	.31**
D4U	-.03	.14*	-.15**	.48**	.06	.20**	.02	.20**	-.15**	-.06	-.16**
D5U	.16**	.24**	.15**	.24**	.55**	.24**	.05	.07	.20**	.19**	.14*
D6U	.12*	.34**	.07	.21**	.31**	.50**	.06	.14**	.18**	.28**	.14**
D7U	.04	.01	.19**	.02	.10*	-.05	.37**	.01	.16**	.10	.22**
D8U	-.07	.06	-.16**	.20**	.02	-.01	-.03	.19**	-.14**	-.07	-.05
D9U	.22**	.20**	.25**	.07	.31**	.16*	.13*	.07	.43**	.32**	.24**
D10U	.19**	.15**	.30**	.05	.30**	.14**	.17**	.12*	.38**	.50**	.25**
D11U	.05	.16**	.08	.16**	.29**	.13*	.19**	.11*	.17**	.18**	.28**

Note. ** $p < .05$. * $p < .01$. *D1*: Knowing how multiple-choice tests work. *D2*: Optimizing time efficiency and effectiveness. *D3*: Avoiding clerical errors. *D4*: Using physical cues. *D5*: Using grammatical or contextual cues. *D6*: Deductive reasoning. *D7*: Guessing. *D8*: Careless guessing. *D9*: Working carefully and thoroughly. *D10*: Staying in control. *D11*: Using recall aids. *U* indicates the use of a dimension. Bolded numbers indicate the relationship of interest.

As conceptualized, the magnitude of the relationships between a dimension of TTS knowledge (measured pretest) and the same dimension of TTS use (measured posttest) was not very strong (i.e., not indicating a perfect relationship). The less-than-perfect relationships might be explained by the fact that test takers might know of these strategies but could not or would not use them on this particular test. For example, participants were highly aware of the strategies in "Using recall aids" dimension (M = 4.01, SD = .73) but they reported only an above-average rate of utilizing these strategies when taking the test (M = 3.21, SD = .77).

There are two implications for this finding. First, the observed discrepancy supports the theoretical distinction of the cognitive aspect (TTS knowledge) from the behavioral aspect (using TTS) in assessing TW. Second, the finding lends support for the need of measuring knowledge of TTS and utilization of TTS separately, prior to and after administering a multiple-choice cognitive ability test. The ratings might provide full diagnostic information about the strengths and weaknesses of a test taker in terms of knowing and using TTS, which is useful for tailoring an intervention to a person's need. For example, if an individual does not seem to know effective TTS well enough, as indicated by his or her low ratings on an empirically effective TTS knowledge scale, the first step of a possible intervention would be to provide the person with a broader range of TTS information. However, if the knowledge aspect is *not* lacking but the test taker reports a low rate of TTS utilization, the question may be to find out what factors interfere with the application of TTS on a test (i.e., motivational, affect or situational factors) and, if necessary and possible, removing these factors through some form of intervention.

As a note of caution, the KOTTS as a self-report measure may be an appropriate instrument for assessing test takers' declarative knowledge of TTS, but its effectiveness in measuring test takers' utilization of TTS is limited because of the retrospective nature of the measure (i.e., tapping long-term memories; measuring how much test takers would think they had used each strategy of interest during test taking, not the actual action of applying a strategy to a test item). Retrospect self-report measures are the most frequently administered instrument in psychology in general (Ruben, 1999) because they are inexpensive to administer and simple to score and interpret, allowing researchers to make rapid inferences about underlying cognitive processes through performance assessments. However, retrospect self-report measures have limited validity because they may be biased by self-presentational concerns or because they rely on people's ability to accurately introspect (Farnham, 1999). If a highly accurate assessment of TTS use is the desirable research goal and if individual-level assessment is feasible, I would suggest employing a different self-report method to measure test takers' use of TTS: the concurrent "think-aloud" technique that helps researchers to keep track of an individual's mental process in test taking and what strategies they actually use (i.e., Tian, 2000; Towns & Robinson, 1993), because this technique assesses cognitions concurrently with their occurrence, suitable for use to tap thought content (Davison, Vogel, & Coffman, 1997). "Think-aloud" methods are inspired from the work of Piaget (1954) who was the first researcher to notice that children have a natural tendency to talk out loud to themselves while solving problems. Educational, cognitive and clinical/counseling psychologists have successfully utilized the think-aloud protocols to assess cognitive and even affective processes such as attention and reading comprehension (i.e., Montague &

Applegate, 1993), problem-solving, self-efficacy and speech anxiety (Davison, Haaga, Rosenbaum, Dolezal, & Weinstein, 1991).

In a standard think-aloud method, participants are asked to verbalize their thoughts while performing some tasks, and the think-aloud responses are recorded for subsequent evaluation. The administering, coding and analyzing protocols for the think-aloud method are developed by Ericsson and Simon (1980, 1993). Based on a theoretical framework of human information processing, the researchers showed how individuals, in response to an instruction to think aloud, would verbalize information that they were attending to in short-term memory with a great level of accuracy (but not with information in long-term memory). In other words, one may be able to accurately and reliably report or verbalize the thought process occurring and being retained in one's short-term memory at the time of one's verbal report.

Relations Between TTS Knowledge, TTS Use and Test Performance (Criterion-related Validity)

Hypotheses 2 and 3 speculated about whether there was any relationship between the dimensions of TTS knowledge or TTS use, that were not linked to test idiosyncrasies, and test takers' CA-T performance. Such relationships, if found, would verify the external validity for the KOTTS measure. Given the theoretical assumption that TTS knowledge and/or use might enhance performance on a multiple-choice test over and above test-takers' knowledge of the subject tested, a positive correlation between one's knowledge or use of such TTS dimensions and one's test score would be interpreted as evidence for the criterion-related validity of the measure. However, given what I had learned so far about the dimensionality of the KOTTS measure, I further expected that a

majority of the dimensions which were adaptive and proactive in nature would be positively linked to test scores, and that there would be negative relationships between test performance and the use of certain dimensions of strategies which were linked to test idiosyncrasies ("Use of grammatical or contextual cues," "Use of physical cues," and "Deductive reasoning"), as well as strategies that had been found ineffective in the literature ("Careless guessing"). Such opposite directions in the relationships of interest, if found, would provide the criterion-related evidence for the measure in general. The correlation matrix on Table 26 shows that the dimensions of TTS (knowledge or use) were valid predictors of test scores as expected.

In terms of the relationship between knowledge of TTS and test scores, the knowledge about certain dimensions (Dimensions 2, 3, 5, 7, 9, 10, and 11) was positively associated with higher verbal, math, and/or total test scores. For example, those who had a higher level of awareness of the importance of working carefully on the test (i.e., highly endorsing Dimension 9, $M = 4.10$, $SD = .66$), as well as avoiding making clerical errors during the CA-T (Dimension 3, $M = 4.51$, $SD = .55$), tended to do better on the test ($r = .22$, $p < .01$, & $r = .28$, $p < .01$, respectively).

Table 26

Means, Standard Deviations, and Correlations of Test Performance, TTS Knowledge, and TTS Usage

Variable	*M*	*SD*	Verbal score	Math score	Total score
Verbal	6.54	2.41	-		
Math	4.14	1.77	.25**	-	
Total	10.69	3.33	.86**	.71**	-
D1	4.12	0.81	.07	.07	.09
D2	2.90	0.80	.09	.09	.10*
D3	4.51	0.55	.20**	.14**	.22**
D4	2.51	0.74	-.12*	-.08	-.13*
D5	3.57	0.64	.11*	.15**	.16**
D6	3.05	0.78	.02	.04	.04
D7	4.21	0.77	.09	.13*	.14**
D8	2.51	0.85	-.10	-.06	-.11*
D9	4.10	0.66	.17**	.11**	.28**
D10	3.63	0.69	.22**	.11*	.22**
D11	4.01	0.73	.23**	.21**	.28**
D1U	3.81	0.74	.11*	.04	.10
D2U	2.67	0.82	-.05	.04	-.02
D3U	3.99	0.66	.10	.05	.10
D4U	2.09	0.71	-.25*	-.23**	-.30*
D5U	3.17	0.65	.05	.08	.08
D6U	3.13	0.74	.13	-.03	.08
D7U	4.02	0.78	.11*	.04	.10*
D8U	2.54	0.98	-.19**	-.27**	-.29**
D9U	3.70	0.73	.10	.01	.08
D10U	3.68	0.63	.16**	.16**	.20**
D11U	3.21	0.77	.08	.13*	.13*

Note. ** $p < .05$. * $p < .01$. *D1*: Knowing how multiple-choice tests work. *D2*: Optimizing time efficiency and effectiveness. *D3*: Avoiding clerical errors. *D4*: Using physical cues. *D5*: Using grammatical or contextual cues. *D6*: Deductive reasoning. *D7*: Guessing. *D8*: Careless guessing. *D9*: Working carefully and thoroughly. *D10*: Staying in control. *D11*: Using recall aids. *U* indicates the reported use of a dimension.

In terms of the relationship between TTS use and test scores, the use of several dimensions (Dimensions 1, 7, 10 and 11) was positively associated with higher verbal, math, and/or total test scores. For example, with math questions, it made sense that the use of strategies in the dimension of "Using recall aids," such as making a drawing or jotting down notes from memory, seemed to be effective for correctly solving math problems ($r = .128, p < .05$). The more test takers reportedly used strategies that helped them to stay in control or concentrate in the task at hand, the more likely they scored high on the test (i.e., $r = .201, p < .01$).

It seems to be that making strategic guesses (i.e., when running out of time; Dimension 7) tended to correspond with higher test scores ($r = .11, p < .05$) but guessing carelessly (Dimension 8) was not only ineffective but also a harmful strategy to employ because it was linked to lower test scores (i.e., $r = -.27, p < .01$ with math test scores). The implication of these findings is that the strategies of Guessing dimension in the original KOTTS were not a uniform construct but consisted of 2 distinct types of guessing behaviors, one detrimental to test performance, and the other one effective in most test-taking situations.

There were a couple of dimensions whose knowledge was practically detrimental to test performance. For example, being knowledgeable about strategies in Dimension 4 (Using physical cues) and Dimension 8 (Careless guessing) was negatively correlated with test scores ($r = -.13, p < .05$, and $r = -.11, p < .05$, respectively). This fact was unsurprising given the nature of these strategies and the high quality of the test administered in this study. In fact, typical participants did not highly endorse knowing these strategies ($M = 2.51$, $SD = .55$ for Dimension 4, and $M = 2.51$, $SD = .85$ for

Dimension 8). This explanation was further evidenced when the relationships between the utilization of these strategies and test performance were examined (i.e., $r = -.30, p < .01$ for Dimension 4; $r = -.29, p < .01$ for Dimension 8). For the use Dimension 4 (Using physical cues), the finding in this study was inconsistent with the results in previous TW studies where a test was deliberately imbedded with faulty test items and test takers had been trained to recognize and take advantage of them to gain points. It appeared that trying to apply the strategies in this category was detrimental to test takers' performance when the test was well developed.

A couple of TTS knowledge dimensions were neither associated with good nor poor performance (Dimensions 1 and 6); so were several TTS use dimensions (Dimensions 2, 3, 5, 6, and 9). Among the dimensions in both TTS knowledge and use measures, only Dimension 6 (Deductive reasoning) was consistently unrelated to test performance. It seemed that knowing about deductive reasoning strategies (Dimension 6) and also using them on the test (knowledge-use $r = .50, p < .01$) might not link to a better performance level if the strategies themselves were not effective given the test characteristics.

Although the knowledge of understanding how multiple-choice tests work (Dimension 1) was not significantly correlated with one's test score, the use of this category of strategies was ($r = .11, p > .05$ between Dimension 1 and verbal test performance). A possible explanation was that the individual strategies in this dimension (i.e., finding out about possible penalty for wrong answers; working on a test section for the time allowed) were generally applicable to multiple-choice tests, facilitating test takers' understanding about the test format. However, the knowledge of these strategies

was distal enough that they might not directly contribute to one's success in optimizing one's test performance even when one was highly aware of these strategies ($M = 4.12$, $SD = .81$ in this case), unless one directly applied them to a specific test.

Test-takers' use of strategies in Dimension 5, "Using grammatical or contextual cues," was not correlated positively with test scores. This fact might be explained by the absence of "cues," such as specific determiners and the resemblance between a question stem and options, in the test participants had taken. Such an absence would render these strategies inapplicable under the circumstances. Traditionally, Millman et al. (1965) and subsequent researchers have often categorized "cue-using strategies" in a broad, consistent construct. Nguyen et al. (2003) were the first researchers to suggest splitting the category into two separate categories, which evolved into the dimensions of "Using physical cues" and "Using grammatical or contextual cues" in this study. The fact that these 2 groups of strategies differentially covaried with the criterion at the knowledge level further supported the argument that these scales measured 2 different constructs and, thus, should be measured separately.

In sum, Hypotheses 2 and 3 were mostly supported. The relationships between test performance and TTS dimensions, either knowledge or use, were fairly complex.

In addition, although the subscale of "Changing answers" was dropped during the reliability analysis stage, I examined the relations of Item 56 ("I do not hesitate to change my answers if I feel I should") with test scores, given that there was empirical evidence for the effectiveness of the strategy of "changing answers" in the literature (i.e., improving test scores) as well the objective nature of this strategy type. Knowledge about this strategy was significantly correlated with verbal and math scores ($r = .13$, $p <$

.05, & $r = .11, p < .05$, respectively). The reported use of this strategy was significantly related to math performance ($r = .14, p < .01$).

Table 27 presents the relationships between TTS dimensions and test takers' demographic variables.

Relations Between Student GPA and TTS Knowledge or TTS Use

Hypotheses 4a and 4b predicted that there would be a positive relationship between participants' cumulative GPA and cue-using and deductive reasoning dimensions of TTS (both knowledge and use) in this study. However, the correlation analysis showed that both dimensions of "Using physical cues" and "Using grammatical or contextual cues," and the dimension of "Deductive reasoning" (knowledge or use) were not significantly linked to participant GPA. Therefore, these hypotheses were not supported. Inconsistent with the literature, how well participants knew these strategies or how much they reported using them did not appear to link to their GPA in the present study. An explanation for this finding is that these strategic test-taking behaviors are very specific to certain classroom examinations; however, the criterion I used to examine the relationships of interest was participants' *cumulative* GPA that might be too general to be able to detect grade differences that were associated with knowledge and use of cue-using strategies and deductive reasoning strategies. It is because GPA is generally based on overall performance on class assignments other than multiple-choice tests (i.e., projects, term papers).

Table 27

Correlations of Test-takers' Demographic Variables with TTS Dimensions

Variable	Age	Gender[a]	GPA	Ability[b]	Testprep	Selftaught
D1	-.15**	-.03	.16**	.14*	.00	.04
D2	-.04	.00	.09	.15**	.09	.08
D3	-.07	-.21**	.17**	.13*	.06	-.03
D4	.04	.05	-.08	-.02	.09	.14*
D5	.00	-.03	.09	.14**	.15**	.03
D6	.02	.03	.01	.06	.09	.10
D7	-.20**	.02	.12*	.19**	-.02	-.01
D8	-.11*	-.04	-.03	-.05	-.02	.06
D9	.01	-.15**	.12*	.07	.07	-.05
D10	.10	.10	.06	.20**	.07	-.02
D11	.01	-.22**	.14*	.18**	.08	.12*
D1U	-.05	-.08	.09	.12*	.04	-.06
D2U	.00	.05	.03	.09	.00	.09
D3U	-.06	-.19**	.13*	.12*	.00	.06
D4U	-.01	.04	-.09	-.14*	.03	.22**
D5U	-.01	.07	.08	.10	.11*	.07
D6U	.03	.01	.02	.09	.06	.06
D7U	-.16**	-.06	.13*	.15**	.09	-.05
D8U	-.01	.07	-.09	-.19**	.02	.12*
D9U	.00	-.17**	.06	.06	.01	.02
D10U	.05	-.08	.14**	.16**	-.05	-.10
D11U	.05	-.08	-.02	.07	.13*	-.01

Note. [a] *Gender*: 1 = Male, 0 = Female. [b] *Ability*: standardized test score (ACT or SAT, centered). [c] *Testprep*: having taken courses with some TTS training: 1= some courses, 0 = no course. ** $p < .05$. * $p < .01$. *D1*: Knowing how multiple-choice tests work. *D2*: Optimizing time efficiency and effectiveness. *D3*: Avoiding clerical errors. *D4*: Using physical cues. *D5*: Using grammatical or contextual cues. *D6*: Deductive reasoning. *D7*: Guessing. *D8*: Careless guessing. *D9*: Working carefully and thoroughly. *D10*: Staying in control. *D11*: Using recall aids. *U* indicates the use of a dimension.

An alternative explanation is that the teacher-made multiple-choice examinations constructed by MSU faculty might be free of these test idiosyncrasies.[1] Another likely reason is that the variance in student GPAs in this sample was not large ($M = 3.23$, $SD = .52$). This restriction in range might have caused the relationship between the TTS dimensions and GPA to be underestimated in this study.

Exploring the relationships between student GPAs and other TTS knowledge dimensions, I found significantly positive relations of several dimensions to the criterion. Specifically, GPA was linked to Dimension 1 (Knowing how multiple-choice tests work, $r = .16, p < .01$), Dimension 3 (Avoiding clerical errors, $r = .17, p < .05$), Dimension 7 (Guessing, $r = .12, p < .05$), Dimension 9 (Working carefully and thoroughly, $r = .12, p < .05$), and Dimension 11 (Using recall aids, $r = .14, p < .05$). In other words, those who were more knowledgeable about these strategies tended to have better GPAs, which made sense.

The lack of significant relationships with the rest of the dimensions might reflect the fact that students' grade was the product of their performance on objective tests (which TTS knowledge might play an influencing role) and their performance on other unrelated types of classroom assignments (i.e., essay tests, term papers). It might also be due to the range restriction in GPAs.

[1] It is possible so but not plausible. I have recently taken two multiple-choice examinations where I found a few apparent faulty test items. When I brought one of them to my professor's attention, he advised me not to read too much into the test idiosyncrasies.

Exploring the relationships between student GPAs and other dimensions of TTS use, I found significantly positive relations of GPAs to 3 dimensions: Dimension 3 ("Avoiding clerical errors," $r = .13, p < .05$), Dimension 7 ("Guessing," $r = .13, p < .05$), and Dimension 10 ("Staying in control," $r = .14, p < .01$). In other words, those who had higher GPAs tended to report using the strategies in these dimensions more when taking the CA-T in this study. Again, the restriction in GPA range might explain the absence of significant relationships for the rest of the TTS use dimensions.

In sum, Hypotheses 4a and 4b were not supported. Inconsistent with previous finding that GPA was positively correlated with the use of cue-using and deductive reasoning strategies (Gentry & Perry, 1993), these types of strategies did not positively relate to student GPA in the present study.

Relations Between Standardized Test Scores and TTS Knowledge or TTS Use

Hypothesis 5a predicted that there were no significant correlations of knowledge or use of cue-using strategies and deductive reasoning strategies with participants' standardized test scores (i.e., ACT or SAT scores). Table 27 (above) shows that the knowledge of "Use of physical cues" (Dimension 4) and "Deductive reasoning" strategies (Dimension 6) was not correlated with participants' standardized test scores. However, the knowledge of grammatical/contextual strategies did (Dimension 5, $r = .14, p < .01$). At first glance, this finding seemed counter-intuitive, assuming that few grammatical or contextual cues are embedded in standardized tests in general. However, knowing grammar is likely related to standardized test scores and one cannot use grammatical clues effectively without knowing grammar. Participants' cognitive ability (as measured by ACT or SAT test scores) was not related to the utilization of the

grammatical/contextual cues and deductive reasoning strategies in this study. There was an inverse relationship between ability and the use of physical-cue strategies: the lower standardized CA-T scores one had, the more likely one reported using physical-cue strategy in taking the test in this study. Therefore, Hypothesis 5a was partially supported. These findings added to our understanding about the general ineffective nature of using these particular types of strategies on standardized multiple-choice tests.

Hypothesis 5b stated that there were positive links between the standardized test scores and other TTS dimensions (knowledge and/or use). The TTS knowledge dimensions that were correlated positively with participants' standardized test scores were Dimension 1 (Knowing how multiple-choice tests work, $r = .14, p < .05$), Dimension 2 (Optimizing time efficiency, $r = .15, p < .01$), Dimension 3 (Avoiding clerical errors, $r = .13, p < .05$), Dimension 7 (Guessing, $r = .19, p < .01$), Dimension 10 (Staying in control, $r = .20, p < .01$), and Dimension 11 (Using recall aids, $r = .18, p < .01$). Those who tended to do better on standardized cognitive ability tests endorsed these strategies more.

Similarly, participants' cognitive ability was positively linked to the reported use of several TTS dimensions in this study: Dimension 1 ($r = .12, p < .05$), Dimension 3 ($r = .12, p < .05$), Dimension 7 ($r = .15, p < .01$), and Dimension 10 (Working carefully and thoroughly, $r = .16, p < .01$). The more cognitively able a test taker was (as indicated by his or her ACT or SAT score), the more likely he or she reportedly used these strategies when working on the test in this study. The inverse relationship between ability and the strategies of Careless guessing (Dimension 8, $r = -.19, p < .01$) could be interpreted that those who were low on cognitive ability were more likely to guess randomly without thinking when taking the test in this study. Therefore, Hypothesis 5b was supported.

Note that a possible implication for the findings regarding Hypotheses 5a and 5b is that the positive relations between some TTS dimensions and participants' ability and GPA are just indicative of the fact that those who are proficient in effective TTS are intelligent students. On the other hand, those who use ineffective strategies are not as intelligent as those who stay away from such strategies. The plausibility of such a causal inference that intelligence leads to TW or proficiency in effective TTS should be further investigated in future research, provided that intelligence is *not* measured with a test in multiple-choice format because it would become impossible to determine whether more intelligent students are more test wise or whether TW help students to score high on the multiple-choice intelligence test.

Taken together, the findings described above and the findings in Hypothesis 1 merged to provide evidence for the criterion-related validity and generalizability of the KOTTS subscales. They also verified the theory that some TTS dimensions (regardless of knowledge or use) were linked to general mental ability whereas other dimensions were not.

Additionally, I conducted a correlation analysis of the relationships between TTS dimensions (knowledge and use) and other external correlates, such as age, gender, and informal or formal training in TTS.

Relations Between Age or Gender and TTS Knowledge or TTS Use

The correlation analysis (Table 27) showed that there were age differences and gender differences in test takers' TTS knowledge or TTS use dimensions.

Younger test takers were more likely than older test takers to endorse the strategies in three categories: Knowing how multiple-choice tests work, Guessing, and

Careless guessing (r = -.15, p < .01; r = -.20, p < .01, and r = -.11, p < .04, respectively). Younger participants also tended to use the strategies of Guessing (r = -.16, p < .01) more than older participants when taking the test. These findings might reflect a stronger tendency to avoid risk-taking behaviors on a test among older test takers. Because there was a positive link between Guessing strategies and participants' ACT or SAT scores, the findings might also explain why older participants tended to have lower standardized test scores than younger test takers (r = -.16, p < .01).

Women test takers were more likely than men to report knowing strategies such as Avoiding clerical errors (r = -.21, p < .01), Working carefully and thoroughly (r = -.15, p < .01) and Using recall aids (r = -.22, p < .01). While taking the CA-T in this study, women were also more likely than men to strategically avoid clerical errors (r = -.19, p < .01), as well as working more carefully and thoroughly than men when taking the test (r = -.17, p < .01). The gender differences in these dimensions might reflect the fact that women had more relative persistence than men, being more committed to a course of action (i.e., test taking) (Huang, 1995).

Relations Between Previous TTS Training and TTS Knowledge or TTS Use

Test takers' acquisition of TTS was assessed by their report of previous exposure to TTS, either via formal training (i.e., taking a course in test preparation that mentioned TTS) or via informal venue (i.e., reading self-taught test preparation materials). Generally, taking at least one formal course where some TTS were taught did not relate to how well participants thought they knew their TTS or how much they reported used most of the TTS. In other words, those who had been exposed to TTS (formal or informal) did not particularly endorse more TTS than those who had not.

The exceptions were the positive relationship between formal acquisition of TTS and the dimension of "Using grammatical/contextual cues" dimension ($r = .15, p < .01$). Those teaching themselves test-taking skills knew more than those who had not done so on two dimensions of strategies: "Using physical cues" ($r = .14, p < .05$) and "Using recall aids" ($r = .12, p < .05$).

Those who had received some formal test-taking training were more likely than those who had not to report using strategies in two dimensions: "Using grammatical or contextual cues" ($r = .11, p < .05$) and "Using recall aids" ($r = .13, p < .05$). Those who taught themselves TTS reported using more "Using physical cues" ($r = .22, p < .01$) and "Careless guessing" ($r = .12, p < .05$) than those who had not.

There are two implications for these findings. First, it seems to be counter-intuitive that having a certain level of exposure to TTS (i.e., acquiring TTS via formal or informal training) was not necessarily linked to a better awareness of most TTS dimensions or test takers' specific use of TTS in this study. There are several possible explanations for this fact. It might mean TTS training is generally ineffective in raising the level of awareness or increasing the use of TTS in a particular test-taking situation. It might also mean that test takers who did not report either formal or informal training of TTS had actually acquired their knowledge of TTS via a different venue that I had not been able to capture by asking them about TTS courses and self-taught materials. It might simply because there was only a small proportion of the sample reported their participating in test-taking preparation in the past (i.e., 16% of test takers reported having some formal training in TTS); therefore, the observed non-significant relations occurred by chance. Another explanation for the fact that self-report TTS training is generally

ineffective for a typical test taker in this sample is that knowledge of TTS merely is not highly effective itself. In other words, training in TW or TTS would not make much difference in terms of improving test performance.

If it were true that training was typically ineffective, the inference is that there would be no need to try and educate test takers about TTS. However, given the fact that participants were college students who would not have survived their multiple-choice examinations throughout the course of their education without some methods of TTS acquisition, and that participants reported a relatively high level of TTS knowledge and use in several TTS dimensions, the other two explanations seemed to make more sense in this study.

Second, examining the strategy categories in which the haves and the have-nots in terms of TTS training significantly differed, I observed a difference due to the two methods of TTS acquisition. For example, those who had attended some formal training of TTS reported using strategies that were better (i.e., more effective empirically) than the strategies used by those who had taught themselves (i.e., "Using recall aids" vs. "Using physical cues" or "Careless guessing). The findings implied that the quality of informal TTS acquisition is questionable.

Table 28 presents means, standard deviations, reliabilities and correlations among TTS dimensions (knowledge and use), and other test-taking skills and attitudes.

Table 28

Means, Standard Deviations, Reliabilities and Correlations of Test-takers' Motivation, Self-efficacy, Personality Dimensions, and Test-Taking Metacognition with TTS Dimensions (Knowledge & Use)

	M	SD	Motivation.	S-efficac	Mct s-effi	Emotiona	Conscien	Metacog
Motiv.	3.81	0.66	(.87)					
S-effica.	3.43	0.58	.40**	(.87)				
Mct s-effi	3.24	0.66	.18**	.33**	(.86)			
Emotiona.	3.14	0.75	.10*	.10	.25**	(.84)		
Conscien.	3.56	0.64	.26**	.10	.08	.08	(.83)	
Metacog.	36.59	5.58	.34**	.24**	.09	.12*	.21**	(.76)
D1			.16**	.15**	.13*	.02	.02	.16**
D2			.09	.14**	.05	.00	.12*	.25**
D3			.31**	.17**	.22**	.06	.18**	.25**
D4			.02	.14**	-.12*	-.02	.07	.06
D5			.22**	.27**	.16**	.13*	.21**	.37**
D6			-.03	.12**	-.09	.03	.06	.11*
D7			.04	.08	.17**	-.01	-.04	.12*
D8			-.05	-.02	-.09	-.02	-.03	.04
D9			.23**	.13**	.18**	.08	.22**	.30**
D10			.20**	.25**	.38**	.31	.22**	.35**
D11			.18**	.12*	.05	-.06	.12*	.30**
D1U			.20**	.11*	.05	.03	.06	.18**
D2U			.03	.02	.01	-.01	.06	.27**
D3U			.34**	.13*	.12*	.08	.21**	.38**
D4U			-.14**	-.05	-.18**	-.02	.07	-.04
D5U			.15**	.18**	.10	.10	.11*	.36**
D6U			.02	.08	.06	.07	.06	.26**
D7U			.08	.09	.10	-.04	.05	.09
D8U			-.17**	-.13*	-.13*	-.05	-.03	-.04
D9U			.29**	.16**	.07	.14**	.23**	.40**
D10U			.24**	.20**	.19**	.17**	.23**	.36**
D11U			.14**	.07	-.01	.01	.13*	.28**

Note. ** $p < .05$. * $p < .01$. *Motiv*: test-taking motivation. *S-effica*: test-taking self-efficacy. *Mct s-effi*: Multiple-choice test-taking self-efficacy. *Emotiona*: Emotional stability. *Conscien*: Conscientiousness. *Metacog*: Test-Taking metacognition. *D1*: Knowing how multiple-choice tests work. *D2*: Optimizing time efficiency and effectiveness. *D3*: Avoiding clerical errors. *D4*: Using physical cues. *D5*: Using grammatical or contextual cues. *D6*: Deductive reasoning. *D7*: Guessing. *D8*: Careless guessing. *D9*: Working carefully and thoroughly. *D10*: Staying in control. *D11*: Using recall aids. *U* indicates the use of a dimension.

Relations Between Test-Taking Metacognition and TTS Use

Hypothesis 6 predicted that there were significant and positive relationships between TTS use and the use of test-taking metacognitive strategies in this study. This hypothesis was partially supported: the more participants reported that they had been aware of the test-taking process and monitored their test-taking behaviors, the more likely they reported using certain TTS categories while taking the CA-T in this study. The relationships ranged from moderate to moderately strong between test-taking metacognitive strategies and the reported utilization of several TTS dimensions: "Knowing how multiple-choice tests work," $r = .18, p < .01$; "Optimizing time efficiency," $r = .27, p < .01$; "Avoiding clerical errors," $r = .38, p < .01$; "Using grammatical/contextual cues," $r = .36, p < .01$; "Deductive reasoning," $r = .26, p < .01$; "Working carefully and thoroughly," $r = .40, p < .01$; "Staying in control," $r = .36, p < .01$, and "Using recall aids," $r = .28, p < .01$. However, there was no significant relationship between participants' monitoring skills and three remaining categories of TTS: "Use of physical cues," "Guessing" and "Careless guessing."

These findings not only supported the important link between one's use of certain types of TTS and that of on-task monitoring skills, but they also shed further light on the role of test-taking metacognition in inferring TW, as laid out in my aforementioned TW framework. Furthermore, the fact that test-taking metacognition—a construct characterized by self-reflection, awareness, monitoring skill, and self-evaluation—was differentially correlated with the strategy categories in this study suggested that the dimensions of TTS might lie on a continuum of metacognitive complexity. In other words, the findings implied that a set of TTS dimensions (all dimensions but "Use of

physical cues," "Guessing" and "Careless guessing") required a certain level of good metacognitive skills, whereas the other three dimensions were likely to be more automatic or less thoughtful in nature.

Relations Between Test-Taking Motivation and TTS Use

Hypothesis 7 predicted that the more motivated test takers were to perform well on this test, the more likely they reported using TTS in taking the test. Again, the relationships between test-taking motivation and the reported use of TTS dimensions differed in terms of directions. More highly motivated test takers reported using more strategies in the category of "Knowing how multiple-choice tests work," $r = .20, p < .01$; "Avoiding clerical errors," $r = .34, p < .01$, "Using grammatical/contextual cues" ($r = .15$, $p < .01$); "Working carefully and thoroughly" ($r = .29, p < .01$); "Staying in control" ($r = .29, p < .01$), and "Using recall aids" ($r = .14, p < .01$). Motivated test takers were less likely to report using the strategies of "Using physical cues" ($r = -.14, p < .01$) and "Careless guessing" ($r = -.17, p < .01$). Therefore, Hypothesis 7 was partially supported.

One might infer from the findings that test-taking motivation might be an influence not only on the intensity in test takers' use of TTS, but also on their conscious effort to stay away from strategies that they perceived as ineffective.

Relations Between Test-Taking Self-efficacy and TTS Use

Hypothesis 8 posited that test takers' use of TTS dimensions was positively correlated with their belief in their ability to perform well on the specific CA-T. This hypothesis was partially supported. The reported use of several TTS dimensions tended to increase when participants felt more self-efficacious toward the upcoming test: "Knowing how multiple-choice tests work" ($r = .11, p < .05$); "Avoiding clerical errors"

(r = .13, p < .05); "Using grammatical or contextual cues" (r = .18, p < .01); "Working carefully and thoroughly" (r = 16, p < .01), and "Staying in control" (r = .20, p < .01). The use of "Careless guessing" strategies was inversely related to test-taking self-efficacy (r = -.13, p < .05). Understandably, those who did not feel particularly confident about their ability to perform well on the upcoming test tended to guess more than those who were self-efficacious whenever they considered a test problem difficult. Therefore, Hypothesis 8 was supported.

The findings implied that a great sense of confidence in oneself as far as handling test problems was concerned might be important in test takers' increasing use effective strategies, as well as reducing a careless attitude toward tackling test questions.

Relations Between Multiple-Choice Test-Taking Self-Efficacy and TTS Use

Hypothesis 9a stated that test takers who endorsed more TTS were more likely to believe in their ability of handling multiple-choice tests in general. This hypothesis was partially supported. Multiple-choice test-taking self-efficacy (MCTTSE) was positively correlated with several TTS categories: "Knowing how multiple-choice tests work" (r = .13, p < .01), "Avoiding clerical errors" (r = .22, p < .01), "Using grammatical and contextual cues" (r = .16, p < .01), "Guessing" (r = .17, p < .01), "Working carefully and thoroughly" (r = .18, p < .01), and "Staying in control" (r = .38, p < .01). The general knowledge of these TTS dimensions seemed to enhance a person's general self-efficacy as far as taking multiple-choice tests were concerned. Those who were less likely to endorse the strategies in "Using physical cues" dimension (r = -.12, p < .05) also tended to have higher multiple-choice test-taking self-efficacy, which made sense given the

ineffective nature of this category of TTS. The implication was that some participants were capable to implicitly judge the effectiveness of TTS dimensions.

The relationships between multiple-choice test-taking self-efficacy and the use of TTS were hypothesized in Hypothesis 9b. Surprisingly, only two categories of TTS use were positively related to the sense of self-efficacy among participants: "Avoiding clerical errors," ($r = .12, p < .05$), and "Staying in control" ($r = .19, p < .01$). Two other TTS use dimensions were negatively linked to multiple-choice test-taking self-efficacy: "Using physical cues" ($r = -.18$, $p < .01$) and "Careless guessing" ($r = -.13$, $p < .05$). It appeared that those having a greater sense of self-efficacy in terms of handling multiple-choice tests in general were better than those who did not in ensuring that their test performance was clerical error-free, as well as refraining themselves in using some ineffective strategies. Therefore, Hypothesis 9b was only partially supported.

Note that the item content of the multiple-choice test-taking self-efficacy used in this study (Miguel-Feruito, 1997) was *not* homogeneous. Among the items expressing a test taker's confidence about his or her ability to handle multiple-choice tests in general (i.e., "I am good at taking multiple-choice tests"), there were also items that described his or her affect about multiple-choice tests (i.e., "I would rather take an essay test than a multiple-choice test"). One might ask, was it possible that a person might feel self-efficacious but still prefer an essay test to a multiple-choice test? In other words, it was possible but not likely that test takers' responses to the affective items in this measure were inversely related to other self-efficacy statements. I further examined the internal consistency and factor analytic structure underlying Miguel-Feruito's (1997) Test-Taking

Self-efficacy Measure. The psychometric properties and factor analysis supported a unidimensional measure.

Relations Between Conscientiousness, Emotional Stability and TTS Use

Hypotheses 10a and 10b predicted that two personality traits, Conscientiousness and Emotional Stability, would be positively linked to test takers' use of TTS dimensions. The hypotheses were partially supported. As expected, those who were higher on Conscientiousness tended to report more utilization of strategies in the categories of "Avoiding clerical errors" ($r = .21, p < .01$); "Using grammatical or contextual cues," ($r = .11, p < .05$); "Working carefully and thoroughly" ($r = .23, p < .01$), "Staying in control" ($r = .23, p < .01$), and "Using recall aids" ($r = .13, p < .05$). Those who were more emotionally stable tended to use more strategies in "Working carefully and thoroughly" ($r = .14, p < .01$), and "Staying in control" ($r = .17, p < .01$). Generally, the personality of test takers was more or less reflected in how they used different TTS dimensions to handle a test at hand: those who were characterized as hard-working, careful, thorough and persevering tended to utilize the strategies whose characteristics corresponded to their dispositions. Having a grip on emotions in general (i.e., not feeling anxious or nervous) was reflected in test takers' use of the strategies that required a mental control of the test-taking situation.

Even though some researchers had previously speculated about and investigated the relations between individual differences in personality and the use of certain TTS, their use of inappropriate personality measures (i.e., the MBTI) to gauge inter-individual differences rendered their findings inconclusive. The findings in this study were the first

appropriate empirical support for the links between personality and TTS use at the scale level.

Table 29 summarizes the results of the main hypotheses in this study.

Relations Among Test-takers' Characteristics

Although not hypothesized, the significant relations among attitudinal measures are also reported here. Table 28 (above) shows that test-specific test-taking motivation was significantly and positively related to test-specific test-taking self-efficacy ($r = .40, p < .01$) and to multiple-choice test self-efficacy ($r = .18, p < .01$). Transforming the two correlation coefficients of interest using Fisher-*z* transformation procedures, and examining the significance level of the *z* difference taking into account the respective sample sizes (see Papoulis, 1990), I found that the relation between motivation and self-efficacy was significantly stronger than the relation between motivation and multiple-choice test-taking self-efficacy ($p < .01$). Test-Taking motivation was also positively linked to test-specific test-taking metacognition ($r = .34, p < .01$), and the two personality traits measured in this study (Conscientiousness: $r = .26, p < .01$, and Emotional stability: $r = .10, p < .05$).

Table 29

Summary of Hypotheses and Results

Hypothesis	Finding
(1) Self-reported TTS knowledge will be positively correlated with self-reported TTS use.	Mostly supported
(2) Knowledge of effective TTS that are not linked to test idiosyncrasies will be positively correlated with test scores.	Mostly supported
(3) Utilization of effective TTS that are not linked to test idiosyncrasies will be positively correlated with test scores.	Mostly supported
(4a) Test takers' GPA will be positively correlated with the cue-using and deductive reasoning dimensions of TTS knowledge.	Not supported
(4b) Test takers' GPA will be positively correlated with the cue-using and deductive reasoning dimensions of TTS use.	Not supported
(5a) Knowledge and/or use of cue-using strategies and deductive reasoning will <u>not</u> be related to participants' standardized test score (ACT or SAT scores).	Partially supported
(5b) Other dimensions of TTS knowledge and/or use will be positively correlated with participants' standardized test score.	Supported
(6) Test-Taking metacognition will be positively related to one's usage of TTS.	Partially supported
(7) Test-Taking motivation will be positively correlated with self-reported use of TTS.	Partially supported
(8) Participants' TTS use will positively correlate with their test-taking self-efficacy.	Partially supported
(9a) Participants' TTS knowledge will positively correlate with their general self-efficacy about taking a multiple-choice test.	Partially supported
(9b) Participants' TTS use will positively correlate with their general self-efficacy about taking a multiple-choice test.	Partially supported
(10a) Test takers' conscientiousness will be positively related to TTS use.	Partially supported
(10b) Test takers' emotional stability will be positively related to TTS use.	Partially supported

Test-Taking self-efficacy was positively and moderately correlated with test-taking motivation, multiple-choice test-taking self-efficacy ($r = .33, p < .01$), and test-taking metacognition ($r = .24, p < .01$). Multiple-choice test-taking self-efficacy, besides its links with motivation and test-specific self-efficacy, was related to Emotional stability ($r = .25, p < .01$). Test-Taking metacognition was linked to test-taking motivation, test-taking self-efficacy (but not with multiple-choice test-taking self-efficacy), Emotional stability ($r = .12, p < .05$), and Conscientiousness ($r = .21, p < .01$).

Table 30 presents the relations between test takers' characteristics and attitudes and participants' performance on the CA-T in this study. Among test-taking attitudes (test-specific or general ones), motivation was not related to test performance but both self-efficacy and multiple-choice test-taking self-efficacy were (i.e., more self-efficacious, higher test scores). Test-Taking metacognition was slightly linked to better performance on the whole test in the positive direction. Regarding test takers' characteristics, there were no relations between personality traits and test performance. Older students tended to perform better than younger ones on the verbal ability section. Male participants had higher scores than females on the whole test or on the verbal section in particular, but there was no gender differences in math performance. Those with higher GPA performed slightly better on the math test than those with lower GPA. Unsurprisingly, participants' ability level (as measured by their ACT or SAT score) was linked to performance on this particular CA-T, the magnitudes being moderately strong. Having some formal training in TTS was neither useful nor harmful as far as the CA-T scores were concerned, but teaching oneself TTS was negatively related to whole test performance and the verbal test score in particular.

Table 30

Correlations Between Test-takers' Characteristics and Attitudes and Test-takers' Test Performance

Variable	Verbal score	Math score	Total test score
Motiv.	.03	.09	.07
S-effica.	.15**	.19**	.21**
Mct s-effi	.28**	.24**	.33**
Emotiona.	.06	.05	.07
Conscien.	-.07	-.02	-.06
Metacog.	.09	.08	.11*
Age	.12*	-.06	.05
Gender[a]	.13*	.02	.11*
GPA	.02	.12*	.08
Ability[b]	.48**	.39**	.56**
Testprep[c]	.02	-.04	-.01
Self-taught	-.12*	-.04	-.11*

Note. ** $p < .05$. * $p < .01$. *Motiv*: test-taking motivation. *S-effica*: test-taking self-efficacy. *Mct s-effi*: Multiple-choice test-taking self-efficacy. *Emotiona*: Emotional stability. *Conscien*: Conscientiousness. *Metacog*: Test-Taking metacognition. [a] *Gender*: 1 = Male, 0 = Female. [b] *Ability*: standardized test score (ACT or SAT, centered). [c] *Testprep*: having taken courses in which some TTS training was imbedded: 1= some courses, 0 = no course.

Chapter 4

DISCUSSION

This study aimed at two main purposes. First, the construct of test wiseness (TW) was redefined to better reflect the multiple psychological mechanisms (cognitive, behavioral, and metacognitive) that underlay the construct. Second, the multidimensional Knowledge of Test-Taking Strategies (KOTTS) measure was developed to help assess individual differences in declarative knowledge of TTS. The internal and external validity of this measure was then evaluated via examining the relations of its subscales with test performance and other test takers' characteristics and attitudes.

This chapter discusses the results of the study for each of the research foci in terms of the contributions and limitations, as well as relevant future research recommendations.

Integrated Theory of Test Wiseness

Limitation of past theories. Given the relatively long history of TW (being first mentioned six decades ago), it was not surprising that there were multiple approaches in the conceptualization of this construct, mainly in the educational and measurement psychology literature (i.e., Millman et al., 1965; Green & Stewart, 1984; Thorndike, 1951) and recently in the industrial-organizational literature (i.e., Guion, 1998; Parham, 1996). It was surprising, however, to learn that most researchers seem to be satisfied with a more or less descriptive definition of TW (i.e., the most recent concept of TW depicting TW as "a broad collections of skills and traits;" Flippo et al., 2000). The tendency of viewing the construct from a static, descriptive perspective has provided invaluable

insights about what the construct might be (i.e., a stable cognitive factor encompassing multiple-choice test characteristics and test takers' capacity of using these characteristics to their advantage; Parham, 1996).

However, the limitation of this tendency is that it overlooks the dynamic and complex psychological mechanisms of the construct, providing an incomplete research framework to help answer pertinent questions about TW. For example, what are exactly the broad "skills and traits" that constitute TW? How is TW established or formulated? How does it operate during a test-taking situation? When and where is TW applicable, dependent or independent of the test characteristics (i.e., test idiosyncrasies)? Does TW exist as a continuum of test-taking ability? Is the construct related to general intelligence; if yes, to what extent, and if not, is TW trainable? How can we operationally measure TW, continuing to assess measurable behaviors of applying strategies to a multiple-choice test item (which are mostly situation-dependent) and inferring the construct from the results, or choosing a different, more stable approach? To I-O psychologists, what would be the implications for employment selection when we understand individual differences in TW better? Would such understanding of TW differential influence on test performance help shed light on our quest of reducing the majority-minority group mean differences in cognitive ability test performance? They are all good questions but a sound framework to begin with is still missing.

Redefining TW. The present study aimed at partly filling the research gap. To conceptualize TW, I adopted an introspection method based on the learned knowledge about TW and other related test-taking constructs. I traced the mental processes of a test taker learning to become generally test-wise as far as taking objective, multiple-choice

tests were concerned, and then applying this learned ability to a specific test-taking situation (i.e., by using strategies in test taking to respond to multiple-choice test items). From such mental processes, I teased the construct of TW into three elements: (a) the cognitive component which encompasses both the strategies of taking a multiple-choice test acquired prior to taking a specific test and the retrieval of these strategies during taking that test, (b) the behavioral component of applying specific strategies to a multiple-choice test item, and (c) the metacognitive component of monitoring the other two elements or being strategic about strategies of test taking.

From a functionalistic perspective, it is useful to conceptualize TW as a broad construct that is constituted by three psychological elements because we might be ultimately able to pinpoint where the root of a test taker's TW deficiency is on the conceptual framework. Part of the cognitive component of TW is expressed in the breadth and depth of a test taker's acquisition of individual strategies, which can be considered the foundation for a test taker being "test wise." An ideal foundation would be a broad knowledge of the multiple facets of strategies that are partly inherent in the format of multiple-choice tests and partly depending on individual judgments, along with an understanding or at least an intuition about the "relative" value or effectiveness of these individual strategies or classes of strategies. I used a quotation mark with the term *relative* because the level of effectiveness (as measured by the criterion of test performance) might vary as a function of specific test characteristics and testing situation (this study has provided some empirical evidence for this fact).

A "test-naïve" test taker might be found lacking in this fundamental area of TW. The theoretical implication is that TW could be visualized as a continuum of adaptivity in

test taking, depending on the characteristics of test takers (i.e., background knowledge and experience in taking tests; general mental ability; attitudes toward test taking); the acquisition of strategic knowledge in taking tests at least gives a test taker the foundation of being test wise and the potential to behave accordingly on a test. That was what I meant when arguing that a basic level of TW was the "prerequisite" to successful test taking. The practical implication would be to design an intervention that helps to increase not only the person's general awareness of the variety of TTS associated with multiple-choice tests, but also a critical view of TTS and test taking in general. How would we go about measuring this component of TW construct? The present study also addressed this question, which I will discuss shortly.

The behavioral component of TW is defined as the application of retrieved TTS (i.e., from the memory explicitly or implicitly) to solving a particular multiple-choice test item. It seems to be straightforward and easy to measure, but deceptively so in my opinion. The prevalent presumption held by most researchers in the past was that we could directly observe test-taking behaviors (i.e., via the self-reported use of strategies, or the correct response of a multiple-choice test item as a result of using some TW principles) and, by the power of deductive reasoning, we would be able to make inferences about the existence of TW at various levels. In other words, the presumption is based on the classic behavioristic principle of stimulus-response and past researchers were interested in the product of TW cognition (i.e., observable behaviors) than the TW cognitive process itself. In fact, this approach was used predominantly in both the construction of TW instruments and the evaluation of TW training effectiveness. However, because test-taking behaviors (including the utilization of TTS) are situation-

specific, this approach might be misleading *if* an important set of assumptions is violated. The major assumptions include the presence (or absence) of standardization in the testing situation and the test itself, and the absence of motivational, affective and dispositional factors (whose relations to the use of TTS were evidenced in past research as well as in the present study). In other words, there might be a discrepancy between the *potential* to behave test-wisely or strategically on a test and the actual test-taking behaviors. This discrepancy was partially supported by the less-than-perfect relationship observed between self-reported knowledge and self-reported use of TTS in this study—what you see is not always what you get. Future research can shed light on the mitigating factors that were indicated in this study as well as in other test-taking studies. Future research can also focus on the operational definition of the behavioral component of TW. Although I modified the instrument developed in this study to measure test takers' self-reported use of TTS on the test administered, I recognized the limitation of self-report assessment in general and the limitation of a TTS use measure in particular (i.e., not being able to capture the mental process leading a test taker to behave as she behaves).

The metacognitive component of the TW framework has been described in detail and discussed at length in a previous section. The idea of an executive system that monitors test-taking behaviors (i.e., the use of TTS) is not entirely new; it was first theorized by Rogers and Bateson (1991) and later tested by Harmon (1997). In the test-taking literature, general test-taking metacognition theories have been proposed and tested (i.e., Pressley et al., 1987; Shraw, 1997). The contribution of this study is that the construct of metacognition in TW was redefined and broadened, from monitoring only the behavior of selecting a correct response based on cue matching (Rogers & Bateson) to

regulating the whole process of conscious retrieval of TTS, judging them and applying them (or not) to the test and test items. The limitation of this study is that no operation definition of TTS metacognition was proposed. Future research should examine the conceptual similarities and differences between general test-taking metacognition (as measured by a modified instrument by Shraw in the present study) and specific TTS metacognition, as well as investigating an appropriate way to measure the construct of TTS metacognition. Future research should also examine the possible link between positive metacognitive factors in applying TTS and negative ones (i.e., those that are conceptually similar to the constructs of off-task thinking and/or negative self-evaluation in test taking; Kirkland & Hollandsworth, 1980). For example, do positive and negative metacognition coexist or are they mutually exclusive? Assuming that the positive cognitive monitoring has a positive effect on the retrieval and application of TTS, what might be the impact of negative metacognition on the process (i.e., interfering with the retrieval and application of effective TTS)?

TW and test anxiety. One may also want to examine the relationship between test anxiety (state) and positive and negative metacognition to further understand the relationship between test anxiety and TW or TTS. Test anxiety is considered a multidimensional problem for an individual facing evaluative test-taking situations. According to Schwarzer, van der Ploeg and Spielberger (1982), anxiety refers to an unpleasant emotional and aversive physical reaction that results from evaluating a particular situation as threatening. Spielberger (1980) defines test anxiety as a situational-specific personality trait with cognitive (worry) and affective (emotionality) components. These emotions are triggered by particular environmental conditions (e.g., a test-taking

situation) and fluctuating over time, as opposed to an individual's disposition to exhibit anxiety responses in a wide variety of situations (state anxiety versus trait anxiety; Spielberger, 1966). Nguyen et al. (2003) found that test-specific anxiety (measured pretest and posttest) was negatively correlated with a component of test-taking metacognitive strategies (Knowledge of Cognition: $r = -.25, p < .01$ for pretest test-anxiety; $r = -.35, p < .01$ for posttest test anxiety); test anxiety was also negatively related to the second component of test-taking metacognition, Regulation of Cognition, although these relationships were not statistically significant. Further, Nguyen et al. found that test-specific test anxiety was positively correlated with a component of off-task thoughts (Attention to performance: $r = .18, p < .05$ for pretest test anxiety; $r = .33, p < .01$ for posttest test anxiety); but there was no link between Offtask attention and test anxiety. It seems that the better test takers were aware about their TTS and about the conditions under which strategies were most useful, the less anxious he or she felt prior to and during the test. However, test anxiety was not associated with how test takers would plan, implement strategies, monitor, correct comprehension errors, and positively evaluate their own test performance during the test. Further, those who felt more anxious during the test were more likely to be distracted by thoughts about other test takers' performance than those who were less anxious.

The causal relationship between test anxiety and TW or TTS is inconclusive. On the one hand, both test anxiety and TW might exist simultaneously (e.g., no significant negative correlation was found between test anxiety and TW among first-year medical students; Harvill, 1984b). On the other hand, training in TW helped to reduce the debilitating effects of high levels of test anxiety (Kirkland & Hollandsworth, 1980).

Given past findings about the relationship between test-taking metacognition and test anxiety, future research may address the possibility that TW (including the metacognitive element) might moderate the negative effects of state or trait test anxiety on test performance (i.e., those who are higher on TW would be less likely than lower test-wise individuals to suffer poor test performance due to a high level of test anxiety).

Effects of TW on cognitive ability test validity. Although it is not in the scope of the present study, it is useful to discuss how TW in general or the actual use of TTS in particular may relate to a CA-T that is used as a predictor of a criterion. Figure 4 demonstrates the hypothetical inter-relationship among TW, CA-T and a criterion (i.e., job performance). In this diagram, the predictive validity of TW is evidenced by its functional relation with the performance on a CA-T (i.e., a positive correlation). The CA-T in turn, may be used as a personnel selection instrument and have a predictive validity indicated by its positive correlation with some aspect of job performance.

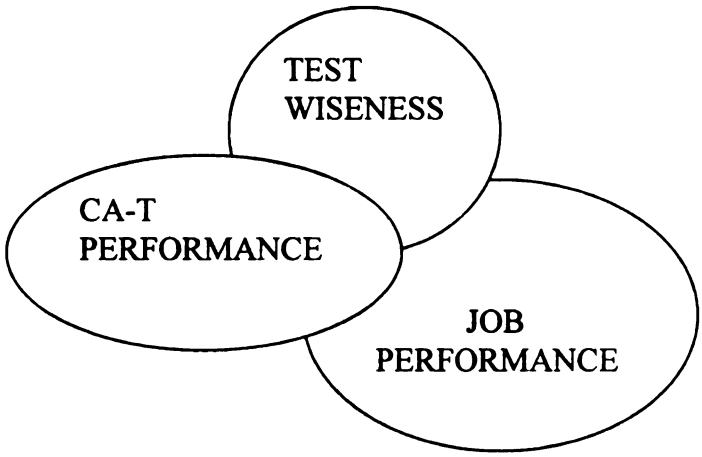

Figure 4. A criterion-related predictor bias

However, it is possible that TW itself may also serve as a predictor of job performance (i.e., relating to job performance). In other words, the criterion of job performance that an objective selection test aims to predict may be systematically influenced by this characteristic of test takers that may undercut the decision-making process in employment selection (see Nunnally & Bernstein, 1994 for a discussion of how factors such as gender and ethnic bias may contribute to criterion-related predictor problem). The overlapping sizes in such predictor-criterion relations are unclear and should be investigated in future research.

Another major concern that has made the construct of TW very controversial among test developers is the issue whether or not TW reduces the validity of a multiple-choice test. As aforementioned, there have been two competing camps of viewpoints, largely depending on how differently researchers define TW. Those who consider TW as nothing more than test takers' ability to take advantage of the idiosyncrasies imbedded in poorly-constructed tests (i.e., Scruggs, 1985) believe that the solution is removing test idiosyncrasies. As a result, the phenomenon of TW affecting test scores would not exist or have unsubstantial influence on test performance. The assumption is that test constructors are able to control for or at least mitigate the effects of TW on the validity of a predictor by developing a "good" test.

Those who define TW as being more than test takers' reliance on faulty test items to gain points (i.e., test-takers' ability to use *legitimate* cognitive and metacognitive strategies to handle a multiple-choice test, even a standardized one) believe that not only the influence of TW on test performance is real but also there are individual differences in TW levels (i.e., differences in knowledge of test-taking strategies and types of

strategies used during a test-taking situation; as evidenced in the present study). Given that knowing how to take a test effectively and strategically is a prerequisite for taking the test successfully, and that the knowledge or the application of such strategies during test taking cannot be controlled or suppressed by test constructors or administrators, individual deficiency in TW would result in the incomplete or misapplication of the true ability being tested in the observed test scores (Flippo & Borthwick, 1982; Gross, 1977; Sarnacki, 1979). In other words, the presence of TW may actually increase test validity. The cross-cultural study that Millman and Setijadi (1966) conducted might provide some indirect evidence for this argument. The researchers compared the performance of American and Indonesian high school students on open-ended and multiple-choice algebra tests. Both groups of students scored similarly well on the open-ended test. But on the multiple-choice version of the algebra test, the American students performed significantly better than the Indonesian students who had just learned how to respond to a multiple-choice question in a pre-test practice session. The implication is that objective tests may be less valid for those who have true ability but score lower on TW. Recently, Maurer, Solomon and Lippstreu (2003) found that coaching interview strategies for job candidates (i.e., orienting interview candidates on what to expect and how to best present answers in a clear/accurate manner, versus giving away answers or teaching candidates to falsely manage impression to distort answers) would reduce measurement error and slightly increase or at least do no harm to the construct and predictive validity of the structured interviews themselves.

In sum, the first contribution of the present study is that a new theoretical framework of TW was proposed, which incorporated components not commonly

acknowledged as an integral element of the TW construct. However, this study did not try to answer many questions related to TW and its possible links to other test-taking constructs, including the relationship between TW and general intelligence, nor did the present study attempt to test the proposed framework. The reason is that there are missing pieces in the puzzle: the operational definitions of the three components in the theory were still lacking. Therefore, the second purpose of this study was to define the most fundamental component of TW construct operationally, developing a measure of knowledge of TTS. As aforementioned, such an instrument would help to facilitate the research of individual differences in the TTS acquisition as an indicator of potential for TW.

The KOTTS Measure

I adopted a similar approach with the one that Parham (1996) used to develop her inventory of TTS: constructing a self-report measure of TTS from a comprehensive pool of strategies extracted from multiple sources. A possible question is the extent to which the KOTTS measure developed in this study might contribute to further the understanding of TTS and TW. In other words, compared with the TTS Checklist that Parham constructed, how different is the KOTTS in terms of scope, theoretical framework, and operational definition?

Scope. Parham established her pool of strategies from four main sources: Bruch (1981a), Millman et al. (1965), Ford (1973), and undergraduate focus groups. While relying on her measure as a first source, I extended my search of strategies to include more than a dozen other sources from both the research literature of TW or TTS and selective self-taught test-preparation materials. Therefore, the original version of the

Knowledge of Test-Taking Strategies (KOTTS) measure was grounded on an updated, comprehensive strategy pool.

Theoretical framework. In her dissertation, Parham initially sorted the strategies she found into six different categories, five of which were based on Millman et al. (1965) logical outline of TW. Trying the reliability analysis approach, she had not satisfactorily reproduced these a priori categories. She later used an exploratory factor analysis to determine the factor analytic solutions underlying her measure. Although I believed in the multi-factorial framework of a TTS instrument, I argued that the dimensions that both Millman et al. and Parham mentioned or tested in their works were neither complete nor totally accurate. Based on the empirical evidence regarding the effectiveness of individual strategies in the literature, as well as the variety in the actual strategies gathered for this study, I revised Millman et al.'s model into an 11-dimesion measure, modifying the existing aforementioned factors (i.e., splitting the cue-using strategy factor into 2 separate cue-using categories), and incorporating new dimensions that the research literature had not discussed, yet existing in some research literature and self-taught test-preparation materials (i.e., the strategies involving test takers' staying in control of the test and testing situation, or test takers' basic knowledge of the test format). Among the original 11 dimensions, one dimension (Changing answers) was later discarded from the final measure because of its poor reliability, and one dimension was split into two separate categories (Guessing, divided into Guessing and Careless guessing) based on the item analysis.

A series of confirmatory factor analyses on a refined version of KOTTS measure provided the fit indices for several alternative measurement models. These fit indices

verified that a multi-factorial measurement model was a reasonably good fit for the data. An exploratory factor analysis further confirmed the fact that an 11-factor structure (corresponding neatly with the content of the 11 a priori dimensions) underlay the KOTTS. Although these factors were related to each other to a certain extent (either positively or negatively), the correlations among them were low enough to make it possible for the categories of strategies to relate differentially to external criteria (Nunally & Bernstein, 1994). Therefore, I had established the discriminant validity for KOTTS subscales, in addition to the measure's content validity (based on expert judgments on the original version of this measure as well as the judgments I made for the item and scale content of the final version).

Operational definition. The operational definition of one's declarative knowledge of TTS at the item level may be debatable because it was the unorthodox combination of two ratings: familiarity and inclination to use. Arguably, these two variables refer to two seemingly different concepts: the extent to which a test taker is familiar with a particular strategy, and the extent to which this person intends to use this strategy in subsequent test taking. At the conceptual level, however, a composite score makes sense because it reflects more than the superficial knowledge of a strategy (i.e., being aware of the existence of the strategy); it also reflects the perceptiveness of a test taker as far as a test-taking strategy is concerned (i.e., judging the usefulness of this strategy). Underlying the need for a composite observation of declarative knowledge of TTS is the need to infer about the functionality of such knowledge. Interpreting a high composite score was straightforward: a test taker who endorsed familiarity highly also rated high on inclination (and vice versa, as indicated by the strong correlations observed for most

items in the KOTTS). However, a high score did not necessarily mean the declared knowledge is functional (i.e., leading to point gains); this is more likely a function of the applicability of a strategy to a given test. A low score on a KOTTS scale, therefore, was not necessarily bad because the low scores on a certain subscale of TTS knowledge might still be positively related to test performance (i.e., by not endorsing ineffective strategies under certain circumstances).

Note that a detail of wording in how each strategy was measured in the KOTTS is also debatable. Participants were asked to report "how inclined you *would* be to use each test-taking strategy," which might yield a hypothetical, future-oriented response based on that one can indirectly assess participants' personal judgment of how good or effective a strategy is. Some studies on hypothetical reference showed that respondents more often responded correctly to future hypothetical questions than to past hypothetical questions (i.e., Kuczaj, 1981). However, an alternative wording of this question might yield different responses. Participants were not asked to report a past behavior (i.e., how much/often they had used a strategy in taking multiple-choice tests in the past) which may yield either a long-term memory-type of responses (i.e., the recollection about the use of TTW on the most recent multiple-choice tests participants had taken prior to the present study), nor were they asked to report a typical behavior (i.e., how much/often participants typically use a strategy in a test-taking situation) which may yield a more general response that encompasses past and future behaviors.

At the scale level, the operation definition of TTS knowledge was sound because it was based on test takers' self-reported declarative knowledge of TTS (as defined above), which was more likely a stable individual characteristic than test takers' TTS

frequency of use or observed behaviors during a test-taking situation, which might rely on extraneous factors such as test-taking motivation, test anxiety, and the characteristics of the test itself.

Reliability. There might be a concern about the reliability of some subscales in KOTTS as indicated by their internal consistencies. Table 25 (above) shows that almost two-thirds of the subscales had Cronbach's α in the .50's or 60's range, which might affect the interpretation regarding the construct validity of these scales (i.e., whether the items in these scales have little in common). There is no fixed rule of thumb as far as how large coefficient α must be because it depends on the purpose of the measure being constructed. On the one hand, Guion (1998) recommends obtaining coefficient α of .70 or greater if possible to satisfy the need for content homogeneity. On the other hand, Nunnally and Bernstein (1994) suggest that it might be fine to have somewhat lower standards of reliability for a construct-validated measure because the measure is more likely to be used to obtain correlations and less likely to be used for making high-stakes decisions about individuals. Given the fact that most of the dimensions in KOTTS were a priori defined conceptually, the relatively low coefficient alphas are attributable to the small number of items in the respective scales. For example, the two subscales with the lowest reliabilities (Dimension 1, α = .59; Dimension 7, α = .58) consisted of two items only. In other words, it is possible to improve the internal consistency for these subscales by adding more strategies whose content adds meaningfully to their respective scale (i.e., *not* presenting the same concept in different wording). The implication is that, the current version of some KOTTS dimensions may be more suitable for use in research than for individual diagnoses of TTS level, unless the diagnostic information is used to make low-

stakes recommendations regarding individuals or groups. For example, an organization that uses CA-Ts as a major selection instrument may want to learn whether there are individual differences in TTS knowledge, or whether there is a need for providing some job applicants with a form of test-taking orientation (including TTS information). In this case, it is appropriate to administer the KOTTS measure to test takers prior to the selection test. However, one still needs to interpret the relationships between these subscales and the criterion of test performance or other criteria with caution.

Cross-validation. Another limitation of the KOTTS is that it had not been cross-validated. Because of the length of the original version of the KOTTS (78 items), I chose to use the whole sample ($N = 369$) to best assess item validity and produce stable statistics for the measure. Therefore, there was no cross-validation sample in this study to make sure the KOTTS items also worked on another sample of college students.[2] Regarding the convergent or divergent validity of KOTTS, I should have included a traditional measure of TTS in this study, such as the 46-item Ferrell Test of Test-wiseness (Ferrell, 1972) which measure two types of TTS (Deductive reasoning and cue-use strategies), or the 5-item multiple-choice test subscale in Dolly and Vick's (1986) measure of TW, which measure time adequacy in test taking. The particular strategies assessed in these traditional instruments might have been correlated positively with the corresponding subscales of the KOTTS measure, but not being related to other subscales that assess different constructs of test takers' TTS knowledge. This fact has provided further direct evidence of the convergent and divergent validity of this instrument.

KOTTS being over-refined. Another limitation, also related to the size of the development sample, is the plausibility that the KOTTS measure with 39 items being

arranged into 11 categories (i.e., an average of 4 items per scale) might have been over-refined because some of the decisions regarding discarding items may have been based on chance levels of relationships. Because the original measure was consisted of 78 items and given the relatively small sample size in the present study (i.e., not meeting the rule-of-thumb ratio of 10:1), it was possible that the validity coefficients of the KOTTS dimensions might have been artificially inflated because they capitalize on chance and may change when validated on a different sample (i.e., the conditions under which the original validity study was conducted would never be exactly reproduced.

Other limitations included the position of the pretest measure of TTS, and the empirically-supported decision of not reverse coding several strategy items based on their content.

Demand characteristics. Because the pretest measure of TTS knowledge was administered immediately prior to the administration of the CA-T and the posttest measures in the present study, one might suspect that participants might have been primed by reading and endorsing the KOTTS items; therefore, their actual use of strategies on the test might have increased, the reported use having been inflated because of their experiencing the pretest measure (the problem of reactions to pretest as discussed in Shadish, Cook, & Campbell, 2002).

[2] I am in the process of cross-validating the KOTTS measure with another sample of MSU students with similar characteristics (as of April 2003). The results may verify the internal validity of the KOTTS, at least for student samples. The KOTTS psychometric properties can also be explored with other operational samples (i.e., high school students, or adult job applicants) in future studies.

Although there was no certain answer for this question, I did go back to the data and check the percentages of missing values for all test items. Because the guessing strategies endorsed making a guess and responding to *all* questions before time was called to maximize the chance of getting additional correct responses, if the pretest measure indeed had primed participants toward employing this relatively simple strategy that was independent of test content, I would have seen a next-to-nil proportion of participants to leave items on their answer sheet blank. However, it was not the case. For the last three items of this short test, the percentages of missing values increased from 2.5% to 5.3% to 8.7%, which were relatively compatible to the test data collected in a different study on the use of TTS (Nguyen et al., 2003). If the actual use of a straightforward, common-sense strategy was any indication, it was possible but not plausible that reading the TTS immediately prior to taking the test would inflate the actual and reported use of more complex TTS, which might require practice and experience (or time) to acquire. Nevertheless, a research design with a time lapse between pre- and post-test administrations of measures should be able to prevent any possible demand characteristics in the future.

Reverse coding. On a different note, my decision for not reverse coding several items on the "inclination to use" scale, based on the empirical evidence (i.e., no consistent negative correlations detected) but not based on the item content, might be debatable. This decision relied on the rationale that by choosing to endorse different levels of their inclination to use a certain strategy participants implicitly expressed their perceptions of the effectiveness of such a strategy. The implication is that this decision may have obscured subgroup relations to TTS, in that a strategy may be ineffective for a specific

group of participants (i.e., minorities), but the same strategy may be good or effective for another group. By not reverse coding items, researchers would not be able to detect such subgroup differentiation except through future research involving subgrouping on some meaningful characteristics that relate to strategy effectiveness, perceived or actual.

Relations Between KOTTS Dimensions and Other Test-takers' Characteristics

Detailed discussions regarding the relations of interest have been presented throughout the section pertaining to the correlation analytic results. Therefore, I will summarize the major findings instead of repeating the aforementioned information. Table 29 (above) provides a summary of the empirical supports for the hypothesized links among the variables of interest. Except for Hypotheses 4a and 4b, which predicted a positive relationship between test takers' achievement levels (as indicated by their GPA) and both general TTS knowledge and specific TTS use, all other hypotheses were supported or partially supported in this study.

Partial support for hypotheses. The reason why most hypotheses were only supported partially is because they were exploratory in nature, not taking into account the varying relationships between each TTS dimension and other variables of interest. For example, Hypothesis 6 predicted that a high level of test-taking metacognition (i.e., giving oneself positive self-evaluation and monitoring one's test-taking process in general) would positively correlate with (all dimensions of) participants' reported TTS use. The results showed that, on the one hand, test-taking metacognition did not correlate with three particular dimensions that might be used without much contemplation ("Using physical cues," "Guessing," and "Careless guessing"); on the other hand, test-taking metacognition was correlated more strongly with four other dimensions whose content

indicated the need for a higher level of thinking about thinking or executing some cognitive monitoring ("Avoiding clerical errors," "Using grammatical cues," "Working carefully and thoroughly" and "Staying in control") than other dimensions in the KOTTS. Similar effects of KOTTS dimensionality were observed throughout other findings from the correlation analyses. The implication was twofold. First, the findings provided further evidence for the multidimensionality of the KOTTS measure. Second, the findings verified the fact that TTS (or at least the TTS used in a standardized CA-T) were not equal in their relationships with other test-taking constructs, particularly with test performance.

Magnitude of predictive validity of KOTTS. One might have some concern about the relatively small magnitudes of the relations between several KOTTS dimensions (both knowledge and use) and participants' test performance. Although the significant relationships (either in the positive or negative direction) supported the criterion-related validity of the KOTTS measure in general, the observed strengths of these links were not substantial (i.e., between |.10| to |.30|[3]). Some researchers would argue that the portion of variance in test scores that these TTS dimensions explained would not be large enough to become practically important (i.e., Scruggs, 1985). However, considering the fact that test takers' cognitive ability (as indicated by their ACT or SAT scores) was correlated r = .56 ($p < .01$) with the test performance, or explaining about 31% of the variance of the test scores in this study, the 8 or 9% of the test performance variance being explained by a single TTS subscale would be noteworthy of research attention.

[3]These correlations were in fact comparable or stronger than the commonly reported relationships between the traditional "TW" and objective test scores in the research literature.

The fact that the sample in the present study consisted of college students who are test-taking veterans might have also restricted the range of the predictor (i.e., dimensions in TTS knowledge), which reduced the strength of any possible predictor-criterion relationships. The fact that the criterion itself was multidimensional (i.e., both the verbal subtest and the mathematical subtest included different types of verbal or mathematical test items that were representative of the original long version; thus the lower internal consistency: Verbal $\alpha = .60$, Math $\alpha = .50$) also deflated the magnitude of relations with KOTTS dimensions.

Implication of TW Theory

An important theoretical implication for the proposed framework of TW in the present study is the importance of conceptually distinguishing TW from the narrower construct of TTS. The tendency is using the 2 terms interchangeably, or equating TTS aspects to TW aspects. This tendency may create a conceptual confusion or misunderstanding about the nature of construct of TW, hindering research in this area. Therefore, I strongly urge future researchers to recognize that the application of TTS is only a subcomponent of TW, and its measure (11-category KOTTS) while multidimensional should not be confused with the multidimensionality of TW (cognitive, metacognitive and behavioral elements).

Practical Use of the KOTTS Measure

I have discussed the theoretical use of the KOTTS measure in terms of testing the theoretical model of TW. What about the practical use of this instrument? Although CA-Ts have long been verified as a sound measure of general mental ability that was empirically linked to job proficiency, job success, or economic gains for organizations

(i.e., Hough, Oswald, & Ployhart, 2001; Hunter & Hunter, 1984; Pearlman, Schmidt, & Hunter, 1980), only a small proportion of organizations have been using some form of standardized CA-Ts for personnel screening purposes.

For example, Terpstra and Rozel (1993) conducted a mail survey with 201 U.S. companies with 200 or more employees; the researchers found that only 20% of their organization respondents reportedly used CA-Ts as a common staff screening practice. If the statistics are accurate, there may be a smaller percentage of organizations that are concerned about increasing the validity of their selection CA-Ts by proactively assessing and/or training their job applicants in terms of test-taking techniques and strategies.

Nevertheless, for those who are interested in or want to be informed about job applicants' differences in levels of TW in general and the knowledge of TTS dimensions in particular, the KOTTS measure should be a good start. Recently, Nguyen et al. (2003) found a racial difference in a general measure of TTS ($r = -.18, p < .05$; African Americans were more likely than Whites to report using TTS in general). They also found that, in one experiment condition, a high level of TTS use significantly mitigated the group mean difference in quantitative reasoning test score between Blacks and Whites. If future studies find more empirical support not only for individual differences in the knowledge base and reported use of TTS, but also for group differences (i.e., minority test takers such as African Americans and Latino Americans having a lower level of TTS knowledge or use, knowing of or using different strategies than Caucasians, or TTS training would have differential influences on subgroup test scores), organizations that are concerned about adverse impact of the CA-Ts but still prefer them to other

selection practices may find the continuing research on TTS and TW of the particular use.

For test developers and test administrators, the results in the present study regarding KOTTS multidimensionality show that it is important to distinguish between dimensions of TTS that can be controlled by improving test validity (i.e., by removing test idiosyncrasies) and dimensions of TTS that their application may actually increase test validity. Specifically, by constructing good test items and standardizing them, test developers can "disarm" test takers who use strategies that mostly capitalize on the presence of faulty test items such as "Using physical cues" (i.e., a correct option that is longer than other options) and "Using grammatical or contextual cues" (i.e., specific determiners such as *always*, *never*; a correct option that is grammatically consistent with the stem; except when these cues are purposefully measured such as in an English grammar test). However, the best test developers cannot control the fact that test takers may use appropriate strategies such as "Knowing how multiple-choice tests work," "Optimizing time efficiency and effectiveness," "Avoiding clerical errors," "Deductive reasoning," "Using recall aids," "Working carefully and thoroughly," and "Staying in control" because these strategies are derived from the format and inherent characteristics of most multiple-choice tests themselves. In other words, these categories of TTS are generally effective or at least not doing test takers any harm across multiple-choice tests (i.e., regardless of the test is a teacher-made one or a standardized CA-T). In fact, test administrators should identify these strategies for test takers prior to a test and encourage them to apply such strategies to ensure that one will not underperform because of any misapplication of the knowledge being tested. A similar approach has been taken by

some test developers or administrators when they have explicitly encouraged test takers to use "Guessing" strategies when running out of time on a test (when incorrect responses are not penalized).

For test takers and test-taking coaches, it is essential that they are able to recognize and acquire or train test takers in effective test-taking skills as well as avoiding "harmful" strategies. I put the term harmful in quotation marks because what is effective may depend largely on a test itself, not on the strategies per se. For example, when taking objective classroom examinations whose constructors are not very experienced, some test takers may find the strategies in the dimensions of "Using physical cues" and "Using grammatical cues" particularly effective, in addition to the aforementioned legitimate strategies. However, as evidenced in the present study, the use of physical cues may be detrimental for test performance if applied to a well-constructed CA-T, and the use of grammatical/contextual cues may not make any difference in standardized test scores. Therefore, I strongly recommend that test takers should not blindly adopt and apply all test-taking tips or tactics provided in test coaching materials but should critically think about what strategies to acquire and under what circumstances to use them.

The only strategy that test takers should definitely avoid under every circumstance is "Careless guessing" (i.e., a random guess as soon as something appears difficult), underlying which there may be a lack of strong motivation to exert efforts and do well on a test.

Test takers should adopt the single strategy of "Changing answers" when necessary because past research as well as the present study supported its effectiveness (i.e., positively relating to higher test scores).

In summary, this study showed that TW could be conceptualized as an integrated psychological phenomenon that occurs in every test-taking situation where a multiple-choice test is used. The foundation for TW, the multidimensional construct of TTS knowledge, can be measured with relative reliability in young adults. The KOTTS measure showed content validity, discriminant validity and criterion-related validity. Future research should cross-validate this instrument with both a college student sample and a sample of adult test takers (i.e., job applicants) in order to establish the reliability and validity of the measure further.

References

Anastasi, A. (1976). *Psychological Testing* (4th Edition). New York: Macmillan.

Anastasi, A., & Cordova, F. (1953). Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York City. *Journal of Educational Psychology, 44,* 1-19.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum.

Anderson, S. B., & Sauser, W. I. Jr. (1995). Measurement of test anxiety: An overview. In C. D. Spielberger, & P. R. Vagg (Eds.), *Test anxiety: Theory, assessment, and treatment* (pp. 15-33). Washington, DC: Taylor & Francis.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695-716.

Bajtelsmit, J. W. (1975). *Development and validation of an adult measure of secondary cue-using strategies on objective examinations: The test of obscure knowledge (TOOK).* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Barbera, K. B. (1994). *Test attitudes and their effect on construct and criterion-related validity of selection tests.* Unpublished doctoral dissertation, Bowling Green University, Ohio.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233-241.

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*(3), 133-141.

Berryman, S. E. (1983). *Who will do science? Minority and female attainment of science and mathematics degrees: Trends and causes.* New York: The Rockefeller Foundation.

Best, J. B. (1979). Item difficulty and answer changing. *Teaching of Psychology, 6,* 228-230.

Bielinski (1999). *Sex difference by item difficulty: An interaction in multiple-choice mathematics achievement test items administered to national probability samples (DIF).* Unpublished doctoral dissertation, University of Minnesota, US.

Bloom, B. S., & Broder, L. J. (1950). *Problem-solving processes of college students: An exploratory investigation.* Chicago: The University of Chicago Press.

Bobrow, J. (1999). *How to prepare for the Law School Admission Test.* Hauppage, NY: Barron's Education Series.

Borrello, G. M., & Thompson, B. (1985). Correlates of selected test-wiseness skills. *Journal of Experimental Education, 53*(3), 124-128.

Brenk, K., & Bucik, V. (1994). Guessing of answers in objective tests, general mental ability and personality traits according to the 16-PF questionnaire. *Review of Psychology, 1*, 11-20.

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology, 88*, 333-340.

Brownstein, S. C., Weiner, M., Green, S. W., & Hilbert, S. (1999). *How to prepare for the Graduate Record Examination.* Hauppage, NY: Barron's Education Series.

Brozo, W. G., Schmelzer, R. V., & Spires, H. A. (1984). *A study of test-wiseness clues in college and university teacher-made tests with implications for academic assistance centers.* (Report No. 84-01). Georgia: College Reading and Learning Assistance Technical Report. (ERIC Document Reproduction Service No. ED 240 928).

Bruch, M. A. (1981a). *Rater's manual for questionnaire of how you take tests* (QHTT) (2nd ed.) Unpublished manuscript.

Bruch, M. A. (1981b). Relationship of test-taking strategies to test anxiety and performance: Toward a task analysis of examination behavior. *Cognitive Therapy and Research, 5*, 41-56.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: A six year study.* (Final Project Report No. PR-73-37). Princeton, NJ: Educational Testing Service.

Carr, J. Z., Bell, B. S., Ryan, A. M., & Kilanowski, D. E. (2002). *An examination of factors that may enhance the retention and success of minorities in the selection process.* A technical report to the City of Toledo.

Cassidy, J. C. (2001). *Self-reported GPA and SAT scores.* (Report No. EDO-TM-01-04). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300-310.

Chang, T. (1979). Test wiseness and passage-dependency in standardized reading comprehension test items. *Dissertation Abstracts International, 39*(4-12), 7-8.

Clause, C. S., Delbridge, K., Schmitt, N., Chan, D., & Jennings, D. (2001).Test preparation activities and employment test performance. *Human Performance,14*, 149-167.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*, 61-75.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Weisman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15-41.

Cook, D. (1957). A comparison of reading comprehension scores obtained before and after a time announcement. *Journal of Educational Psychology, 48*, 440-446.

Crehan, K. D., Gross, L. J., Koehler, R. A., & Slakter, M. J. (1978). Developmental aspects of test-wiseness. *Educational Research Quarterly, 3*, 40-44.

Cronbach, L. J. (1946). Response sets and test validity. *American Psychologist, 1*, 247-248.

Davison, G. C., Vogel, R. S., & Coffman, S. G., (1997). Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. *Journal of Consulting and Clinical Psychology, 65*, 950-958.

Davison,G. C., Haaga, D. A., Rosenbaum, J., Dolezal, S. L., & Weisntein. (1991). Assessment of self-efficacy in articulated thoughts: "States of mind" analysis and association with speech-anxious behavior. *Journal of Cognitive Psychotherapy, 5*(2), 83-92.

Dearman, N. B., & Plisko, V. W. (1981). *The condition of education.* Washington, DC: National Center for Education Statistics.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement, 9*, 145-150.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Dolly, J. P., & Vick, D. S. (1986). An attempt to identify predictors of testwiseness. *Psychology Reports, 58,* 663-672.

Doverspike, D., & Miguel-Feruito, R. (2001, April). *Test-wiseness training as a mechanism for reducing adverse impact: Effects on responses to reading comprehension test items with omitted passages.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Dressel, E. P., & Jensen, G. C. (1955). *How to study and read successfully.* Longmeadow, MA: Teachers National Information Service.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L., & Damrin, D. E. (1960). Tests and examinations. In C. Harris (Ed.), *Encyclopedia of educational research* (pp. 1502-1517). New York: Macmillan.

Ellis, A., & Ryan, A. M. (in press). Race and cognitive ability test performance: The mediating effects of test-taking strategy use, test preparation, and test-taking self-efficacy. *Journal of Applied Social Psychology.*

Engelhardt, D. F. (1979). Motivation and test-wiseness. In C. Banks, D. Penfield, & P. Miller (Eds.), *Director's handbook: Topics in testing measurement and evaluation* (pp. 1-13). Trenton, NJ: New Jersey State Department of Education.

English, H. B., & English, A. C. (1970). *A comprehensive dictionary of psychological and psychoanalytical terms.* New York: David McKay.

Erickson, M. E. (1972). Test sophistication: An important consideration. *Journal of Reading, 16,* 140-144.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge: MIT Press.

Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology, 79,* 95-97.

Fagley, N. S., Miller, P. M., & Downing, R. (1990, August). *Convergent and discriminant validity of the Experimental Test of Testwiseness.* Paper presented at the annual meeting of the American Psychological Association, Boston, MA.

Farnham, S. D. (1999). From implicit self-esteem to in-group favoritism. *Dissertation Abstracts International: Section-B:The Sciences and Engineering, 60*(4-B), 1912.

Ferrell, G. M. (1972). *The relationship of scores on a measure of test-wiseness to performance on teacher-made objective achievement examinations, and on standardized ability and achievement tests, to grade point average, and to sex for each of five high school samples.* Unpublished doctoral dissertation, University of Southern California.

Flippo, R. F., & Borthwick, P. (1982). Should testwiseness curriculum be a part of undergraduate teacher education? In G. H. McNinch (Ed.), *Reading in the disciplines* (pp. 117-120), Athens, GA: American Reading Forum.

Flippo, R. F., Becker, M. J., & Wark, D. M. (2000). Preparing for and taking tests. In R. F. Flippo & D. C. Caverly (Eds.), *Handbook of college reading and study strategy research* (pp. 221-260). Mahwah, NJ: Lawrence Erlbaum.

Flynn, J., & Anderson, B. (1977, Summer). The effects of test item cue sensitivity on IQ and achievement test performance. *Educational Research Quarterly, 2*(2), 32-39.

Ford, V. A. (1973). *Everything you wanted to know about test-wiseness.* (ERIC Document Reproduction Service No. ED 093 912).

Frederick, R. W. (1938). *How to study handbook.* New York: Appleton-Century.

Frederickson, C. G. (1999). Multiple-choice answer changing: A type connection? *Journal of Psychology Type, 51*, 40-46.

Gael, S., & Grant, D. L. (1972). Employment test validation for minority and non minority telephone company service representatives. *Journal of Applied Psychology, 56*, 135-139.

Geiger, M. A. (1991). Changing multiple-choice answers: Do students accurately perceive their performance? *Journal of Experimental Education, 59*, 250-257.

Gentry, J. M., & Perry, J. S. (1993). *Test-wiseness, memory, and academic performance in university students.* ERIC, URL: http://orders.edrs.com/members/sp.cfm?AN=ED375351.

Gibb, B. G. (1964). *Test-wiseness as secondary cue response.* Unpublished doctoral dissertation, Stanford University, CA.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Gottfredson, L. S. (1986). Societal consequences of the *g* factor in employment. *Journal of Vocational Behavior, 29*, 379-410.

Green, D. S., & Stewart, O. (1984). Test wiseness: The concept has no clothes. *College Student Journal, 18*, 416-424.

Gross, L. J. (1975). *The effects of three selected aspects of test-wiseness on the standardized test performance of eight grade students*. Unpublished doctoral dissertation, State University of New York at Buffalo.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Harmon, M. G. (1997). *The relationship between metastrategic knowledge and testwiseness. Unpublished doctoral dissertation*, Missouri State University, Missouri.

Harmon, M. G., Morse, D. T., & Morse, L. W. (1996). Confirmatory factor analysis of the Gibb Experimental Test of Testwiseness. *Educational and Psychological Measurement, 56*, 276-286.

Harvill, L. M. (1984a, April). *Measuring the test-wiseness of medical students*. Paper presented at the 68th annual meeting of the American Educational Research Association, New Orleans, LA.

Harvill, L. M. (1984b, November). *Test anxiety, previous test taking experience, and test-wiseness in entering medical students*. Paper presented at the 13th annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Hennessy, J. J., & Merrifield, P. R. (1978). Ethnicity and sex distinctions in patterns of aptitude factor scores in a sample of urban high school seniors. *American Educational Research Journal, 15*, 385-389.

Hernstein, R. J., & Murray, C. (1994). *The Bell Curve: Intelligence and class structure in American life*. New York: Free Press.

Hoffman, B. (1962). *The tyranny of testing*. New York: Collier.

Holden, L. M. (1996). *The effectiveness of training in improving test performance*. Poster presented at the meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hollandsworth, J. G., Glazeski, R. C., Kirkland, K., Jones, G., & van Norman, L. R. (1979). An analysis of the nature and effects of test anxiety: Cognitive, behavioral, and physiological components. *Cognitive Therapy and Research, 3*, 165-180.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues,

evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1-2), 152-194

Huang, H. J. (1995). Commitment to a chosen course of action: Individual differences in decisional persistence. *Dissertation Abstracts International Section A: Humanities & Social Sciences, 56*(1-A), 0268.

Huck, S. (1978). Test performance under the condition of known item difficulty. *Journal of Educational Measurement, 15*, 53-58.

Huff, D. (1961). *Score: The strategy of taking tests*. New York: Appleton-Century-Crofts.

Hunter, F. L. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U. S. Employment Service, U. S. Department of Labor.

Hunter, F. L. (1981). *The economic benefits of personnel selection using ability tests: A state of the art review including a detailed analysis of the dollar benefit of U.S. employment Service placements and a critique of the low cutoff method of test use*. Report prepared for the U.S. Employment Service, U. S. Department of Labor, Washington, D.C.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6*, 151-158.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721-735.

Hunter, J., & Hunter, R. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Jacobs, S. S. (1975, April). *Test-wiseness: Several methodological problems*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Jaffe, E. D., & Hilbert, S. (1998). *How to prepare for the Graduate Management Admission Test*. Hauppage, NY: Barron's Education Series.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, S. T. (April, 1984). *The test, the tested, and the test-taking: A model to better understand test performance.* Paper presented at the 68th meeting of the American Educational Research Association, New Orleans, LA.

Johnston, J. J. (1975). Sticking with first responses on multiple choice exams: For better or for worse? *Teaching of Psychology, 2*, 178-179.

Johnston, J. J. (1977). Exam-taking speed and grades. *Teaching of Psychology, 4*, 148-149.

Johnston, J. J. (1978). Answer-changing behavior and grades. *Teaching of Psychology, 5*, 44-45.

Jones, P., & Kaufman, G. (1975). The differential formation of response sets by specific determiners. *Educational and Psychological Measurement, 35*, 821-833.

Kalechstein, P., Kalechstein, M., & Docter, R. (1981). The effects of instruction on test-taking skills in second grade black children. *Measurement and Evaluation in Guidance, 13*(4), 198-202.

Kim, Y. H., & Goetz, E. T. (1993). Strategic processing of test questions: The test marking responses of college students. *Learning and Individual Differences, 5*, 211-218.

Kirkland, K., & Hollandsworth, J. G. Jr. (1980). Effective test taking: Skills acquisition versus anxiety-reduction techniques. *Journal of Consulting and Clinical Psychology, 48*, 431-438.

Kuczaj, S. E. (1981). Factors influencing children's hypothetical reference. *Journal of Child Language, 8*, 131-137.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139-161.

Linn, R. L. (1978). Single-group validity, differential validity, and differential predictions. *Journal of Applied Psychology, 63*, 507-514.

Loehlin, J. C., Lindzey, G., & Spuhler, J. M. (1975). *Race differences in intelligence.* San Francisco: Freeman.

Lynch, D., & Smith, B. (1975). Item response changes: Effects on test scores. *Measurement and Evaluation in Guidance, 7*, 220-224.

Maller, J. B., & Zubin, J. (1933). The effect of motivation upon intelligence test scores. *Journal of Genetic Psychology, 41*, 136-151.

Maurer, T. J., Solomon, J. M., & Lippstreu, M. (2003, April). *Structured interviews: Effects of coaching interviewees on performance and validity.* Paper presented at the meeting of Society for Industrial-Organizational Psychology, Orlando, FL.

McClain, L. (1983). Behavior during examinations: A comparison of A, C, and F students. *Teaching of Psychology, 10*(2), 69-71.

McMorris, R., & Leonard, G. (1976, April). *Item response changes and cognitive style.* (ED 129 918)

Miguel-Feruito, R. F. (1997). *Explaining passage independence: An analysis of the ability to respond to reading comprehension test items when the passages are omitted.* Unpublished doctoral dissertation, University of Akron, OH.

Miller, P. M., Fagley, N. S., & Lane, D. S., Jr. (1988). Stability of the Gibb (1964) Experimental Test of Testwiseness. *Educational and Psychological Measurement, 48*, 1123-1127.

Miller, P. M., Fagley, N. S., Downing, R., Jones, R. N., Campbell, J. L., & Knox-Harbour, P. (1989, August). *Convergent validity of the Gibb (1964) Experimental Test of Testwiseness.* Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

Millman, J. (1966). *Test-wiseness in taking objective achievement and aptitude examinations.* Final Report, College Entrance Examination Board.

Millman, J., & Setijadi (1966). A comparison of American and Indonesian students on three types of test items. *Journal of Educational Research, 59,* 273-275.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726.

Montague, M., & Applegate, B. (1993). Middle school students' mathematical problem solving: An analysis of think-aloud protocols. *Learning Disability Quarterly, 16,* 19-32.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14*, 9-13.

Mueller, D., & Schwedel, A. (1975). Some correlates of net gain resulting from answer changing on object achievement test items. *Journal of Educational Measurement, 12*(4), 251-254.

Nguyen, H. -H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16,* 261-293.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nunnally, J. C., & Berntstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Oakland, T. (1972). The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. *Journal of School Psychology, 10*, 355-360.

Ogbu, J. U. (1991). Immigrant and involuntary minorities in comparative perspective. In M. A. Gibson & J. U. Ogbu (Eds.), *Minority status and schooling: A comparative study of immigrant and involuntary minorities* (pp. 3-33). New York: Garland Publishing.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress Mathematics performance. *Educational Assessment, 3*, 135-157.

Papoulis, A. (1990) *Probability and Statistics*. New Jersey: Prentence-Hall International Editions.

Parham, S. E. (1996). *The relationships between test-taking strategies and cognitive ability test performance*. Unpublished doctoral dissertation, Bowling Green State University, Ohio.

Parrish, B. W. (1982). A test to test test-wiseness. *Journal of Reading, 25*, 672-675.

Passbook (1998). *Rudman's questions and answers on the Dental Admission Test*. Syosset, NY: National Learning.

Paul, C. A., & Rosenkoetter, J. S. (1980). The relationship between the time taken to complete an examination and the test score received. *Teaching of Psychology, 7*, 108-109.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 64, 373-406.

Penfield, D., & Mercer, M. (1980). Answer changing and statistics. *Educational Research Quarterly, 5*(5), 50-57.

Piaget, J. (1954). *The construction of reality in the child*. Oxford, England: Basic Books.

Pike, L. (1978). Short term instruction, test-wiseness, and the Scholastic Aptitude Test: A literature review with research recommendations. Princeton, NJ: ETS.

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33-40.

Powers, D. E., & Leung, S. W. (1995). Answering the new SAT reading comprehension questions without the passages. *Journal of Educational Measurement, 32*, 105-129.

Presley, M., Borkowski, J. G., & Schneider, W. (1987). *Cognitive strategies: Good strategies users coordinate metacognition and knowledge.* In R. Vasta & G. Whitehurst (Eds.), Annals of Child Development (Vol. 5, pp. 89-129). Greenwich, CT: JAI Press.

Preston, R. (1964). Ability of students to identify correct responses before reading. *Journal of Educational Research, 58*, 181-182.

Psychological Services, Inc. (1985). *PSI basic skills tests for business, industry, and government.* Glendale, CA: Test Publications Division, author.

Quinn, K. D. (1993). *The effects of test-preparation training on intelligence test score composites.* Poster presented to the 8th meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Rindler, S. E. (1980). The effects of skipping over more difficult items on time-limited tests: Implications for test validity. *Educational and Psychological Measurement, 40*, 989-998.

Rogers, W. T., & Bateson, D. J. (1991). Verification of a model of test-taking behavior of high school seniors. *Journal of Experimental Education, 59*, 331-350.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S. III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple choice tests? *Journal of Educational Measurement, 11*, 15-23.

Ruben, D. H. (1999). Diagnosing alcoholism and its addictive patterns using self-report rating scales. *Alcoholism Treatment Quarterly,17*, 37-46.

Sackett, P. R., & Wilk, S. L. (1994). Within group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929-954.

Samson, G. E. (1985). Effects of training in test-taking skills on achievement test performance: A quantitative synthesis. *Journal of Educational Research, 78*, 261-266.

Sarnacki, R. E., (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49*, 252-279.

Schmidt, F. L. (1988). The problem of ability test scores in employment selection. *Journal of Vocational Behavior, 33*, 272-292.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33*, 705-724.

Schmit, M. J., & Ryan, A. M. (1992). Test-Taking dispositions: A missing link? *Journal of Applied Psychology, 77*, 629-637.

Schraw, G. (1997). The effect of generalized metacognitive knowledge of test performance and confidence judgments. *The Journal of Experimental Education, 65*, 135-146.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement, 28*(2), 163-171.

Schwarzer, R., van der Ploeg, H. M., & Spielberger, C. D. (Eds.) (1982). *Advances in test anxiety research, Vol 1*. Hillsdale, NJ: Lawrence Erlbaum.

Scruggs, T. E. (1985). *The administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school children. Year Two final report.* Washington, DC: Special Education Programs.

Scruggs, T. E., & Lifson, S. A. (1985). Current conceptions of test-wiseness: Myths and realities. *School Psychology Review, 14*, 339-350.

Scruggs, T. E., & Mastropieri, M. A. (1992). *Teaching test-taking skills: Helping students show what they know.* Purdue University: Brookline Books.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shipley, W. C. (1986). *Shipley Institute of Living Scale*. CA: Western Psychological Corporation.

Shuller, S. M. (1979). *A large-scale assessment of an instructional program to improve test-wiseness in elementary school students*. New York: Educational Solutions.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O.* (AFHRL-TR-86-68). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Skinner, N. F. (1983). Switching answers on multiple-choice questions: Shrewdness or shibbaleth? *Teaching of Psychology, 10*, 220-222.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement, 19*, 211-220.

Spielberger, C. D. (1966). The effects of anxiety on complex learning and academic achievement. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press.

Spielberger, C. D. (1980). *Test anxiety inventory: Preliminary professional manual.* Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., & Vagg, P. R. (1995). *Test anxiety: Theory, assessment, and treatment.* Washington, DC: Taylor & Francis.

Stanley, J. C. (1971). Predicting college success of the educationally disadvantaged. *Science, 171*, 640-647.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and intellectual test performance of African Americans. *Attitudes and Social Cognition, 69*, 797-811.

Stock, W. A., Kulhavy, R. W., Pridemore, D. R., & Krug, D. (1992). Responding to feedback after multiple-choice questions: The influence of response confidence. *Quarterly Journal of Experimental Psychology, 45A*, 649-667.

Terpstra, D. E., & Rozel, E. J. (1993). The relationship of staffing practices to organizational level measures of performance. *Personnel Psychology, 46*, 27-48.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York: John Wiley & Sons.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagan, E. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York: Macmillan.

Tian, S. P. (2000). *TOEFL reading comprehension: Strategies used by Taiwanese students with coaching-school training.* Unpublished dissertation.
Towns, M. H., & Robinson, W. R. (1993). Student use of test-wiseness strategies

in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching, 30*, 709-722.

Towns, M. H., & Robinson, W. R. (1993). Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching, 30*, 709-722.

Wahlstrom, M., & Boersma, F. J. (1968). The influence of test-wiseness upon achievement. *Educational and Psychological Measurement, 28*, 413-420.

Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology, 19*, 251-265.

Wigdor, A. K., & Garner, W. R. (1982). *Ability testing: Uses consequences, and controversies*. Washington, DC: Academy Press.

Woodley, K. K. (1975). *Test-wiseness: A cognitive function?* Paper presented at National Council on Measurement in Education, Washington, DC.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*, 51-59.

APPENDICES

APPENDIX A

An Outline of Test-Wiseness Principles

(Source: Millman, Bishop, & Ebel, 1965; pp. 711-713)

I. Elements independent of test constructor or test purpose.

A. Time-using strategy
   1. Begin to work as rapidly as possible with reasonable assurance of accuracy.
   2. Set up a schedule for progress through the test.
   3. Omit or guess at items, which resist a quick response.
   4. Mark omitted items, or items that could use further consideration, to assure easy relocation.
   5. Use time remaining after completion of the test to reconsider answers.

B. Error-avoidance strategy
   1. Pay careful attention to directions, determining clearly the nature of the task and the intended basis for response.
   2. Pay careful attention to the items, determining clearly the nature of the question.
   3. Ask the examiner for clarification when necessary, if it is permitted.
   4. Check all answers.

C. Guessing strategy
   1. Always guess if right answers only are scored.
   2. Always guess if the correction for guessing is less severe than a "correction for guessing" formula that gives an expected score of zero for random responding.
   3. Always guess even if the usual correction or a more severe penalty for guessing is employed, whenever elimination of options provides sufficient chance of profiting.

D. Deductive reasoning strategy
   1. Eliminate options which are known to be incorrect and choose from among the remaining options.
   2. Choose neither or both of two options which imply the correctness of each other.
   3. Choose neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other.
   4. Restrict choice to those options which encompass all of two or more given statements known to be correct.
   5. Utilize relevant content information in other test items and options.

II. Elements dependent upon the test constructor or purpose.

A. Intent consideration strategy
   1. Interpret and answer questions in view of previous idiosyncratic emphases of the test constructor or in view of the test purpose.
   2. Answer items as the test constructor intended.
   3. Adopt the level of sophistication that is expected.
   4. Consider the relevance of specific details.

B. Cue-using strategy
   1. Recognize and make use of any consistent idiosyncrasies of the test constructor which distinguish the correct answer from incorrect options.
      a. He makes it longer (shorter) than the incorrect options.
      b. He qualifies it more carefully, or makes it represent a higher degree of generalization.
      c. He includes more false (true) statements.
      d. He places it in certain physical positions among the options (such as in the middle).
      e. He places it in certain logical positions among an ordered set of options (such as the middle of the sequence).
      f. He includes (does not include) it among similar statements, or makes (does not make) it one of a pair of diametrically opposite statements.
      g. He composes (does not compose) it of familiar or stereotyped phraseology.
      h. He does not make it grammatically inconsistent with the stem.
   2. Consider the relevancy of specific detail when answering a given item.
   3. Recognize and make use of specific determiners.
   4. Recognize and make use of resemblances between the options and an aspect of the stem.
   5. Consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.

APPENDIX B

Test-Taking Strategy Checklist

(Source: Parham, 1996)

Check to make sure you have answered every question.
Answer all the questions in order without skipping any.
First answer questions you are sure are right.
Pay attention to how much time is left so you can finish in the allotted time.
Always choose "All of the above" when you are not sure of the correct answer.
Look at the total number of questions and calculate how long you have to answer each question.
Choose neither or both of two choices that imply the correctness of each other.
Recognize and make use of specific determiners (e.g., always, never)
Use the first answer that looks correct and go on to the next question.
Make sure the ovals on the answer sheet are completely filled in.
Always guess if right answers only are scored.
If you don't know the correct answer right away, skip that question and come back to it later.
Use time remaining after completion of the test to reconsider answers.
Go with your first instinct and don't change your answer.
Watch for overlapping answer choices in which the truth of one choice implies the correctness of several others.
Choose answers that will make a design on your answer sheet.
Pay close attention to directions concerning allotted time and scoring procedures.
Pick one letter (A, B, C, D, or E) and use that same letter for the answer to all the questions you are not sure of.
Choose answer that is one of a pair or opposite statements.
Eliminate similar answer choices, i.e., choices that imply the correctness of each other.
Choose answer that is one of several similar statements.
Look for correct answers that are more precise and specific in meaning than other choices.
Take your time working through the test.
Look at your neighbor's answers if you are not sure of the answer.
Work as rapidly as possible with reasonable assurance of accuracy.
Pay careful attention to directions.
Remember to mark the answer you've chosen.
Select the answer choice that resembles some aspect of the question, such that a name or phrase is repeated in both the question and the answer.
Recheck your answers if there is a series of questions that all have the same letter answer choice.
Define vague terms in your own words.
Read deeply into the question and ask yourself what the question is really asking.

When reading the questions, cover-up the answer choices and look only at the question.
Choose answer composed of familiar or stereotyped phrases.
Use each answer choice as an answer to the question and see which choice works best.
Identify the choice code and question number and carefully transfer this information from the test booklet to the answer sheet.
Beware of one item "giving away" the answer to another question occurring in a later part of the test.
Always choose the most complicated answer.
Recognize and make use of similarities between the answer and the question.
Underline key terms and clue words in question.
Choose answer that represents a higher degree of generalization.
Always underline phrases that make the question negative such as "not," "except," or "false."
Use scrap paper to figure answers before looking at answer choices.
Examine carefully all alternatives before attempting to choose the correct answer.
Consider the subject matter and difficulty of neighboring questions when interpreting and answering a give question.
When guessing, pick an answer choice at random.
Read the questions carefully.
Do not hesitate to change answer if you feel you should.
Read all of the answer choices before you read the question.
Rule out choices that contradict the question.
Choose answer that is grammatically consistent with the question.
Try to think of answer before reading choices.
Consider the relevance of specific detail when answering a given question.
If you are unsure of the answer, guess "C."
Read the answer choices from the bottom up (i.e., D, C, B, A).
Read and understand question before answering.
Ask examiner for clarification when necessary, if it is permitted.
Check answers during any remaining time to assess correctness and avoid careless mistakes.
Try to work backwards from each answer, especially with math questions.
Choose answer placed in a certain logical position among an ordered set of choices (such as in the middle of a sequence).
Skip ahead to questions you know more about.
Eliminate answer choices known to be incorrect and choose from among the remaining options.
Read through the entire test before you start working.
Understand the proper way to mark responses.
Guess only after an honest attempt has been made to answer the question.
Don't read too much into a question.
Guess, especially if one or more wrong alternatives can be identified.
Take your time thinking about each question.
If all else fails, choose "D."
Choose answer that is obviously longer than the other choices.
Determine clearly the nature of the question.

With "all of the above" type questions, treat each choice as a True/False question.
Concentrate your attention to test-relevant variables while directing attention away from self-evaluative thoughts.
Mark questions you are not sure of to go back and review when you are finished.
If you are unsure, guess "B."
Set up a schedule for progress through the test.
Always guess whenever elimination of some choices provides sufficient chance of guessing correctly.
Compare each answer choice with other choices and determine any differences.
Choose answer placed in a certain physical position among the answer choices (such as in the middle).

APPENDIX C

Strategy Source References

ARCO (1999). *Mechanical aptitude and spatial relations tests* (4th Edition). New York: Mcmillan General Reference.

Bobrow, J. (1999). *How to prepare for the Law School Admission Test.* Hauppage, NY: Barron's Education Series.

Brownstein, S. C., Weiner, M., Green, S. W., & Hilbert, S. (1999). *How to prepare for the Graduate Record Examination.* Hauppage, NY: Barron's Education Series.

Budd, T., Frade, J. P., Craven, H. Goldberg, B., et al. (1999). *The best test preparation for the Miller Analogies Test.* Piscataway, NJ: Research & Education Association.

Educational Testing Service. (1992). *Practicing to take the GRE General test – No. 9.* Princeton, NJ: Educational Testing Service.

Hassan, A. S. (1998). *Complete preparation for the Pharmacy College Admission Test.* Baltimore, Maryland: Williams & Wilkins.

Hassan, A. S. (1999). *Complete preparation for the Allied Health Professions Admission Test.* Baltimore, Maryland: Lippincott Williams & Wilkins.

Jaffe, E. D., & Hilbert, S. (1998). *How to prepare for the Graduate Management Admission Test.* Hauppage, NY: Barron's Education Series.

Michigan State Board of Education. (1998). *Michigan test for teacher certification* (Volume 1: Early, elementary, and middle level education). Amherst, MA: National Evaluation Systems.

Nguyen, H. -H. D., O'Neal, A., & Ryan, A. M. (in press). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance.*

Parham, S. E. (1996). *The relationships between test-taking strategies and cognitive ability test performance.* Unpublished doctoral dissertation, Bowling Green State University, Ohio.

Passbook (1998). *Rudman's questions and answers on the Dental Admission Test.* Syosset, NY: National Learning.

Powers, D. E., & Leung, S. W. (1995). Answering the new SAT reading comprehension questions without the passages. *Journal of Educational Measurement, 32*, 105-129.

Scruggs, T. E., & Marsing, L. (1988). Teaching test-taking skills to behaviorally disordered students. Behavioral Disorders, 13, 240-244.

Scruggs, T. E., & Mastropieri, M. A. (1992). *Teaching test-taking skills: Helping students show what they know*. Purdue University: Brookline Books.

Seibel, H. R., Guyer, K. E., Mangum, A. B., Conway, C. M., Conway, A. F., & Shanholtzer, W. L. (1997). *How to prepare for the MCAT*. Hauppauge, NY: Barron's Educational Series.

APPENDIX D

List of 85 Original Strategies

1. I know that all multiple-choice questions are really TRUE or FALSE decisions on a statement.
2. On the answer sheet (or in the test booklet if writing is permitted), I mark different signs for items that can use further consideration (e.g., a plus (+) sign) and for very unfamiliar or difficult questions (e.g., a minus (-) sign); then I skip these items, and go back to work on them (items with a + first) after I have answered all easier questions.
3. I use relevant content information in other test items and options to answer an item (e.g., one item "giving away" the answer to another question).
4. I go with my first instinct and don't change my answers.
5. I know that the test constructor may place the correct answer in a certain logical position among the ordered set of options (e.g., the middle of the sequence).
6. I should answer all test questions under any circumstances.
7. I know that questions for which I mark no answer or more than one answer are not counted in scoring.
8. I mark my answers on the test booklet (if allowed) as I work with the questions and periodically transfer a block of answers to the answer sheet to reduce clerical errors.
9. I seldom second-guess myself, except when I have good reasons to believe the second answer is definitely more correct than the first one.
10. I do not rush too much: there is no prize for finishing early.
11. I know if I immediately choose an answer (e.g., out of four options), I'll have 25% chance of getting it correct unless I am definitely certain about my choice.
12. I answer all the questions in order without skipping any.
13. If a question is taking too long, I guess and go on to the next question, then come back to it at the end if I have time.
14. If I draw a temporary blank, I recite the question to myself, or even write it out to help recall the material.
15. I become thoroughly familiar with test directions.
16. I vary my test-taking speed from section to section, depending on how much time I have for a particular section.
17. I scan the test for more time-consuming/difficult questions and leave them for last.
18. I know what to expect (e.g., being familiar with all types of test questions).
19. When responding to a multiple-choice test item, I should first eliminate apparent incorrect answer choices after careful scrutiny.
20. I restrict my choice to those options which encompass all of two or more given statements known to be correct.
21. I am careful not to make or leave any stray marks on my (machine-scored) answer sheet.

22. To save time, I memorize the (unchanged) directions for each type of questions in advance and only skimming through them when taking tests.
23. I pay attention to how much time is left so I can finish (e.g. a section) in the allotted time.
24. I use time remaining after completion of the test to reconsider answers.
25. I figure out how many minutes per question on average and spend the same amount of time on each question.
26. I am familiar with the columns and ovals on a multiple-choice answer sheet and how to completely fill in the ovals.
27. At the end of a test section, I go back to the test questions that gave me difficulty and verify my work on them.
28. I periodically check my answers to catch careless mistakes.
29. When I skip a question, I remember to skip the corresponding row of answer choices on my answer sheet.
30. I choose neither of two options which imply the correctness of each other.
31. I find out if writing in the test booklet is permitted.
32. If there are only a few minutes left for a test/test section and there is no penalty for wrong answers, I will fill in the remaining problems with guesses (randomly or picking a particular answer choice such as B or C) before the time is called.
33. I work as rapidly as possible with reasonable assurance of accuracy.
34. I note that the test constructor may make the correct answer an unusual option, different from those of the incorrect options (e.g., unduly longer or shorter, more complex or simpler than the incorrect ones).
35. I first scan the test for question types (e.g., certain types of questions require more thoughts and processing) and plan strategy (e.g., budgeting my time) accordingly.
36. I avoid internal distractions by directing attention away from self-evaluative thoughts.
37. I look for "short-cuts" to save time (e.g., doing a quick estimate to quickly eliminate some answer possibilities, or dividing an unfamiliar word into its prefix, suffix and root).
38. I check the general accuracy of my work at the end of a test section.
39. I know the test constructor may compose the correct answer of familiar or stereotyped phrases.
40. If there is a series of questions that all have the same letter choice (e.g., all Bs or Cs), there must be something wrong and I should recheck my answers.
41. I consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.
42. I tend to guess a particular choice (e.g., A or C) as soon as something looks unfamiliar or difficult.
43. I know the test constructor may qualify the correct answer more carefully (e.g., more precise and specific in meaning), or make it represent a higher degree of generalization.
44. I give thought to what the answer should include before reading the answer choices.
45. I pay particular attention to negatives.

46. In order to save time, I put slash marks through clearly wrong answers on the test booklet (if allowed) so that I would not waste time rereading them.
47. I eliminate answer choices which have some similarities.
48. I recognize and make use of resemblance between the options and an aspect of the stem.
49. I read the test items carefully, determining clearly the nature of the question.
50. I try to work backwards from the answers to the stem, especially with math questions.
51. Before answering test questions, I note in the margins of my test booklet (if allowed) certain short-term memory items.
52. I interpret questions at face value (e.g., don't read meanings into them) because I know test questions are designed to be straightforward, not tricky.
53. I choose one of two statements, which, if correct, would imply the incorrectness of the other.
54. If I am uncertain of what a word or phrase means, I try to resolve the vagueness or ambiguity by defining the word or phrase according to what I think it might be, then use the definition to solve the problem.
55. I eliminate the choice "all of the above" when there are opposite choices in the answers.
56. I guess as soon as something looks unfamiliar or difficult.
57. I do not hesitate to change my answers if I feel I should.
58. I read all instructions/directions carefully to make sure I understand them, determining clearly the nature of the task and the intended basis for response.
59. I rule out choices that contradict the question.
60. When I do not understand something about directions, I ask the examiner for clarification.
61. If I get annoyed for some reason (e.g., an "off-the-wall" question) and my concentration lapses, I will take a quick break, relaxing and regrouping for a minute.
62. I eliminate options which are known to be incorrect and chose from among the remaining options.
63. During breaks, I do not talk to other test takers: I keep my energies focused on the test.
64. I underline important key words or important qualifiers in the stem (e.g., almost, except, not) and double-checking my answer against these highlighted words.
65. I don't panic if I cannot answer a question: I keep calm and move on.
66. I look for the dimensions that underlie the options for multiple-choice questions and try to figure out how these dimensions relate to the dimensions in the stem: Only the correct answer converges all the dimensions in the stem.
67. I do not read into or think too much about a question I feel I have answered *correctly*.
68. I do not become impatient or discouraged if I cannot start at once effectively.
69. I know that time element is a factor in taking tests, but accuracy should not be sacrificed for speed.
70. If my annoyance persists, I try looking over the material I have successfully completed to recover and remotivate myself.

71. If I draw a temporary blank, I try to visualize the place in the book where this material appeared—this might lead me to relate facts and increase chances of recall.
72. I am aware that some exams may penalize you for wrong answers (e.g., deducting points from your total score).
73. I always ask myself what is the "hidden" meaning in a test item.
74. During the test, I divorce myself from everything (e.g. the proctor, my surrounding), and let no disturbance influence my concentration.
75. I sometimes make a quick drawing or a formula to clarify the questions and initiate and aid in recall.
76. I recognize and make use of specific determiners (e.g., always, never), disclaimers (e.g., best, all, or none), and/or "hedging" words (e.g., probably, most likely).
77. I erase the old answer *completely* if I change my mind about an answer.
78. I guess only after eliminating as many wrong answers as I can.
79. I take the same time I might spend on the single hardest question to answer three easier questions.
80. I don't waste time worrying during the test but work efficiently.
81. I read all information provided, even when I see an immediate answer.
82. If an item looks unfamiliar or difficulty, I brainstorm about what I know about the problem topic, even if it seems only tangentially related, to get enough information to answer the question correctly.
83. I know that I may work only on the section the test administrator designates and only for the time allowed (e.g., I may not go back to an earlier test section).
84. I know that the test constructor may make the correct answer grammatically consistent with the stem.
85. For difficult questions, I use my personal knowledge about the topic, or my common sense (e.g., choosing the answer that makes the best impression on me at that point) to make an educated guess.

APPENDIX E

Instructions for Expert Categorization

DIRECTIONS: When responding to a multiple-choice test, test takers may try to optimize their test performance over and above their knowledge of subject content tested by applying some tactics or strategies in taking the test. The list starting from Page 3 is my compilation of general test-taking strategies presented in previous studies and in standardized test guidebooks.

1. Please first **classify each strategy** on this list into 11 *a priori* categories by marking a category number on in Column A on your Expert Rating Sheet on Pages 7-9. The content and/or characteristics of these *a priori* categories are entailed on Page 2.

2. After you categorize a strategy, please indicate to what extent this strategy can be specifically applied to a multiple-choice test or test items by checking a rating number in Column B on your Expert Rating Sheet.

Please use the following scale to rate each test-taking strategy in terms of specificity:

1	2	3	4	5	6	7
Very general strategy, uniformly applicable to any multiple-choice test						Very specific strategy, only applicable to certain type of test items

(Note: An E-version of the test-taking strategies and Expert Rating Sheet has been sent to you via email.)

APPENDIX F

Informed Consent Form

Investigator's name & affiliation: Hannah-Hanh Nguyen, Graduate student, Industrial/Organizational Psychology Program, Michigan State University.

Summary: This is my Master's Thesis project. My goal is to develop a good scale that measures how much students are aware of strategies or tactics for taking multiple-choice tests. This study has three parts: You will (1) complete a set of pre-test surveys, (2) take a short test, and (3) complete another set of post-test surveys.

Estimated time required: 02 hours.

Risks: There are no risks associated with this study.

Compensation: (1) You are entitled to receive 2 hours of course credits for your participation. As an additional benefit, (2) you may be entered in a drawing for one of three cash prizes ($50.00 each) if you score in the top 20% on the short test.

Please note that your participation is *voluntary*. That means you may choose not to participate at all, may refuse to answer certain questions, or may discontinue the study at any time without penalty or loss of benefits to which you are otherwise entitled. That means you can still be awarded credits for participation; however, you *might not* be eligible for the additional prize.

Confidentiality: Your privacy will be protected to the maximum extent allowable by law. Although you will need to provide your name, email address, and student ID number, your name and email will be *only* used to identify prize winners and will be stripped from the data once the winners are determined. All your responses will be kept confidential. You will not be identified in any way by your responses in any report of research findings. Only my advisor and I will have access to the data collected in this study. Upon your request and within these restrictions, my study results may be made available to you.

Contact Information: If you have any questions about this study, please contact Hannah-Hanh Nguyen, 129 Psychology Research Building, East Lansing, MI 48824, E-mail: nguyen67@msu.edu. If you have questions or concerns regarding your rights as a study participant, or are dissatisfied at any time with any aspect of this study, you may contact – anonymously, if you wish – Ashir Kumar, M.D., Chair of the University Committee on Research Involving Human Subjects (UCRIHS) by phone: (517) 355-2180, fax: (517) 432-4503, e-mail: ucrihs@msu.edu, or regular mail: 202 Olds Hall, East Lansing, MI 48824. *A copy of this Informed Consent form will be provided to you upon your request.*

Your signature below indicates your voluntary agreement to participate in this study.

_____ _____

Participant's Signature Date

_____

Participant's Name (printed)

APPENDIX G

Debriefing Form

Thank you very much for participating in our study! Please DO NOT discuss the purpose and method of this study with others who have not participated in the study yet, but might participate some time later.

Michigan State University's college students are recruited for this study via the Department of Psychology' Subject Pool. They receive 2-hour credits for participation. To motivate participants to do their best on a cognitive ability test, the experimenter will enter the top 20% scorers on the test in a drawing for one of three cash prizes ($50 each). The winners will be determined and contacted via email.

Students as well as job applicants are often required to take a cognitive ability test for selection purposes in both academic and employment settings, because these tests measure general mental ability of applicants and provide a good predictor of success. Most cognitive ability tests are constructed objectively with a multiple-choice format. Past research has evidenced that, for multiple-choice tests, some test takers are able to apply several test-taking tactics or strategies in a certain fashion to optimize their test performance over and above their true knowledge of the subject matter tested, while some others are not—this ability is termed *testwiseness*. In order to measure the individual differences in testwiseness, researchers first need to know the extent to which test-taking strategies are acquired. Although there is a taxonomy of testwiseness available (Millman, Bishop, & Ebel, 1965) and several measures of testwiseness or test-taking strategies have been constructed (e.g., Gibb, 1964; Parham, 1996), a good instrument for assessing test-takers' knowledge of these strategies is still lacking.

For this study, the researcher aims to develop and validate a measure of test-taking strategies that could be used to assess test takers' self-report knowledge of test-taking strategies associated with standardized, paper-and-pencil multiple-choice tests. The researcher has reviewed the literature on this topic and come up with a list of 79 strategies that are administered to participants in this study, along with a short cognitive ability test and several test-taking attitudinal measures (e.g., motivation, self-efficacy, personality, metacognition). The relationships between test-taking strategy knowledge and test-takers' ability and attitudes will also be examined. (Those who are interested in learning more about this study and test-taking strategies can send the researcher a request for a report of the findings via email at nguyen67@msu.edu.)

References

Gibb, B. G. (1964). Test-wiseness as secondary cue response. Unpublished doctoral dissertation, Stanford University, CA.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. Educational and Psychological Measurement, 25, 707-726.

Parham. S. E. (1996). The relationships between test-taking strategies and cognitive ability test performance. Unpublished doctoral dissertation, Bowling Green State University, Ohio.

APPENDIX H

Pretest Surveys

TEST-TAKING MOTIVATION (Arvey et al., 1990)
I am extremely motivated to do well on this test.
I just don't care how I do on this test.
I will try my best on this test.
While taking this test, I will try to concentrate and do well.
Doing well on this test is important to me.
I want to be among the top scorers on this test.
I will not put much effort into this test.

TEST-TAKING SELF-EFFICACY (Pintrich & DeGroot, 1990)
I believe I will have no problems on this test.
I think I will do very well on this test.
Compared with other applicants taking this test, I expect to do well.
I am confident that I will receive a high score on this test.
I'm confident I can solve the problems presented in this test.

MULTIPLE-CHOICE TEST-TAKING SELF-EFFICACY (Miguel-Feruito, 1997)
I do NOT worry before taking multiple-choice tests.
Multiple-choice tests are usually tricky.
I would rather take an essay test than a multiple-choice test.
I usually perform well on multiple-choice tests.
I am NOT very comfortable taking multiple-choice tests.
I am good at taking multiple-choice tests.
Most people are probably better than I am at taking multiple-choice tests.
I feel at ease taking multiple-choice tests.
Multiple-choice tests are usually difficult.

CONSCIENTIOUSNESS (Goldberg, 1999)
I am always prepared.
I leave my belongings around.
I pay attention to details.
I make a mess of things.
I get chores done right away.
I often forget to put things back in their proper place.
I like order.
I shirk my duties.
I follow a schedule.

I am exacting in my work.

EMOTIONAL STABILITY (Goldberg, 1999)
I am relaxed most of the time.
I get stressed out easily.
I seldom feel blue.
I worry about things.
I am easily disturbed.
I get upset easily.
I change my mood a lot.
I have frequent mood swings.
I get irritated easily.
I often feel blue.

TEST-TAKING STRATEGY (Nguyen, 2003)	FAMILIAR		INCLINED TO USE	
	M	*SD*	*M*	*SD*
KNOWING HOW MULTIPLE-CHOICE TESTS WORK				
1. I know that all multiple-choice questions are really TRUE or FALSE decisions on a statement.	3.34	1.25	3.30	1.13
2. I try to familiarize myself with test directions prior to taking a test.	4.16	1.06	4.18	.95
3. I know what to expect (e.g., being familiar with all types of test questions).	3.74	.98	3.66	1.01
4. I find out if writing in the test booklet is permitted.	3.92	1.16	4.04	1.05
5. I interpret questions at face value (e.g., don't read meanings into them) because I know test questions are designed to be straightforward, not tricky.	2.73	1.19	2.61	1.16
6. I know that questions for which I mark no answer or more than one answer are not counted in scoring.	3.84	1.55	3.63	1.58
7. I am aware that some exams may penalize you for wrong answers (e.g., deducting points from your total score).	4.50	.94	3.66	1.42
8. I know that I may work only on the section the test administrator designates and only for the time allowed (e.g., I may not go back to an earlier test section).	4.44	.90	3.87	1.26
9. I am familiar with the columns and ovals on a multiple-choice answer sheet and how to completely fill in the ovals.	4.87	.46	4.78	.60
OPTIMIZING TIME EFFICIENCY AND EFFECTIVENESS				
10. If a question is taking too long, I guess and go on to the next question, then come back to it at the end if I have time.	4.35	1.00	4.01	1.18
11. I vary my test-taking speed from test section to	3.89	1.06	3.57	1.09

section, depending on how much time I have for a particular section.				
12. To save time, I scan the test for more time-consuming/difficult questions and leave them for last.	3.12	1.36	2.75	1.22
13. To get all of the easier questions answered before time runs out, I mark questions that can use further consideration and those that are very unfamiliar or difficult, skip them and go back to work on them later.	4.13	1.10	3.93	1.17
14. To save time, I memorize the (unchanged) directions for each type of questions in advance and only skimming through them when taking tests.	2.93	1.29	2.95	1.23
15. I pay attention to how much time is left so I can finish (e.g. a test section) in the allotted time.	4.41	.82	4.18	.97
16. I figure out how many minutes per question on average and spend the same amount of time on each question.	2.92	1.39	2.58	1.16
17. I work as rapidly as possible with reasonable assurance of accuracy.	3.75	1.08	3.42	1.17
18. I first scan the test for question types (e.g., certain types of questions require more thoughts and processing) and plan strategy accordingly (e.g., budgeting my time).	3.12	1.28	2.86	1.16
19. I look for "short-cuts" to save time (e.g., doing a quick estimate to quickly eliminate some answer possibilities, or dividing an unfamiliar word into its prefix, suffix and root).	3.54	1.26	3.46	1.22
20. In order to save time, I put slash marks through clearly wrong answers on the test booklet (if allowed) so that I would not waste time rereading them.	4.39	1.03	4.41	.97
21. I know that time element is a factor in taking tests, but accuracy should not be sacrificed for speed.	4.18	.92	4.01	.97
22. I take the same time I might spend on the single hardest question to answer three easier questions.	3.77	1.11	3.59	1.08
AVOIDING CLERICAL ERRORS				
23. I am careful not to make or leave any stray marks on my (machine-scored) answer sheet.	4.56	.76	4.46	.87
24. I periodically check my answers to catch careless or clerical mistakes.	4.31	.86	4.26	.91
25. When I skip a question, I remember to skip the corresponding row of answer choices on my answer sheet.	4.64	.74	4.50	.86
26. I check the general accuracy of my work at the end of a test section.	4.25	.90	4.07	.95
27. I erase the initial answer completely (when I change my mind about an answer).	4.71	.65	4.63	.70
28. To reduce clerical errors, I mark my answers on the test booklet (if allowed) as I work with the questions, and	3.44	1.41	3.22	1.43

periodically transfer a block of answers to the answer sheet.				
USING PHYSICAL CUES				
29. Among the options, I choose the answer that is longer than other options.	2.46	1.21	2.15	1.13
30. Among the options, I choose the answer that is shorter than the rest of the options.	2.17	1.07	1.95	1.04
31. I do not answer a series of questions with the same letter choice (e.g., all As).	3.01	1.29	2.65	1.22
32. I know that the test constructor may place the correct answer in a certain physical position among the options (e.g., in the middle).	3.08	1.23	2.64	1.22
USING GRAMMATICAL OR CONTEXTUAL CUES				
33. I use relevant content information in other test items and options to answer an item (e.g., one item "giving away" the answer to another question).	4.43	.79	4.40	.78
34. I consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.	3.49	1.11	3.42	1.06
35. I notice that the test constructor may qualify the correct answer more carefully (e.g., more precise and specific in meaning), or make it represent a higher degree of generalization.	3.74	1.04	3.55	1.00
36. I pay particular attention to negatives.	3.46	1.15	3.39	1.08
37. I recognize and make use of resemblance between the options and an aspect of the stem.	3.23	1.13	3.25	1.05
38. I recognize and make use of specific determiners (e.g., always, never), disclaimers (e.g., best, all, or none), and/or "hedging" words (e.g., probably, most likely).	4.13	1.00	3.98	1.01
39. I know that the test constructor may make the correct answer grammatically consistent with the stem.	3.50	1.16	3.43	1.08
DEDUCTIVE REASONING				
40. When responding to a multiple-choice test item, I should first eliminate apparent incorrect answer choices after careful scrutiny.	4.58	.71	4.58	.72
41. I restrict my choice to those options which encompass all of two or more given statements known to be correct.	3.87	1.03	3.80	1.00
42. I choose neither of two options which imply the correctness of each other.	3.08	1.05	3.06	1.00
43. I eliminate answer choices which have some similarities.	2.87	1.14	2.79	1.08
44. I try to work backwards from the answers to the stem, especially with math questions.	3.05	1.24	2.89	1.19

45. I choose one of two statements, which, if correct, would imply the incorrectness of the other.	3.32	1.06	3.19	1.04
46. I eliminate the choice "all of the above" when there are opposite choices in the answers.	4.00	1.21	3.95	1.24
47. I rule out choices that contradict the question.	4.05	1.00	4.07	.94
48. I eliminate options which are known to be incorrect and choose from among the remaining options.	4.67	.61	4.61	.69
49. Among the choices, I look for the answer that converges ALL the dimensions in the stem.	3.75	1.08	3.66	1.03
GUESSING				
50. I know if I randomly choose an answer out of four options, I'll have 25% chance of getting it correctly.	4.57	.79	3.77	1.39
51. If there are only a few minutes left for a test/test section and there is no penalty for wrong answers, I will fill in the remaining problems with guesses (randomly or picking a particular answer choice such as B or C) before time is called.	4.45	.85	4.02	1.22
52. I tend to guess a particular choice (e.g., A or C) as soon as something looks unfamiliar or difficult.	3.04	1.27	2.45	1.14
53. I guess as soon as something looks unfamiliar or difficult.	2.59	1.23	1.97	.96
54. I guess only after eliminating as many wrong answers as I can.	4.44	.76	4.26	.90
CHANGING ANSWERS				
55. I go with my first instinct and don't change my answers.	3.67	1.09	3.11	1.07
56. I do not hesitate to change my answers if I feel I should.	3.94	.95	2.89	1.07
57. I seldom second-guess myself, except when I have good reasons to believe the second answer is definitely more correct than my initial answer.	3.79	1.09	3.45	1.10
WORKING CAREFULLY AND THOROUGHLY				
58. I will reconsider my answers if I finish the test before time is called.	4.07	1.01	3.71	1.15
59. At the end of a test section, I go back to the test questions that gave me difficulty and verify my work on them.	4.45	.74	4.33	.86
60. I give thought to what the answer should include before reading the answer choices.	3.90	1.05	3.85	.98
61. I read the test items carefully, determining clearly the nature of the question.	4.19	.84	4.07	.85
62. I read all instructions/directions carefully to make sure I understand them, determining clearly the nature of the	4.18	.89	4.07	.89

task and the intended basis for response.				
63. When I do not understand something about directions, I ask the examiner for clarification.	4.11	1.00	3.80	1.15
64. I read all information provided, even when I see an immediate answer.	4.11	1.02	3.95	1.07
STAYING IN CONTROL				
65. I avoid internal distractions by directing attention away from negative self-evaluative thoughts.	3.79	1.00	3.60	.99
66. If I get annoyed for some reason (e.g., an "off-the-wall" question) and my concentration lapses, I will take a quick break, relaxing and regrouping for a minute.	3.36	1.21	3.16	1.16
67. During breaks (if any), I do not talk to other test takers: I keep my energies focused on the test.	3.37	1.30	3.08	1.29
68. I don't panic if I cannot answer a question: I keep calm and move on.	4.04	.91	3.74	1.11
69. I do not become impatient or discouraged if I cannot start at once effectively.	3.62	1.09	3.23	1.11
70. If I get annoyed for any reason during the test, I try looking over the material I have successfully completed to recover and remotivate myself.	3.07	1.20	3.00	1.11
71. During the test, I divorce myself from everything (e.g. the proctor, my surrounding), and let no disturbance influence my concentration.	3.56	1.11	3.27	1.14
72. I don't waste time worrying during the test but work efficiently.	3.81	1.04	3.61	1.12
TROUBLE-SHOOTING AND USING RECALL AIDS				
73. If I draw a temporary blank, I recite the question to myself, or even write it out to help recall the material.	3.73	1.04	3.66	1.04
74. Before answering test questions, I note in the margins of my test booklet (if allowed) certain short-term memory items.	3.83	1.16	3.88	1.10
75. If I am uncertain of what a word or phrase means, I try to resolve the vagueness or ambiguity by defining the word or phrase according to what I think it might be, then use the definition to solve the problem.	4.05	.90	3.90	.90
76. If I draw a temporary blank, I try to visualize the place in the book where this material appeared—this might lead me to relate facts and increase chances of recall.	4.19	.94	4.08	.96
77. I sometimes make a quick drawing or a formula to clarify the questions and initiate and aid in recall.	4.05	.98	4.00	.98
78. If an item looks unfamiliar, I brainstorm about what I know about the problem topic, even if it seems only tangentially related, to get enough information to answer the question correctly.	3.94	.95	3.80	.94

APPENDIX I

Cognitive Ability Test

PART 1

VERBAL REASONING

This part of the test begins with Question 1. You will have exactly 5 minutes to answer 12 questions. If you complete all the questions before the allotted time has elapsed, you may go back over this part of the test. However, YOU MAY NOT TURN TO THE NEXT PART OF THE TEST.

1. It was once believed that there was a controlling planet, the motion of which was so strong that all of the other planets were ruled by it. Disruption in the orbit of the central planet caused corresponding disruption in the orbits of the lesser planets. In like manner, the belief has existed that the uncertain, contradictory actions of a head of state result in

A changes in the orbits of the planets.
B conflict between its citizens.
C organized opposition to the leader.
D solidification of the government.
E ascendance of great people.

2. DEMENTED

A criminal
B insane
C adolescent
D melancholy
E measured

3. IMMINENT

A famous
B dangerous
C immense
D emitted
E impending

4. Sigmund Freud, the founder of psychoanalysis, repeatedly emphasized the pervasive importance of the acts and gestures that individuals unrealizingly perform. To understand the underlying truth of anyone's intentions and attitudes, place little credence in what the individual proclaims openly; seek, rather, real meanings.

A. in unconscious gestures and words which escape the individual.
B. by probing the person's earliest childhood memories.

C. from the person's psychoanalyst (if he has such).
D. by insisting that the person say what is really meant.
E. in the dreams which the person remembers.

5. FRUGALITY
A emptiness
B poverty
C thrift
D refuse
E cleanliness

6. ANIMOSITY
A hate
B liveliness
C zoo
D festival
E ambition

7. BUFFOON
A polish
B tool
C slipper
D ape
E clown

8. It is a universal desire to have poetry written in a straight-forward way so that its meaning is unambiguous. At least, we all like to think that we would want to have poets make themselves clear. The truth is that an utterly clear poet would be glaring, which would put the poet's work into opposition to the principles of art. For art cannot be obvious. Only when the imagination of the audience is provoked can aesthetic experience occur. So the poet can be just so clear and no more, and cannot afford to be diverted from his/her mysterious ways. By withholding a little meaning, the mystification is increased. Aesthetics demands that the poet be
A prophetic.
B original.
C lucid.
D incomprehensible.
E subtle.

9. VORACIOUS
A rebellious
B courageous
C ravenous
D beauteous
E courteous

10. RESILIENT

A relaxing
B resounding
C elastic
D quiet
E porous

11. ELUCIDATE

A confuse
B arrange
C explain
D avoid
E decorate

12. Of all the existing barriers to the transmission of ideas and knowledge, perhaps the most subtle is the individual's unwillingness to listen to what is being communicated and an unwitting distortion of what is heard. The ideological aspects of a culture give the individual a frame of reference into which ideas and knowledge fit. In general, those communications which do not fit neatly into this pre-existing frame of reference are

A immediately and violently rejected.
B regarded as intellectually incomprehensible.
C disregarded or unconsciously altered.
D accepted without equivocation.
E the impetus to alternation of this frame.

PART 2

MATHEMATICAL REASONING

This part of the test begins with Question 13. You will have exactly 10 minutes to answer 8 questions. If you complete all the questions before the allotted time has elapsed, you may go back over this part of the test or continue with the Post-test surveys. However, YOU MAY NOT GO BACK TO THE PREVIOUS PART OF THE TEST.

13. One field, 70 yards by 50 yards, is three times as large in area as another field. The second field is 40 yards long. How wide is the second field?

A 19 yards
B 24 yards
C 29 yards
D 34 yards
E 39 yards

14. If $x + y = 3$ and $3x - y = 3$, then

A $x = 1.5, y = 1.5$

B $x = 2, y = 1$
C $x = .5, y = 2.5$
D $x = 4.5, y = -1.5$
E $x = -2.5, y = 5.5$

Questions 15 through 17 are based on the following table:

Time of Daylight, Sunrise, Sunset, and Darkness on January 1 and May 1

North Latitude	Daylight A.M.	Sunrise A.M.	Sunset P.M.	Darkness P.M.
		January 1		
0°	5:09	6:00	6:08	6:58
10°	5:25	6:17	5:50	6:41
20°	5:40	6:35	5:32	6:26
		May 1		
0°	5:07	5:54	6:00	6:48
10°	4:54	5:43	6:11	7:01
20°	4:38	5:31	6:23	7:17

15. As the latitude increases, the sun
A rises later on Jan 1.
B rises later on May 1.
C rises earlier on Jan 1.
D sets earlier on May 1.
E sets later on Jan 1.

16. How many hours of daylight are there at a point at 0 degrees north latitude on May 1?
A 10 hours, 10 minutes
B 12 hours, 6 minutes
C 12 hours, 7 minutes
D 13 hours, 41 minutes
E 13 hours, 49 minutes

17. According to this table, the period between sunset and darkness always
A increases as the length of the day increases.
B increases as the length of the day decreases.
C increases as the latitude increases.
D increases as the latitude decreases.
E remains the same throughout the year at any given latitude.

18. Given triangle ABC, Angle bisector AD of angle A is perpendicular to BC. Which of these is true?
A AB = AC
B Angle BAD = angle ACD
C Angle CAD = angle ADC
D AD = CD
E AD = BC

19. A local movie theater counts the number of passes issued to adults, students, and children. Thirteen percent of the passes were issued to adults and 27% were issued to students. If the remaining 480 tickets were issued to children, how many total tickets were there?

A 720
B 756
C 800
D 900
E 960

20. If 15.382 were multiplied by .0106, the product would be between

A .001 and .002
B .015 and .02
C .01 and .05
D .15 and .20
E 1.0 and 2.0

APPENDIX J

Posttest Surveys

TEST-TAKING STRATEGY (UTILIZATION)	*M*	*SD*
KNOWING HOW MULTIPLE-CHOICE TESTS WORK		
1. I know that all multiple-choice questions are really TRUE or FALSE decisions on a statement.	3.19	1.18
2. I tried to familiarize myself with test directions prior to taking a test.	3.97	.83
3. I knew what to expect (e.g., being familiar with all types of test questions).	2.89	1.15
4. I found out if writing in the test booklet is permitted.	3.61	1.26
5. I interpreted questions at face value (e.g., don't read meanings into them) because I knew test questions are designed to be straightforward, not tricky.	3.15	1.16
6. I know that questions for which I mark no answer or more than one answer are not counted in scoring.	3.57	1.32
7. I was aware that this exam did not penalize you for wrong answers (e.g., deducting points from your total score).	3.00	1.36
8. I worked only on the section the test administrator designated and only for the time allowed (e.g., I may not go back to an earlier test section).	4.61	.63
9. I was familiar with the columns and ovals on a multiple-choice answer sheet and how to completely fill in the ovals.	4.69	.57
OPTIMIZING TIME EFFICIENCY AND EFFECTIVENESS		
10. If a question was taking too long, I guessed and go on to the next question, then came back to it at the end.	3.79	1.17
11. I varied my test-taking speed from test section to section, depending on how much time I had for a particular section.	3.84	1.00
12. To save time, I scanned the test for more time-consuming/difficult questions and left them for last.	2.77	1.28
13. To get all of the easier questions answered before time runs out, I marked questions that could use further consideration and those that were very unfamiliar or difficult, skipped them and went back to work on them later.	3.25	1.30
14. To save time, I memorized the (unchanged) directions for each type of questions in advance and only skimming through them when taking the test.	3.11	1.20
15. I paid attention to how much time was left so I could finish (e.g. a test section) in the allotted time.	3.59	1.14
16. I figured out how many minutes per question on average and spent the same amount of time on each question.	2.15	1.06
17. I worked as rapidly as possible with reasonable assurance of accuracy.	3.93	.90
18. I first scanned the test for question types (e.g., certain types of questions require more thoughts and processing) and planned strategy accordingly (e.g., budgeting my time).	2.65	1.17

19. I looked for "short-cuts" to save time (e.g., doing a quick estimate to quickly eliminate some answer possibilities, or dividing an unfamiliar word into its prefix, suffix and root).	3.43	1.14
20. In order to save time, I put slash marks through clearly wrong answers on the test booklet (if allowed) so that I would not waste time rereading them.	3.47	1.33
21. I know that time element is a factor in taking tests, but accuracy should not be sacrificed for speed.	3.99	.84
22. I took the same time I might spend on the single hardest question to answer three easier questions.	3.43	1.12
AVOIDING CLERICAL ERRORS		
23. I was careful not to make or leave any stray marks on my (machine-scored) answer sheet.	3.41	1.16
24. I periodically checked my answers to catch careless or clerical mistakes.	4.14	.87
25. When I skipped a question, I remembered to skip the corresponding row of answer choices on my answer sheet.	2.48	1.33
26. I checked the general accuracy of my work at the end of a test section.	1.92	.85
27. I erased the initial answer completely (when I changed my mind about an answer).	1.87	.79
28. To reduce clerical errors, I marked my answers on the test booklet as I worked with the questions, and periodically transferred a block of answers to the answer sheet.	2.44	1.13
	2.15	.97
USING PHYSICAL CUES		
29. Among the options, I chose the answer that was longer than other options.	3.16	1.05
30. Among the options, I chose the answer that was shorter than the rest of the options.	3.09	1.03
31. I did not answer a series of questions with the same letter choice (e.g., all As).	2.72	.94
32. I knew that the test constructor placed the correct answer in a certain physical position among the options (e.g., in the middle).	2.95	1.05
	3.24	1.01
USING GRAMMATICAL OR CONTEXTUAL CUES		
33. I used relevant content information in other test items and options to answer an item (e.g., one item "giving away" the answer to another question).	3.37	1.02
34. I considered the subject matter and difficulty of neighboring items when interpreting and answering a given item.	3.20	.84
35. I noticed that the test constructor qualified the correct answer more carefully (e.g., more precise and specific in meaning), or made it represent a higher degree of generalization.	4.14	.78
36. I paid particular attention to negatives.	3.63	.90
37. I recognized and made use of resemblance between the options and an aspect of the stem.	3.18	.93

38. I recognized and made use of specific determiners (e.g., always, never), disclaimers (e.g., best, all, or none), and/or "hedging" words (e.g., probably, most likely).	3.02	1.00
39. The test constructor made the correct answer grammatically consistent with the stem.	3.10	1.18
	3.21	.95
DEDUCTIVE REASONING		
40. When responding to a multiple-choice test item, I first eliminated apparent incorrect answer choices after careful scrutiny.	3.06	1.12
41. I restricted my choice to those options which encompassed all of two or more given statements known to be correct.	3.84	.90
42. I chose neither of two options which implied the correctness of each other.	4.15	.78
43. I eliminated answer choices which had some similarities.	3.60	.97
44. I tried to work backwards from the answers to the stem, especially with math questions.	4.33	.80
45. I chose one of two statements, which, if correct, would imply the incorrectness of the other.	3.71	1.18
46. I eliminated the choice "all of the above" when there were opposite choices in the answers.	2.70	1.18
47. I ruled out choices that contradicted the question.	2.39	1.06
48. I eliminated options which were known to be incorrect and chose from among the remaining options.	3.95	.95
49. Among the choices, I looked for the answer that converged ALL the dimensions in the stem.	3.43	1.09
	3.57	.93
GUESSING		
50. I know if I randomly choose an answer out of four options, I'll have 25% chance of getting it correctly.	3.56	.96
51. When there were only a few minutes left for a test/test section, I filled in the remaining problems with guesses before time was called.	3.45	1.08
52. I tended to guess a particular choice (e.g., A or C) as soon as something looked unfamiliar or difficult.	3.46	1.15
53. I guessed as soon as something looked unfamiliar or difficult.	3.39	1.03
54. I guessed only after eliminating as many wrong answers as I could.	3.77	.87
	3.76	.90
CHANGING ANSWERS		
55. I went with my first instinct and didn't change my answers.	2.80	1.15
56. I did not hesitate to change my answers when I felt I should.	3.58	1.08
57. I seldom second-guessed myself, except when I had good reasons to believe the second answer was definitely more correct than my initial answer.	3.61	1.00
	2.84	1.08
WORKING CAREFULLY AND THOROUGHLY		
58. I reconsidered my answers if I finish the test before time is called.	3.17	1.23

59. At the end of a test section, I went back to the test questions that had given me difficulty and verified my work on them.	3.90	.80
60. I gave thought to what the answer should include before reading the answer choices.	3.63	.92
61. I read the test items carefully, determining clearly the nature of the question.	2.80	1.08
62. I read all instructions/directions carefully to make sure I understood them, determining clearly the nature of the task and the intended basis for response.	3.42	1.06
63. When I did not understand something about directions, I asked the examiner for clarification.	3.84	.83
64. I read all information provided, even when I saw an immediate answer.	3.36	1.05
	2.88	1.13
STAYING IN CONTROL		
65. I avoided internal distractions by directing attention away from negative self-evaluative thoughts.	3.86	.84
66. If I got annoyed for some reason (e.g., an "off-the-wall" question) and my concentration lapses, I took a quick break, relaxing and regrouping for a minute.	2.94	1.10
67. If I had a break, I would not talk to other test takers.	3.81	.97
68. I didn't panic if I couldn't answer a question: I kept calm and moved on.	3.69	.93
69. I did not become impatient or discouraged if I could not start at once effectively.	3.19	1.18
70. If I got annoyed for any reason during the test, I tried looking over the material I had successfully completed to recover and remotivate myself.	3.97	.83
71. During the test, I divorced myself from everything (e.g. the proctor, my surrounding), and let no disturbance influence my concentration.	2.89	1.15
72. I didn't waste time worrying during the test but worked efficiently.	3.61	1.26
	3.15	1.16
TROUBLE-SHOOTING AND USING RECALL AIDS		
73. If I drew a temporary blank, I recited the question to myself, or even wrote it out to help recall the material.	3.57	1.32
74. Before answering test questions, I noted in the margins of my test booklet (if allowed) certain short-term memory items.	3.00	1.36
75. If I was uncertain of what a word or phrase means, I tried to resolve the vagueness or ambiguity by defining the word or phrase according to what I thought it might be, then used the definition to solve the problem.	4.61	.63
76. If I drew a temporary blank, I tried to visualize the place in a book where this material appeared.	4.69	.57
77. I sometimes made a quick drawing or a formula to clarify the questions and aid in recall.	3.79	1.17
78. If an item looked unfamiliar, I brainstormed about what I knew about the problem topic, even if it seemed only tangentially related, to get enough information to answer the question correctly.	3.84	1.00

TEST-TAKING METACOGNITIVE STRATEGIES (Schraw, 1997)

1. I asked myself periodically if I was doing well.
2. I consciously focused my attention on important parts of the problem.
3. I had a specific purpose for each test-taking strategy I used.
4. I am a good judge of how well I understand something.
5. I found myself pausing regularly to check my comprehension.
6. I knew when each strategy I used was most effective.
7. I stopped and went back over answers that were not clear.
8. I was aware of what strategies I used when I solved problems.
9. I changed strategies when I failed to understand a problem.
10. I stopped and reread when I got confused.

Demographic Questionnaire

1. Age _____

2. Gender (please circle ONE) Male Female

3. Ethnicity (please circle ONE)

a. African American	c. Caucasian	e. Native American
b. Asian	d. Latino/Chicano	f. Other

4. Level of academic standing (please circle ONE)

a. Freshman	c. Junior	e. Senior (5 years+)
b. Sophomore	d. Senior	

5. Current grade point average _____ (Max GPA = 4)
[*For Freshmen*: If you do not have a college GPA yet, please provide your high school cumulative GPA _____ (Max GPA = ____)]

6. Major _____
[If you have not declared a major, please put "Undeclared".]

7. Standardized test score: ACT _____ *OR* SAT_____

8. How many formal courses did you have which taught test-taking strategies? (please circle ONE)

a. 4 or more	c. 2	e. 0 (none)
b. 3	d. 1	

9. If you answered *1 course or more* for Question 9, please list up to 3 such courses and how long the course(s) took you to finish (i.e., 8 hours):

_____

_____

_____

10. If you answered *1 or more* for Question 9, how would you rate the course(s)? (please circle ONE)

Course 1

a. Much worse than average
b. Worse than average
c. Somewhat worse than average
d. About average
e. Somewhat better than average
f. Better than average
g. Much better than average

Course 2

a. Much worse than average
b. Worse than average
c. Somewhat worse than average
d. About average
e. Somewhat better than average
f. Better than average
g. Much better than average

Course 3

a. Much worse than average
b. Worse than average
c. Somewhat worse than average
d. About average
e. Somewhat better than average
f. Better than average
g. Much better than average

11. How would you agree with this statement, "I have studied self-taught materials involving test-taking strategies"? (please circle ONE)

a. Strongly disagree
b. Somewhat disagree
c. Neither disagree nor agree
d. Somewhat agree
e. Strongly agree

APPENDIX K

Modification Indices of λ_X and Completely Standardized Solution Matrices for the Correlated 11-Factor Model

Modification Indices for LAMBDA-X

	mctwork	time	clerror	physical	grammar	dereason	guess	cguess	careful	control	recaid
d1_1	--	0.41	0.02	0.24	0.10	0.00	0.97	0.07	0.06	0.14	0.10
d1_2	--	0.41	0.02	0.24	0.10	0.00	0.97	0.07	0.06	0.14	0.10
d2_1	3.73	--	3.98	2.97	12.37	0.98	4.67	4.63	3.76	4.28	7.31
d2_2	0.65	--	0.76	0.11	3.30	3.52	0.16	0.41	2.42	0.08	0.68
d2_3	0.10	--	0.30	1.77	0.88	2.92	2.21	0.07	3.54	1.56	0.08
d2_4	4.06	--	6.15	3.43	0.16	0.27	13.98	4.19	1.64	4.49	3.41
d2_5	0.76	--	1.13	0.25	5.28	10.89	3.13	0.01	0.44	0.90	2.23
d2_6	2.34	--	7.02	7.42	9.05	0.08	7.26	6.25	0.27	0.00	5.00
d3_1	0.01	0.29	--	1.40	0.85	1.32	1.20	0.01	0.00	0.08	1.33
d3_2	0.38	1.94	--	2.30	0.50	0.30	0.00	2.17	0.81	3.83	0.01
d3_3	1.07	0.00	--	3.81	0.01	0.02	0.58	0.29	0.29	0.08	0.01
d3_4	0.36	3.20	--	0.38	2.31	0.78	3.61	0.41	0.06	3.01	1.47
d4_1	0.94	1.14	1.38	--	0.05	1.16	0.01	2.61	1.13	0.01	0.34
d4_2	0.25	0.37	4.97	--	7.68	5.58	6.03	0.00	3.04	0.30	3.16
d4_3	1.77	0.30	0.21	--	3.66	1.30	5.05	4.70	0.01	0.84	0.02
d4_4	5.06	8.60	3.65	--	16.07	3.22	12.76	3.44	2.18	2.42	5.90
d5_1	0.31	8.05	3.26	12.22	--	4.87	0.58	0.66	4.39	0.81	0.62
d5_2	0.68	1.76	0.00	2.38	--	0.10	2.94	1.53	1.03	0.00	0.66
d5_3	2.30	0.71	0.57	0.42	--	0.07	4.64	0.11	0.07	0.00	0.32
d5_4	1.33	1.67	6.45	0.05	--	0.57	0.91	1.11	2.46	0.06	8.99
d5_5	4.49	2.85	32.51	12.63	--	1.91	6.54	1.10	16.68	2.70	20.87
d5_6	2.90	1.51	1.63	0.03	--	1.74	0.24	0.00	2.59	1.59	2.97
d5_7	3.33	1.97	0.03	0.10	--	3.52	0.20	0.01	2.36	0.29	0.02
d6_1	0.07	0.56	4.06	1.10	4.62	--	2.80	4.26	3.81	0.68	2.93
d6_2	19.20	4.85	22.36	9.31	11.92	--	8.78	6.37	11.63	9.09	15.75
d6_3	0.45	0.13	0.74	0.05	0.02	--	0.78	2.25	1.65	1.23	2.49
d6_4	5.88	1.67	22.85	6.80	6.97	--	16.72	0.79	19.14	9.55	18.36
d6_5	9.39	1.53	35.04	11.67	23.42	--	13.36	10.92	24.66	15.14	35.40
d7_1	4.91	11.23	1.17	12.42	3.59	5.17	--	1.84	0.10	0.02	3.05
d7_2	4.91	11.23	1.17	12.42	3.59	5.17	--	1.84	0.10	0.02	3.05
d8_1	3.04	0.03	6.66	5.71	1.64	0.57	15.27	--	4.10	0.66	1.94
d8_2	3.04	0.03	6.66	5.71	1.64	0.57	15.27	--	4.10	0.66	1.94

d9_1	0.16	5.21	1.76	3.43	17.78	14.03	1.03	1.23	--	1.23	0.78
d9_2	0.08	1.65	0.02	0.13	4.50	2.90	3.72	3.25	--	0.51	6.79
d9_3	0.88	0.04	0.04	0.40	2.38	1.14	6.01	5.65	--	6.18	0.59
d9_4	2.70	0.79	3.17	1.00	0.59	1.39	1.93	2.39	--	0.21	1.23
d10_1	0.77	0.63	1.21	1.57	4.84	2.93	0.37	0.30	10.11	--	7.79
d10_2	3.41	0.32	5.23	1.65	0.21	0.25	11.37	0.08	0.18	--	1.02
d10_3	5.06	1.52	4.05	4.01	0.10	0.00	5.10	6.79	0.10	--	0.00
d10_4	3.95	0.00	4.45	0.16	0.26	0.18	7.77	0.15	0.00	--	0.07
d10_5	5.66	1.37	7.02	0.02	4.27	1.68	7.54	2.73	10.58	--	8.33
d11_1	0.16	5.28	0.07	0.01	0.00	0.33	0.05	0.00	0.64	0.00	--
d11_2	0.20	5.46	1.19	0.65	0.31	3.52	1.58	0.78	0.98	0.47	--
d11_3	0.58	0.00	0.54	0.39	0.29	1.24	1.72	0.52	0.02	0.34	--

Completely Standardized Solution

LAMBDA-X

	mctwork	time	clerror	physical	grammar	dereason	guess	cguess	careful	control	recaid
d1_1	0.50	--	--	--	--	--	--	--	--	--	--
d1_2	0.84	--	--	--	--	--	--	--	--	--	--
d2_1	--	0.44	--	--	--	--	--	--	--	--	--
d2_2	--	0.49	--	--	--	--	--	--	--	--	--
d2_3	--	0.66	--	--	--	--	--	--	--	--	--
d2_4	--	0.38	--	--	--	--	--	--	--	--	--
d2_5	--	0.59	--	--	--	--	--	--	--	--	--
d2_6	--	0.51	--	--	--	--	--	--	--	--	--
d3_1	--	--	0.71	--	--	--	--	--	--	--	--
d3_2	--	--	0.60	--	--	--	--	--	--	--	--
d3_3	--	--	0.67	--	--	--	--	--	--	--	--
d3_4	--	--	0.67	--	--	--	--	--	--	--	--
d4_1	--	--	--	0.82	--	--	--	--	--	--	--
d4_2	--	--	--	0.80	--	--	--	--	--	--	--
d4_3	--	--	--	0.30	--	--	--	--	--	--	--
d4_4	--	--	--	0.44	--	--	--	--	--	--	--
d5_1	--	--	--	--	0.52	--	--	--	--	--	--
d5_2	--	--	--	--	0.58	--	--	--	--	--	--
d5_3	--	--	--	--	0.55	--	--	--	--	--	--
d5_4	--	--	--	--	0.58	--	--	--	--	--	--
d5_5	--	--	--	--	0.62	--	--	--	--	--	--
d5_6	--	--	--	--	0.55	--	--	--	--	--	--
d5_7	--	--	--	--	0.50	--	--	--	--	--	--
d6_1	--	--	--	--	--	0.60	--	--	--	--	--
d6_2	--	--	--	--	--	0.61	--	--	--	--	--
d6_3	--	--	--	--	--	0.64	--	--	--	--	--

d6_4	--	--	--	--	--	0.46	--	--	--	--	--	
d6_5	--	--	--	--	--	0.43	--	--	--	--	--	
d7_1	--	--	--	--	--	--	0.58	--	--	--	--	
d7_2	--	--	--	--	--	--	0.70	--	--	--	--	
d8_1	--	--	--	--	--	--	--	0.69	--	--	--	
d8_2	--	--	--	--	--	--	--	0.76	--	--	--	
d9_1	--	--	--	--	--	--	--	--	0.66	--	--	
d9_2	--	--	--	--	--	--	--	--	0.66	--	--	
d9_3	--	--	--	--	--	--	--	--	0.60	--	--	
d9_4	--	--	--	--	--	--	--	--	0.61	--	--	
d10_1	--	--	--	--	--	--	--	--	--	0.63	--	
d10_2	--	--	--	--	--	--	--	--	--	0.61	--	
d10_3	--	--	--	--	--	--	--	--	--	0.61	--	
d10_4	--	--	--	--	--	--	--	--	--	0.61	--	
d10_5	--	--	--	--	--	--	--	--	--	0.75	--	
d11_1	--	--	--	--	--	--	--	--	--	--	0.60	
d11_2	--	--	--	--	--	--	--	--	--	--	0.57	
d11_3	--	--	--	--	--	--	--	--	--	--	0.69	

PHI

	mctwork	time	clerror	physical	grammar	dereason	guess	cguess	careful	control	recaid
mctwork	1.00										
time	0.40	1.00									
clerror	0.44	0.31	1.00								
physical	0.14	0.40	-0.12	1.00							
grammar	0.37	0.55	0.37	0.26	1.00						
dereason	0.21	0.47	0.07	0.37	0.51	1.00					
guess	0.43	0.25	0.42	-0.01	0.27	0.19	1.00				
cguess	0.09	0.32	-0.10	0.35	0.15	0.21	0.23	1.00			
careful	0.36	0.30	0.59	-0.12	0.46	0.30	0.23	0.08	1.00		
control	0.29	0.31	0.46	0.04	0.42	0.33	0.28	0.09	0.62	1.00	
recaid	0.25	0.41	0.51	0.04	0.50	0.30	0.30	0.04	0.49	0.37	1.00

THETA-DELTA

d1_1	d1_2	d2_1	d2_2	d2_3	d2_4	d2_5	d2_6	d3_1	d3_2	d3_3	d3_4
0.75	0.30	0.81	0.76	0.57	0.85	0.66	0.74	0.50	0.65	0.55	0.55
d4_1	d4_2	d4_3	d4_4	d5_1	d5_2	d5_3	d5_4	d5_5	d5_6	d5_7	d6_1
0.33	0.36	0.91	0.81	0.73	0.66	0.70	0.67	0.62	0.70	0.75	0.65
d6_2	d6_3	d6_4	d6_5	d7_1	d7_2	d8_1	d8_2	d9_1	d9_2	d9_3	d9_4
0.63	0.58	0.79	0.81	0.66	0.51	0.52	0.42	0.56	0.56	0.64	0.62
d10_1	d10_2	d10_3	d10_4	d10_5	d11_1	d11_2	d11_3				
0.61	0.62	0.63	0.63	0.44	0.64	0.68	0.53				

APPENDIX L

Completely Standardized Solution Matrix for the Zero-correlation 11 Factor Model

LAMBDA-X

	mctwork	time	clerror	physical	grammar	dereason	guess	cguess	careful	control	recaid
d1_1	0.65	--	--	--	--	--	--	--	--	--	--
d1_2	0.65	--	--	--	--	--	--	--	--	--	--
d2_1	--	0.49	--	--	--	--	--	--	--	--	--
d2_2	--	0.46	--	--	--	--	--	--	--	--	--
d2_3	--	0.62	--	--	--	--	--	--	--	--	--
d2_4	--	0.34	--	--	--	--	--	--	--	--	--
d2_5	--	0.66	--	--	--	--	--	--	--	--	--
d2_6	--	0.48	--	--	--	--	--	--	--	--	--
d3_1	--	--	0.71	--	--	--	--	--	--	--	--
d3_2	--	--	0.58	--	--	--	--	--	--	--	--
d3_3	--	--	0.68	--	--	--	--	--	--	--	--
d3_4	--	--	0.67	--	--	--	--	--	--	--	--
d4_1	--	--	--	0.87	--	--	--	--	--	--	--
d4_2	--	--	--	0.77	--	--	--	--	--	--	--
d4_3	--	--	--	0.28	--	--	--	--	--	--	--
d4_4	--	--	--	0.42	--	--	--	--	--	--	--
d5_1	--	--	--	--	0.51	--	--	--	--	--	--
d5_2	--	--	--	--	0.56	--	--	--	--	--	--
d5_3	--	--	--	--	0.55	--	--	--	--	--	--
d5_4	--	--	--	--	0.62	--	--	--	--	--	--
d5_5	--	--	--	--	0.59	--	--	--	--	--	--
d5_6	--	--	--	--	0.59	--	--	--	--	--	--
d5_7	--	--	--	--	0.49	--	--	--	--	--	--
d6_1	--	--	--	--	--	0.64	--	--	--	--	--
d6_2	--	--	--	--	--	0.65	--	--	--	--	--
d6_3	--	--	--	--	--	0.64	--	--	--	--	--
d6_4	--	--	--	--	--	0.40	--	--	--	--	--
d6_5	--	--	--	--	--	0.37	--	--	--	--	--
d7_1	--	--	--	--	--	--	0.64	--	--	--	--
d7_2	--	--	--	--	--	--	0.64	--	--	--	--
d8_1	--	--	--	--	--	--	--	0.73	--	--	--
d8_2	--	--	--	--	--	--	--	0.73	--	--	--
d9_1	--	--	--	--	--	--	--	--	0.68	--	--
d9_2	--	--	--	--	--	--	--	--	0.72	--	--
d9_3	--	--	--	--	--	--	--	--	0.54	--	--
d9_4	--	--	--	--	--	--	--	--	0.58	--	--
d10_1	--	--	--	--	--	--	--	--	--	0.59	--
d10_2	--	--	--	--	--	--	--	--	--	0.60	--
d10_3	--	--	--	--	--	--	--	--	--	0.58	--
d10_4	--	--	--	--	--	--	--	--	--	0.61	--
d10_5	--	--	--	--	--	--	--	--	--	0.80	--
d11_1	--	--	--	--	--	--	--	--	--	--	0.59
d11_2	--	--	--	--	--	--	--	--	--	--	0.58
d11_3	--	--	--	--	--	--	--	--	--	--	0.69

PHI
Note: This matrix is diagonal.

mctwork	time	clerror	physical	grammar	dereason	guess	cguess	careful	control	recaid
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

THETA-DELTA

d1_1	d1_2	d2_1	d2_2	d2_3	d2_4	d2_5	d2_6	d3_1	d3_2	d3_3	d3_4
0.58	0.58	0.76	0.79	0.61	0.88	0.56	0.77	0.49	0.67	0.53	0.55

THETA-DELTA

d4_1	d4_2	d4_3	d4_4	d5_1	d5_2	d5_3	d5_4	d5_5	d5_6	d5_7	d6_1
0.25	0.41	0.92	0.83	0.74	0.69	0.70	0.62	0.66	0.65	0.76	0.59

THETA-DELTA

d6_2	d6_3	d6_4	d6_5	d7_1	d7_2	d8_1	d8_2	d9_1	d9_2	d9_3	d9_4
0.57	0.59	0.84	0.87	0.59	0.59	0.47	0.47	0.53	0.49	0.70	0.67

THETA-DELTA

d10_1	d10_2	d10_3	d10_4	d10_5	d11_1	d11_2	d11_3
0.65	0.64	0.66	0.62	0.36	0.65	0.67	0.52

APPENDIX M

Completely Standardized Solution Matrix for the Correlated 2-Factor Model

LAMBDA-X

	gen1	gen2
d1_1	0.28	--
d1_2	0.43	--
d2_1	0.15	--
d2_2	0.26	--
d2_3	0.39	--
d2_4	0.32	--
d2_5	0.29	--
d2_6	0.35	--
d3_1	0.56	--
d3_2	0.51	--
d3_3	0.50	--
d3_4	0.49	--
d4_1	--	0.33
d4_2	--	0.26
d4_3	--	0.25
d4_4	--	0.37
d5_1	--	0.52
d5_2	--	0.55
d5_3	--	0.52
d5_4	--	0.52
d5_5	--	0.57
d5_6	--	0.50
d5_7	--	0.48
d6_1	--	0.36
d6_2	--	0.36
d6_3	--	0.46
d6_4	--	0.42
d6_5	--	0.46
d7_1	0.30	--
d7_2	0.31	--
d8_1	--	0.20
d8_2	--	0.19
d9_1	0.53	--
d9_2	0.50	--
d9_3	0.52	--
d9_4	0.50	--
d10_1	0.54	--
d10_2	0.52	--
d10_3	0.52	--
d10_4	0.43	--
d10_5	0.51	--
d11_1	0.41	--
d11_2	0.38	--
d11_3	0.45	--

PHI

	gen1	gen2
gen1	1.00	
gen2	0.58	1.00

THETA-DELTA

d1_1	d1_2	d2_1	d2_2	d2_3	d2_4
0.92	0.81	0.98	0.93	0.84	0.90

THETA-DELTA

d2_5	d2_6	d3_1	d3_2	d3_3	d3_4
0.92	0.88	0.69	0.74	0.75	0.76

THETA-DELTA

d4_1	d4_2	d4_3	d4_4	d5_1	d5_2
0.89	0.93	0.94	0.87	0.73	0.70

THETA-DELTA

d5_3	d5_4	d5_5	d5_6	d5_7	d6_1
0.73	0.73	0.68	0.75	0.77	0.87

THETA-DELTA

d6_2	d6_3	d6_4	d6_5	d7_1	d7_2
0.87	0.78	0.82	0.79	0.91	0.91

THETA-DELTA

d8_1	d8_2	d9_1	d9_2	d9_3	d9_4
0.96	0.97	0.71	0.75	0.73	0.75

THETA-DELTA

d10_1	d10_2	d10_3	d10_4	d10_5	d11_1
0.71	0.73	0.73	0.81	0.74	0.83

THETA-DELTA

d11_2	d11_3
0.86	0.80