INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

University of Alberta

COMPARISON OF ABILITY ESTIMATES FROM DICHOTOMOUSLY AND NOMINALLY SCORED TESTWISE SUSCEPTIBLE AND NON-SUSCEPTIBLE ITEMS

BY

JOANNA TERESA TOMKOWICZ 

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Department of Educational Psychology

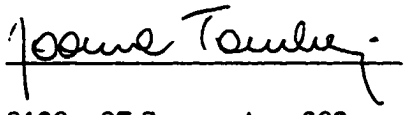
Edmonton, Alberta

Fall 2000

University of Alberta

Library Release Form

Name of Author:	JOANNA TERESA TOMKOWICZ
Title of Thesis:	COMPARISON OF ABILITY ESTIMATS FROM DICHOTOMOUSLY AND NOMINALLY SCORED TESTWISE SUSCEPTIBLE AND NON-SUSCEPTIBLE ITEMS
Degree:	DOCTOR OF PHILOSOPHY
Year this Degree Granted:	2000



8120 – 97 Street, Apt. 302
Edmonton, Alberta T6E 3K5

August 15, 2000

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommended to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled COMPARISON OF ABILITY ESTIMATES FROM DICHOTOMOUSLY AND NOMINALLY SCORED TESTWISE SUSCEPTIBLE AND NON-SUSCEPTIBLE ITEMS submitted by JOANNA TERESA TOMKOWICZ in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

W. T. Rogers

M. J. Gierl

H. L. Janzen

E. W. Romaniuk

A. B. Nielsen

R. E. Traub

August 11, 2003
Date

ABSTRACT

Different scoring methods are used to produce ability estimates. Among them, the number right and the binary IRT models are commonly used. However, these methods are perceived as failing to incorporate examinees' partial knowledge in ability estimation processes. Hence, the model that includes information from all item alternatives in ability estimation was proposed (Bock, 1997). Another factor affecting ability estimates is testwiseness. It was found that examinees who posses both partial knowledge and test-taking skills can obtain a higher test score (e.g., Rogers & Yang, 1996). Therefore, the purpose of this study was to compare ability estimates yielded by the number right (NR), one- (1PL), two- (2PL), and three-parameter (3PL) models, and nominal response model (NRM) using items not susceptible to testwiseness (NTW) and items susceptible to the ID1 testwiseness strategy ("eliminate options that are known to be incorrect and choose from the remaining alternatives"). These comparisons were conducted for high, middle and low ability examinees.

The initial and replication studies conducted using responses of 4,000 high school students to multiple-choice items in the Social Studies 30 and Chemistry 30 examinations yielded essentially the same results. The differences between the subtest of NTW items and the ID1 items were not found for the social studies but were observed for the chemistry examination. The correlations and root mean square deviations (RMS) for pairs of scores yielded by the five models were comparable for both subtests on the social studies examination. For the chemistry examination, lower correlations and larger RSMs were for the pairs of scores yielded by the NR and 1PL, and the 2PL, 3PL, and NRM

models for the ID1 subtest than for the NTW subtest. These differences were greater for the middle and low ability examines than for high ability examinees.

The psychometric characteristics of the four subtests partially explain the differences between the two subject areas. The ID1 items were easier than the NTW items for the chemistry examination. The difference between the NTW and ID1 subtests was smaller for the social studies. It appears that subtest difficulty influences agreements among scores yielded by different models. Also, it seems that the influence of information from incorrect responses on ability estimates is weak. The discrepancies between the scores yielded by the NRM and 2PL and 3PL were small and consistent for both the NTW and the ID1 subtests. Given these findings, implications for educational testing practices and suggestions for future research are discussed.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to many people who have contributed to the completion of this thesis. I would like to thank my supervisor, **Dr. W. Todd Rogers** for his neverending support, encouragement, and thoughtful advice on many academic and non-academic issues. I am also thankful to **Drs. Mark Gierl, Hank Janzen, Gene Romaniuk**, and **Brian Nielsen** form the University of Alberta as well as to **Dr. Ross Traub** from the University of Toronto for their willingness to serve on my thesis committee. Their comments and suggestions contributed greatly to the final product of my work.

My very special thanks to **Dr. Mark Gierl** who introduced me to the world of item response theory. His influence is obvious in this project as well as my career choice.

I would also like to express my appreciation to the Alberta Education, Student Evaluation Branch for providing the data sets used in this study. My special thanks to the Strathcona Composite High School teachers, **Les Ferrar, Clayton Kinasevich**, and **Dan McDirmid**, and my CRAME colleagues, **Fernando Cartwright** and **Don Klinger**, for their help in analyzing materials used in my research. I would also like to recognize the assistance provided by **Dr. David Thissen** from the University of North Carolina in solving problems related to the use of the Multilog computer program.

I would also like to take this opportunity to thank the many CRAMErs, whom I was fortunate to work with in the past few years. I look forward to working with all of you as colleagues in the future.

Table of Content

List of Tables

List of Figures

CHAPTER I - INTRODUCTION

Background of the Study

Tests are used to select, classify, screen, and promote students. Developed to assess students' knowledge and skills, test results are used to make inferences and decisions about student performance. Increasing emphasis on local and provincewide accountability has produced the situation where students in elementary and secondary schools are frequently required to write achievement tests designed to assess their knowledge, skills, and attitudes in relation to objectives in their programs of studies. How well a student performs on a test is often considered critical to success in these settings. Given the significance of a test score, a question arises about how well the score reflects an individual's competency in a particular subject area or field (Salvia & Ysseldyke, 1995).

The multiple-choice test is one of the more popular test forms used to assess student academic achievement. This is particularly true at the higher grade levels and in large scale testing programs. Among the numerous advantages of the multiple-choice test, quick and objective scoring are often cited as a factors leading to preference of these tests over other forms of assessment (e.g., performance assessment) (Aiken, 1987; Bennet & Ward, 1993; Hambleton & Murphy, 1992). However, the way in which multiple-choice items are scored has also been criticized for failing to provide adequate information about the knowledge or skills possessed by students. The multiple-choice items on achievement tests are most often scored dichotomously: one point is awarded for the correct response and no point is given for any incorrect response. Several researchers perceive this type of conventional scoring as deficient because it implicitly assumes that examinees act according to a "knowledge-or-random guessing" principle (Lord, 1980). That is, examinees either have the knowledge to answer an item correctly or simply randomly select their answers from among the alternatives provided. It is reasonable to presume, however, that some examinees may possess only part of the knowledge necessary to select the correct answer and that they may use this incomplete knowledge to choose a particular incorrect alternative (De Ayala, 1989, 1993; Lord, 1980). As Tatsuoka (1983) and others (e.g., Brown & Burton, 1978; Jacobs & Vandeventer, 1970; Lane, Stone, & Hsu, 1990) have found in the analyses of student performance on a variety of academic and non-academic problem solving tasks, "...wrong responses can be more than just one kind, although the binary scoring procedure uniformly assigns a score of zero to all the wrong responses" (Tatsuoka, 1983, p.346). The number right or conventional type of scoring does not take into account whether the correct response was chosen on the basis of a student's total knowledge, partial knowledge, misinformation, or guessing (Frary, 1989; Rogers & Bateson, 1991a).

This problem is often linked to the quality of the answer choices on a multiple-choice item (Bock, 1972; Rogers & Ndalichako, 2000; Thissen, Steinberg, & Fitzpatrick, 1989). An ideal, properly functioning multiple-choice item is an item with equally plausible distracters or foils. Such an item would lead to a testing situation in which test-takers who do not possess sufficient knowledge to answer it correctly are equally attracted to each of the item alternatives. Studies of testwiseness reveal, however, that

this is not always true (Diamond & Evans, 1972; Millman, 1966; Rogers & Bateson, 1991b).

Testwiseness has been defined as a person's cognitive capacity to utilize the characteristics and formats of the test and/or the test-taking situation to improve a test scores. If an examinee possesses relevant partial knowledge and knowledge of the testwiseness strategies, and if the test contains testwise susceptible items, then the combination of these elements may result in improved or higher test scores. In contrast, an examinee with little testwiseness will likely be disadvantaged whenever the test involves susceptible items (Millman, 1966; Rogers & Bateson, 1991).

Several researchers have demonstrated that many tests contain testwise susceptible items. The most common types of testwiseness clues include absurd options, stem-option association, similar and opposite options, and options containing specific cues (Crehan, Koehler, & Slakter, 1974; Diamond & Evans, 1972; Hughes, Salvia, & Bott, 1991; Millman, 1966; Rogers & Bateson, 1991a; Sarnacki, 1979). Absurd options are options known by most examinees to be incorrect. In consequence, students knowledgeable of this strategy and possessing relevant partial knowledge avoid absurd options and choose from among the remaining ones. Stem-option association allows examinees to recognize and make use of a resemblance between an option and an aspect of the stem. Similar options tend to be considered by an examinee simultaneously and, given that there is only one correct response, none of them are chosen. In contrast, opposite options will guide a skilled test-taker toward choosing neither or one (but not both) of two options, one of which, if correct, would imply the incorrectness of the other. Recognizing and making use of a specific determiner included in an option has also been found helpful in distinguishing the correct answer from incorrect alternatives (Millman, 1996). The complete taxonomy of testwiseness principles is presented in Appendix A.

Given these findings a question arises about the relevance of partial knowledge and test-taking skills when determining ability estimates. If the partial knowledge is considered not to be relevant to the individual ability estimate, then dichotomous scoring of students' responses to multiple-choice items may be warranted. If partial knowledge, however, is considered relevant, then a total test score which takes into account the additional information coming from incorrect responses may be a more valid indicator of accomplishment (Messick, 1989). For example, Levine and Drasgow (1983) demonstrated that at least for some items, examinees at different levels of ability tended to have different patterns of wrong responses (i.e., very able examinees will differ from low ability examinees in their pattern of wrong responses). Further, if a test contains testwise susceptible items and if examinees employ both partial knowledge and test-taking skills to respond to testwise susceptible items, then the distribution of incorrect responses to these items may differ from the distribution of incorrect responses to the items that are not testwise susceptible (Levine & Drasgow, 1983; Nedelsky, 1954; Rogers & Ndalichako, 1999; Rogers & Yang, 1996; Thissen et. al., 1989).

Consequently, Bock (1972; 1997) and Thissen and Steinberg (1997) have suggested that incorrect responses contain information that may be useful in estimating the latent ability of a test-taker. They argue that the dichotomization of an examinee's response ignores any partial knowledge that the examinee may have and, as a result, this information is not used for ability estimation. They have suggested that because simple right-wrong scoring does not provide sufficient information on the latent source of

ability, a thorough analysis of students' responses using all the item alternatives should be conducted in order to assess the probability of response to each alternative of the item. Furthermore, Levine and Drasgow (1983) demonstrated that an item's incorrect alternatives may improve the estimate of an examinee's ability level by providing information about the examinee's level of understanding (i.e., provide diagnostic information).

It has also been shown that increased precision of measurement may result when information from the incorrect alternatives on multiple-choice test items is included in the ability estimate (Bock, 1972; Levine & Drasgow, 1983; Sympson, 1983; Thissen, 1976; Thissen & Steinberg, 1984; Thissen, Steinberg, & Fitzpatrick, 1989). Bock (1972) and Thissen (1976) found that for examinees in the lower half of the ability range, where wrong responses occur with greater frequency, analysis of the incorrect responses using item response theory increased information (i.e., reduced measurement error) by one-third to twice the information yielded by conventional right-wrong scoring. In the upper range of ability, the two scoring methods yielded approximately equal information. In terms of test length, this means that analyzing all response categories of multiple-choice items will, for half of the examinee population, increase in the precision of an ability estimates as compared to dichotomously scored tests (De Ayala, 1993).

In the search for the best scoring and ability estimation methods, a number of techniques have been developed. By far, the dominant methods for estimation of item and ability parameters come from the classical test theory and item response theory. Classical test theory postulates the relationship between the true score and the observed score (Gulliksen, 1950). The true score of an individual is defined as this individual's mean observed score on an infinite number of parallel or interchangeable tests. Since the construction and administration of parallel tests to an examinee is not feasible, the true score is a theoretical concept. Thus, the observed score of an examinee represents the ability of that particular examinee on a particular sample of items (Suen, 1990). All dichotomously scored correctly answered multiple-choice items, regardless of their parameters, equally contribute to the estimated ability of an individual. Item response theory (IRT) provides an alternative method of ability estimation. Although IRT consists of family of models, the most commonly used are the one-, two- and three-parameter models. These models require that individual responses be categorized as either correct or incorrect. In this respect they are similar to conventional right-wrong scoring. The item response models, however, allow for estimation of individual latent ability as a function of the joint probability of a response pattern to a set of items with specific characteristics (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980).

Not all examinee-item interactions may be appropriately modeled by dichotomous scoring models, however. Given the finding that incorrect option choice is related to ability, it has been suggested that applying a polychotomous test model may be more appropriate for estimation of the ability of individuals than a dichotomous model (Bock, 1972; De Ayala, 1992, 1993; Thissen & Steinberg, 1984, 1997). In response to this, several IRT models for use with polychotomous scored items have been proposed. Among them, the best known are the graded response model (Samejima, 1969), the nominal response model (Bock, 1972), the rating scale model (Andrich, 1978), and the partial credit model (Master, 1982).

The most common type of multiple-choice items that appear on aptitude and achievement tests are items with unordered response options and one correct answer. These items can be best analyzed using the nominal response model (NRM) proposed by Bock (1972). In this model it is assumed that item alternatives represent responses measured at a nominal level of measurement. An elaboration of the simple IRT models used to score a dichotomously scored test, the nominal response model allows for examination of the relationship between each option and the cognitive ability measured by the test. Similar to other latent-trait IRT models, the probability of a given pattern of item responses, in which the actual responses of the examinees are considered, is expressed as the product of the corresponding category characteristics conditional on ability. However, while binary IRT models allow for analyzing the relationship between the probability of selecting a correct response and ability, the NRM uses the relationship between the probabilities of selecting the correct and each incorrect response option and ability across all options of all test items to produce ability estimates (Bock, 1972; De Ayala, 1993; Thissen et. al., 1989).

Ability estimation methods that are capable of recovering maximum information from examinees are still being evaluated in the light of factors that influence ability estimates (Bock, Thissen, & Zimowski, 1997; De Ayala, 1995). Bock (1997), De Ayala (1993) and Thissen et. al. (1989), for example, perceive polychotomous item response models to be a promising trend in the area of measurement and strongly advocate employment of polychotomous measures to gain greater precision of individual ability estimation and better understanding of the examinee's performance on a test item.

Purpose of Research

The purpose of this study, therefore, was to investigate the comparability of ability estimates obtained using the number right conventional scoring model and the one-, two-, and three-parameter item response binary models and ability estimates obtained from nominally scored multiple-choice items using a nominal item response model in the presence of testwiseness. Two types of item were analyzed: testwise susceptible items and items not susceptible to testwiseness. Since several researchers (Bock, 1972; Levine & Drasgow, 1983; Thissen, 1976; Thissen & Steinberg, 1984) have agreed that examinee ability levels may likely be associated with the selection of a particular option of a multiple-choice item, the comparability of these estimates was investigated at high, middle, and low ability levels.

More specifically, this study was designed to answer the following research questions:

1. To what extent are ability estimates obtained from right-wrong scoring model, the one-, two-, and three-parameter item response theory binary scoring models, and the two-parameter item response theory nominal scoring model applied to responses of Grade 12 students to a subtest of testwise susceptible and non-susceptible items similar to each other?
2. Do ability estimates obtained from the different estimation procedures across the two types of items yield different results for Grade 12 students at high, middle, and low proficiency levels?

3. Are differences in ability estimates yielded by different estimation methods across the two different types of items and three proficiency level consistent across two subjects areas: social studies and chemistry?

Prior to addressing these questions, an item analysis within each scoring model was conducted for each subtest. The results of these analyses allowed for comparison of the two subtests at the item and test level. In addition, an examination of item alternatives was performed. Examination of item response distributions using the classical item analysis allowed for determination whether testwise susceptible items display a specific response pattern that was different from that obtained from items not susceptible to testwiseness. As the presented earlier research findings suggest, the pattern of incorrect responses differs for examinees at different ability levels (Levine & Drasgow, 1983; Thissen et. al., 1989). Item parameters obtained in the nominal response model were used to describe and analyze the relationship between each alternative and the cognitive proficiency measured by the test. It was hypothesized that unequal attractiveness of response options in testwise susceptible items would result in a lower probability of selection of some alternatives by middle and low ability students who possess partial knowledge of the item content. For these groups of examinees, the probability of selecting the correct response or an incorrect response option that was most often chosen by high ability students was anticipated to be higher for testwise susceptible items. Consequently, it was expected that the response patterns for examinees at different proficiency levels would be different for testwise susceptible and non-susceptible items.

It was also hypothesized that ability estimates yielded by the five scoring models at different levels of performance would be quite different. Using nominal scoring rather than binary scoring would likely intensify these differences, particularly at the middle and low ability levels, where incorrect responses occur with greater frequency. In addition, it was hypothesized that these differences, if found, would be greater for the testwise susceptible items than for items not susceptible to testwiseness. Also, if differences in ability estimates yielded by different estimation methods across the two different types of items and three proficiency levels were found, it was expected that they would be consistent across the two subject areas.

Definition of Terms

Dichotomous scoring: Scoring that categorizes an individual's response as either correct or incorrect (i.e., 0,1). Commonly used to score multiple-choice items designed to measure aptitude or ability (Lord & Novick, 1968).

Nominal scoring: Scoring that reflects an individual's original response to a test item. The number of scoring categories equals the number of non-ordered options in a multiple-choice item (Bock, 1972).

Number right scoring model: Model that is based on scoring of multiple-choice items by awarding one point for a correct response and zero for any other response. The sum of item scores constitutes the observed score of an examinee and reflects the ability of that particular examinee on a particular sample of items (Lord, 1952).

One-parameter logistic model: Model that specifies the probability of a correct (or incorrect) response to a dichotomously scored item as a function of an examinee's latent ability and item difficulty. It requires the assumption that all items are equally discriminating and no allowance is made for the possibility that some questions may be correctly answered by guessing. In this model the amount of a latent trait or ability possessed or achieved by an examinee is associated with a joint probability between response pattern and ability. The expression for the joint probability is called the likelihood function and the value of θ at which the likelihood function reaches its maximum represents the maximum likelihood estimate of θ or ability estimate for that examinee (Hambleton et. al., 1991; Lord, 1980)

Two-parameter logistic model: Model that specifies the probability of a correct (or incorrect) response to a dichotomously scored item as a function of an examinee's latent ability and two item parameters: difficulty and discrimination. Similar to the one parameter model the total score or ability possessed by an examinee is associated with a joint probability of a response pattern and ability and is reflected by the maximum likelihood estimate of θ or ability estimate for that examinee (Hambleton et. al., 1991; Lord, 1980).

Three-parameter logistic model: Model that specifies the probability of a correct (or incorrect) response to a dichotomously scored item as a function of an examinee's latent ability and three item parameters: difficulty, discrimination, and guessing. Similar to the one- and two-parameter logistic models the amount of a latent trait or ability possessed or achieved by an examinee is reflected by the maximum likelihood estimate of ability (Hambleton et. al., 1991; Lord, 1980).

Nominal response model: Model that specifies the probability of a response to a polychotomously scored item with k options as a function of an examinee's latent ability and two item parameters: slope or discrimination parameter and intercept parameter indicating the overall popularity of an alternative k. This model is suitable for categorically scored items and allows for description of the relationship between each alternative and the cognitive ability measured by the test. It defines operating characteristics for each response category such that the probability of response, conditional on ability, is restricted to sum to unity. The total score or ability estimate is related to the joint probability of an individual's response pattern and is reflected by the maximum likelihood estimate of ability (Bock, 1972; Lord, 1980).

Organization of the Thesis

The introduction of the issues related to different ability estimation methods using multiple-choice testwise susceptible and non-susceptible items and presentation of specific research questions in Chapter I is followed by a review of the relevant literature in Chapter II. This chapter is organized thematically to address the different aspects of the research problem. It presents and discusses the ability estimation from the perspective of classical test model and item response models. The review of studies involving testwiseness and methods of ability estimation from multiple-choice items concludes the

chapter. Chapter III describes the procedures used for the study. The topics dealt with include the data sets used in this study and the procedures involved in analyzing the data. Chapters IV and V present research results for the social studies and chemistry examinations, respectively. Chapter VI provides a discussion of the results and comparison of the two subject areas, and Chapter VII includes the summary of the study, conclusion, and recommendations.

CHAPTER II - LITERATURE REVIEW

Multiple-choice achievement and aptitude tests are frequently used to measure various abilities of a person. An individual's performance on any of these tests is viewed as the reflection of his or her accomplishment in an area or domain a test is supposed to measure. From an institutional perspective, the results of these tests often have an important influence on a student's placement in school, admission to university, and choice of and participation in various professions and activities. At the personal level, the results of these tests are likely to influence a student's view of himself or herself and how others view this person. Given the importance of testing, it is essential that the derived numerical score that reflects the quantity of the trait the test is designed to measure be as truthful and accurate as possible (Suen, 1990).

In a typical paper-and-pencil multiple-choice examination an examinee is given a set of items with a number of response options and asked to select the correct or best response. Given the responses to the set of items, an attempt is made to describe the examinee's ability by summarizing the responses in some way. There are a number of methods for aggregating responses to multiple-choice items. This chapter presents an overview of the conventional scoring method that is rooted in the classical test theory, the one-, two-, and three-parameter item response theory binary scoring methods, and the nominal scoring method of the item response theory. This overview is followed by a discussion of selected factors that, if present, may affect ability estimates yielded by these scoring methods. The chapter concludes with review of studies in which these scoring methods have been compared. As will be noted, there has been a poverty of research involving the nominal response model in comparison to the other four scoring models.

Classical Test Theory

The classical test theory is the earliest theory of measurement (Gulliksen, 1950). It is also referred to as the classical reliability theory for one of its major tasks is to estimate the reliability of an individual's observed score on the test. That is it attempts to estimate the strength of the relationship between the observed score and the true score of a person writing a particular test. The classical test theory is also sometimes referred to as the true score theory because its theoretical derivations are based on a mathematical model known as the true score model (Suen, 1990).

In a classical test theory, a common approach to arrive at an examinee's observed score on a multiple-choice test is to give each correct or best response a score of 1 and each incorrect response score of 0. In this case a score for the examinee is the number of correct responses. This observed score for the individual represents the ability of that individual estimated from his or her responses to a particular set of items administered at a particular occasion under a particular set of conditions. Many factors, however, can affect a person's performance on the test. An examinee may perform differently on different sets of items, at different times, and under different personal and environmental conditions. In other words, multiple administrations of the test, including different, but parallel sets of items to the same individual under different conditions, would likely produce many different observed scores for that person. Given an infinite number of

observations, the mean of all these observed scores would be the unbiased estimate of that person's true score on the test. This mean is defined as (Suen, 1990):

$$\xi_f (X_{jf}) = \tau_j,$$

where

X_{jf} is the observed score on a person j on the f form of the test, and

τ_j is the true score of a person j.

The observed score from any single test administration will likely be different from the estimated true score and this difference is defined as measurement error. Consequently, the mathematical relationship between the observed score on one form of the test and the true score for a person j is expressed as:

$$X_{jf} = \tau_j + \varepsilon_{jf}$$

where

ε_{jf} is the error score of a person j on the form f.

Since

$\xi_f (X_{jf}) = \tau_j$, it follows that

$$\xi_f (\varepsilon_{jf}) = 0.$$

Hence, X_{jf} is an unbiased estimate of τ_j (Gulliksen, 1950). The numerical score on a test is considered to be an examinee's true score (Gulliksen, 1950).

For most aptitude and achievement tests, it has been found that the number of items answered correctly is a highly satisfactory score. Since the majority of these tests are designed in the way that each correctly answered item is awarded 1 point and each incorrectly answered item is given 0 point, the number right score is considered to be an adequate true score for an examinee (Gulliksen, 1950; Haladyna, 1999; Suen, 1990).

Although the classical theory of measurement maintains a strong influence among testing and measurement practitioners today, the use of conventional methods of item analysis and true score estimation is often perceived as having a number of important limitations. This is particularly true when a sample of people used is relatively small (Suen, 1990).

In classical test theory an examinee's score on the test is said to be test-dependent and his or her ability is defined only in terms of a particular test. When a test contains many difficult items, the examinees will appear to have low ability. When a test contains many easy items, the same examinee will seem to have high ability. Whether an item is hard or easy is determined by its difficulty (p-value), that is the proportion of examinees in a group of interest who answered the item correctly. Such a concept of item difficulty does not take into account examinee differences in ability and ignores the fact that for individuals at different ability levels, the probability of responding correctly to an item will not be the same. Moreover, the item parameters are sample-dependent and cannot be compared to different item indices obtained from different populations of examinees. Given the lack of a standardized scale that applies across different tests, examinees' scores are test-dependent. Consequently, the classical test theory does not permit

examinees to be compared when they have not taken the same items (Crocker & Algina, 1986; Hambleton, et. al, 1991; Suen, 1990).

Another criticism of the classical test theory is associated with the standard error of measurement that is assumed to be the same for all examinees. As Lord (1984) pointed out, for a given test, some observed scores contain more measurement error than others.

Item Response Theory

Lord (1952, 1980) introduced item response theory in an attempt to overcome some of the limitations of classical test theory. The desirable features of item response theory include the possibility of obtaining item characteristics that are not group-dependent and test-scores describing examinees ability that are not test-dependent. IRT also permits determination of how well an examinee or a group of examinees at different proficiency levels respond to a given item. Finally, IRT provides models that allow for a measure of precision for each ability score (Hambleton and Swaminathan, 1985; Hambleton et. al., 1991; Suen, 1990).

Item response theory rests on two basic postulates: (a) the performance of an examinee on a test item can be predicted by a set of factors called latent traits or abilities; and (b) the relationship between examinee item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC). This function describes the relationship between an examinee's trait or ability level being measured by the test and the probability that the examinee will answer the item correctly (Hambleton & Swaminathan, 1985).

Many possible item response models exist. They rely on several common assumptions. First, an assumption common to the IRT models most widely used is the assumption of unidimensionality, that is, only one ability is measured in common by a set of items in a test. Second, the assumption of local independence requires that when the abilities affecting test performance are held constant, examinees' responses to any pair of test items be statistically independent. Another assumption made in all IRT models is that the item characteristic function reflects the true relationship between the unobservable variable (ability) and a response to an item (Hambleton et al., 1991; Lord, 1980).

Item response models differ in the mathematical form of the item characteristic curve and/ or the number of parameters specified in the model. In cognitive testing through multiple-choice items several item response models can be employed to estimate examinees' ability underlying their performance on the test. Among them, the one-, two-, and three-parameter logistic models can be applied to dichotomously scored multiple-choice items, and the two-parameter nominal response model can be applied to nominally scored multiple-choice items. The latter model, takes into consideration examinees' responses to all item response options.

One-Parameter Logistic Model

The one-parameter model (1PL) specifies the probability of a correct response to a dichotomously scored item as a function of an examinee's latent ability and item difficulty. The item difficulty, b_i, is the point on the ability scale where the probability of a correct answer is .5. Although, theoretically, the values of the b_i parameter can range from $-\infty$ to $+\infty$, typically they vary from about –2.0 to +2.0. The greater the value of the b_i parameter, the greater the ability that is required for an examinee to have a 50% chance

of getting the item right. In this model it is assumed that item difficulty is the only item characteristic that influences examinee performance. Consequently, use of the one-parameter model requires the assumption that all items are equally discriminating. Moreover, no allowance is made for the possibility that some questions may be correctly answered by guessing.

The probability of a correct response to item i for a randomly chosen person j with ability θ_j for the one-parameter model is given by the equation:

$$P_i(\theta_j) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}, \quad i = 1, 2, \ldots, n,$$

where

$P_i(\theta_j)$ is the probability that a randomly chosen examinee with ability θ_j answers item i correctly,

b_i is the difficulty parameter for item i,

n is the number of items on the test, and

e is the base for natural logarithms.

Four examples of item characteristic curves for the one-parameter model are presented in Figure 1. Given the restrictive assumptions of the one-parameter model, the item characteristic curves differ only in respect to their location (b_i) on the ability scale. As shown, less ability is required to answer item 1 than to answer item 2. The slopes of the curves are equal indicating equal discrimination of items, and the lower asymptote for each curve is zero indicating zero probability of correctly answering an item by low ability examinees.

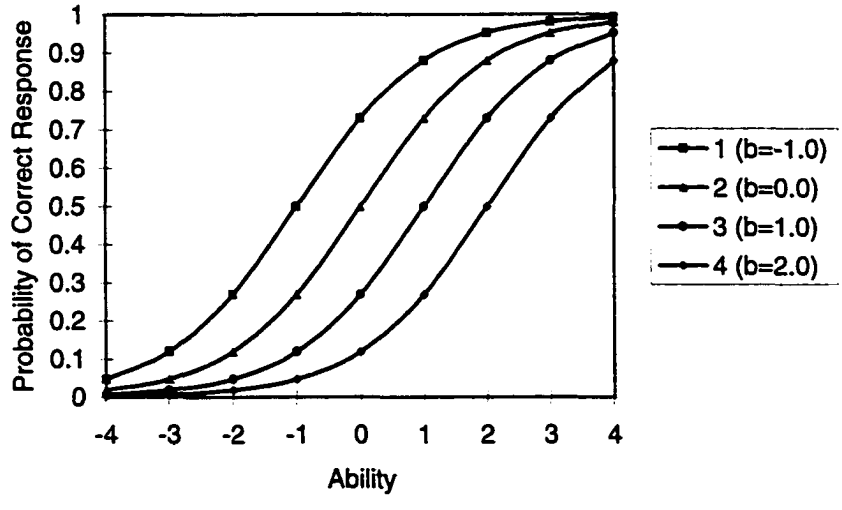


Figure 1. One-Parameter Item Characteristic Curves for Four Items (Hambleton et. al., 1991, p.14).

Two-Parameter Logistic Model

The two-parameter model (2PL) specifies the probability of a correct (or incorrect) response to a dichotomously scored item as a function of an examinee's latent ability and two item parameters: difficulty (b_i) and discrimination (a_i). The parameter a_i is

proportional to the slope of the item characteristic curve at the point b_i on the ability scale. It allows for identification of the items that best separate examinees into different ability groups. Although, theoretically, the values of the a_i parameter can range from -∞ to +∞, typically they vary from about 0.0 to +2.0. The higher the value of a_i parameter, the greater discrimination power. Similar to the one-parameter model, this model makes no allowance for guessing behavior.

The probability of a correct response to item i for a randomly chosen person j with ability θ for the two-parameter model is given by the equation:

$$P_i(\theta_j) = \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}, \quad i = 1, 2, \ldots, n,$$

where the parameters $P_i(\theta_j)$ and b_i and the elements e and n are defined as in the one-parameter model and

a_i is the item discrimination parameter, and

D is a scaling factor with the value of 1.7 used to make the logistic function as close as possible to the normal ogive function (Hambleton et. al., 1991; Lord, 1980).

Four samples of item characteristic curves in the two-parameter model are displayed in Figure 2. As shown, the item characteristic curves in this model are not parallel indicating different discrimination power (a_i) of the items. The curve for item 1 has the steepest slope indicating discrimination power greater than that of other items. Item 3 is least discriminating. Similar to the one-parameter model, the lower asymptote of each curve is zero, indicating no guessing behavior in low ability students.

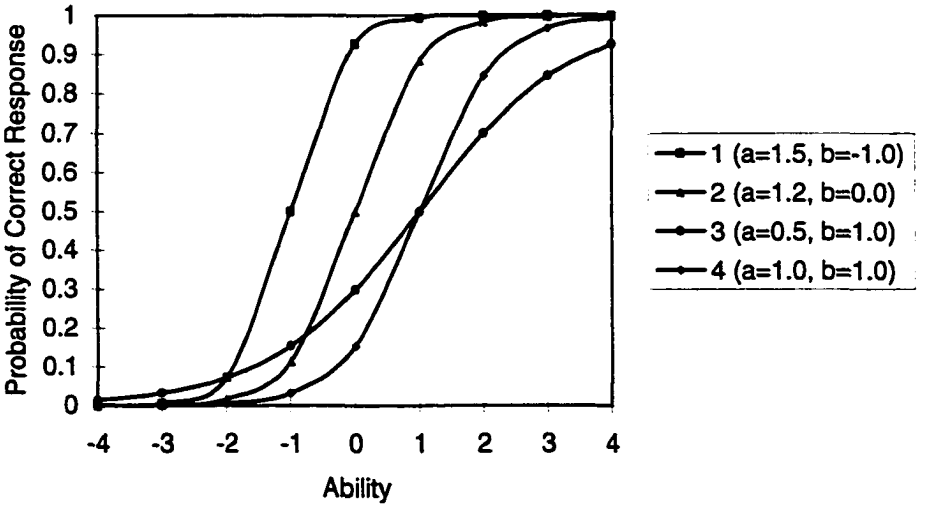


Figure 2. Two-Parameter Item Characteristic Curves for Four Items (Hambleton et. al., 1991, p.15).

Three-Parameter Logistic Model

The three-parameter model (3PL) specifies the probability of a correct (or incorrect) response to a dichotomously scored item as a function of an examinee's latent ability and three item parameters: difficulty (b_i), discrimination (a_i), and guessing (c_i). In

addition to item difficulty and discrimination, the guessing parameter (also called the pseudo-chance-level parameter) is incorporated into the model to take into account behavior of the examinees at the low ability levels where guessing is a factor believed to affect test performance on selected-response (i.e., multiple-choice) items. This parameter provides a non-zero lower asymptote for the item characteristic curve and represents the probability of examinees with low ability level answering the item correctly. Typically, the c_i parameter assumes values that are smaller than the value that would result if examinees guessed randomly on the item. The probability of a correct response to item i for a randomly chosen person j with ability θ for the three-parameter model is given by the equation:

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}, \quad i = 1, 2, \ldots, n,$$

where the parameters $P_i(\theta_j)$, a_i, and elements D, e, and n are defined as in the two-parameter model and

b_i is the difficulty parameter for item i given a value of the lower asymptote of item characteristic curve. When $c_i > 0$, the probability associated with a correct response at a given b value on ability scale will exceed 50 percent and will equal $p_i = (1 + c_i)/2$, and

c_i is the guessing parameter (Hambleton et. al., 1991; Lord, 1980).

Item characteristic curves in the three-parameter model are presented in Figure 3.

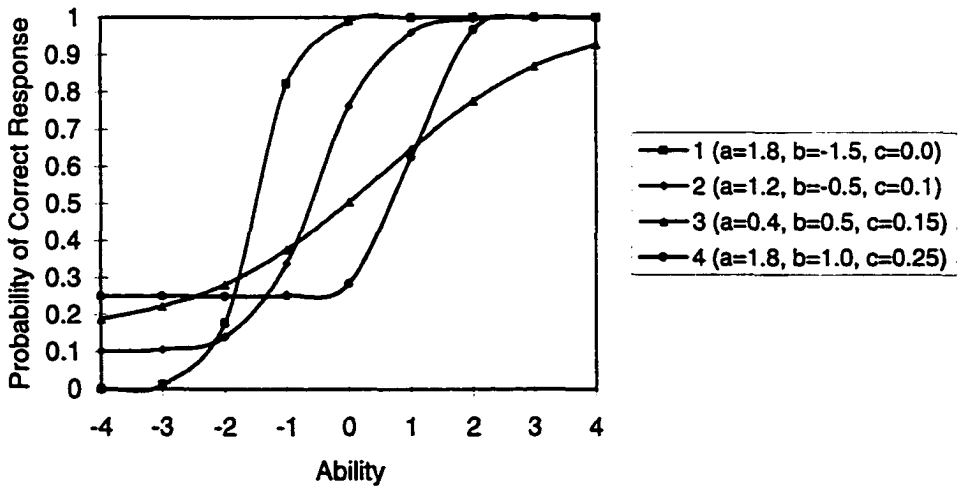


Figure 3. Three-Parameter Item Characteristic Curves for Four Items (Hambleton et. al., 1991, p.16).

As shown in Figure 3, the four item characteristic curves differ in terms of their location on the ability scale (b_i), steepness of their slopes (a_i), and values of their lower asymptotes (c_i). The ICC curves of more difficult items (items 3 and 4) are shifted to the higher end of the ability scale while the curves of items 1 and 2 are allocated at the lower end of the ability scale. The steepness of the item 3 indicates its lower discrimination

power as compared to other items. The different lower asymptotes provide information on the probability of responding correctly to an item by the low ability examinees.

Two-Parameter Nominal Response Model

In contrast to the one-, two-, and three-parameter logistic models, the nominal response model (NRM) is applied to non-dichotomous test data. It is assumed that item alternatives represent responses measured at a nominal level of measurement and that the alternatives are not ordered (Bock, 1972). The nominal response model is suitable for categorically scored items and allows for the description of the relationship between each alternative and the cognitive ability measured by the test. The purpose of this model is to maximize the precision of ability estimates by using all the information contained in the examinees' responses, not just the correct response.

According to Bock (1972, 1997), the probability of a response to a nominally scored item with k options can be defined as a function of an examinee's latent ability and two item option parameters: slope or discrimination (a_k) and intercept indicating the overall popularity of an alternative k (c_k). The nominal response model employs the following two equations:

$$z_{ik} = a_{ik}\theta_j + c_{ik}$$

$$P_i(\theta_j) \mid (x_i = k \mid \theta; a, c) = \frac{e^{(a_{ik}\theta_j + c_{ik})}}{\sum_{k=1}^{m_i} e^{(a_{ik}\theta_j + c_{ik})}}, \quad i = 1, 2, \ldots, n,$$

where

$P_i(\theta_j) \mid (x_i = k \mid \theta; a, c)$ is the probability that a test-taker of ability θ_j will respond to category k of item i,

x_i is a response to item category

k is the item response category ($k = 1, 2, \ldots, m$),

m_i is the number of response categories,

z_{ik} is a linear function of the latent proficiency θ,

a_{ik} is the slope, or regression coefficient relating z_{ik} to latent ability, and

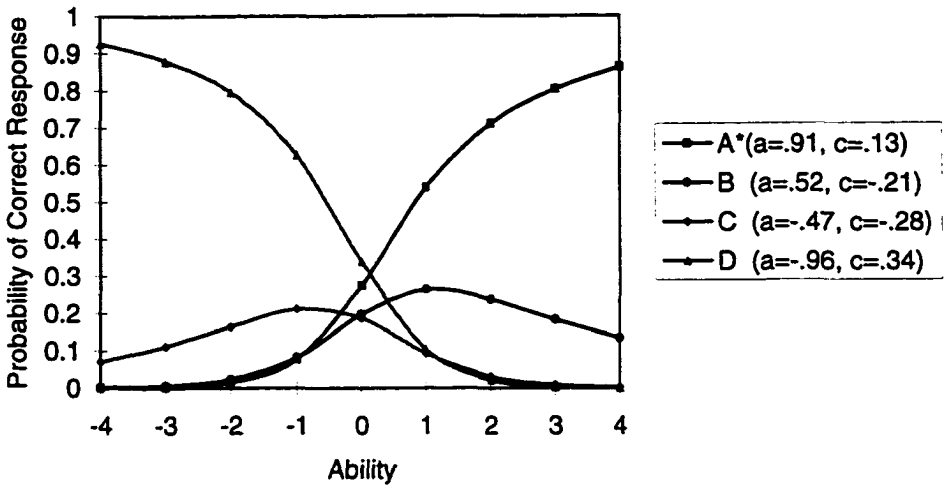
c_{ik} is an intercept parameter indicating the overall popularity of an alternative k.

Large positive values of z_k are associated with likely responses while smaller or negative values are associated with less likely responses. The $a_{correct}$ should have a positive slope for the correct option and the $a_{incorrect}$ should have a negative slope for each wrong alternative if an item is functioning in the correct manner. The item parameters are the vectors $\vec{a}$ and $\vec{c}$, with imposed linear constraints, yielding $2(m_i - 1)$ free parameters. The linear constraints require that the sum of the a_{ik} parameters and the sum of c_{ik} parameters equal zero:

$$\sum_{k=1}^{m_i} a_{ik} = \sum_{k=1}^{m_i} c_{ik} = 0.$$

In addition, the nominal response model defines operating characteristics for each response category such that the probability of response, conditional on ability, is restricted to sum to unity (Bock, 1972, 1997).

In the nominal response model, the item characteristic curves describe the simple relationship between an examinee's response to any item option and his or her ability (Bock, 1972; Lord, 1980). The four characteristic curves for a four-option multiple-choice item in the nominal response model are shown in Figure 4.



Note: * indicates the correct item option.

Figure 4. Two-Parameter Nominal Response Model Item Characteristic Curves (Bock, 1997, p.39).

As shown, the curve with the smallest discrimination parameter (option D) is monotonically decreasing from 1 at $-\infty$, while the curve with the largest discrimination parameter (option A) is monotonically increasing as ability increases. The curves with the intermediate values of discrimination parameter (options B and C) reach their maximum at finite value of ability.

Ability Estimation

The values of item and ability parameters that characterize each item and examinee can be determined given a specific item response model. Typically, a random sample of examinees from a target population is selected and their responses to a set of items are obtained. Given the item responses, the item and ability parameters are estimated (Hambleton et. al., 1991).

Several estimation procedures are available (e.g., Bayesian estimation, maximum likelihood estimation, and approximate estimation). Among them the maximum likelihood procedure (MLE) is the one most frequently used. Maximum likelihood estimators have been found to be a) consistent (i.e., as the sample size and the number of items increase, the estimators converge to the true value); b) efficient (i.e., asymptotically the maximum likelihood estimators have the smallest variance); and c) asymptotically normally distributed. Moreover, maximum likelihood estimators are known to be functions of sufficient statistics when sufficient statistics exist (Hambleton & Swaminathan, 1985; Swaminathan, 1983).

Ability estimates in binary logistic models. When an examinee responds to a set of items that are scored either 1 (a correct response) or 0 (an incorrect response), then under the assumption of local independence, the joint probability of observing the particular response pattern is the product of the probabilities of observing each item response, and is given by the following equation:

$$P(U_1, U_2, \ldots, U_n \mid \theta) = \prod_{i=1}^{n} P(U_i \mid \theta), \quad i = 1, 2, \ldots, n,$$

where

U_i is a response to item i, and

$P(U_i \mid \theta)$ is a probability of a response to item i.

Since U_i is either 1 or 0, the formula can be rewritten as:

$$P(U_1, U_2, \ldots, U_n \mid \theta) = \prod_{i=1}^{n} P(U_i \mid \theta)^{U_i} [1 - P(U \mid \theta)]^{1-U_i},$$

or more simply as

$$P(U_1, U_2, \ldots, U_n \mid \theta) = \prod_{i=1}^{n} P_i^{U_i} Q_i^{1-U_i},$$

where

$$P_i = P(U_i \mid \theta) \text{ and } Q_i = 1 - P(U_i \mid \theta).$$

When the response pattern is observed, $U_i = u_i$, then the probabilistic interpretation is no longer appropriate. In this case the likelihood function, denoted as $L(u_1, u_2, \ldots, u_n \mid \theta)$, is used. Given this, the likelihood function for an examinee with a particular response pattern is expressed as:

$$L(u_1, u_2, \ldots, u_n \mid \theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}$$

where

P_i is the probability of responding correctly to item i, and

Q_i is the probability of incorrect response to item i.

The value of θ that makes the likelihood function for an examinee a maximum is defined as the maximum likelihood estimate of θ for that examinee. The likelihood function ranges from $+\infty$ for examinees who answer all items correctly to $-\infty$ for individuals who answer all items incorrectly. In practice, the maximum likelihood estimation fails when a perfect score or a zero score is encountered. The θ values typically range from –4 to +4 (Hambleton & Swaminathan, 1985; Hambleton et. al., 1991; Swaminathan, 1983).

Ability estimates in the nominal response model. When examinees respond to a multiple-choice item with unordered response options, their original response patterns can be used to compute their ability scores. It has been shown that an increase in precision of measurement may be obtained when information in the incorrect alternatives on multiple-choice items is included in the ability estimation (Bock, 1972, 1997; De Ayala, 1995; Thissen & Steinberg, 1997). For the most part, the additional information

obtained from incorrect responses is limited to examinees at the lower level of ability, as most of the higher ability students choose few incorrect alternatives (Bock, 1972, 1997; Thissen, 1976; Levine & Drasgow, 1983; Thissen & Steinberg, 1984, 1997).

Similar to the binary logistic models, the joint probability of observing the particular response pattern in the nominal response model is the product of the probabilities of observing each item response, and is given by the following equation:

$$L_\ell(\theta) = P(U_\ell|\theta) = \prod_{i=1}^{n} P_{ik}(\theta), \quad i = 1, 2, \ldots, n,$$

where

$L_\ell(\theta)$ is the likelihood of θ

U_ℓ is the response pattern denoted as $[U_{1\ell}, U_{2\ell}, \ldots, U_{n\ell}]$,

$k = U_{i\ell}$ is the item score designating the category to which the response to item i in pattern ℓ corresponds, and

P_{ik} is probability of an examinee responding to an item i in category k.

As with the binary logistic models, the value of θ that makes the likelihood function for an examinee a maximum is defined as the maximum likelihood estimate of θ for that examinee. The values for θ range from $+\infty$ for examinees whose responses to the items are all in the category with the largest a value to $-\infty$ for individuals whose responses to the items are all in the category with the smallest a value. In practice, however, θ values typically range from –4 to +4 (Bock, 1972, 1997; Hambleton et. al., 1991; Thissen & Steinberg, 1984, 1997).

Influence of Selected Factors on Ability Estimates

Research evidence shows that a test score does not depend entirely on an examinee's "full" knowledge of the item content. Other factors include partial knowledge of the information sampled by test items, guessing, and testwiseness. These factors have been found in varying degrees in individuals of all ages and at all grade levels (Crehan, Gross, Koehler, & Slakter, 1978; Diamond & Evans, 1972; Thissen & Steinberg, 1997).

Partial Knowledge

In a typical multiple-choice testing situation, an examinee is presented with a set of questions with a number of possible responses and asked to select the correct or best response. As indicated earlier, the usual scoring of multiple-choice items is based on awarding one point to each correctly answered question and no points to each incorrectly answered test item. This scoring formula is based on the assumption that examinees select the correct response, either because they know the correct response or because they guess successfully. Selecting an incorrect alternative is seen as a result of the examinee's lack of knowledge or unsuccessful guessing. The examinee's total score is then taken to be a simple sum of the test item scores. Although convenient and efficient, this simple "knowledge-or-random guessing" model (Lord, 1952, 1980) rarely yields precise scores. Examinees who posses partial knowledge about an item and those who are misinformed about an item do not respond to the item randomly. For these two types of examinees the incorrect alternatives are not equally attractive (Lord & Novick, 1968).

Recovering information from the incorrect responses to the multiple-choice items has received considerable attention. Over three decades ago, Lord and Novick (1968) and Shuford, Albert, and Massengill (1966) criticized the conventional scoring formula for its failure to extract all of the potentially available information from students' responses to multiple-choice items. Similarly, the one-, two-, and three-parameter item response models for dichotomously scored items make no allowance for examinees having partial information about the question asked. These models are unable to predict what might happen in situations had the partial information been used (Bock, 1972, 1997; Thissen & Steinberg, 1997). The basic logic behind the critique of dichotomous item scoring is that among examinees who earn identical item scores on conventionally scored multiple-choice items, there may be varying degrees of knowledge or partial knowledge about that item (Crocker & Algina, 1986).

Bock (1972, 1997) and Thissen, Steinberg, and Fitzpatrick (1984) suggested that distractors should be considered an important part of the item. Several studies demonstrated the existence of a relationship between distractor choice and total test score (Bock, Thissen, & Zimowski, 1997; De Ayala, 1995; Levine & Drasgow, 1883; Thissen et. al., 1989). One of the earliest studies attempting to investigate the patterns of incorrect responses was conducted by Sigel (1963) who analyzed the items on the *Raven's Progressive Matrices*. He found no relationship between error pattern and total score on the test. Contrary to this finding, Jacobs and Vandeventer (1970) arrived at a systematic method for *a priori* ordering the distractors on the *Raven's Coloured Progressive Matrices* test as to degree of correctness. They categorized the incorrect responses on Progressive Matrices as "superior" (that is partially correct) and "other". Finding that the proportion of "superior" wrong responses correlated positively with the total score on the test, the authors concluded that additional information may be available in the incorrect choices and the relationship between type of error and total score does exist when information on type of error is summed across items (Jacobs & Vandeventer, 1970).

Thissen (1976) continued the investigation of extracting information from wrong responses on the *Raven's Coloured Progressive Matrices* and using this information to improve the accuracy of ability estimation. He demonstrated that the scoring model that uses information in the patterns of wrong responses as well as in the correct responses in the process of estimating ability provides from one third more to nearly twice the information of the binary model for the lower half of the ability range.

This finding was similar to that obtained earlier by Bock (1972) who investigated the response patterns to vocabulary multiple-choice items on the *Cooperative Achievement Test*. Bock (1972) concluded that multiple category scoring improves the accuracy of ability estimation mainly in the lower part of ability distribution because examinees of lower ability respond incorrectly more often than do individuals of higher proficiency.

Support for these findings is provided by the results of the study conducted by Levine and Drasgow (1983), who analyzed the responses of large groups of examinees to multiple-choice items of the *Graduate Record Examination* – Verbal section and the *Scholastic Aptitude Test* – Verbal section (n=9,900 and n=49,470 respectively). They demonstrated that very able examinees have characteristic patterns of wrong responses. Moreover, for some items, high scoring examinees were found to select the same one or two options when they answered the item incorrectly; the other distractors were unlikely

to be selected. Some of these rarely selected distractors were found to be popular among middle ability examinees. Examinees in the lower ability strata had different patterns of wrong responses when compared to the patterns found at the high ability level (Levine & Drasgow, 1983).

Despite rather sparse research done in this area, information in incorrect responses to multiple-choice items proved to be an important factor in ability estimation. It has been demonstrated that examinees do not select their responses at random but instead they use partial knowledge of the item content. Therefore, it seems that the multiple category scoring model may be a more appropriate model for obtaining ability estimates.

Guessing

An intrinsic characteristic of a multiple-choice item is the probability that an examinee without the necessary knowledge can correctly guess the answer. This leads to the general postulate that item and test score are affected by guessing. That is,

$$X_{j,tot} = X_{kd,j} + X_{g,j}$$

where

$X_{j,tot}$ is the total number of correct responses for a person j,
$X_{kd,j}$ is the number of correct knowledge derived responses, and
$X_{g,j}$ is the number of correct random responses (Rogers & Yang, 1996).

It is frequently assumed that random guessing is an adequate representation of guessing behavior. Based on the random guessing model, the probability that an examinee can answer an item correctly through guessing is equal to $1/k$ for the multiple-choice item with k options (Suen, 1990).

There are however reasons to believe that the random guessing model does not match reality. It has been found that the probability of a correct guess by an examinee who cannot identify the correct answer may be higher or lower than chance and it mainly depends on the quality of the distractors (Gulliksen, 1950; Lord, 1952; Suen, 1990). There are two main views regarding the probability of a correct guess. Some researchers suggest that examinees who do not know the correct answer generally possess some partial knowledge and based on this partial knowledge they are able to eliminate some distractors. For such examinees the probability of choosing the correct response is enhanced. For example, on a four-option multiple-choice item, an examinee who has some partial knowledge and is able to eliminate two of the three distractors as implausible will choose from only two remaining options. If the examinee guesses at random at this point, the probability of a correct answer increases form 25% to 50%. With this view, the probability of guessing right is higher than $1/k$ (Crocker & Algina, 1986; Diamond & Evans, 1972; Hughes, Salvia, & Bott, 1991).

An alternative, somewhat less popular view goes in the opposite direction. According to Lord (1974), when constructing a test, item writers deliberately generate distractors that are not only plausible, but attractive. He suggested that examinees without the necessary knowledge to select the correct response option would be attracted to these incorrect options. Consequently, the probability of correctly guessing for these examinees will be lower than $1/k$.

Although these two views on examinee guessing behavior differ with respect to the probability of a correct guess, they agree that whenever examinees possess partial knowledge about an item or are guided by misinformation about that item, the item

response options are not equally attractive to them. Therefore, the random guessing model cannot be considered as an adequate explanation of guessing behavior (Lord, 1952, 1974; Suen, 1990).

Testwiseness

Testwiseness has been defined as a person's cognitive capacity to utilize the characteristics and formats of the test and/or the test-taking situation to improve a test score. If the examinee possesses relevant partial knowledge and knowledge of the testwiseness strategies, and if the test contains susceptible items, then the combination of these elements may result in an improved or higher test score. In contrast, an examinee with little testwiseness will likely be disadvantaged whenever the test involves susceptible items (Millman, 1966; Rogers & Bateson, 1991a). Rogers and Bateson (1991b) proposed a model of testwise test taking behavior of skilled high school test takers. This model reflects various routes a test-taker may choose to determine what option to select on the multiple-choice item. It suggests that the cognitions of skilled test takers consist of: 1) a cognitive monitor that controls which abilities and skills are going to be engaged to answer the item under consideration; 2) abilities and skills relevant to the content or trait being measured (knowledge); 3) testwiseness strategies; and 4) the response (selection and record of choice). According to this model, other characteristics of a test taker in addition to knowledge or partial knowledge of the content being tested contribute to a testwise person' test score (Rogers & Bateson, 1991b).

When a multiple-choice test item contains testwise elements, examinees who do not possess complete knowledge about the item content can employ their test-taking skills to answer the item. There are several strategies of testwiseness that can be applied. Millman, Bishop, and Egel (1965) classified these strategies into two sets as presented in Appendix A. The first set contains strategies that are independent of test constructor or test purpose. Among them, the deductive reasoning strategies: eliminating options, which are known to be incorrect and choosing from among the remaining options; choosing neither or both of two options which imply the correctness of each other; choosing neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other; restricting choice to those options, which encompass all of two or more given statements known to be correct; and using relevant content information in other test items and options, are the most commonly found in tests and used by testwise examinees to gain points beyond what they would have received on the basis of full knowledge of what is being tested (Allen, 1992; Millman, 1966; Rogers & Wilson, 1993; Slakter et al., 1972). The testwise elements dependent on the test constructor are listed in the Part II of the taxonomy. Among them, the strategies that rely on the presence of specific cues in an item will help examinees gain points. These strategies include: recognizing and making use of any consistent idiosyncrasies of the test constructor, which distinguish the correct answer from incorrect options; considering the relevancy of specific detail when answering a given item; recognizing and making use of specific determiners; recognizing and making use of resemblance between the options and an aspect of the stem; and considering the subject matter and difficulty of neighboring items when interpreting and answering a given item (Diamond & Evans, 1972; Millman,1966; Rogers & Wilson, 1993).

The presence of testwise elements in multiple-choice items results in situations where skilled examinees who do not possess knowledge of the item content are not equally attracted to all item distractors. It has been demonstrated that incorrect or absurd response options of a multiple-choice item are eliminated by examinees who possess the necessary partial knowledge. This decreases the number of possible correct options. On the other hand, response options that contain certain cues about the answer are those on which skilled test-takers focus (Rogers & Bateson, 1991a; Rogers & Wilson, 1993; Towns & Robinson, 1993). Presented below are examples of testwise susceptible items (Rogers & Bateson, 1991a).

1. Which of the following is classified as an organ?
   A. Bone
   B. Skin
   C. Blood
   D. Muscle

This item contains an absurd option C easily eliminated by testwise examinees who will choose their response from the remaining alternatives.

2. A substance that, in its pure form, is the best conductor of electricity is
   A. Water
   B. Deuterium
   C. H_2O
   D. Silver

Item 2 contains two options A and C that imply correctness of each other. If the item has only one correct response, a skilled test-taker will chose neither of these options and focus on the remaining ones.

3. A spherical triangle is the triangle on the surface of a sphere. What name is given to the number of degrees in a spherical triangle minus 180?

   A. The arc of the triangle
   B. The size of the triangle
   C. The spherical excess of the triangle
   D. The polar measurement of the triangle

In item 3, both option C and the stem contain word "spherical". Recognizing the resemblance between the option and the aspect of the stem will allow a testwise examinee to select option C as the one most likely correct (Rogers & Bateson, 1991a).

Thissen et al. (1989) suggested that multiple-choice items with non-equivalent and informative distractors can be of particular interest when extracting information from incorrect responses. He argued that if the answer choices divide the examinees into distinct ability groups, the advantage of the information obtained from differentially attractive distractors could be used in the process of ability estimation (Thissen et al., 1989).

There has been an abundance of research on testwiseness involving high-school and college students (Morse, 1985; Rogers & Bateson, 1991a, 1991b; Samson, 1985; Towns & Robinson, 1993). For the purpose of this study, research focusing on the relationship between testwiseness and ability is of particular interest. Bangert-Drowns, Kulik, and Kulik (1983), Hughes, Schumaker, Deshler, and Mercer (1988), Rogers and Wilson (1993), Rowley (1974), Slakter, Koehler, and Hampton (1970b) and others have presented evidence that students can be taught test-taking stills and that the acquired testwiseness skills would improve scores on standardized, large scale tests as well as classroom tests. Fagley (1987) concluded that using the secondary cues on difficult multiple-choice history items by testwise freshmen university students resulted in an increased test scores as compared to the those obtained by test-naïve students. Roznowski and Basset (1992) investigated the effect of coaching practices used in training testwiseness for analogy items on the *Scholastic Aptitude Test* and found that students who received training in testwiseness were able to answer more items correctly than students who did not receive such preparation. Rogers and Bateson (1991a, 1991b) and Rogers and Wilson (1993) suggested that the effective application of testwise reasoning strategies depends upon the partial knowledge that examinees possess. Based on the results of a study of high-school seniors' responses to a multiple-choice test of testwiseness and multiple-choice school leaving examinations, they argued that an examinee who possesses both partial knowledge about item content and test-taking skills would have a greater probability of correctly answering a testwise susceptible item than a student low in testwiseness. Rogers and Wilson (1993) found that a test-sophisticated examinee is able to obtain a test score 10% to 15 % higher than a person who lacks test-taking skills. Similar results were obtained by Towns and Robinson (1993), who showed that students who used a variety of testwiseness strategies on chemistry examination gained points beyond those gained for specific content knowledge.

Several researchers (Morse, 1994; Rogers & Wilson, 1993; Rogers & Yang, 1996; Samson, 1985; Slakter, Koehler, & Hampton, 1970a) have also found that not all testwiseness skills are of equal difficulty and some of them can be taught to and learned by students as young as those in upper elementary school grades. As students increase in age, they are able to improve and extend their test-taking skills. Due to experience or cognitive strategies that have evolved over time, the majority of older students will possess considerable test-taking stills.

Summary

Different scoring methods are used to score multiple-choice items. The major shortcoming of the right-wrong conventional scoring model and the one-, two-, and three-parameter item response models for dichotomously scored items is that they fail to consider examinees' partial knowledge about an item content. The nominal scoring model overcomes this disadvantage and allows for information coming from correct as well as incorrect response options. Research that has been done (e.g., Bock, 1997; Levine & Drasgow, 1983; Thissen, 1976) strongly suggests that examinees who do not posses the necessary knowledge to answer an item do not randomly select an answer, but rather use partial knowledge about the item content to select their response. Therefore, partial knowledge appears to be an important factor that should be taken into consideration when

estimating examinees' abilities. It has been shown that using this additional information yields more precise ability estimates, particularly for examinees in the lower half of ability distribution.

Another factor that has been known to affect ability estimates is testwiseness. Testwise examinees who are able to take advantage of secondary cues in multiple-choice items are able to improve their test scores (e.g., Diamond & Evans, 1972; Millman, 1966; Towns & Robinson, 1993). It has also been shown that effective application of testwiseness strategies requires partial knowledge. Rogers and Bateson (1991b) argued that examinees who posses both partial knowledge and test-taking skills have a greater probability of selecting a correct response then their peers who possess partial knowledge but are not testwise or those who have knowledge of testwiseness principles but have low partial knowledge.

Although there have been several research studies in which testwiseness has been examined (e.g., Diamond & Evans, 1972; Millman, 1966; Rogers & Yang, 1996; Sarnacki, 1979), it appears that none of these studies considered the use of the nominal response model to produce ability estimates. Considering that the majority of high-school students who write multiple-choice achievement tests possess some partial knowledge as well as some test-taking skills (Rogers & Wilson, 1993), it seems worthwhile to consider the use of nominal response model to investigate the influence of information in wrong responses on ability estimates. Therefore, the purpose of the present study was to compare the ability estimates from dichotomously and nominally scored testwise susceptible and non-susceptible multiple–choice items.

CHAPTER III - METHOD

The purpose of this study was to examine the differences in ability estimates obtained using dichotomous and nominal scoring of 4-option multiple-choice items. The objectives of the study were the following:

1. Examination of the comparability of ability estimates yielded by the number right scoring model, the one-, two-, and three parameter item response binary scoring models, and the two-parameter item response nominal scoring model obtained from a subtest of testwise susceptible items and a subtest of non-susceptible items;
2. Investigation of the differences in ability estimates, if any, obtained from the five different scoring procedures for the two types of items for examinees at high, middle, and low proficiency levels; and
3. Determination of whether or not the differences, if any, in ability estimates yielded by different estimation methods across the two different types of items at the three proficiency level are consistent across two subjects areas: social studies and chemistry.

The method used to address these objectives as well as the procedures and computer programs involved in the data analysis are described in this chapter.

Identification of Testwise Susceptible Items and Subtests

Two data sets were used in this study. Each set consisted of the responses of high school students to the multiple-choice items contained in school-leaving examinations for social studies and chemistry (Alberta Education, 1999c; 1999e). These two subjects were selected to represent a humanities course and a science course. These diploma examinations are "high stakes" tests, which count 50% of a student's final grade, and are intended for students who are planning or wish to leave open the opportunity to pursue some form of tertiary education.

Prior to conducting data analyses, the items in each test were analyzed for the presence of testwise cues. Two current high school social study teachers and two graduate students with expertise in the social studies content completed the task for social studies. One current high school chemistry teacher, one former high school chemistry teacher, one graduate student familiar with the chemistry content area, and one expert in testwiseness completed the task for chemistry. All raters were familiar with the concept of testwiseness. The raters worked separately to identify which, if any, items on the test were testwise susceptible and, for each identified item, what testwise element(s) were present. Once they had completed this task, each person worked with the results of conventional item analyses and explained, using their subject matter expertise, the profile of results for each item. Given that the responses constituting the data sets used in this study were collected by Alberta Education, the item evaluation criteria for the *p*-values and item discrimination indices used by Alberta Education were employed by the raters. These criteria include:

a) Minimum and maximum acceptable difficulty levels of respectively, 0.30 and 0.85;
b) Minimum acceptable corrected point-biserial of 0.20;
c) Negative point-biserials for distractors; and
d) Difficulty level of at least 0.05 for the distractors (Alberta Education, 1999a).

The individual findings for each item were then discussed in a meeting of the raters for each subject area. For both the social studies and chemistry examinations the agreement among judges as to what items were susceptible to particular testwiseness strategies was reasonably strong. Across the two examinations at least three out of four judges agreed on the presence of testwiseness on 86 of 114 items. However, there was less agreement among the judges on the specific testwise elements within items identified as susceptible to the ID1 testwiseness strategy for the social studies examination. In the chemistry examination, there was strong agreement as to what item alternatives are absurd options. For both panels, consensus identification was reached for each item. The ratings and the list of final testwise cues identified for each item are presented in Appendix B for the social studies test and in Appendix C for the chemistry test. The summary of these ratings is reported in Table 1.

Table 1
Classification of Items by Type of Testwiseness Cue.

Diploma Examination	Number of Item Containing Testwise Cues						Total
	ID1	ID2	ID3	IIB4	Other	NTW	
Social Studies	23	3	2	3	4	35	70
Chemistry	14	0	4	0	5	21	44

Note: NTW refers to items not susceptible to testwiseness

As shown in Table 1, of the 70 items contained in the Social Studies 30 Diploma Examination, 35 were judged as not containing testwise elements. The remaining 35 items were found to be susceptible to one or more testwiseness strategy. Among them, 23 items were susceptible to the ID1 strategy (eliminate option(s) that are known to be incorrect and choose from the remaining alternatives); three items were susceptible to the ID2 strategy (choose neither or both of two options that imply the correctness of each other); two items were susceptible to the ID3 strategy (choose neither or one of two statements, one of which, if correct, implies the incorrectness of the other); and three items were susceptible to the IIB4 strategy (recognize and make use of the resemblance between the options and an aspect of the stem). Three items were found to be susceptible to both the ID1 and IIB4 strategies and one item was sensitive to both the ID1 and ID2 strategies.

Of the 44 items included in the Chemistry 30 Diploma Examination, 21 were judged to be non-susceptible to testwiseness. The remaining 23 items were found to be susceptible to at least one of the testwiseness strategies. Among them, 14 items were found to be susceptible to the ID1 strategy; four items were susceptible to the ID3 strategy; one item was susceptible to the ID5 strategy of testwiseness (utilize relevant content information in other test items and options); and one item was susceptible to the IIB3 strategy (recognize and make use of specific determiners). In addition, one item was found to be susceptible to both the ID1 and ID2 strategies; one item was sensitive to both

the ID3 and IIB4 strategies; and one item was susceptible to both the ID1 and IIB1a strategies.

Based on this item classification, two subtests were identified within each diploma examination for further analyses. For the Social Studies 30 Diploma Examination, one subtest consisted of the 35 non-susceptible to testwiseness items and the second contained the 23 items susceptible to the ID1 strategy. For convenience, these subtests are referred to as the SS-NTW and the SS-ID1, respectively. In case of the Chemistry 30 Diploma Examination, one subtest contained the 21 items that were not susceptible to any of the testwise strategies and the other subtest contained the 14 items susceptible to the presence of the ID1 testwise elements. These subtests are referred to as the CH-NTW and the CH-ID1, respectively. Since the numbers of items sensitive to other testwiseness strategies on both diploma examinations were not sufficient to obtain stable and reliable ability estimates, these items were not included in the further analyses.

Comparison of Subtests

According to Alberta Education standards, the questions on the diploma examinations require students to demonstrate knowledge of subject content and to apply cognitive skills to that knowledge base. Thus, in order to examine whether or not the subtests of items not susceptible to testwiseness and the subtests of items susceptible to the ID1 testwiseness strategy were similar to each other in regard to the knowledge and skills being tested, the distributions of the two types of items across the Social Studies 30 and the Chemistry 30 examination blueprints (Alberta Education, 1999b; 1999d) were compared. The results for items not susceptible to testwiseness are reported in regular print and the results for items susceptible to the ID1 testwiseness strategy are reported in **bold** print.

Social Studies 30

Each item on the social studies examination was designed to measure one of the two curricular content areas (topics) and one of the three knowledge or skill objectives required to answer the question (Alberta Education, 1999d). The distributions of the SS-NTW and the SS-ID1 items across the examination blueprint are presented in Table 2.

Table 2
Classification of the SS-NTW and the SS-ID1 Items According to the Social Studies 30, June 1999 Examination Blueprint

Question Classification by Knowledge and Skill Objectives	Question Classification by Topic		Total Number of Items
	Topic A: Political and Economic Systems	Topic B: Global Interaction in the 20th Century	
Comprehension of Information and Ideas	3 (8.57%) **5 (21.74%)**	7 (20.00%) **3 (13.04%)**	10 (28.57%) **8 (34.78%)**
Interpretation and Analysis of Information and Ideas	5 (14.29%) **5 (21.74%)**	6 (17.14%) **5 (21.74%)**	11 (31.42%) **10 (43.48%)**
Synthesis and Evaluation of Information and Ideas	6 (17.14%) **3 (13.04%)**	8 (22.86%) **2 (8.70%)**	14 (40.00%) **5 (21.74%)**
Total Number of Items	14 (40.00%) **13 (56.52%)**	21 (60.00%) **10 (43.48%)**	35 (100.00%) **23 (100.00%)**

Note: Numbers in regular print refer to non-susceptible to testwiseness items; numbers in **bold** print refer to items susceptible to the ID1 strategy of testwiseness.

As shown in Table 2, the two subtests were slightly different in the content and level of thinking assessed. Approximately 29% of the non-susceptible to testwiseness items (SS-NTW) and **35%** of items susceptible to the ID1 testwiseness strategy (SS-ID1) assessed examinees' comprehension of information and ideas. About 31% of the SS-NTW and about **43%** of the SS-ID1 items measured examinees' proficiency in the interpretation and analysis of information an ideas. Forty percent of the SS-NTW and close to **22%** of the SS-ID1 items assessed examinees' competence in synthesizing information and ideas. Two topics were assessed. Forty percent of the SS-NTW and about **57%** of the SS-ID1 items were referenced to political and economic systems (Topic A), while 60 % of the SS-NTW and about **43%** of the SS-ID1 items referenced to global interaction in the 20th century (Topic B).

Chemistry 30

Items on the chemistry examination were categorized by general learner expectations (GLE). Each item was classified in two ways: by the knowledge of one of two areas being tested and by none, one, or two types of skills: scientific processes and understanding of links between science and society (Alberta Education, 1999b). The results of this classification are presented in Table 3.

Table 3
Classification of the CH-NTW and the CH-ID1 Items According to the Chemistry 30, June 1999 Examination Blueprint

Classification of Questions by Skills and Understanding	Classification of Questions by Knowledge of Relationships in Transfer of Energy, Electron and Proton, and Equilibrium Systems		Total Number of Items
	Quantitatively Predicting Outcomes	Qualitatively Analyzing Systems	
Scientific Process and Communication (SPC)	2 (9.52%) **3 (21.43%)**	10 (47.62%) **1 (7.14%)**	12 (57.14%) **4 (28.57%)**
Science Technology and Society (STS)	1 (4.76%) **0 (0.00%)**	0 (0.00%) **1 (7.14%)**	1 (4.76%) **1 (7.14%)**
SPC and STS	5 (23.81%) **4 (28.57%)**	2 (9.52%) **4 (28.57%)**	7 (33.33%) **6 (42.86%)**
None	0 (0.00%) **0 (0.00%)**	1 (4.76%) **1 (7.14%)**	1 (4.76%) **1 (7.14%)**
Total Number of Items	8 (38.10%) **7 (50.00%)**	13 (61.90%) **7 (50.00%)**	21 (100.00%) **14 (100.00%)**

Note: Numbers in regular print refer to non-susceptible to testwiseness items (CH-NTW); values in **bold** print refer to items susceptible to the ID1 strategy of testwiseness (CH-ID1).

As shown in Table 3, the two subtests differed in regard to type of skills required to answer the items. Fifty seven percent of the CH-NTW items and about **29%** of the CH-ID1 items measured students' scientific process and communication skills (SPC). Approximately 33% of the CH-NTW items and about **43%** of the CH-ID1 items assessed both SPC and STS types of scientific skills. One item on each subtest measured examinees' understanding of connections among science, technology, and society (STS) and one item in each subtest assessed only knowledge but no scientific skills. The two subtests were similar in regard to the knowledge of chemistry content. Approximately 38% of the CH-NTW items and **50%** of the CH-ID1 items were related to the quantitatively predicted outcomes. About 62% of the CH-NTW items and half of the CH-ID1 items referenced qualitatively analyzed systems.

Initial and Replication Samples

The total number of students that completed the social studies and chemistry examinations was 10,905 and 8,594, respectively. For the purposes of this study, two samples of 4,000 students each were randomly drawn from the total number of examinees for each test. The initial data analyses were performed using one sample for each test.

These are labeled as Social Studies-Sample 1 and Chemistry-Sample 1. In order to examine the stability of the results obtained from the initial analyses, the analyses were replicated using the remaining two samples - Social Studies-Sample 2 and Chemistry-Sample 2.

Formation of Ability Groups

The external measure of examinees' ability was not available. Therefore, given that the number right score is the most often used indicator of ability, the examinees' total raw scores on the social studies and chemistry tests were used to create three ability groups. Using the guidelines provided by Kelley (1927), students who scored at or above 74^{th} percentile were classified as the high ability examinees, students who scored between 36.5^{th} and 63.5^{th} percentiles were labeled as the middle ability examinees, and students who scored at or below the 27^{th} percentile were described as the low ability group. These groups were formed separately for each the social studies and chemistry sample.

Data Analyses

Prior to conducting any analyses each of the obtained data sets was recoded twice. First, examinees' original responses were recoded as 1 for option A, 2 for option B, 3 for option C, and 4 for option D. Second, examinees' original responses were dichotomized and recoded as 1 for a correct answer and 0 for an incorrect answer. The psychometric characteristics of the items not susceptible to testwiseness and items susceptible to the ID1 testwiseness strategy in each diploma examination were examined using conventional item analysis procedures and item response models for dichotomously and nominally scored items.

Classical Item Analysis

The LERTAP (Nelson, 1983) program was used to conduct item analyses. The Alberta Education item evaluation criteria presented earlier (see p. 24) were employed to determine whether or not the items in each subtest functioned properly. Moreover, examination of the percentage of examinees who chose each option for each item on the four subtests was conducted. This analysis permitted determination whether items susceptible to the ID1 strategy of testwiseness produced different distribution of examinees' responses compared to properly functioning items. It was expected that this information would be helpful in interpretation of potential differences in ability estimates produced by different scoring methods.

Item Response Item Analysis

The item response models considered in the present study are predicated on five basic assumptions. Thus, before the analyses were conducted, the tenability of the assumptions of unidimensionality, local independence, equal discrimination, non-speededness, and non-guessing were tested.

Unidimensionality. It is assumed that only one ability is measured by the set of test items. This assumption cannot be strictly met because several cognitive, personality, and test-taking factors (i.e., level of motivation, test anxiety, ability to work quickly) always affect, at least to some extent, an examinee's performance on the test. Therefore,

it was assumed that in order to satisfy the assumption of unidimensionality, a test must measure a single dominant ability (Hambleton & Swaminathan, 1985; Nanadakumar, 1994). Although counting only dominant dimension violates Lord and Novick's (1968) definition of unidimensionality, it is commonly accepted that in order to apply the IRT unidimensional models, it is sufficient to show that there is one dominant ability underlying examinees responses to a set of items (Nanadakumar, 1994).

To determine the unidimensionality of the data sets, three procedures were employed. First, principal components extraction was conducted. It was expected that a) the first component would be fairly large thus accounting for a large proportion of variance; b) the difference between the first and the second component would be large enough to support an inference about the first component being a dominant component; and c) the differences among successive pairs of neighboring components, beginning with the second component, would be insignificant (Gorsuch, 1983). Second, scree plots were employed to clarify and confirm the results from the first method (Cattell, 1952). The third method involved the use of the Stout's T statistic (Nandakumar & Stout, 1993; Stout, 1987, 1990). Stout's T statistic has been found to discriminate well between unidimensional and multidimensional sets of test scores for both simulated and real data (Nandakumar, 1993, 1994; Nandakumar & Stout, 1993; Stout, 1987).

The DIMTEST (Stout, 1993) program was used to compute the Stout's T statistic and to test the null hypothesis that each set of items was essentially unidimensional. The automatic procedure available in the DIMEST was employed. In this procedure, the test is divided into three subtests AT1 (assessment subtest 1), AT2 (assessment subtest 2) and PT (partitioning subtest) using the factor loadings of the items. These subtests are formed on the basis that the items within each of them represent a dominant dimension. The purpose of the PT subtest is to group examinees into subgroups. The purpose of the AT1 is to compute the Stout's T statistic, and the purpose of the AT2 is to reduce the statistical bias in AT1 arising from the short test length and/ or extremely high or low difficulty level of the AT1 items. The computed Stout's T value is referred to the upper tail of the standard normal distribution to obtain the significance level. The obtained p-values associated with unidimensional tests are expected to be large ($p > .05$), while the p-values associated with multidimensional tests are expected to be within the margin of the specified level of significance ($p < .05$). In the present study, it was anticipated that the items of AT1, AT2 and PT would be of the same dominant dimension indicating that the model underlying the test responses is essentially unidimensional. Thus, the value of the computed Stout's T would be small, leading to the tenability of the null hypothesis at the significance level of .05 (Nandakumar, 1994).

Unidimensionality of each of the analyzed subtests was assessed separately. It was expected that the results of the principal components factor analysis, the scree test, and the Stout's T statistic would lead to the same decision about the dimensionality of the item sets.

Local independence. Another main assumption of the IRT is that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent (i.e., no relationship exists between examinees' responses to different items). The assumption of local independence is met when the latent trait specified in the model is the only factor influencing examinees' responses to test items. In such case, the conditional on ability covariances of responses to any pair of

items are zero (Lord, 1980; Lord & Novick, 1968). The weaker form of local independence that satisfies the essential unidimensionality is essential independence (Nanadakumar, 1994). A data set is said to be essentially independent with respect to the latent dominant ability trait when the average absolute conditional on ability covariances of responses to item pairs approach zero as the test length increases. When the assumption of essential independence holds, the essential unidimensionality is obtained.

The presence of testwise susceptible items on the test may lead to a concern about meeting the assumption of essential independence. As Hambleton et al. (1991) pointed out, item independence may not hold when "a test item contains a clue to the correct answer, or provides information that is helpful in answering another item" (p.12). As explained earlier, testwise susceptible items contain certain clues about correct answers, which may be detected by some examinees but not by others. If this ability to detect a clue is considered to be a factor other than the ability being measured by a test, then the complete latent space will not be specified and the item independence will not hold (Hambleton et al., 1991). However, previous research suggests that the presence of testwise elements in test items does not affect its dimensionality. Ndalichako (1997) found that there was a dominant component underlying the subtest of testwise susceptible items selected from the English 30 Diploma Examination indicating its essential unidimensionality and, thus, essential independence. Therefore, it was expected that the dimensionality analyses conducted in the present study would yield comparable results, that is, no dominant dimension other than proficiency in a particular subject matter would be found.

Equal discrimination. The assumption of equal item discrimination is required in case of the one-parameter model. To determine whether the items are homogeneous in regard to their discrimination indices, the values of the point-biserial correlations obtained in the classical item analysis were examined. If the difference between the lowest and the highest point-biserial correlations were found to be lower than 0.15, the assumption of equal item discrimination would be considered to be tenable (Hambleton & Murray, 1983).

Non-speededness. When a test has a time limit such that some examinees finish but others do not, an examinee's working pace will systematically influence his or her performance on a test (Crocker & Algina, 1986). To ensure that the rates at which examinees worked did not constitute an additional source of variance that is irrelevant to the latent trait being measured, the test data were analyzed by reviewing the percentage of examinees completing each test. The percentage of examinees who completed the last three items on each test was calculated. If 95% of the examinees completed these items, then the pace of work would be considered to be an unimportant factor in test performance (Lord, 1980).

Guessing. Minimal guessing is a desirable characteristic of examinees' performance on a test. Since the one- and two-parameter models as well as the nominal response model do not include a guessing parameter, an assumption that guessing has minimal influence on individuals' performance on a test is required. Hambleton et al. (1991) suggested that this assumption be checked by examining the performance of low ability students on the most difficult items. If their performance levels are close to zero, the assumption is viable. In this study, low-achieving examinees were operationally defined as examinees whose number right score on the full social studies or chemistry

test fell at or below the chance level. The chance level score was calculated using the following formula:

$$Y = X_c + 2SE_c,$$

where

$$X_c = N_i / k,$$

$$SE_c = \sqrt{N_i p_i q_i}, \text{ and}$$

Y is the calculated chance level score given two standard errors of chance,
X_c is a score obtained by guessing on a set of multiple-choice items,
N_i is a number of test items,
k is a number of response options on a test item,
SE_c is a standard error of chance,
p_i is a probability of guessing a correct answer given k, and
q_i is a probability of guessing an incorrect answer given k.

The items identified as the most difficult were three items with the lowest p-values. Close to zero performance of low-ability examinees on these items was considered to be an indication of tenability of the minimal guessing assumption.

Item parameter estimation. Given that the assumptions of the item response theory models were satisfied, item parameters for each subtest were estimated. The MULTILOG program (Thissen, 1991) was used to produce marginal maximum likelihood estimates of item parameters for the unrestrained one-, two-, and three-parameter logistic models and for the two-parameter nominal response model.

Analysis of item trace lines. Since the use of information from the incorrect responses in a process of ability estimates was of particular interest in this study, analyses of the examinees' responses to the not susceptible items and the items susceptible to the ID1 testwiseness strategy were performed using the nominal response model. The basis for the comparison between the two types of items were the c_{ik} and a_{ik} item option parameters. Larger positive values of c_{ik} indicate the higher popularity of a particular alternative among examinees and lower negative values indicate less likely responses. The a_{ik} parameter, a regression coefficient or a slope, reflects the relationship between choice a particular response and the latent proficiency of an examinee. For the correct response, a_{ik} is expected to be positive and high. The remaining a_{ik} parameters are expected to have low positive or negative values associated with the decreasing probability of choosing the incorrect alternatives as the test-takers ability (θ) increases.

The item option parameters, c_{ik} and a_{ik}, were used to calculate the probability of selecting a particular response option across ability levels. The obtained values were used to produce the trace lines for each item, which, in turn, allowed for graphical representation of the distribution of probabilities of selecting each response option across ability levels. Visual inspection of the item trace lines for non-susceptible and susceptible to the ID1 testwiseness strategy items was performed to determine if there were differences between the two types of items in each of the two subject areas. This information was used in subsequent interpretation of comparisons of ability estimates produced by different scoring models.

Ability Estimation

Following estimation of the item parameters, the ability estimates using the five scoring methods were obtained for each subtest.

Conventional ability estimation. Ability estimation in classical test theory relies on the principles of number right scoring model. The conventional scoring model is based on awarding one point for a correct response and zero for any other response. Thus, the sum of item scores constitutes the observed score of an examinee and reflects the ability of that particular examinee on a given set of items (Crocker & Algina, 1986).

Ability estimation in item response models. The number right true score τ and θ are monotonically related and the true score may be considered a non-linear transformation of θ. Since a person's number right true score on a test is defined as the expectation of his or her observed score, it follows that every person at ability level θ has the same number right true score. This relationship is expressed in the following equation

$$\tau_j = \sum_{i=1}^{n} P_i(\theta_j),$$

where

τ_j is defined as a true score of a person j, and

$P_j(\theta_j)$ is the probability that a test-taker of ability θ_j will respond correctly to item i, i = 1,2, ... n.

Thus, the true score of an examinee with ability θ is the sum of the item characteristic curves or the test characteristic curve. This relationship holds when item response model fits the data (Lord, 1980; Hambleton, et al., 1991).

When the values of item parameters are known from the prior calibration, it is possible to estimate the ability, θ, of an examinee from a knowledge of the examinees' responses to a set of test items (Swaminathan, 1983). Given that, the obtained sets of item parameter estimates were used in a subsequent run to produce ability estimates. The MULTILOG program (Thissen, 1991) was used to produce maximum *a posteriori* (MAP) estimates of individual ability using a Gaussian prior distribution within each IRT model using item parameters from each analyzed subtest. An advantage of the MAP estimation is that it is defined for all response patterns, including patterns in which all items were answered correctly or all items were answered incorrectly (Thissen & Steinberg, 1997; Thissen, personal communication, August 9th, 1999).

Comparison of Ability Estimates

Ability estimates produced by the number right scoring model, the one-, two-, and three-parameter binary item response models, and the nominal response model were compared with each other for both subtests within each of the two subject areas. The purpose of these comparisons was to determine whether or not these scoring models provided similar ability estimates for the two types of items. As it was explained earlier, differences among ability estimates yielded by the five scoring models were expected to be greater for items susceptible to testwiseness than for items not susceptible to testwiseness. Using nominal scoring rather than dichotomous scoring would likely intensify these differences particularly for examinees at the low and middle levels of ability, where incorrect responses occur with greater frequency. Therefore, separate comparisons were conducted for students at high, middle, and low proficiency levels.

Since the ability estimates produced by the number right scoring model and the item response models were expressed in different metrics, they were converted to T-scores ($\mu = 50$, $\sigma = 10$) prior to the comparative analyses. The transformed ability estimates were compared using two different procedures. First, to examine if the students were ranked differently by the different scoring methods, the Pearson-product moment correlations among the ability estimates were compared. Second, to examine if the scores for examinees were equal in an absolute sense, the root mean square deviations (RMS) among pairs of ability estimates were calculated. The following formula was used:

$$RMS_{xy} = \sqrt{\frac{\sum_{j=1}^{N}(T_x - T_y)^2}{N-1}},$$

where

RMS_{xy} is the root mean square between the transformed scores produced by models x and y,

T_x is the transformed score produced by model x,

T_y is the transformed score produced by model y, and

N is the number of examinees.

The hypotheses tested were that the correlations between pairs of scores yielded by different scoring models would be 1.0 (i.e., H_0: $\rho_{xy} = 1.0$) and that the root mean square deviations between pairs of ability scores would be zero (i.e., H_0: $RMS_{xy} = 0.0$). These hypotheses were not tested statistically due to the large sample sizes. However, as shown in next two chapters, the correlations were lower than 1.0 and the root mean square deviations were greater than zero. Due to the lack of statistical guidance, relative comparisons of correlation coefficients and root mean square deviations were conducted, using an arbitrary decision that the correlations equal to or greater than 0.95 would be 1.0, with the difference from 1.0 due to sampling error. Correlations less than 0.95 were considered to denote systematic differences in examinees ranking by different scoring models. In the case of root mean square deviations, values greater than 2.0 were considered as indicators of systematic differences between pairs of ability estimates yielded by different scoring models.

CHAPTER IV – SOCIAL STUDIES 30

The results of the analyses conducted for the social studies examination are reported in the present chapter. The corresponding results for the chemistry examination are presented in the next chapter. Both chapters are organized in two major sections. In the first section, results for the initial sample are presented. The corresponding results for the replication sample are reported in the second section. Each of the sections starts with the results of conventional and item response theory item analyses, followed by the comparisons of ability estimates for examinees at different ability levels. To facilitate the comparisons, these outcomes are presented simultaneously for the subtest of items not susceptible to testwiseness and the subtest of items susceptible to the ID1 testwiseness strategy. The results for the non-susceptible to testwiseness items are reported in regular print and the results for items susceptible to the ID1 strategy are reported in **bold** print. Both chapters conclude with a discussion of results obtained from the initial and replication data analyses.

Initial Data Analyses: Social Studies - Sample 1

The distribution of the total test scores for the initial sample (Sample 1, N = 4,000) was similar to the distribution of total test scores obtained for the entire population of grade 12 students who wrote the Social Studies 30 Diploma Examination in June 1999. The means of the total test scores for the population and Sample 1 were, respectively, 46.88 (66.97%) and 46.96 (67.09%); the corresponding standard deviations were 11.87 (16.96%) and 11.79 (17.01%). As outlined in the method section in the previous chapter, three distinct ability groups were identified for further analyses. Employing the cut-off scores listed there (see p. 29), 988 students were classified as high ability examinees, 1,141 were classified as middle ability examinees, and 1,086 students were identified as low ability examinees.

Classical Item Analysis

The test statistics for both subtests are reported in Table 4. For convenience and to differentiate the initial sample from the replication sample, the non-susceptible and ID1 susceptible subtests are labeled SS-NTW-S1 and SS-ID1-S1, respectively, where S1 denoted the initial sample or Sample 1. The obtained difficulty index (p-value), point-biserial (r_{pb}), and corrected point-biserial correlations (r_{cpb}) for each item are presented in Table 20 for the SS-NTW-S1 and Table 21 for the SS-ID1-S1 of Appendix D.

Table 4
Summary of Test Statistics for the SS-NTW-S1 and the SS-ID1-S1

Subtest	Number of Items	Mean Raw Score	Standard Deviation	Reliability and SEM
SS-NTW-S1	35	22.85 (65.29%)	6.20 (17.71%)	0.83 (2.50)
SS-ID1-S1	**23**	**15.54 (67.57%)**	**4.11 (17.88%)**	**0.75 (2.00)**

Note: SEM refers to the standard error of measurement.

As shown in Table 4, the mean number right score for the SS-NTW-S1 was 22.85 (65.29%) and the corresponding standard deviation was 6.20 (17.71%). The internal consistency index (Cronbach's α) was 0.83 and the standard error of measurement was 2.50.

The item difficulties for the correct option ranged from 0.30 (item 49) to 0.87 (item 28) and yielded a mean *p*-value of 0.65 with the corresponding standard deviation of 0.13 for the SS-NTW-S1 items. Adopting the evaluation criteria used by Alberta Education, all items met the minimum difficulty standard of 0.30 for the correct option. One item (item 28) exceeded the maximum difficulty standard of 0.85. The point-biserial correlation coefficient values for the correct option ranged from 0.20 (item 49) to 0.51 (item 28). The corrected point-biserial item discrimination index for the correct option ranged from 0.13 (item 49) to 0.45 (item 27). Items 49 and 51 did not meet the Alberta Education standards of corrected point biserial correlation being at least 0.20.

The proportion of examinees who selected the incorrect options ranged from 0.03 (items 18 and 28) to 0.36 (item 49) across the 35 items. Of the 105 foils, six did not meet the Alberta Education criterion for distractors: *p*-value for the distractors less than 0.05. Items with these foils were generally found to be easy items with the remaining incorrect options selected by less than 9 % of examinees. The point biserial correlations for the incorrect alternatives were all less than zero and varied from -0.05 (item 33) to -0.32 (item 26).

The mean number right score for the SS-ID1-S1 was **15.54** (**67.57%**) and the corresponding standard deviation was **4.11** (**17.88%**) (cf. Table 4). The internal consistency index (Cronbach's α) was **0.75** and the standard error of measurement was **2.00**.

The item difficulties for the correct option varied from **0.40** (item 12) to **0.84** (items 11, 24, and 37) yielding a mean difficulty of **0.68** and standard deviation of **0.11** for the SS-ID1 items. Employing the evaluation standards used by Alberta Education, all items met the minimum difficulty standard of **0.30** and none of the items exceeded the maximum difficulty standard of **0.85**. The point-biserial correlation coefficient values for the correct option ranged from **0.24** (item 38) to **0.53** (item 65). The corrected point-biserial item discrimination index for the correct option ranged from **0.13** (item 38) to **0.44** (item 65). For two items (11 and 13), the corrected point-biserial correlation was lower than Alberta Education's standard of 0.20.

The proportion of examinees who selected the incorrect options ranged from **0.01** (item 11) to **0.52** (item 12). Of the 69 foils, 26 did not meet the Alberta Education criterion that at least 5% of examinees select the incorrect option. The point-biserial correlations for the incorrect alternatives were all less than zero and varied from **-0.06** (item 13) to **-0.33** (item 65).

Taken together, the difference between the SS-NTW-S1 and the SS-ID1-S1 in terms of their item difficulty values was small and can be considered non-significant. The mean *p*-values for these subtests were respectively, 0.65 and **0.68**. The distribution of incorrect responses across item foils was more proportional for the SS-NTW-S1 than for the SS-ID1-S1 items. In the subtest of 35 items non-susceptible to testwiseness, six (5.71%) out of 105 incorrect alternatives were found to have *p*-values lower than 0.05 while in the subtest of items susceptible to the ID1 testwiseness strategy, **26** (**37.68%**) of 69 foils were found to have *p*-values less than 0.05. The distribution of item discrimination indices was approximately the same for both subtests. On each subtest two items yielded the corrected point biserial lower than 0.20 that is lower than standard set by Alberta Education. Lastly, for both subtests, the point-biserial correlations for the foils were all less than zero.

Item Response Item Analysis

Assumptions of Item Response Theory

As pointed out in the method section, the use of the item response theory is based on five assumptions: unidimensionality, item independence, equal item discrimination, non-speededness, and non-guessing. The results of the test of each of these assumptions are provided below.

Unidimensionality. Principal component factor analysis yielded five components with eigenvalues greater than 1.0 for the SS-NTW-S1. The eigenvalue for the first component, 5.41, was 4.36 times greater than the eigenvalue of the second component, 1.24. Further, the successive differences between remaining components were small (0.15, 0.04, and 0.05). The scree plot displayed in Figure 5 confirms the dominance of the first principal component. The value of Stout's T statistic, computed using the automatic execution procedure was 1.12 yielding a *p*-value less than 0.13. The obtained *p*-value suggests that the hypothesis of essential unidimensionality is tenable for the SS-NTW-S1.

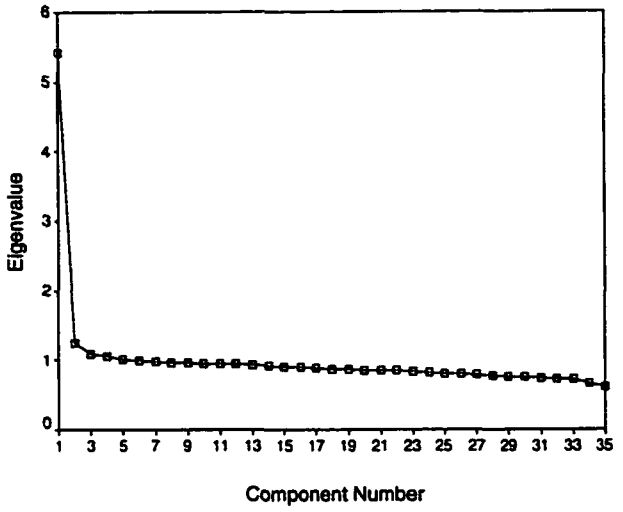


Figure 5. Scree Plot for the SS-NTW-S1

Similarly, five principal components with eigenvalues greater than 1.0 were extracted for the SS-ID1-S1. The eigenvalue for the first component, **3.71**, was 3.40 times greater than the eigenvalue for the second component, **1.09**. The successive differences between the remaining eigenvalues were small (**0.04**, **0.01**, and **0.05**). The shape of scree plot confirms these results (see Figure 6). Stout's T statistic was **1.17**; the associated *p*-value was less than **0.12**. These results indicate essential unidimensionality of the SS-ID1-S1.

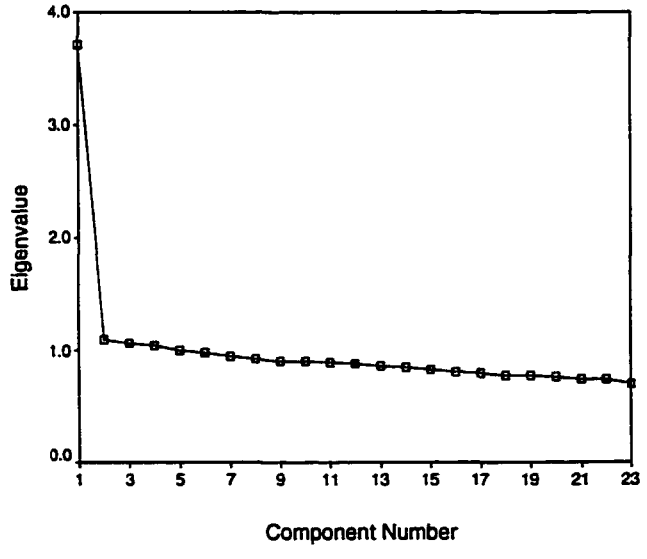


Figure 6. Scree Plot for the SS-ID1-S1

Taken together, the results of principal component analysis, the scree plots, and the Stout's T statistics suggest that there was a dominant component underlying examinees responses to the set of items contained in the SS-NTW-S1 and SS-ID1-S1 (Cattell, 1952; Gorsuch, 1983; Nandakumar, 1993).

Local independence. Given that the assumption of essential unidimensionality was met for both subtests, the assumption of essential item independence in both cases was tenable (Hambleton, et al., 1991; Lord, 1980).

Equal discrimination. The differences between the lowest and the highest values of item point-biserial correlations were examined for both subtests. As reported in the classical item analysis section, these values ranged from 0.20 to 0.51 for the SS-NTW-S1 and from **0.24** to **0.53** for the SS-ID1-S1. Due to large differences between these values, it was concluded that the assumption of equal discrimination indices was not met for both subtests (Hambleton & Murray, 1983). Therefore, the one-parameter item response model, which partly relies on this assumption, may not be appropriate for the SS-NTW-S1 or the SS-ID1-S1.

Non-speededness. The percentages of individuals who did not complete the last three items on the SS-NTW-S1 and the SS-ID1-S1 were calculated. In both cases, out of 4000 examinees, only two students did not complete one of the last three questions. Therefore, it was concluded that the assumption of non-speededness was tenable for both subtests (Hambleton, et al., 1991).

Non-guessing. The performance of low-achieving examinees on the three most difficult items on each subtest was examined to test the tenability of the assumption of non-guessing. In this study, the low achieving examinees were defined as those whose number right score on the full Social Studies test of 70 items fell at or below 23. Out of 4000 examinees, 123 (3.07%) students obtained a score of 23 or lower.

The three most difficult items in the SS-NTW-S1 were: item 5 ($p = 0.46$), item 33 ($p = 0.36$), and item 49 ($p = 0.30$). Examination of the performance of the low-achieving students on these items revealed that 66 (53.65%) did not answer any of the three items correctly, 48 (39.02%) examinees correctly answered one of these questions, 8 (6.50%) examinees provided the correct answer to two of the questions, and only one (less than 1%) examinee correctly answered all 3 questions.

The three most difficult items in the SS-ID1-S1 were: item 12 ($p = \mathbf{0.40}$), item 16 ($p = \mathbf{0.57}$), and item 34 ($p = \mathbf{0.54}$). Examination of the performance of low achievers on these items revealed that **58 (47.15%)** did not answer any of the three items correctly, **57 (46.34%)** correctly answered one question, **8 (6.50%)** examinees correctly answered two questions, and no examinee correctly answered all 3 questions.

Given these findings, it was concluded that the effect of guessing was minimal. Additional evidence supporting this conclusion is provided by the fact that only 35 (0.9%) out of 4000 examinees scored at or below the chance level (score of 17).

Summary. The results presented above indicate that with the exception of the assumption of equal item discrimination, the assumptions of item response theory were tenable for both subtests. The SS-NTW-S1 and the SS-ID1-S1 were found to be essentially unidimensional. Consequently, local independence was obtained for both subtests. In both cases it was concluded that speed was not a factor affecting students' test performance and that the effect of guessing was minimal. Taken together, the results of testing item response theory model assumptions permit the use of the two-, three-

parameter, and nominal response models. The results of item and ability estimation using the one-parameter models should be interpreted with caution due to the failure to meet the assumption of equal discrimination.

Item Parameter Estimation

The SS-NTW-S1 and SS-ID1-S1 items were calibrated separately using the MULTILOG program (Thissen, 1991). The full sample of 4000 examinees was used for item calibration. Marginal maximum likelihood estimation was employed to estimate item parameters within each IRT model. The results of the binary models item analysis are presented in Table 22 for SS-NTW-S1 and in Table 23 for the SS-ID1-S1 of Appendix D. The results of the nominal response model item analysis are reported in Table 24 for SS-NTW-S1 and in Table 25 for the SS-ID1-S1

One-parameter model. The values of item difficulty parameter for the SS-NTW-S1 items ranged from –2.44 (item 28) to 1.14 (item 49), with a mean difficulty of -0.89 and a standard deviation of 0.79. The difficulty estimates for the SS-ID1-S1 items varied from **–2.16** (items 11, 24 and 37) to **0.55** (item 12), with a mean difficulty of **–1.05** and standard deviation of **0.73**. The results reveal that, according to the one-parameter model, the items on the SS-NTW-S1 were on average more difficult than the items contained in the SS-ID1-S1.

Two-parameter model. The values of item difficulty parameter for the SS-NTW-S1 items varied from –2.03 (item 28) to 2.53 (item 49) yielding a mean of -0.80 and a standard deviation of 0.89. The estimates of item discrimination ranged from 0.35 (item 49) to 1.34 (item 27) with a mean of 0.92 and a standard deviation of 0.25.

For the SS-ID1-S1, the difficulty estimates ranged from **–2.92** (item 11) to **0.78** (item 12), with a mean of **-1.06** and a standard deviation of **0.81**.The values of the item discrimination estimates varied from **0.31** (item 38) to **1.47** (item 65) yielding a mean of **0.89** and a standard deviation of **0.28**.

Taken together, the mean difficulty of the SS-NTW-S1 items was higher than the mean of the SS-ID1-S1 items. The item discrimination estimates were comparable for both subtests.

Three-parameter model. For the SS-NTW-S1, the values of item difficulty parameter varied from –1.71 (item 28) to 2.07 (item 49) with a mean of -0.24 and a standard deviation of 0.78. The values of the item discrimination estimates ranged from 0.30 (item 51) to 1.14 (item 4) yielding a mean of 0.72 and a standard deviation of 0.20. Lastly, the values of pseudo-guessing parameter estimates ranged from 0.14 (items 5 and 52) to 0.35 (item 20) producing a mean of 0.23 and a standard deviation of 0.06.

For the SS-ID1-S1, the values of item difficulty estimates varied from **–2.04** (item 11) to **1.30** (item 12) with a mean of **-0.44** and a standard deviation of **0.80**. The item discrimination parameter estimates ranged from **0.21** (item 38) to **1.15** (item 65) yielding a mean of **0.67** and a standard deviation of **0.21**. Finally, the values of pseudo-guessing parameter ranged from **0.13** (item 23) to **0.35** (item 36) with a mean of **0.23** and a standard deviation of **0.06**.

Looking at the three-parameter model, the SS-NTW items were on average more difficult that the SS-ID1 items. Both subtests were found to have comparable mean discrimination parameter estimates, and, since the prior normal distribution of the

pseudo-guessing parameter was specified for each subtest, both subtests had similar mean pseudo-guessing estimates.

Nominal response model. Similar to the classical item analysis and unlike the one-, two-, and three-parameter models, the nominal response model provides information about incorrect response options. For the SS-NTW-S1, the values of the $c_{ik\text{-}correct}$ parameter estimates for the correct answer ranged from 0.27 (item 49) to 2.65 (item 28) with a mean of 1.52 and a standard deviation of 0.52. The values of the popularity parameter for the 105 foils across all items varied from –1.3 (item 19) to 0.50 (item 49) with a mean $c_{ik\text{-}incorrect}$ value of -0.51 and a standard deviation of 0.45. The values of item discrimination estimates for the correct response varied from 0.30 (item 49) to 1.15 (item 27) with a mean of 0.74 and standard deviation of 0.19. The values of the $a_{ik\text{-}incorrect}$ parameter for the incorrect alternatives for the 35 items ranged from -0.87 (item 45) to 0.25 (item 45) yielding a mean of -0.25 and a standard deviation of 0.23.

For the SS-ID1-S1, the values of the $c_{ik\text{-}correct}$ parameter for the correct response varied from **1.09** (item 34) to **2.55** (item 11) with a mean of **1.78** and a standard deviation of **0.46**. The values of the popularity parameter for the 69 incorrect alternatives varied from **–2.38** (item 11) to **1.51** (item 12) yielding a mean of **-0.59** and a standard deviation of **0.86**. The values of item discrimination parameter for the correct option ($a_{ik\text{-}correct}$) ranged from **0.35** (item 38) to **1.13** (item 65) yielding a mean of **0.76** and a standard deviation of **0.19**. The values of the $a_{ik\text{-}incorrect}$ parameter across all items ranged from **-0.79** (item 60) to **0.23** (item 12) with a mean of **-0.26** and a standard deviation of **0.24**.

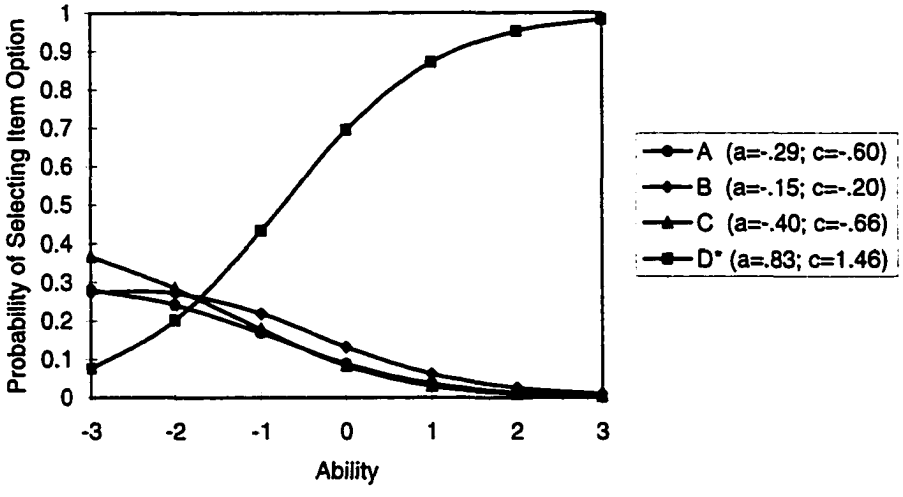
As indicated by the values of the popularity parameter, the correct options in the SS-NTW-S1 were selected on average by fewer examinees than the correct options in the SS-ID1-S1, suggesting that the SS-NTW-S1 contained more difficult items than the SS-ID1-S1 subtest. In addition, the range of the popularity parameter estimates for incorrect item alternatives indicate more equal distributions of incorrect responses across the foils for the SS-NTW-S1 than for the SS-ID1-S1. The mean discrimination parameter estimates for the correct and incorrect options were similar across both subtests.

Item option trace lines. The item option parameters obtained from the nominal response models were used to calculate the probabilities of selecting each option across different ability levels. These values, reported in Table 26 for the SS-NTW-S1 and Table 27 for the SS-ID1-S1 in Appendix D, were subsequently used to produce item option characteristic curves (trace lines).

Inspection of item option characteristic curves revealed that for both SS-NTW-S1 and SS-ID1-S1 items the probability of selecting the correct option increased as examinee ability increased. However, behavior of incorrect options was found to be different for the two types of items. For the SS-NTW-S1 items, the values of $c_{ik\text{-}incorrect}$ parameter indicated a reasonably equal distribution of incorrect responses across item foils. The corresponding trace lines suggested that the incorrect alternatives had similar probabilities of being selected across different ability levels. Two examples of items not susceptible to testwiseness are presented in Figures 7 and 8 (Alberta Education, 1999e). As shown in Figure 7, incorrect options A, B, and C of Item 17 became more attractive to examinees at the lower and middle ability levels (θ in the neighborhood or below 0.0). The trace lines for Item 59 (see Figure 8) indicate similar behavior for incorrect options A and D. The trace lines for these two alternatives indicate uniform, decreasing, probabilities of selecting these distractors as examinees' ability increases. The trace line

for the alternative C in Item 59 (see Figure 8) indicates a moderate probability of selecting this option across ability levels, with the greatest probability of selecting C by middle ability examinees.

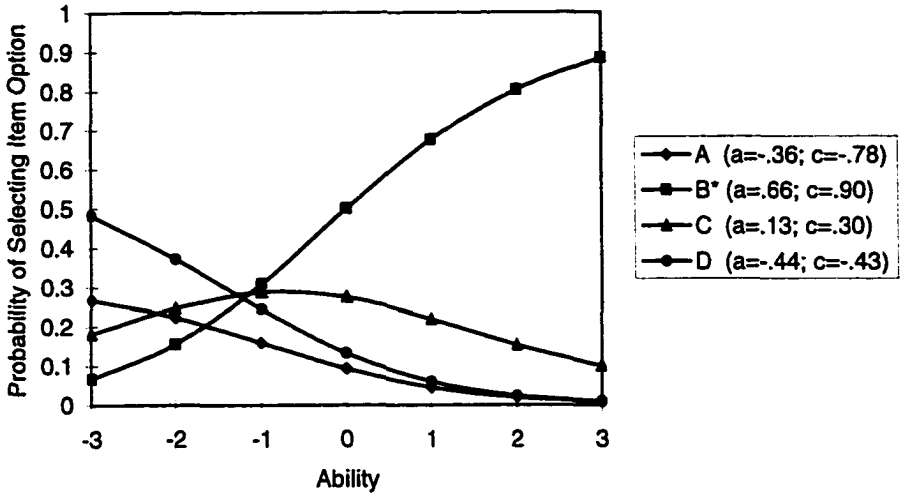
17. Historically, the primary reason for establishing Crown corporations in Canada was to:
    A. encourage increased entrepreneurship in the private sector
    B. attract entrepreneurs from the private sector to the public sector
    C. decrease the size of government bureaucracies through decentralization
    D. provide services and products generally unavailable from the private section



Note: D* indicates the correct option; A, B, and C are item foils.

Figure 7. Non-Susceptible to Testwiseness Item 17, Social Studies 30, Sample 1.

59. With the ending of the Cold War, which of the following terms best describe the issues facing nations regarding the future of nuclear weapons capability and technology?
    A. Escalation and brinkmanship
    B. Proliferation and disarmament
    C. Détente and peaceful coexistence
    D. Mutual deterrence and containment

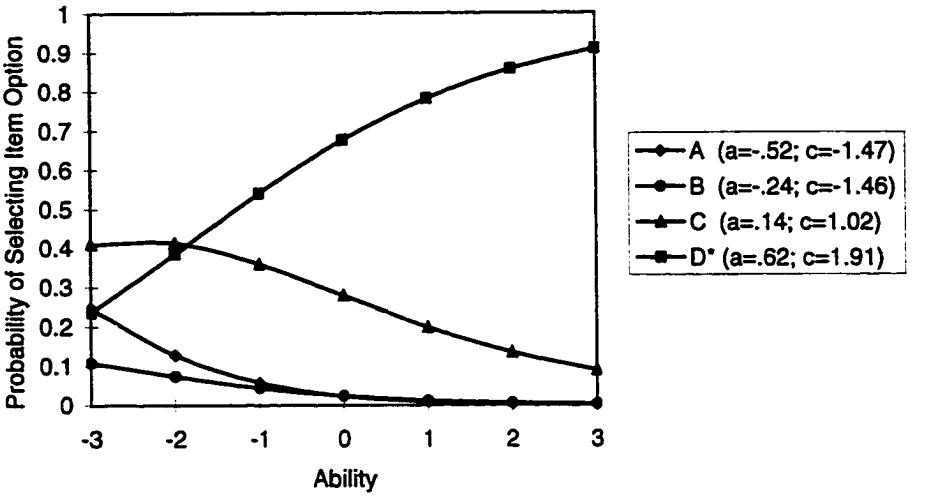


Note: B* indicates the correct option; A, C, and D are item foils.

Figure 8. Non-Susceptible to Testwiseness Item 59, Social Studies 30, Sample 1.

For the SS-ID1-S1 items, the high negative values of the $c_{ik\text{-}incorrect}$ parameters that were found for absurd options indicated their rare selection by examinees. The corresponding item option trace lines indicated that the rarely chosen distractors were either most attractive to lower ability examinees or were disregarded by examinees at all ability levels. Two examples of items susceptible to the ID1 strategy of testwiseness and item option characteristic curves are presented in Figure 9 and Figure 10 (Alberta Education, 1999e). As shown in Figure 9, the probabilities of choosing absurd options A and B of Item 1 approach zero for examinees with ability levels at or higher than one standard deviation below the mean ($\theta \geq -1.0$). The trace lines for item 23 (see Figure 10) reveal that the probability of selecting an absurd option C was close to zero for examinees at all ability levels. The probabilities of selecting non-absurd distractors in items susceptible to the ID1 testwiseness strategy indicate their greater attractiveness across lower and middle ability levels (see Figure 9, Option C; and Figure 10, Options A and B). The probabilities of choosing properly functioning foils were found to be gradually decreasing as examinee ability increased.

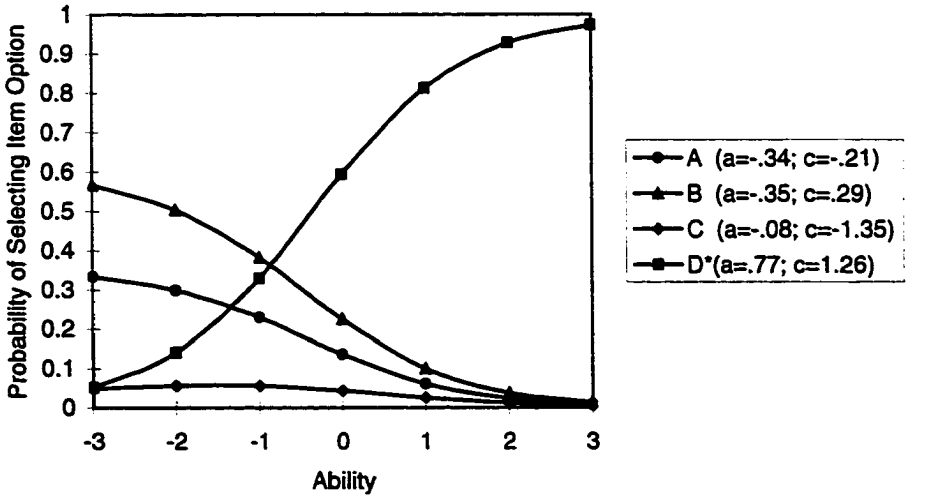
1. In a centrally planned economy, state regulation of supply has the greatest restrictive effect on
   A. social control
   B. class mobility
   C. property ownership
   D. consumer sovereignty



Note: D* indicates the correct answer; A and B are absurd options; C is a properly functioning item foil.

Figure 9. Susceptible to ID1 Testwiseness Strategy Item 1, Social Studies 30, Sample 1

23. In a parliamentary system, the principle of responsible government is most clearly demonstrated by
    A. party discipline
    B. cabinet solidarity
    C. patronage appointments
    D. a vote of non-confidence



Note: D* indicates the correct option; C is an absurd option; A and B are properly functioning foils.

Figure 10. Susceptible to ID1 Testwiseness Strategy Item 23, Social Studies 30, Sample 1

Taken together, the trace lines for the incorrect options of items not susceptible to testwiseness and the trace lines for the incorrect option of items susceptible to the ID1 testwiseness strategy were different. While item trace lines did not reveal clear elimination of any of the incorrect alternatives by the examinees for the items contained in the SS-NTW-S1, the trace lines for alternatives identified as absurd options in the SS-ID1-S1 items were lower than the remaining one or two incorrect options in these items. This observation is congruent with the finding that the values of the c_{ik} parameter for absurd options were considerably lower than the corresponding values for the remaining distractors indicating that the majority of examinees were able to eliminate the absurd options before choosing from among the remaining options. In addition, absurd options were found to have either high discrimination power at the low ability level or essentially no discrimination power. Consequently, it appears that the absurd option(s) either attracted only low proficiency students who did not posses sufficient partial knowledge of the subject content to be able to eliminate these options or were rejected by examinees at all proficiency levels.

Summary of Item Analysis

The classical item analysis and the item response theory item analyses conducted using the one-, two-, three-parameter, and nominal response models provide comparable information about items contained in the SS-NTW-S1 and SS-ID1-S1. The mean item difficulty was found to be higher for the SS-NTW-S1 than for the SS-ID1-S1 across all five models (e.g., p-value of 0.65 and **0.68** in classical item analysis, or b_i value of -0.89 and **-1.05** in one-parameter model, respectively). These differences, however, were small.

The distributions of discrimination indices for items on both subtests were approximately the same (e.g., the values of point biserial correlations varied from .20 to 0.51 for the SS-NTW-S1 and from **0.24** to **0.53** for the SS-ID1-S1; the mean item discrimination parameters were respectively, 0.93 and **0.89** in the two-parameter model for the SS-NTW-S1 and the SS-ID1-S1).

As indicated by the results of the classical and nominal response model item analyses, the SS-NTW-S1 and the SS-ID1-S1 differed in terms of the distribution of wrong responses. The distribution of incorrect responses was found to be more uniform across item foils for the SS-NTW-S1 than for the SS-ID1-S1 (e.g., 5.71% vs. **37.68%** of foils were selected by less than five percent of examinees, respectively). The analysis of item option trace lines supported these findings. In the SS-NTW-S1, the distractors appeared to be fairly equally attractive for examinees at a given proficiency level. In case of the SS-ID1-S1 items, the trace lines indicated either very low probability of selecting an absurd option by majority of students or its attractiveness to low ability examinees only.

Comparison of Ability Estimates

The ability estimates expressed as T-scores ($\mu = 50$, $\sigma = 10$) for each of the scoring methods for each subtest were separately compared within the high, middle, and low ability sub-samples. The results of these comparisons are presented in Tables 5, 6, and 7, respectively. The first two rows of each table contain the means and standard deviations of the five sets of ability estimates. The correlations between pairs of estimates

are reported in the upper triangle while the root mean square deviations between pairs of estimates are shown in the lower triangle. As explained in the previous chapter, correlations less than 0.95 were considered to denote systematic differences in examinees ranking by different scoring models. Values of the root mean square deviations that were greater than 2.0 were considered to be indicators of systematic differences between pairs of ability estimates yielded by different scoring models.

High Ability Group

As shown in Table 5, the means and standard deviations of transformed ability scores yielded by the five scoring methods were approximately the same for both the SS-NTW-S1 and the SS-ID1-S1. The means for the SS-NTW-S1 ranged from 61.88 to 62.72, and for the SS-ID1-S1 from **61.10** to **61.72**. The corresponding standard deviations varied from 3.56 to 5.01 for the SS-NTW and from **3.96** to **5.15** for the SS-ID1-S1, indicating comparable variability of the scores yielded by the five models across two sets of items.

Table 5
Comparison of Ability Estimates Obtained from the SS-NTW-S1 and SS-ID1-S1 in the High Ability Group (N=988)

	θ_{nr} (number right)	θ_{1} (1PL)	θ_{2} (2PL)	θ_{3} (3PL)	θ_{4} (NRM)
Means and Standard Deviations	61.88 (3.56)	62.54 (4.94)	62.70 (5.01)	62.39 (4.97)	62.72 (4.95)
	61.10 (3.96)	**61.67 (5.08)**	**61.77 (5.11)**	**61.68 (5.15)**	**61.72 (5.04)**
θ_{nr} (number right)		0.99	0.97	0.96	0.94
		0.99	**0.96**	**0.95**	**0.94**
θ_{1} (1PL)	1.60		0.97	0.97	0.94
	1.33		**0.96**	**0.96**	**0.94**
θ_{2} (2PL)	2.00	1.27		0.99	0.98
	1.85	**1.41**		**0.99**	**0.98**
θ_{3} (3PL)	1.92	1.26	0.77		0.97
	1.94	**1.52**	**0.55**		**0.98**
θ_{4} (NRM)	2.18	1.69	1.07	1.30	
	2.04	**1.79**	**0.97**	**1.10**	

Note: Values in regular print are for the SS-NTW-S1 subtest and in **bold** print for the SS-ID1-S1 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

The correlations among the estimates yielded by the five scoring models were essentially unchanged for the SS-ID1-S1 as compared to the SS-NTW-S1. With the exception of correlations between scores produced by the number right and nominal response models, and by the one-parameter and nominal models, the correlations between the remaining pairs of ability estimates were equal or greater than 0.95. For both subtests, the lowest correlation of 0.94 was observed for the pairs of ability estimates produced by the nominal response model and the number right or one-parameter models. Taken together, although there are slight differences among the correlations obtained in both subtests, the students' ranked position will in effect be the same across the five scoring methods, with the possible exception of some differences in rankings that may occur when number right or one-parameter scores are used in a place of nominal response model scores.

The absolute agreements between pairs of estimates were also found to be fairly similar across the SS-NTW-S1 and SS-ID1-S1. The closest agreements for both subtests were found between the scores produced by the two- and three parameter models, and by the two-parameter and nominal models. The values of the root mean square deviations for these pairs were, respectively, 0.77 and 1.07 for the SS-NTW-S1 and **0.55** and **0.97** for the SS-ID1-S1. There was slightly less agreement between the pairs of scores yielded by the one-parameter model and two-, three-parameter, and nominal models, and by the three-parameter and nominal response models. The root mean square deviations for these comparisons were found to be, respectively, 1.27, 1.26, 1.69, and 1.30 for the SS-NTW-S1 and **1.41**, **1.52**, **1.79**, and **1.10**. Turning to the number right model, the root mean square deviation values for the pairs of transformed ability scores yielded by the number right and the four IRT models were larger, varying from 1.60 to 2.18 for the SS-NTW-S1 and from **1.33** to **2.04** for the SS-ID1-S1. In both cases, the largest root mean square deviations, 2.18 and **2.04,** were found for the pairs of scores yielded by the number right and nominal models.

Therefore, it seems that for the high ability group, only the number right model will likely produce different ability estimates when used in place of the nominal model. Since this finding was consistent for both subtests, it also appears that, for high ability examinees, no significant differences in ability estimates occur when a test contains items susceptible to the ID1 testwiseness strategy as compared to a test without such items.

Middle Ability Group

As shown in Table 6, the means of transformed ability scores yielded by the five scoring methods for the SS-NTW-S1 for the middle proficiency group were essentially the same as their counterparts for the SS-ID1-S1. The means for the SS-NTW-S1 ranged from 50.06 to 51.10 and for the SS-ID1-S1 from **50.42** to **51.04**. The standard deviations of the scores yielded by the five scoring models were slightly greater for the SS-ID1-S1 subtest than for the SS-NTW-S1 subtest (3.77 vs. **4.92**, 3.63 vs. **4.89**, 3.70 vs. **4.82**, 3.59 vs. **4.75**, and 3.77 vs. **4.80**). The differences in each case were slightly greater than one point on the T-score metric.

The correlations among pairs of scores yielded by the five scoring methods were found to be consistent across the SS-NTW-S1 and SS-ID1-S1. For both subtests, the correlations between the number right and one-parameter model scores, and between the two-, three-parameter, and nominal model scores were equal or greater than 0.95. The

correlations between scores produced by the three-parameter model and the number right or one-parameter models were found to be 0.94 for both subtests. The lowest correlations, 0.91 for the SS-NTW-S1 and **0.92** for SS-ID1-S1 were observed for the pairs of ability estimates yielded by the nominal response model and the number right or one-parameter models. Given these findings, it appears that there will be some differences among the students' ranked position across the five scoring methods, particularly when the number right or one-parameter model scores are used in a place of the nominal response model scores.

Table 6
Comparison of Ability Estimates Obtained from the SS-NTW-S1 and the SS-ID1-S1 in the Middle Ability Group (N=1141)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and	51.10	50.31	50.12	50.43	50.06
Standard	(3.77)	(3.63)	(3.70)	(3.59)	(3.77)
Deviations	**51.04**	**50.44**	**50.42**	**50.49**	**50.48**
	(4.92)	**(4.89)**	**(4.82)**	**(4.75)**	**(4.80)**
θ_{nr} (number right)		0.99	0.95	0.94	0.91
		0.99	**0.95**	**0.94**	**0.92**
θ_1 (1PL)	0.81		0.95	0.94	0.91
	0.69		**0.95**	**0.94**	**0.92**
θ_2 (2PL)	1.51	1.13		0.99	0.96
	1.67	**1.51**		**0.99**	**0.97**
θ_3 (3PL)	1.45	1.26	0.45		0.95
	1.76	**1.64**	**0.38**		**0.97**
θ_4 (NRM)	1.93	1.62	1.10	1.18	
	2.04	**1.93**	**1.20**	**1.23**	

Note: Values in regular print are for the SS-NTW-S1 subtest and in **bold** print for the SS-ID1-S1 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

The closest agreements between the transformed ability estimates were found for the scores produced by the number right and one-parameter models, and by the two- and three-parameter model scores for both subtests. The values of the root mean square deviations for these comparisons were 0.81 and 0.45 for the SS-NTW-S1 and **0.69** and **0.38** for the SS-ID1-S1. There was less agreement between the scores yielded by the one- and two-parameter models, and by the nominal response model and two- and three-parameter model scores. The root mean square deviations for these three pairs varied

from 1.10 to 1.18 for the SS-NTW-S1 and from **1.20** to **1.51** for the SS-ID1-S1. The differences between ability estimates were somewhat greater for the remaining pairs of scores. The root mean square deviations for pairs of ability estimates obtained from the number right and the two-parameter, three-parameter, or nominal models, and from the one-parameter and three-parameter or nominal models ranged from 1.26 to 1.93 for the SS-NTW-S1 and from **1.64** to **2.04** for the SS-ID1-S1.

When all of these results are considered, with the exception of the comparison between ability estimates yielded by the number right and one-parameter model, the root mean square deviations were greater for the comparison in which the number right and the one-parameter scores were involved than for the comparisons involving the two-, three-parameter, and nominal response models for both the SS-NTW-S1 and SS-ID1-S1. However, with the exception of differences between the number right and nominal response model scores for the SS-ID1-S1 (root mean square deviation **2.04**), differences between other pairs of scores were relatively small (root mean square deviations less than 2.00). Thus, it seems that, for middle ability students, only replacement of number right ability scores with the nominal model scores will likely result in significant differences.

Comparison of the SS-NTW-S1 and SS-ID1-S1 revealed that, the lack of agreement between scores yielded by the number right and other IRT models and by the one-parameter and other IRT models were slightly greater for the SS-ID1-S1 than for the SS-NTW-S1 subtest. The root mean square deviations increased from 1.51, 1.45, and 1.93 for the SS-NTW-S1 to **1.67**, **1.76**, and **2.04** for the SS-ID1-S1 for the number right and two-, three-parameter, and nominal model scores comparisons. Likewise, for the one-parameter model and two-, three-parameter, and nominal model scores comparisons, the root mean square deviations increased from 1.13, 1.26, and 1.62 for the SS-NTW-S1 to **1.51**, **1.64**, and **1.93** for the SS-ID1-S1, respectively. However, since these differences between the two subtests were small it appears that, for middle ability examinees, no significant differences in ability estimates occur when a test contains items susceptible to the ID1 testwiseness strategy as compared to a test without such items.

Low Ability Group

As shown in Table 7, the means of the transformed ability scores yielded by the five scoring methods were consistent for both subtests for the low ability group. The means ranged from 37.63 to 38.38 for the SS-NTW-S1 and from **38.20** to **38.62** for the SS-ID1-S1. With the exception of the three-parameter model scores, the standard deviations were on average one point lower for the SS-NTW-S1 than for the SS-ID1-S1 (5.49 vs. **6.62**, 4.85 vs. **5.88**, 4.69 vs. **5.61**, and 4.58 vs. **5.59**).

The correlations among pairs of ability estimates yielded by the five scoring methods were essentially the same across the two subtests. With the exception of correlations between scores produced by the number right and nominal response models, and by the one-parameter and nominal models, the correlations between the remaining pairs of ability estimates were equal or greater than 0.95. The lowest correlations, 0.94, were observed for the pairs of ability estimates produced by nominal response model and number right or one-parameter model on both the SS-NTW and SS-ID1-S1.

Taken together, the correlations among the estimates yielded by the five scoring models suggest that low ability examinees' ranked position would be similar across these

models. For both subtests there may be some differences in rankings when nominal model scores is used in a place of the number right or one-parameter model scores.

Table 7
Comparison of Ability Estimates Obtained from the SS-NTW-S1 and SS-ID1-S1 in the Low Ability Group (N=1086)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and	37.63	38.23	38.36	38.18	38.38
Standard	(5.49)	(4.85)	(4.69)	(5.24)	(4.58)
Deviations	**38.20**	**38.62**	**38.54**	**38.50**	**38.46**
	(6.62)	**(5.88)**	**(5.61)**	**(5.81)**	**(5.59)**
θ_{nr} (number right)		1.00	0.98	0.97	0.94
		1.00	**0.97**	**0.96**	**0.94**
θ_1 (1PL)	0.89		0.98	0.97	0.94
	0.86		**0.97**	**0.96**	**0.94**
θ_2 (2PL)	1.51	0.99		0.98	0.96
	1.83	**1.43**		**0.99**	**0.96**
θ_3 (3PL)	1.53	1.39	1.06		0.95
	1.98	**1.69**	**0.87**		**0.96**
θ_4 (NRM)	2.06	1.62	1.26	1.70	
	2.43	**2.08**	**1.52**	**1.69**	

Note: Values in regular print are for the SS-NTW-S1 subtest and in **bold** print for the SS-ID1-S1 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

For both subtests, the closest agreements were found between the scores produced by the number right and one-parameter models, and by the two- and three parameter models. The values of the root mean square deviations for these comparisons were 0.89 and 1.07 for the SS-NTW-S1 and **0.86** and **0.87** for the SS-ID1-S1, respectively. Also for both subtests, the greatest differences occurred when the number right model was used as an alternative to the nominal response model (e.g., root mean square deviations were 2.06 for the SS-NTW-S1 and **2.43** for the SS-ID1-S1). The difference between ability estimates were also found for the pairs of scores produced by one-parameter and nominal model in the case of SS-ID1-S1 (root mean square deviation of **2.08**). The differences in ability estimates yielded by other scoring models were smaller with root mean square deviations less than 2.0 across both subtests.

With the exception of the difference among the scores yielded by the number right and one-parameter models, the root mean square deviations for the pairs in which

the number right or the one parameter model scores were involved were slightly greater for the SS-ID1-S1 than for the SS-NTW-S1. For example, the root mean square deviations for comparisons involving the number right and two- or three-parameter models were 1.51 and 1.53 for the SS-NTW-S1 and **1.83** and **1.98** for the SS-ID1-S1, the root mean square deviations for the comparisons between ability scores yielded by the one- and three-parameter models were 1.39 for the SS-NTW-S1 and **1.69** for the SS-ID1-S1, and the root mean square deviations between the nominal model scores and the number right, and one- parameter models scores were 2.06 and 1.62 for the SS-NTW-S1 and **2.43**, **2.08** for the SS-ID1-S1. However, despite consistently larger differences between ability scores in the SS-ID1-S1 as compared to the corresponding values in the SS-NTW-S1, the differences between the two subtests were small. Therefore, it also appears that, for the low ability examinees, no significant differences in ability estimates occur when a test contains items susceptible to the ID1 testwiseness strategy as compared to a test without such items.

Comparison of Groups

While the pattern of correlations was consistent across the SS-NTW-S1 and the SS-ID1-S1 within each ability group, the magnitude of some correlations varied by the type of scoring models and by group. The correlations between the number right and one- and two-parameter model scores, and between the two- and three-parameter and nominal model scores were all relatively high across all groups and both subtests. The values of these correlations varied from 0.95 to 1.00. Correlations for the pairs of scores yielded by the number right and three-parameter models, and by the one- and three-parameter models were found to be lower in the middle ability groups for both subtests. These correlations ranged from 0.95 to 0.97 for the high and low ability groups, and dropped to 0.94 in the middle ability group. The lowest correlations across all ability groups were observed for the pairs of scores obtained from the nominal response model and number right or one-parameter models. Again, the correlations found in the middle ability group were slightly lower that the corresponding correlations in the high and low ability groups (0.91 and **0.92** vs. 0.94). These results suggest, that the greatest changes in students' ranks will likely occur if the number right or one-parameter model scores are used in place of nominal response model scores across all ability levels. Since these findings were consistent for both subtests, it appears that the presence of items susceptible to the ID1 strategy will not affect students' ranking position.

Turning to agreements between scores, although the extent of agreement between the pairs of scores yielded by different scoring models varied, the patterns of root mean square deviations were generally consistent across the three ability groups. The closest agreements for all three ability groups and both subtests were found between the scores yielded by the two- and three-parameter models. The root mean square deviations for these comparisons varied from 0.38 to 1.06. Less agreement was found for the pairs of scores produced by the number right and one-parameter models and by the nominal response model and the two-, and three-parameter models. The values of root mean square deviations for these comparisons varied from 0.97 to 1.70 and were approximately the same for the SS-NTW-S1 and SS-ID1-S1.

The greatest discrepancies between ability estimates for all groups and both subtests were found for the pairs of scores involving the number right or one-parameter

models and the two-, three-parameter, and nominal model scores. The root mean square deviations for these comparisons varied from 0.99 to 2.43, with the least agreement for scores produced by the number right and nominal response models (root mean square deviations greater than 2.00). These results suggest that for all students, number right scores will likely produce different ability estimates when used in place of the nominal model scores.

For the middle and low ability groups, with the exception of comparisons between the number right and one-parameter model scores, the differences between pairs of scores yielded by the number right or one-parameter models and other IRT models were consistently greater for the SS-ID1-S1 than for SS-NTW-S1. The values of root mean square deviations for these comparisons ranged from **1.41** (one- and two-parameter model scores, middle ability group) to **2.43** (number right and nominal model scores, low ability group) for the SS-ID1-S1. For the SS-NTW-S1, the corresponding values of root mean square deviations varied from 1.13 (one- and two-parameter model scores, middle ability group) and 2.06 (number right and nominal model scores, low ability group). However, given that the differences between the SS-NTW-S1 and SS-ID1-S1 were small, it appears that there will be no significant discrepancies in ability estimates yielded by the two types of items.

Replication Data Analyses: Social Studies - Sample 2

The distribution of the total test scores for the Sample 2 (N = 4000) was similar to that obtained for the whole population of grade 12 students who wrote the Social Studies 30 Diploma Examination in June 1999 and for the Sample 1 (see p. 36). The mean test score for the Sample 2 was 46.82 (66.89%) and the standard deviation was 11.87 (16.96%). The number of students in each ability group were 1,087 high ability examinees, 1,083 middle ability examinees, and 1,112 low ability students.

Classical Item Analysis

For convenience and to differentiate the subtests from those described earlier in the initial data analyses section, the subtest containing non-susceptible items and the subtest comprised of items susceptible to the ID1 testwiseness strategy in the cross-validation sample are labeled SS-NTW-S2 and SS-ID1-S2, respectively. The results of test characteristics for the SS-NTW-S2 and the SS-ID1-S2 are presented in Table 8.

Table 8
Summary of Test Statistics for the SS-NTW-S2 and the SS-ID1-S2

Subtest	Number of Items	Mean Raw Score	Standard Deviation	Reliability And SEM
SS-NTW-S2	35	22.76 (65.03%)	6.22 (17.77%)	0.83 (2.56)
SS-ID1-S2	**23**	**15.50 (67.39%)**	**4.15 (18.04%)**	**0.76 (2.03)**

Note: SEM refers to the standard error of measurement.

As shown in Table 8, mean number right scores were 22.76 for the SS-NTW-S2 and **15.50** for the SS-ID1-S2 with the corresponding standard deviations of 6.22 and **4.15**, respectively. The internal consistency index (Cronbach's α) was 0.83 for the SS-NTW-S2 and **0.76** for the SS-ID1-S2. Comparison of these results with those reported in Table 4 for the initial sample, S1, revealed that the initial and replication samples were essentially the same. Likewise, the item characteristics for the SS-NTW-S2 and the SS-ID1-S2 subtests were very similar to those obtained from the initial study. The difficulty index (*p*-value), point-biserial (r_{pb}), and corrected point-biserial correlations (r_{cpb}) for each item are presented in Table 28 for SS-NTW-S2 and Table 29 for SS-ID1-S2 in Appendix D. The mean difficulty of the SS-NTW-S2 and SS-ID1-S2 subtests were 0.65 and **0.67**, with standard deviations of 0.13 and **0.11**, respectively. The distribution of item discrimination indices was approximately the same for both subtests, with the point biserial correlations ranging from 0.21 to 0.52 for the SS-NTW-S2 and from **0.27** to **0.53** for the SS-ID1-S2. Two items on each subtest yielded a corrected point-biserial correlation lower then 0.20. Similar to the results for the initial sample, the distribution of incorrect responses across item foils was more proportional for the SS-NTW-S2 than for the SS-ID1-S2. Across all items, 5.71% of incorrect alternatives in the SS-NTW-S2 and 36.23 % of incorrect options in the SS-ID1-S2 were selected by fewer than 5 percent of examinees.

Item Response Item Analysis

The results of the IRT assumption tests were comparable to those obtained in the initial study for both the SS-NTW-S2 and SS-ID1-S2. Both subtests were found to be essentially unidimensional. Consequently, in both subtests the assumption of local independence was tenable. The large ranges of the item point-biserial correlations indicated that the assumption of equal discrimination indices was not met for both subtests suggesting that the use of one-parameter item response model may not be appropriate. The assumption of non-speededness was tenable for both subtests. Lastly, the effect of guessing was found to be minimal for both subtests. The results of the assumption tests are presented in Tables 30 through 33 and Figures 17 and 18 of Appendix D.

The SS-NTW-S2 and SS-ID1-S2 items were calibrated separately using MULTILOG program (Thissen, 1991). The results of item analyses using the four IRT models are presented in Appendix D. Tables 34 and 35 contain the results for the one-, two-, and three-parameter models and Tables 36 and 37 – the results for the nominal response model, for the SS-NTW-S2 and SS-ID1-S2, respectively. The values of the probabilities of selecting each item option across ability levels are reported in Table 38 for the SS-NTW-S2 and Table 39 for the SS-ID1-S2. Since the obtained item parameter estimates for both subtests in Sample 2 were comparable to their counterparts yielded by the four IRT models in Sample 1, only a brief summary of these results is presented here.

In one-parameter model the mean item difficulty estimates were found be -0.87 for the SS-NTW-S2 and **–1.03** for the SS-ID1-S2. In the two-parameter model, the mean difficulty parameters for these subtests were -0.80 and **–1.04**, respectively. Both subtests yielded comparable mean discrimination parameters, 0.94 and **0.81**, respectively. Looking at the three-parameter model, the mean difficulty parameter estimates were -0.19 and **-0.37**, and the mean discrimination parameter estimates were 0.76 and **0.71** for the SS-NTW-S2 and SS-ID1-S2, respectively. The mean pseudo-guessing parameter was 0.24 for both subtests. Turning to the nominal response model, the mean values of the correct option popularity parameter estimates were 1.52 and **1.79**, and the mean discrimination parameter estimates for the correct option were 0.75 and **0.79** for the SS-NTW-S2 and SS-ID1-S2, respectively. Similar to the initial study findings, the values of popularity parameter estimates for incorrect item alternatives indicate more equal distributions of incorrect responses across foils for the SS-NTW-S2 ($c_{ik\text{-}incorrect}$ ranged from 0.50 to –1.70) than for the SS-ID1-S2 ($c_{ik\text{-}incorrect}$ ranged from **1.50** to **–2.68**).

Further, analysis of the SS-NTW-S2 and SS-ID1-S2 item option trace lines indicated their similarity to item option trace lines obtained in S1 for both subtests. In the SS-NTW-S2 items, the distractors seemed to be approximately equally attractive to the lower and middle ability students. In case of the SS-ID1-S2 items, the item trace lines indicated that the majority of test-takers were able to eliminate one or more incorrect options (see Figures 19 through, 22 in Appendix D).

Comparison of Ability Estimates

Following the sequence of procedures employed to analyze data from Sample 1, the comparisons of transformed ability scores were conducted simultaneously for the SS-NTW-S2 and for the SS-ID1-S2. The results of these comparisons for the high, middle, and low ability groups are reported in Tables 9, 10 and 11, respectively.

Table 9
Comparison of Ability Estimates Obtained from the SS-NTW-S2 and SS-ID1-S2 in the High Ability Group (N=1087)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and	61.60	62.15	62.27	61.93	62.28
Standard	(3.68)	(5.06)	(5.13)	(5.08)	(5.13)
Deviations	**60.90**	**61.42**	**61.56**	**61.41**	**61.54**
	(4.22)	**(5.41)**	**(5.35)**	**(5.35)**	**(5.24)**
θ_{nr} (number right)		0.99	0.97	0.96	0.95
		0.99	**0.96**	**0.95**	**0.94**
θ_1 (1PL)	1.57		0.97	0.97	0.95
	1.37		**0.96**	**0.96**	**0.94**
θ_2 (2PL)	1.94	1.23		0.99	0.98
	1.82	**1.41**		**0.99**	**0.98**
θ_3 (3PL)	1.92	1.30	0.98		0.96
	1.89	**1.54**	**0.64**		**0.98**
θ_4 (NRM)	2.13	1.60	1.02	1.41	
	1.96	**1.75**	**0.95**	**1.08**	

Note: Values in regular print are for the SS-NTW-S2 subtest and in **bold** print for the SS-ID1-S2 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

Table 10

Comparison of Ability Estimates Obtained from the SS-NTW-S2 and the SS-ID1-S2 in the Middle Ability Group (N=1083)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and Standard Deviations	50.79 (3.50) **50.89 (4.47)**	50.00 (3.34) **50.22 (4.38)**	49.83 (3.40) **50.09 (4.36)**	50.20 (3.28) **50.25 (4.28)**	49.80 (3.48) **50.15 (4.35)**
θ_{nr} (number right)		0.99 **0.99**	0.95 **0.95**	0.93 **0.94**	0.89 **0.91**
θ_1 (1PL)	0.82 **0.73**		0.95 **0.95**	0.93 **0.94**	0.89 **0.91**
θ_2 (2PL)	1.48 **1.67**	1.09 **1.42**		0.99 **0.99**	0.95 **0.96**
θ_3 (3PL)	1.45 **1.70**	1.29 **1.53**	0.55 **0.42**		0.94 **0.96**
θ_4 (NRM)	1.92 **2.06**	1.61 **1.86**	1.13 **1.21**	1.25 **1.27**	

Note: Values in regular print are for the SS-NTW-S2 subtest and in **bold** print for the SS-ID1-S2 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

Table 11
Comparison of Ability Estimates Obtained from the SS-NTW-S2 and SS-ID1-S2, Low Ability Group (N=1112)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and	37.72	38.33	38.44	38.25	38.47
Standard	(5.37)	(4.74)	(4.55)	(5.20)	(4.49)
Deviations	**38.31**	**38.78**	**38.77**	**38.68**	**38.66**
	(6.45)	**(5.69)**	**(5.48)**	**(5.72)**	**(5.49)**
θ_{nr} (number right)		1.00	0.97	0.95	0.94
		1.00	**0.97**	**0.95**	**0.94**
θ_1 (1PL)	0.88		0.98	0.95	0.94
	0.89		**0.97**	**0.95**	**0.94**
θ_2 (2PL)	1.56	1.07		0.98	0.96
	1.75	**1.31**		**0.98**	**0.96**
θ_3 (3PL)	1.84	1.71	1.23		0.94
	2.05	**1.78**	**1.06**		**0.95**
θ_4 (NRM)	2.09	1.66	1.29	1.88	
	2.38	**2.03**	**1.54**	**1.82**	

Note: Values in regular print are for the SS-NTW-S2 subtest and in **bold** print for the SS-ID1-S2 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

Comparison of the results provided in these tables with the results reported in Tables 5, 6, and 7 reveals that the results for the initial and replication samples are consistently the same. Consequently only the set of results for the high ability group in Sample 2 is discussed here, with the comparisons drawn with the results for the high ability group in the initial sample to highlight the stability of these results.

High Ability Group

As shown in Table 9, the means, standard deviations, and the patterns and magnitudes of the correlations and the root mean square deviations among pairs of ability scores are very similar to the corresponding results and patterns from the initial sample reported in Table 5.

The means and standard deviations of transformed ability scores were fairly similar across the five scoring models for the two subtests. The means ranged from 61.60 to 62.28 for the SS-NTW-S2 and from **61.10** to **61.77** for the SS-ID1-S2. The corresponding values in the initial study varied from 61.88 to 62.72 for the SS-NTW-S1, and for the SS-ID1-S1 from **61.10** to **61.72**. The corresponding standard deviations varied

from 3.68 to 5.13 for the SS-NTW-S2 (3.56 to 5.01 for the SS-NTW-S1) and from **4.22** to **5.41** for the SS-ID1-S2 (**3.96** to **5.15** for the SS-ID1-S1).

The patterns and values of the correlations were essentially the same for the SS-ID1-S2 and the SS-NTW-S2. With the exception of correlations yielded by the pairs of scores produced by the number right or one-parameter models and the nominal response model in the SS-ID1-S2, all correlations were equal or greater than 0.95 for both subtests. The lowest correlations of **0.94** for the SS-ID1-S2 and 0.95 for the SS-NTW-S2 were observed for pairs of scores yielded by the number right or one-parameter models and nominal response models. Again, these patterns and values were essentially the same as the pattern and values for the initial sample.

Consistent with the findings from the initial sample, it appears that although there are slight differences among the correlations obtained in both subtests, the students' ranked position will be in essentially the same across the five scoring methods, with the exception of some differences in rankings that will likely occur when number right or one-parameter scores are used in a place of the nominal response model scores.

The agreements between pairs of ability estimates were also found to be relatively similar for both subtests. The closest agreements between pairs of transformed ability estimates for the SS-NTW-S2 and SS-ID1-S2 were found for the scores produced by the two- and three parameter models, and by the two-parameter and nominal models. The values of the root mean square deviations for these pairs ranged from **0.64** to 1.02. In the initial study, the values of root mean square deviation for the corresponding comparisons ranged form **0.55** to 1.07.

As for the initial sample, there was less agreement between the pairs of scores yielded by the remaining models for both subtests. The root mean square deviations for these comparisons in the cross-validation sample varied from 1.23 (one- and two-parameter model scores) to 2.13 (number right and nominal model scores) for the SS-NTW-S2 and from **1.08** (three-parameter and nominal model scores) to **1.96** (number right and nominal model scores) for the SS-ID1-S2 subtest. The corresponding values in the initial sample ranged from 1.27 to 2.18 for the SS-NTW-ID1 and from **1.10** to **2.04** for the SS-ID1-S1.

Again, the greatest discrepancies for between ability estimates were found for comparisons in which the number right scores, and to some extent, the one-parameter model scores were involved. However, with the exception of differences in ability estimates yielded by the number right and nominal response models for the SS-NTW-S2 (root mean square deviation 2.13), these differences were relatively small with the values of root mean square deviations varying form 1.23 to 1.92 for the SS-NTW-S2 and from **1.37** to **1.96** for the SS-ID1-S2. Likewise, in the initial study, the root mean square deviations for comparisons of ability scores yielded by the number right and nominal response model were 2.18 for the SS-NTW-S1 and **2.04** for the SS-ID1-S1, while the remaining differences were relatively small with root mean square deviations varying from 1.26 to 2.00 across both subtests.

Thus, it appears that for the high ability group, only number right model may yield different ability estimates when used in place of nominal response model scores. Moreover, it seems that for the high ability examinees no significant differences in ability estimates will take place when a test contains items susceptible to the ID1 strategy of testwiseness. This finding was again consistent for both samples.

Argument for Not Combining Sample 1 and Sample 2

When the results of the initial and replication studies are found to be essentially the same, the analyses are often expected for the combined samples to yield the final set of results based on the larger size sample. It is expected that replication of the data analyses using the pooled sample would yield essentially the same values for the means, standard deviations, correlation coefficients, and root mean square deviations, but with reduced standard errors of these statistics. However, since reasonably large samples were used in both initial and replication studies, combining these samples would not bring significant changes in terms of reducing standard error (Glass & Stanley, 1970). Consequently, the two samples for the social studies examination were not combined.

Discussion

As expected, the mean level of performance decreased with the decreasing ability in both samples. The standard deviations of scores yielded by the five methods were found to be on average one point lower on the T-score metric in the middle ability group as compared to the corresponding values in the high and low ability groups (cf. Table 5 with Table 9, Table 6 with Table 10, and Table 7 with Table 11).

In both subtests, the lowest correlations (0.89 to 0.94) in each ability group were for the pairs of scores obtained from the nominal response model and the number right or one-parameter models. Correlations for the pairs of scores yielded by the three-parameter model and the number right or one- parameter models were found to be lower in the middle ability group than in the high and low ability groups (0.93 and 0.94 vs. 0.95, 0.96, and 0.97). The remaining correlations were relatively high (0.95 to 1.00). The similar values of correlations for both subtests in each ability group suggest that the presence of items susceptible to the ID1testwisenes strategy on a test will not affect students' ranking position.

The patterns of root mean square deviations for three ability groups were fairly consistent with the patterns of correlations and also indicated comparability of both subtests. The largest root mean square deviations were found for scores yielded by the number right and nominal response models for students in all ability groups (root mean square deviations greater than 2.00). The values of root mean square deviations obtained for the remaining pairs of scores were on average smaller. Although the pattern of differences between ability estimates yielded by the number right or one-parameter models and the remaining IRT models were slightly greater for the SS-ID1 than for the SS-NTW (**1.31** to **2.43** vs. 0.99 to 2.09) for the middle and low ability examinees, these differences were relatively small and can be considered non-significant.

The expected differences between items not susceptible to testwiseness and items susceptible to the ID1 testwiseness strategy were smaller than anticipated. Although the pattern of differences between scores yielded by the number right or one-parameter and other ITR models suggested that these differences were slightly greater for the SS-ID1 than the SS-NTW (**1.31** to **2.43** vs. .99 to 2.09) for the middle and low ability examinees, these differences were relatively small and can be considered non-significant.
This finding may be due to the observed small differences between the two subtests in terms of their difficulty. As pointed earlier, although the SS-NTW was found to be more

difficult then the SS-ID1 (e.g., the mean b_i parameters in the one-parameter model were -0.89 and -0.87 for the SS-NTW and **–1.05** and **–1.03** for the SS-ID1), these differences were small. The summary of the mean item parameters across the five scoring models and two types of items for both samples is presented in Table 40 of Appendix D.

It was anticipated that the use of information from incorrect responses would affect ability estimates and result in the differences between the SS-NTW and the SS-ID1. However, despite the fact that differences in the incorrect item option trace lines were found for the SS-NTW as compared to the SS-ID1, it appears that the influence of information from incorrect responses on ability estimates obtained from these subtests is weak. The discrepancies between the pairs of scores yielded by nominal response model and the two- and three-parameter scoring models were small and consistent for both subtests.

CHAPTER V – CHEMISTRY 30

The results of the analyses conducted for the chemistry examination are presented in the present chapter. Like the previous chapter, this chapter is organized in two major sections in which the results for the initial and replication samples are reported and discussed. The results for the subtest of items not susceptible to testwiseness are reported in regular print and the results for the subtest of items susceptible to the ID1 testwiseness strategy are reported in **bold** print.

Initial Data Analyses: Chemistry - Sample 1

The distribution of the total test scores for the initial sample (Sample 1, N = 4,000) was comparable to the distribution of total test scores obtained for the entire population of grade 12 students who wrote the Chemistry 30 Diploma Examination in June 1999. The mean of the total test scores for both the population and Sample 1 was 30.70 (69.77%) and the corresponding standard deviations were 6.78 (15.42%) and 6.76 (15.36%). As outlined in Chapter III, three ability groups were identified. The number of examinees in each ability group were 1,066 high ability examinees, 1,255 middle ability examinees, and 1,117 low ability examinees.

Classical Item Analysis

The test statistics for the subtest of items not susceptible to testwiseness and the subtest of items susceptible to the ID1 testwiseness strategy are reported in Table 12. For convenience and to differentiate the initial sample from the replication sample, these subtests are identified as the CH-NTW-S1 and CH-ID1-S1, respectively, where S1 represents the initial sample or Sample 1. The obtained difficulty index (p-value), point-biserial (r_{pb}), and corrected point-biserial correlations (r_{cpb}) for each item are presented in Table 41 for the CH-NTW-S1 and Table 42 for the CH-ID1-S1 in Appendix E.

Table 12
Summary of Test Statistics for the CH-NTW-S1 and the CH-ID1-S1

Subtest	Number of Items	Mean Raw Score	Standard Deviation	Reliability and SEM
CH-NTW-S1	21	13.58 (64.67%)	3.65 (17.38%)	0.71 (1.93)
CH-ID1-S1	**14**	**10.53 (75.21%)**	**2.29 (16.35%)**	**0.58 (1.48)**

Note: SEM refers to the standard error of measurement.

As shown in Table 12, the mean number right score for the CH-NTW-S1 was 13.58 (64.67%) and the corresponding standard deviation was 3.65 (17.38%). The

internal consistency index (Cronbach's α) was 0.71 and the standard error of measurement was 1.93.

The item difficulties for the correct option varied from 0.15 (item 39) to 0.89 (item 2), and yielded a mean *p*-value of 0.65 and a standard deviation of 0.18 for the CH-NTW-S1 items. Adopting the evaluation criteria used by Alberta Education, one item (item 39) did not meet the minimum difficulty standard of 0.30 for the correct option, and three items (items 1, 2, and 32) exceeded the maximum difficulty standard of 0.85. The point-biserial correlations for the correct option varied from 0.28 (item 39) to 0.51 (item 29). The corrected point-biserial correlation coefficient values for the correct option ranged from 0.18 (item 39) to 0.39 (item 29). Item 39 did not meet the Alberta Education standards of corrected point-biserial correlation being at least 0.20.

The proportion of examinees who selected the incorrect options ranged from 0.02 (items 1 and 2) to 0.64 (item 39) across the 21 items. Of the 63 foils, ten did not meet the Alberta Education criterion for distractors: *p*-value for distractors less than 0.05. Items with these foils were generally found to be easy items with the remaining incorrect options selected by less than 11% of examinees. The point-biserial correlations for the incorrect alternatives were all less than zero and varied from -0.04 (item 39) to -0.34 (item 17).

The mean number right score for the CH-ID1-S1 was **10.53** (**75.21%**) and the corresponding standard deviation was **2.29** (**16.35%**) (cf. Table 12). The internal consistency index (Cronbach's α) was **0.58** with the standard error of **1.48**.

The item difficulties for the correct option ranged from **0.47** (item 20) to **0.90** (items 28) producing a mean difficulty of **0.75** and a standard deviation of **0.11** for the CH-ID1-S1 items. Using the evaluation standards used by Alberta Education, all items met the minimum difficulty standard of **0.30**. Four items (8, 12, 19 and 28) exceeded the maximum difficulty standard of **0.85**. The point-biserial correlation coefficient values for the correct option varied from **0.28** (item 33) to 0.51 (item 18). The corrected point-biserial item discrimination indices for the correct option ranged from **0.07** (item 33) to **0.34** (item 18). For four items (7, 30, 33, and 41), the corrected point-biserial correlation was lower than Alberta Education's standard of 0.20.

The proportion of examinees who selected the incorrect options ranged from **0.01** (items 19 and 33) to **0.52** (item 12). Of the 42 foils, 13 did not meet the Alberta Education criterion that at least 5% of examinees select the incorrect option. The point-biserial correlations for the incorrect alternatives were all less than zero and varied from **-0.08** (item 19) to **-0.32** (items 18, 25, and 36).

Taken together, the CH-NTW-S1 items were on average more difficult than the CH-ID1-S1 items (mean *p*-value of 0.65 vs. **0.75**, respectively). The distribution of incorrect responses across item foils was more proportional for the CH-NTW-S1 than for the CH-ID1-S1 items. In the subtest of 21 non-susceptible to testwiseness items, 10 (15.87%) out of 63 incorrect alternatives were found to have *p*-values lower than 0.05 while in the subtest of items susceptible to the ID1 testwiseness strategy, **13** (**30.95%**) of 42 foils were found to have *p*-values less than 0.05. Although the distributions of item point-biserial correlations for the correct options were approximately the same for both subtests, the CH-NTW-S1 and the CH-ID1-S1 were slightly different in regard to the distributions of the corrected point-biserials. One item on the CH-NTW-S1 and four items on the CH-ID1-S1 did not meet the Alberta Education standard of the corrected

point-biserial being at least 0.20. Lastly, for both subtests, the point-biserial correlations for the foils were all less than zero.

Item Response Item Analysis

Assumptions of Item Response Theory

Following the sequence of presentation of results for the social studies examination, the results of the tests of the item response theory assumptions for the CH-NTW-S1 and CH-ID1-S1 are provided below.

Unidimensionality. Principal component factor analysis yielded three components with eigenvalues greater than 1.0 for the CH-NTW-S1. The eigenvalue for the first component, 3.32, was 2.91 times greater than the eigenvalue of the second component, 1.14. Further, the difference between the second and third components was small, 0.09. The scree plot displayed in Figure 11 confirms the dominance of the first principal component. Further, Stout's T statistic was 0.33 yielding a *p*-value of 0.37. These results indicate the tenability of the hypothesis of essential unidimensionality of the CH-NTW-S1.

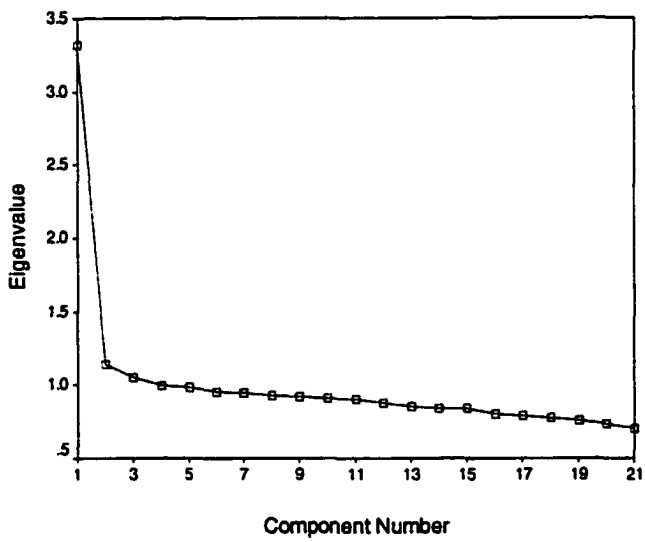


Figure 11. Scree Plot for the CH-NTW-S1

Three principal components with eigenvalues greater than 1.0 were extracted for the CH-ID1-S1. The eigenvalue for the first component, **2.35**, was **2.21** times greater than the eigenvalue for the second component, **1.06**. The difference between the second and third components was small, **0.06**. The shape of scree plot confirms these results (see Figure 12). Since one of the requirements of the DIMTEST is having at least 20 items in the test, the Stout's T statistic was not computed for the CH-ID-S1 (Nandakumar, 1993; Stout, Douglas, Junker, Roussos, 1993).

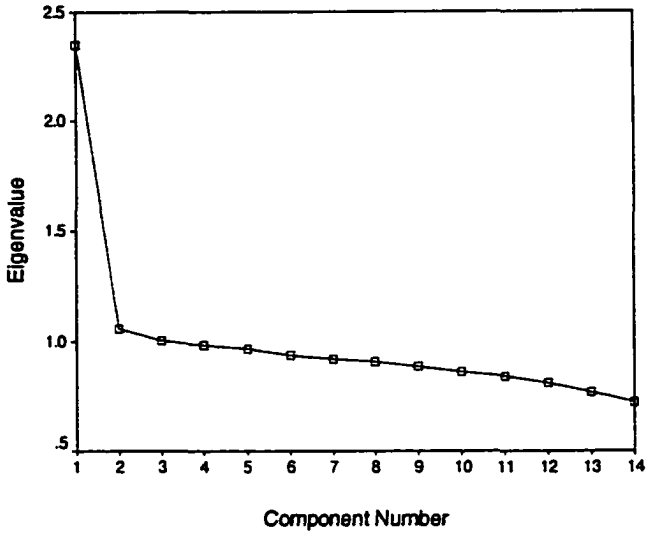


Figure 12. Scree Plot for the CH-ID1-S1

Taken together, the results of principal component analyses, the scree plots, and Stout's T statistic suggest that there was a dominant component underlying examinees responses to the set of items contained in the CH-NTW-S1. The results of the classical factor analysis indicated that the CH-ID1-S1 was unidimensional (Cattell, 1952; Gorsuch, 1983; Nandakumar, 1993).

Local independence. Given that the assumption of essential unidimensionality was met in both cases, the assumption of item independence for both subtests was tenable (Hambleton, et al., 1991; Lord, 1980).

Equal discrimination. The differences between the lowest and the highest values of item point-biserial correlations were examined for the CH-NTW-S1 and the CH-ID1-S1. As reported in the conventional item analysis section, these values varied from 0.28 to 0.51 for both subtests. Due to large difference between these values, it was concluded that the assumption of equal discrimination indices was not met for both subtests (Hambleton & Murray, 1983). Therefore, the one-parameter item response model, which partly relies on this assumption, may not be appropriate for the CH-NTW-S1 or the CH-ID1-S1.

Non-speededness. The percentages of individuals who did not complete the last three items on the CH-NTW-S1 and the CH-ID1-S1 were calculated. Out of 4000 examinees, 13 examinees (less than 1%) did not complete one the last three items on the CH-NTW-S1 and 8 (less than 1%) examinees did not complete one the last three items on the CH-ID1-S1. Therefore, it was concluded that the assumption of non-speededness was tenable for both subtests (Hambleton, et al., 1991).

Non-guessing. The performance of low-achieving examinees on the three most difficult items on each subtest was examined to test the tenability of the assumption of non-guessing. In this study, the low achieving examinees were defined as those whose

number right score on the full Chemistry test of 44 items fell at or below 14. Out of 4000 examinees, 46 (1.15%) students obtained a score of 14 or lower.

The three most difficult items in the CH-NTW-S1 were: item 39 ($p = 0.15$), item 9 ($p = 0.43$), and item 17 ($p = 0.49$). Examination of the performance of the low-achieving students on these items revealed that 24 (52.17%) did not answer any of the three items correctly, 22 (47.82%) examinees correctly answered one of these questions, and no examinee correctly answered more than one of these questions.

The three most difficult items in the CH-ID1-S1 were: item 20 ($p = \mathbf{0.47}$), item 41 ($p = \mathbf{0.63}$), and item 33 ($p = \mathbf{0.65}$). Examination of the performance of low achievers on these items revealed that **15** (**32.61%**) did not answer any of the three items correctly, **22** (**47.82%**) correctly answered one question, **9** (**19.56%**) examinees correctly answered two questions, and no examinee correctly answered all 3 questions.

Given these findings, it was concluded that the effect of guessing was minimal. Additional evidence supporting this conclusion is provided by the fact that only 8 (less than 1%) out of 4000 examinees scored at or below the chance level (score of 11).

Summary. The results presented above suggest that with the exception of the assumption of equal item discrimination, the assumptions of item response theory were met for both subtests. The CH-NTW-S1 and the CH-ID1-S1 were found to be essentially unidimensional. Consequently, local independence was obtained for both subtests. In both cases it was found that speed was not a factor affecting examinees' test performance and that the effect of guessing was minimal. Taken together, the results of testing item response theory assumptions permit the use of the two-, three-parameter, and nominal response models. The item and ability estimates obtained from the one-parameter models should be interpreted with caution due to the failure to meet the assumption of equal discrimination.

Item Parameter Estimation

The CH-NTW-S1 and CH-ID1-S1 items were calibrated separately using the MULTILOG program (Thissen, 1991). The full sample of 4000 examinees was used for item calibration. Marginal maximum likelihood estimation was employed to estimate item parameters within each IRT model. The results of the binary models item analyses are presented in Tables 43 and 44 and the results of the nominal model item analysis are reported in Tables 45 and 46 in Appendix E for the CH-NTW-S1 and the SS-ID1-S1, respectively.

One-parameter model. The values of item difficulty parameter for the CH-NTW-S1 items ranged from –2.77 (item 2) to 2.30 (item 39), with a mean difficulty of -0.91 and a standard deviation of 1.19. The difficulty estimates for the CH-ID1-S1 items varied from **–3.08** (item 28) to **0.17** (item 20), with a mean difficulty of **–1.71** and standard deviation of **0.89**. The results reveal that, according to the one-parameter model, the items on the CH-NTW-S1 were more difficult than the items contained in the CH-ID1-S1.

Two-parameter model. The values of item difficulty parameter for the CH-NTW-S1 items varied from –2.90 (item 2) to 2.85 (item 39) yielding a mean of -0.82 and a standard deviation of 1.24. The estimates of item discrimination ranged from 0.54 (item 19) to 1.43 (item 3) with a mean of 0.90 and a standard deviation of 0.24.

For the CH-ID1-S1, the difficulty estimates ranged from **–3.13** (item 33) to **0.16** (item 20), with a mean of **-1.76** and a standard deviation of **0.85**. The values of the item discrimination estimates varied from **0.20** (item 33) to **1.42** (item 19) yielding a mean of **0.88** and a standard deviation of **0.39**. Taken together, the mean difficulty of the CH-NTW-S1 was higher than the mean of the CH-ID1-S1. The item discrimination estimates were similar for both subtests.

Three-parameter model. For the CH-NTW-S1, the values of item difficulty parameter varied from –2.36 (item 2) to 2.42 (item 39) with a mean of -0.24 and a standard deviation of 1.19. The values of the item discrimination estimates ranged from 0.46 (item 16) to 1.35 (item 6) yielding a mean of 0.75 and a standard deviation of 0.23. The values of pseudo-guessing parameter estimates ranged from 0.07 (item 39) to 0.36 (items 6 and 38) producing a mean of 0.24 and a standard deviation of 0.07.

For the CH-ID1-S1, the values of item difficulty estimates varied from **–1.93** (item 28) to **0.73** (item 20) with a mean of **-0.91** and a standard deviation of **0.69**. The item discrimination parameter estimates ranged from **0.15** (item 33) to **1.16** (item 20) yielding a mean of **0.65** and a standard deviation of **0.30**. Lastly, the values of pseudo-guessing parameter ranged from **0.18** (item 44) to **0.37** (item 5) with a mean of **0.26** and a standard deviation of **0.04**.

Looking at the three-parameter model, the CH-NTW-S1 items were on average more difficult that the CH-ID1-S1 items. Both subtests were found to have comparable mean discrimination parameter estimates, and, since the prior normal distribution of the pseudo-guessing parameter was specified for each subtest, both subtests had similar mean pseudo-guessing estimates.

Nominal response model. For the CH-NTW-S1, the values of the $c_{ik\text{-}correct}$ parameter estimates for the correct answer ranged from -0.17 (item 39) to 2.68 (item 2) with a mean of 1.55 and a standard deviation of 0.75. The values of the popularity parameter for the 63 foils across 21 items varied from –1.64 (item 21) to 1.43 (item 39) with a mean $c_{ik\text{-}incorrect}$ value of -0.52 and a standard deviation of 0.55. The values of item discrimination estimates for the correct response varied from 0.42 (item 38) to 1.12 (item 4) with a mean of 0.71 and standard deviation of 0.19. The values of the $a_{ik\text{-}incorrect}$ parameter for the incorrect alternatives for the 35 items ranged from -0.73 (item 39) to 0.26 (item 16) yielding a mean of -0.24 and a standard deviation of 0.24.

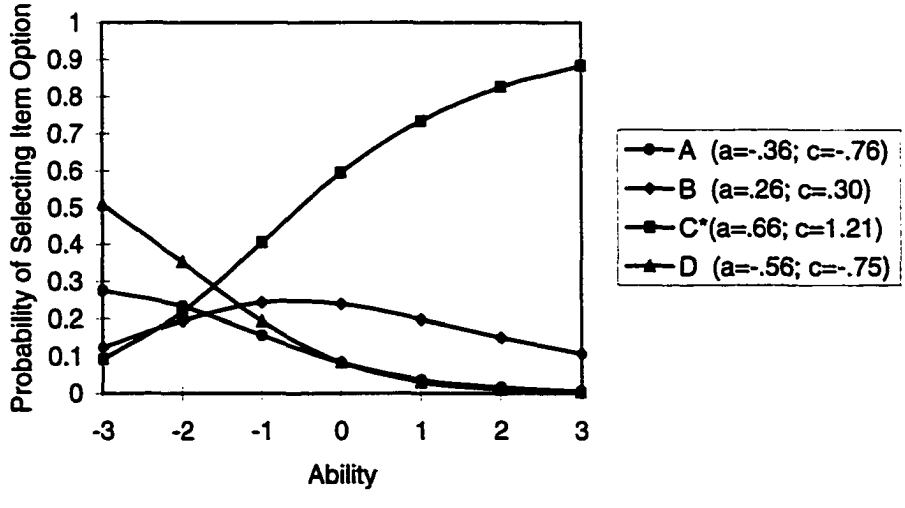
For the CH-ID1-S1, the values of the $c_{ik\text{-}correct}$ parameter for the correct response varied from **0.98** (item 20) to **2.93** (item 19) with a mean of **2.11** and a standard deviation of **0.60**. The values of the popularity parameter for the 42 incorrect alternatives varied from **–3.09** (item 33) to **0.92** (item 33) yielding a mean of **-0.70** and a standard deviation of **0.87**. The values of item discrimination parameter for the correct option ($a_{ik\text{-}correct}$) ranged from **0.28** (item 30) to **1.03** (item 8) yielding a mean of **0.72** and a standard deviation of **0.25**. The values of the $a_{ik\text{-}incorrect}$ parameter across the 14 items ranged from **-0.91** (item 33) to **0.28** (item 33) with a mean of **-0.24** and a standard deviation of **0.27**.

As indicated by the values of the popularity parameter, the CH-NTW-S1 contained more difficult items than the CH-ID1-S1. Further, the range of the popularity parameter estimates for incorrect item alternatives indicated more equal distributions of incorrect responses across the foils for the CH-NTW-S1 than for the CH-ID1-S1. The mean discrimination parameter estimates for the correct and incorrect were similar across both subtests.

Item option trace lines. The item option parameters obtained from the nominal response models were used to calculate the probabilities of selecting each option across different ability levels. These values, reported in Table 47 for the CH-NTW-S1 and Table 48 for the CH-ID1-S1 in Appendix E, were subsequently used to produce item option characteristic curves (trace lines).

Inspection of item option characteristic curves revealed that for both the CH-NTW-S1 and CH-ID1-S1 items the probability of selecting the correct option increased as examinee ability increased. However, behavior of the incorrect options was found to be different for the two types of items. For the CH-NTW-S1 items, the values of $c_{ik\text{-}incorrect}$ parameter indicated a fairly equal distribution of incorrect responses across item foils. The corresponding trace lines suggested that the incorrect alternatives had similar probabilities of being selected across different ability levels. Two examples of items not susceptible to testwiseness are presented in Figures 13 and 14 (Alberta Education, 1999c). As shown in Figure 13, incorrect options A and D of Item 16 became more attractive to examinees at the lower and middle ability levels (θ in around or below 0.0). The trace line for the alternative B indicates a moderate probability of selecting this option across ability levels, with the greatest probability of selecting it by middle ability examinees. The trace lines for Item 27 (see Figure 14) indicate uniform, decreasing, probability of selecting distractors A, C, and D as examinee ability increases.

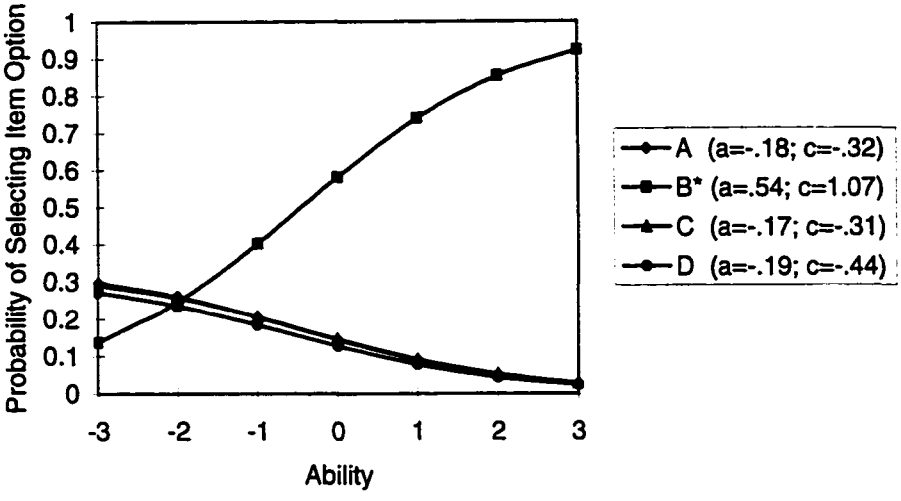
16. The voltage of an electrochemical cell is +0.20 V. If one of the half-reactions is the reduction of $Cu^{2+}_{(aq)}$, then the other half-reaction that occurs could be
    A. $2I^-_{(aq)} \rightarrow I_{2(s)} + 2e^-$
    B. $S_{(s)} + 2H^+_{(aq)} + 2e^- + H_2S_{(aq)}$
    C. $H_2S_{(aq)} \rightarrow S_{(s)} + 2H^+_{(aq)} + 2e^-$
    D. $I_{2(s)} + 2e^- \rightarrow 2I^-_{(aq)}$



Note: C* indicates the correct option; A, B, and D are item foils.

Figure 13. Non-Susceptible to Testwiseness Item 16, Chemistry 30, Sample 1.

27. A property that is **not** consistent with the behavior of water is that water is able to
    A. act both as an acid and a base in proton transfer reaction
    B. absorb 241.8 kJ when one mole of water vapour is formed from its elements
    C. act as an oxidizing agent or reducing agent in electrolytic cell
    D. react with acids to produce hydronium ions

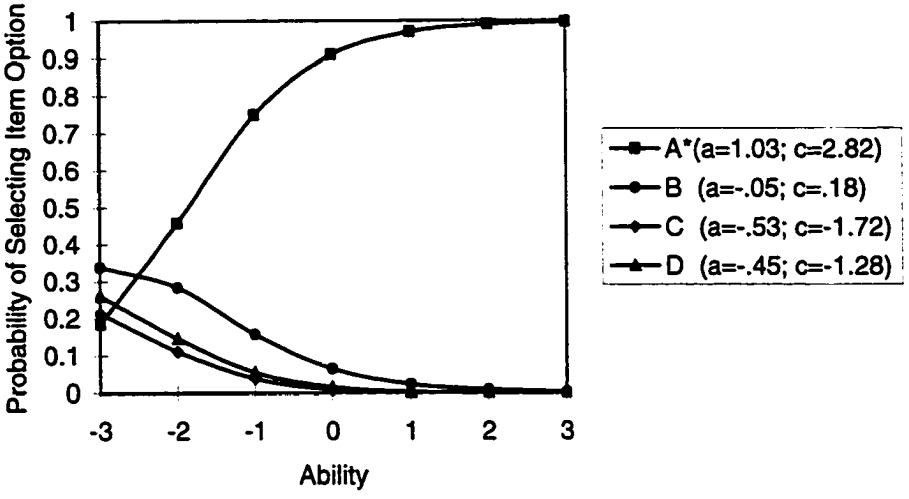


Note: B* indicates the correct option; A, C, and D are item foils.

Figure 14. Non-Susceptible to Testwiseness Item 27, Chemistry 30, Sample 1

For the CH-ID1-S1 items, the high negative values of the $c_{ik\text{-}incorrect}$ parameters indicated absurd options that were rarely selected by examinees. The corresponding item option trace lines indicated that the seldom chosen options were either most attractive to lower ability examinees or were disregarded by the majority of examinees at all ability levels. Two examples of items susceptible to the ID1 strategy of testwiseness and item option characteristic curves are presented in Figure 15 and Figure 16 (Alberta Education, 1999c). As shown in Figure 15, the probabilities of choosing absurd options C and D of Item 8 approach zero as examinee ability increases beyond one standard deviation below the mean ($\theta \geq -1.0$). The trace lines for item 33 (see Figure 16) reveal that the probability of selecting an absurd option A was close to zero for the majority of examinees (θ between –2.0 and 3.0). The probabilities of selecting non-absurd distractors in items susceptible to the ID1 testwiseness strategy indicate their greater attractiveness to lower and middle ability examinees (see Figure 9, Option B) and in some case, to examinees at all ability levels (see Figure 16, Options B and C). Overall, the probabilities of choosing properly functioning foils were found to be gradually decreasing as examinee ability increased.

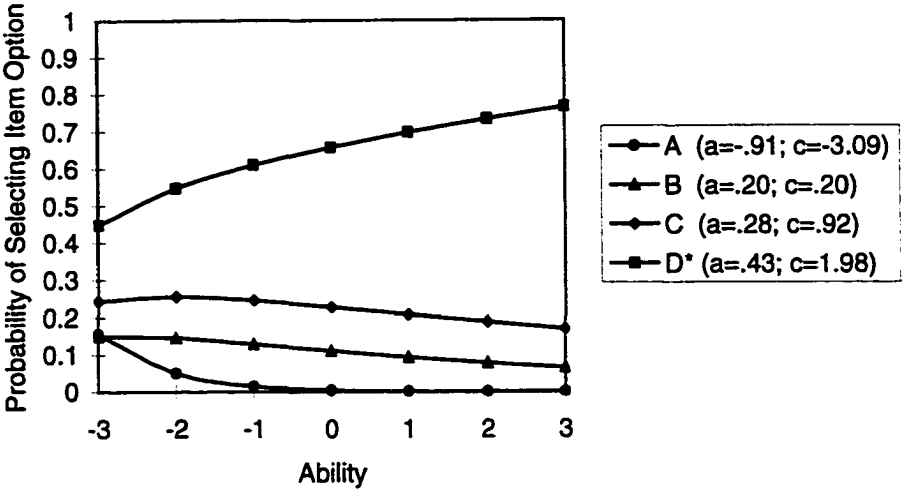
8. The molar heat of solution for $NaOH_{(s)}$ is –44.6 kJ/mol. If 25.0 g of $NaOH_{(s)}$ is dissolved in water in a calorimeter, the heat released inside the calorimeter is
   A. 27.9 kJ
   B. 71.4 kJ
   C. 1.12 MJ
   D. 1.78 MJ



Note: A* indicates the correct answer; C and D are absurd options; B is a properly functioning item foil.

Figure 15. Susceptible to ID1 Testwiseness Strategy Item 8, Chemistry 30, Sample 1.

33. Most plants grow best in soil with a pH between 6 and 7. Higher or lower pH values prevent them from absorbing essential nutrients. Plants can absorb phosphorous in the form of $H_2PO_4^-{}_{(aq)}$. In basic soil, $H_2PO_4^-{}_{(aq)}$.
   A. $P_{4(s)}$
   B. $PO_4^{3-}{}_{(aq)}$
   C. $H_3PO_4^-{}_{(aq)}$
   D. $HPO_4^{2-}{}_{(aq)}$



Note: D* indicates the correct option; A is an absurd option; B and C are properly functioning foils.
Figure 16. Susceptible to ID1 Testwiseness Strategy Item 33, Chemistry 30, Sample 1.

Taken together, the item trace lines for the incorrect options of items not susceptible to testwiseness and the trace lines for the incorrect option of items susceptible to the ID1 testwiseness strategy were different. While item trace lines did not reveal clear elimination of any of the incorrect alternatives by the examinees for the items contained in the CH-NTW-S1, the trace lines for alternatives identified as absurd options in the CH-ID1-S1 items were lower than the remaining one or two options in these items. This observation is congruent with the finding that the values of the c_{ik} parameter for absurd options were considerably lower than the corresponding values for the remaining distractors indicating that the majority of examinees were able to eliminate the absurd options before choosing from among the remaining options. In addition, absurd options were found to have either high discrimination power at the low ability level or essentially no discrimination power. Consequently, it appears that the absurd option(s) either attracted only low proficiency students who did not posses sufficient partial knowledge of the subject content to be able to eliminate these options or were rejected by examinees at all proficiency levels.

Summary of Item Analysis

The classical item analysis and the item response theory item analyses conducted using the one-, two-, three-parameter, and nominal response models provide comparable information about items contained in the CH-NTW-S1 and CH-ID1-S1. The mean item difficulty was found to be higher for the CH-NTW-S1 than for the CH-ID1-S1 across all five models (e.g., p-value of 0.65 and **0.75** in classical item analysis, or b_i value of -0.89 and **-1.71** in one-parameter model, respectively). The distributions of discrimination indices for items on both subtests were approximately the same (e.g., the values of point biserial correlations varied from .28 to 0.51 both subtests; the mean a_i in the two-parameter model were 0.90 for the CH-NTW-S1 and **0.88** for the CH-ID1-S1).

As indicated by the results of the classical and nominal response model item analyses, the distribution of wrong responses was more uniform across item foils in the CH-NTW-S1 than in the CH-ID1-S1 (e.g., 15.87% vs. **30.95%** of foils were selected by less than five percent of examinees, respectively). The analysis of item option trace lines supported these findings. In the CH-NTW-S1, the distractors appeared to be fairly equally attractive for examinees at a given proficiency level. In case of the CH-ID1-S1 items, the trace lines indicated either very low probability of selecting an absurd option by majority of students or its attractiveness to low ability examinees only.

Comparison of Ability Estimates

The ability estimates expressed as T-scores ($\mu = 50$, $\sigma = 10$) for each of the scoring methods for each subtest were separately compared within the high, middle and low ability sub-samples. The results of these comparisons are presented in Table 13, 14, and 15, respectively. The first two rows of each table contain the means and standard deviations of the five sets of ability estimates. The correlations between pairs of estimates are reported in the upper triangle while the root mean square deviations between pairs of estimates are shown in the lower triangle. Using the guideline presented in Chapter III, correlations less than 0.95 were considered to indicate systematic differences in examinees ranking by different scoring models. Root mean square deviations greater than 2.0 were considered to be indicators of systematic differences between pairs of scores yielded by different scoring models.

High Ability Group

As shown in Table 13, the means of transformed ability scores yielded by the five scoring methods were on average two points on the T-score scale greater for CH-NTW-S1 than for the CH-ID1-S1. The means for the CH-NTW-S1 ranged from 61.40 to 62.85, and for the CH-ID1-S1 from **59.54** to **60.02**. The corresponding standard deviations varied from 4.24 to 5.20 for the CH-NTW-S1 and from **4.55** to **5.38** for the CH-ID1-S1, indicating comparable variability of the scores yielded by the five models for the two sets of items.

The correlations between pairs of estimates yielded by the number right and one-parameter models, and between the pairs of estimates yielded by the two-, three-parameter, and nominal models were essentially unchanged for the CH-ID1-S1 as compared to the CH-NTW-S1. These correlations were equal or greater than 0.95. The correlations between the remaining pairs of scores were substantially lower for the CH-ID1-S1 than for the CH-NTW-S1. These correlations varied from **0.86** to **0.89** for the CH-ID1-S1 and from 0.94 to 0.97 for the CH-NTW-S1.

Given these findings, although there were slight differences among the correlations obtained in the CH-NTW-S1, the students' ranked position will in effect be the same across the five scoring methods, with the possible exception of some differences in rankings that may occur when number right or one-parameter scores are used in a place of nominal response model scores (correlation of 0.94). In a case of the CH-ID1-S1, the high ability examinees' ranked position will be quite different when the number right or one-parameter scores are used in place of the two-, three-parameter or nominal model scores.

Table 13
Comparison of Ability Estimates Obtained from the CH-NTW-S1 and CH-ID1-S1 in the High Ability Group (N=1066)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and Standard Deviations	61.40 (4.24) **59.54 (4.55)**	61.78 (5.20) **59.98 (5.38)**	61.85 (5.13) **59.90 (4.90)**	61.66 (5.20) **60.02 (5.28)**	61.78 (5.13) **59.85 (4.75)**
θ_{nr} (number right)		0.99 **0.99**	0.97 **0.89**	0.96 **0.86**	0.94 **0.87**
θ_1 (1PL)	1.09 **0.98**		0.97 **0.89**	0.97 **0.86**	0.94 **0.86**
θ_2 (2PL)	1.48 **2.26**	1.18 **2.45**		0.98 **0.99**	0.97 **0.97**
θ_3 (3PL)	1.65 **2.75**	1.36 **2.83**	0.96 **0.91**		0.95 **0.96**
θ_4 (NRM)	1.91 **2.44**	1.82 **2.71**	1.28 **1.17**	1.60 **1.55**	

Note: Values in regular print are for the CH-NTW-S1 subtest and in **bold** print for the CH-ID1-S1subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

The absolute agreements between pairs of estimates were fairly consistent with the correlations among ability scores. The closest agreements for both subtests were found between the scores produced by the number right and one-parameter models, and by the two- and three-parameter models. The values of the root mean square deviations for these pairs were, respectively, 1.09 and 0.96 for the CH-NTW-S1 and **0.98** and **0.91** for the CH-ID1-S1. Also for both subtests, there was slightly less agreement between the pairs of scores yielded by the two-parameter and nominal models, and by the three-parameter and nominal models. The root mean square deviations for these comparisons were 1.28 and 1.60 for the CH-NTW-S1 and **1.17** and **1.55** for the CH-ID1-S1. The lack of agreement between the number right and the two-, three-parameter, and nominal model scores and between the one-parameter and other IRT model scores was greater for the CH-ID1-S1 than for the CH-NTW-S1 (root mean square deviations **2.26** vs.1.48, **2.75** vs.1.65, **2.44** vs. 1.91, **2.45** vs. 1.18, **2.83** vs. 1.36, and **2.71** vs. 1.82).

When the correlations and root mean square deviations are considered together, it seems that if a test contains items susceptible to the ID1 testwiseness strategy, for the high ability group, the number right or one-parameter models will likely produce substantially different ability estimates when used in place of the two-, and three-

parameter models, and the nominal response model. In the case of a test containing items not susceptible to testwiseness, likely no significant differences in ability estimates will occur when ability scores yielded by the five scoring models are used in place of one another.

Middle Ability Group

As shown in Table 14, the means and standard deviations of transformed ability scores yielded by the five scoring methods for the CH-NTW-S1 for the middle proficiency group were approximately one point on the T-score metric less than for the CH-ID1-S1. The means for the CH-NTW-S1 ranged from 50.41 to 50.99 and for the CH-ID1-S1 from **51.23** to **51.64**. The standard deviations of the scores yielded by the five scoring models varied from 4.42 to 4.66 for the CH-NTW-S1 and from **5.64** to **6.04** for the CH-ID1-S1.

Table 14

Comparison of Ability Estimates Obtained from the CH-NTW-S1 and the CH-ID1-S1 in the Middle Ability Group (N=1255)

	θ_{nr} (number right)	θ_{1} (1PL)	θ_{2} (2PL)	θ_{3} (3PL)	θ_{4} (NRM)
Means and Standard Deviations	50.99 (4.49) **51.64** **(5.64)**	50.49 (4.42) **51.23** **(5.88)**	50.41 (4.55) **51.52** **(5.91)**	50.52 (4.53) **51.34** **(6.04)**	50.49 (4.66) **51.55** **(5.85)**
θ_{nr} (number right)		0.99 **0.99**	0.95 **0.88**	0.93 **0.85**	0.90 **0.85**
θ_{1} (1PL)	0.55 **0.59**		0.95 **0.88**	0.93 **0.86**	0.90 **0.85**
θ_{2} (2PL)	1.54 **2.86**	1.40 **2.90**		0.99 **0.99**	0.95 **0.97**
θ_{3} (3PL)	1.81 **3.20**	1.73 **3.21**	0.69 **0.83**		0.95 **0.96**
θ_{4} (NRM)	2.11 **3.13**	2.03 **3.19**	1.48 **1.50**	1.64 **1.69**	

Note: Values in regular print are for the CH-NTW-S1 subtest and in **bold** print for the CH-ID1-S1subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

The correlations among pairs of scores yielded by the number right and one-parameter model, and by the two-, three-parameter and nominal models were found to be consistent across the CH-NTW-S1 and CH-ID1-S1. For both subtests, these correlations

were equal to or greater than 0.95. The correlations between scores obtained from the number right and the two-, three-parameter, and nominal response models, and between scores yielded by the one-parameter model and the two-, three-parameter, and nominal response models were lower for the CH-ID1-S1 than for the CH-NTW-S1 (**0.88** vs. 0.95, **0.85** vs. 0.93, **0.85** vs. 0.90, **0.88** vs. 0.95, **0.86** vs. 0.93, and **0.85** vs. 0.90).

Given these findings, it appears that there will be some differences among the students' ranked position across the five scoring methods, particularly when a test contains items susceptible to the ID1 testwiseness strategy. In such cases examinees' ranked positions will be different when the number right or one-parameter scores are used in place of the two-, three-parameter or nominal model scores. Differences in middle ability examinees ranking position are also likely to occur in the case of a test not containing testwise susceptible items when the number right or one-parameter model scores are used in place of the three-parameter or nominal model scores.

The closest agreements between the transformed ability estimates for both subtests were found for the scores produced by the number right and one-parameter models, and by the two- and three parameter models. The values of the root mean square deviations for these comparisons were 0.55 and 0.69 for the CH-NTW-S1 and **0.59** and **0.83** for the CH-ID1-S1. Also for both subtests, there was slightly less agreement between the scores yielded by the two-parameter and the nominal response models and by the three-parameter and nominal models. The root mean square deviations for these two pairs were 1.48 and 1.64 for the CH-NTW-S1 and from **1.50** and **1.69** for the CH-ID1-S1.

The differences between the remaining pairs of ability estimates were greater for the CH-ID1-S1. The root mean square deviations for the comparison in which the number right and the one-parameter scores were involved were substantially greater for the CH-ID1-S1 than for the CH-NTW-S1. The root mean square deviations increased from 1.54, 1.81, and 2.11 for the CH-NTW-S1 to **2.86**, **3.20**, and **3.13** for the CH-ID1-S1 for the number right and the two-, three-parameter, and nominal model scores comparisons. Likewise, for the one-parameter model and the two-, three-parameter, and nominal model scores comparisons, the root mean square deviations increased from 1.40, 1.73, and 2.03 for the CH-NTW-S1 to **2.90**, **3.21**, and **3.19** for the CH-ID1-S1, respectively.

Thus, it seems that for the CH-ID1-S1, for middle ability examinees, the replacement of the number right or one-parameter model ability scores with the scores yielded by the other IRT models will likely result in significant differences. In the case of the CH-NTW-S1, with the exception of differences between the number right or one-parameter scores and the nominal response model scores (root mean square deviations 2.11 and 2.03, respectively), differences between pairs of scores yielded by the five methods were relatively small (root mean square deviations less than 2.00). These results suggest that if a test does not contain testwise susceptible items, the middle ability examinees, will likely receive a different score when the number right or one-parameter models are used in place of the nominal model.

Low Ability Group

As shown in Table 15, for the low ability group, the means of the transformed ability scores yielded by the five scoring methods were fairly comparable for both subtests. The means ranged from 38.32 to 38.75 for the CH-NTW-S1 and from **38.97** to

39.39 for the SS-ID1-S1. With the exception of the three-parameter model scores, the standard deviations were on average two points lower for the CH-NTW-S1 than for the CH-ID1-S1 (6.06 vs. **8.37**, 5.52 vs. **7.47**, 5.40 vs. **7.25**, and 5.34 vs. **7.36**). The standard deviation of the three-parameter model scores was approximately one point on the T-score scale lower for the CH-NTW-S1 than for the CH-ID1-S1 (5.69 vs. **6.97**)

Table 15
Comparison of Ability Estimates for the CH-NTW-S1 and CH-ID1-S1, Low Ability Group (N=1117)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and Standard Deviations	38.32 (6.06) **39.15** **(8.37)**	38.69 (5.52) **39.39** **(7.47)**	38.75 (5.40) **39.04** **(7.25)**	38.73 (5.69) **39.19** **(6.97)**	38.73 (5.34) **38.97** **(7.36)**
θ_{nr} (number right)		1.00 **0.99**	0.97 **0.93**	0.92 **0.91**	0.93 **0.91**
θ_1 (1PL)	0.66 **0.99**		0.97 **0.93**	0.92 **0.91**	0.93 **0.91**
θ_2 (2PL)	1.63 **3.17**	1.36 **2.79**		0.97 **0.99**	0.95 **0.97**
θ_3 (3PL)	2.39 **3.54**	2.23 **3.08**	1.39 **1.00**		0.94 **0.96**
θ_4 (NRM)	2.36 **3.45**	2.12 **3.12**	1.67 **1.50**	2.14 **2.06**	

Note: Values in regular print are for the CH-NTW-S1 subtest and in **bold** print for the CH-ID1-S1 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

With the exception of correlation between the three-parameter and the nominal model scores for the CH-NTW-S1, the correlations among pairs of scores yielded by the number right and one-parameter model, and the correlations among pairs of scores yielded by the two-, three-parameter and nominal models were found to be equal to or greater than 0.95 for both the CH-NTW-S1 and CH-ID1-S1. The lowest correlations for both subtests were found for pairs of scores obtained from the number right and the three-parameter and nominal response models, and between scores yielded by the one-parameter model and the three-parameter, and nominal response models. These correlations ranged from 0.92 to 0.93 for the CH-NTW-S1 and were **0.91** the CH-ID1-S1. The correlations between the ability estimates obtained form the number right or one-

parameter models and the two-parameter model were lower for the CH-ID1-S1 than for the CH-NTW-S1 (**0.93** vs. 0.97).

Given these findings, it appears that there will be some differences among the students' ranked position across the five scoring methods for both subtests. For both CH-NTW-S1 and CH-ID1-S1, low ability examinees' ranking position will be different when the number right or one-parameter scores are used in place of the three-parameter or nominal model scores. Some differences in examinees ranking position are also likely to occur in for the CH-ID1-S1 when the number right or one-parameter model scores are used in place of the two-parameter model scores.

For both subtests, the closest agreements were found between the scores produced by the number right and one-parameter models, and by the two- and three parameter models. The values of the root mean square deviations for these comparisons were 0.66 and 1.39 for the CH-NTW-S1 and **0.99** and **1.00** for the CH-ID1-S1, respectively. Also consistent for both subtests were the differences between ability scores yielded by the two-parameter and nominal models and by the three-parameter and nominal models. The root mean square deviations for these comparisons were 1.67 and 2.14 for the CH-NTW-S1 and **1.50** and **2.06** for the CH-ID1-S1. The differences between the remaining pairs of ability estimates were larger for the CH-ID1-S1 than for the CH-NTW-S1. The root mean square deviations increased from 1.63, 2.39, and 2.36 for the CH-NTW-S1 to **3.17**, **3.54**, and **3.45** for the CH-ID1-S1 for comparisons involving the number right and the two-, three-parameter, or nominal models. Likewise, for the one-parameter model and the two-, three-parameter, or nominal model scores comparisons, the root mean square deviations increased from 1.36, 2.23 and 2.12 for the CH-NTW-S1 to **2.79**, **3.08**, and **3.12** for the CH-ID1-S1.

Considering these results, it appears that the low ability examinees will likely receive a different score when the number right or one-parameter models are used in place of the three-parameter or nominal models, and when the three-parameter model is used instead of the nominal model. For both subtests, the root mean square deviations for these comparisons were greater than 2.00. Comparison of the differences among pairs of ability estimates for the both subtests suggests that these differences will be greater for the CH-ID1-S1 than for the CH-NTW-S1 (root mean square deviations from **2.76** to **3.54** vs. 1.36 to 2.39, respectively).

Comparison of Groups

The pattern and the magnitudes of correlations between pairs of ability estimates yielded by the number right and one-parameter models, and between scores yielded by the two-, three-parameter and nominal models were found to be consistent for the CH-NTW-S1 and CH-ID1-S1. For all three ability groups these correlations were equal to or greater than 0.94. The pattern and magnitudes of correlations between the remaining pairs of scores varied by the subtest, by the type of scoring models and by group. The correlations between the pairs of ability estimates obtained from the number right and the two-, three-parameter, and nominal response models, and between scores yielded by the one-parameter model and the two- and three-parameter and nominal response models were higher for the CH-NTW-S1 than for the CH-ID1-S1 for both the high and middle ability groups. These correlations ranged from 0.90 to 0.97 for the CH-NTW-S1 and from **0.85** to **0.89** for the CH-ID1-S1. In the low ability group, with the exception of

correlation between pairs of scores yielded by the number right or one-parameter models and the two-parameter models for the CH-NTW-S1, the correlations between the pairs of ability estimates obtained from the number right and the two-, three-parameter, and nominal response models, and between scores yielded by the one-parameter model and the two- and three-parameter and nominal models were similar for both subtests and ranged from 0.92 to 0.93 for the CH-NTW-S1 and from **0.91** to **0.93** for the CH-ID1-S1.

For both subtests, regardless their susceptibility to testwiseness, the lowest correlations were observed for the pairs of scores obtained from the number right and the nominal response models for all ability groups (**0.85** to 0.94), followed by the correlations between the three-parameter model scores and the number right or one-parameter model scores in the middle and low ability groups (**0.85** to 0.93). These correlations were slightly lower in the middle ability group than that the corresponding correlations in the high and low ability groups (e.g. correlations between the number right and nominal model scores were 0.90 and **0.85** in the middle ability group, 0.94 and **0.87** in the high ability group, and 0.93 and **0.91** in the low ability group).

Given these findings, it appears that the greatest changes in all examinees' ranks will likely occur if the test contains items susceptible to the ID1 testwiseness strategy and when the number right or one-parameter model scores are used in place of the two- and three-parameter and nominal response model scores. These differences will be slightly greater for the middle ability examinees than for the low and high ability examinees. Some differences in ranking may also occur in a case of a test not containing items susceptible to testwiseness. These differences will be greatest for the middle and low ability groups when the number right or one-parameter model scores are used in place of the three-parameter or nominal model scores.

Turning to the absolute agreements between scores, although the extent of agreement between the pairs of scores yielded by different scoring models varied, the patterns of root mean square deviations were generally consistent across the three ability groups and corresponded to the patterns of correlations. The closest agreements in all three ability groups and for both subtests were found between the scores yielded by the number right and one-parameter models and by the two- and three-parameter models. The root mean square deviations for these comparisons varied from 0.55 to 1.39. Less agreement was found for the pairs of scores produced by the nominal response model and the two- or three parameter models. The values of root mean square deviations for these comparisons varied from 1.28 to 2.14 and were approximately the same for the CH-NTW-S1 and CH-ID1-S1.

Differences between the CH-NTW-S1 and CH-ID1-S1 were found in case of agreements between ability estimates involving the number right or one-parameter model scores and the two- and three-parameter and nominal model scores. The root mean square deviations for these comparisons varied from 1.18 to 2.39 for the CH-NTW-S1 and from **2.26** to **3.54** for the CH-ID1-S1. These results suggest that for all students, if a test contains items susceptible to the ID1 testwiseness strategy, the number right or one-parameter models will produce different ability estimates when used in place of the two-, three-parameter or nominal models.

It should be noted that some discrepancies in absolute ability estimates also occurred for the CH-NTW-S1. For the middle ability examinees, the root mean square deviations for the pairs of scores produced by the number right or one-parameter models

and the nominal model were, respectively, 2.11 and 2.03. For the low ability examinees, the root mean square deviations for the pairs of estimates yielded by the number right or one-parameter models and the three-parameter or nominal models varied from 2.12 to 2.39. These values were substantially smaller than the corresponding values found for the CH-ID1-S1 (**3.13** and **3.19** for the number right and one-parameter models and the nominal model scores comparisons, middle ability group; 3.08 to 3.54 for the number right or one-parameter models and the three-parameter or nominal model scores comparisons, low ability group).

Replication Data Analyses: Chemistry - Sample 2

The distribution of the total test scores for the Sample 2 (N = 4000) was similar to that obtained for the whole population of grade 12 students who wrote the Chemistry 30 Diploma Examination in June 1999 and for the Sample 1 (see p. 61). The mean test score for the Sample 2 was 30.65 (69.66%) and the standard deviation was 6.77 (15.37%). The numbers of students in each ability group were 1,084 high ability examinees, 1,266 middle ability examinees, and 1,094 low ability students.

Classical Item Analysis

In order to differentiate the subtests from those described earlier in the initial data analyses section, the subtest containing non-susceptible items and the subtest comprised of items susceptible to the ID1 testwiseness strategy in the replication sample are labeled CH-NTW-S2 and CH-ID1-S2, respectively. The results of test characteristics for the SS-NTW-S2 and the SS-ID1-S2 are presented in Table 16.

Table 16
Summary of Test Statistics for the CH-NTW-S2 and the CH-ID1-S2

Subtest	Number of Items	Mean Raw Score	Standard Deviation	Reliability And SEM
CH-NTW-S2	21	13.61 (64.80%)	3.65 (17.38%)	0.73 (1.90)
CH-ID1-S2	**14**	**10.53 (75.21%)**	**2.28 (16.29%)**	**0.57 (1.50)**

Note: SEM refers to the standard error of measurement.

As shown in Table 16, the mean number right scores were 13.61 for the CH-NTW-S2 and **10.53** for the CH-ID1-S2 with the corresponding standard deviations of 3.65 and **2.28**, respectively. The internal consistency index (Cronbach's α) was 0.73 for the CH-NTW-S2 and **0.57** for the CH-ID1-S2. Comparison of these results with those reported in Table 12 for the initial sample, S1, revealed that the initial and replication samples were essentially the same.

Likewise, the item characteristics for the CH-NTW-S2 and the CH-ID1-S2 were very similar to those obtained from the initial study. The difficulty index (p-value), point-biserial (r_{pb}), and corrected point-biserial correlations (r_{cpb}) for each item are presented in Table 49 for CH-NTW-S2 and Table 50 for CH-ID1-S2 in Appendix E. The mean difficulty of the CH-NTW-S2 and CH-ID1-S2 were 0.65 and **0.75**, with the standard deviations of 0.19 and **0.11**, respectively. The distribution of item discrimination indices was approximately the same for both subtests, with the point biserial correlations ranging from 0.28 to 0.51 for the CH-NTW-S2 and from **0.29** to **0.51** for the CH-ID1-S2. Two items on the CH-NTW-S2 and six items on the CH-ID1-S2 yielded a corrected point-biserial correlation lower then 0.20. Similarly to the results for the initial sample, the distribution of incorrect responses across item foils was more proportional for the CH-NTW-S2 than for the CH-ID1-S2. Across all items, 14.29% of incorrect alternatives in the CH-NTW-S2 and **30.95** % of incorrect options in the CH-ID1-S2 were selected by fewer than 5 percent of examinees.

Item Response Item Analysis

The results of the IRT assumption tests were comparable to those obtained in the initial study for both the CH-NTW-S2 and the CH-ID1-S2 subtests. Both subtests were found to be essentially unidimensional. Consequently, the assumption of local independence was tenable for both subtests. The large ranges of the item point-biserial correlations indicated that the assumption of equal discrimination indices was not met for both subtests suggesting that the use of one-parameter item response model may not be appropriate. The assumption of non-speededness was tenable for both subtests. Lastly, the effect of guessing was found to be minimal for both subtests. The results of the assumption tests are presented in Tables 51 through 54 and Figures 23 and 24 in Appendix E.

The CH-NTW-S2 and CH-ID1-S2 items were calibrated separately using MULTILOG program (Thissen, 1991). The results of the item analyses using the four IRT models are presented in Appendix E. Tables 55 and 56 contain the results for the one-, two-, and three-parameter models and Tables 57 and 58 contain the results for the nominal response models, for the CH-NTW-S2 and CH-ID1-S2, respectively. The values of probabilities of selecting each item option across ability levels are reported in Table 59 for the CH-NTW-S2 and Table 60 for the CH-ID1-S2. Since the obtained item parameter estimates for both subtests in Sample 2 were comparable to their counterparts yielded by the four IRT models in Sample 1, only a brief summary of these results is presented here.

In the one-parameter model the mean item difficulty estimates were found be -0.92 for the CH-NTW-S2 and **–1.71** for the CH-ID1-S2. In the two-parameter model, the mean difficulty parameters were -0.86 and **–1.72**, and the mean discrimination parameters were, respectively, 0.91 and **0.85** for the CH-NTW-S2 and CH-ID1-S2. In the three-parameter model, the mean difficulty parameter estimates were -0.27 and **-0.81**, and the mean discrimination parameter estimates were 0.76 and **0.66** for the CH-NTW-S2 and CH-ID1-S2, respectively. The mean pseudo-guessing parameter was 0.24 for the CH-NTW-S2 and **0.29** for the CH-ID1-S2. Turning to the nominal response model, the mean values of the correct option popularity parameter estimates were 1.55 and **2.08**, and the mean discrimination parameter estimates for the correct option were 0.72 and **0.70** for the

CH-NTW-S2 and CH-ID1-S2, respectively. Similar to the initial study findings, the values of popularity parameter estimates for incorrect item alternatives indicate more equal distributions of incorrect responses across foils for the CH-NTW-S2 ($c_{ik\text{-}incorrect}$ ranged from 0.40 to –1.69) than for the CH-ID1-S2 ($c_{ik\text{-}incorrect}$ varied from **0.97** to **–2.71**).

Further, analysis of the CH-NTW-S2 and CH-ID1-S2 item option trace lines indicated their similarity to item option trace lines obtained in Sample 1 for both subtests. In the CH-NTW-S2 items, the distractors seemed to be approximately equally attractive to the lower and middle ability students. In case of the CH-ID1-S2 items, the item trace lines indicated that the majority of test-takers were able to eliminate one or more incorrect options (see Figures 25 through 28 in Appendix E).

Comparison of Ability Estimates

Following the sequence of procedures employed to analyze data from Sample 1, the comparisons of transformed ability scores were conducted simultaneously for the CH-NTW-S2 and for the CH-ID1-S2. The results of these comparisons for the high, middle, and low ability groups are reported in Tables 17, 18, and 19, respectively.

Table 17
Comparison of Ability Estimates Obtained from the CH-NTW-S2 and CH-ID1-S2 in the High Ability Group (N=1084)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and Standard Deviations	61.36 (4.43) **59.68 (4.46)**	61.74 (5.43) **60.10 (5.27)**	61.80 (5.43) **60.08 (4.84)**	61.61 (5.42) **60.22 (5.30)**	61.78 (5.35) **59.94 (4.70)**
θ_{nr} (number right)		0.99 **0.99**	0.98 **0.91**	0.96 **0.90**	0.95 **0.90**
θ_1 (1PL)	1.14 **0.94**		0.98 **0.91**	0.96 **0.90**	0.95 **0.90**
θ_2 (2PL)	1.55 **2.02**	1.17 **2.03**		0.99 **0.99**	0.98 **0.97**
θ_3 (3PL)	1.76 **2.34**	1.50 **2.28**	0.87 **0.86**		0.97 **0.97**
θ_4 (NRM)	1.88 **2.09**	1.73 **2.31**	1.15 **1.10**	1.40 **1.46**	

Note: Values in regular print are for the CH-NTW-S2 subtest and in **bold** print for the CH-ID1-S2 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

Table 18
Comparison of Ability Estimates Obtained from the CH-NTW-S2 and the CH-ID1-S2 in the Middle Ability Group (N=1266)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and Standard Deviations	50.86 (4.50) **51.32** **(6.10)**	50.33 (4.41) **50.92** **(6.25)**	50.25 (4.49) **51.09** **(6.22)**	50.43 (4.44) **50.91** **(6.22)**	50.31 (4.55) **51.18** **(6.26)**
θ_{nr} (number right)		0.99 **0.99**	0.95 **0.91**	0.93 **0.89**	0.91 **0.88**
θ_1 (1PL)	0.58 **0.59**		0.96 **0.91**	0.93 **0.89**	0.91 **0.88**
θ_2 (2PL)	1.49 **2.64**	1.34 **2.63**		0.99 **0.99**	0.95 **0.97**
θ_3 (3PL)	1.78 **2.91**	1.72 **2.85**	0.73 **0.74**		0.95 **0.96**
θ_4 (NRM)	2.00 **3.06**	1.90 **3.08**	1.40 **1.63**	1.57 **1.80**	

Note: Values in regular print are for the CH-NTW-S2 subtest and in **bold** print for the CH-ID1-S2 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

Table 19
Comparison of Ability Estimates for the CH-NTW-S2 and CH-ID1-S2, Low Ability Group (N=1094)

	θ_{nr} (number right)	θ_1 (1PL)	θ_2 (2PL)	θ_3 (3PL)	θ_4 (NRM)
Means and Standard Deviations	38.34 (6.06) **39.25** **(8.17)**	38.75 (5.48) **39.46** **(7.30)**	38.76 (5.31) **39.19** **(7.17)**	38.68 (5.63) **39.36** **(6.90)**	38.72 (5.30) **39.12** **(7.30)**
θ_{nr} (number right)		1.00 **0.99**	0.97 **0.93**	0.93 **0.91**	0.94 **0.91**
θ_1 (1PL)	0.71 **0.96**		0.97 **0.93**	0.93 **0.91**	0.94 **0.91**
θ_2 (2PL)	1.61 **2.94**	1.31 **2.61**		0.96 **0.99**	0.96 **0.97**
θ_3 (3PL)	2.32 **3.41**	2.16 **2.99**	1.61 **1.08**		0.94 **0.96**
θ_4 (NRM)	2.19 **3.41**	1.93 **3.13**	1.56 **1.80**	2.15 **2.11**	

Note: Values in regular print are for the CH-NTW-S2 subtest and in **bold** print for the CH-ID1-S2 subtest. Means and standard deviations are reported in the upper row. Correlations are above the principal diagonal; root mean square deviations between transformed scores are below the principal diagonal.

Comparison of the results provided in these tables with the results reported in Tables 13, 14, and 15 reveals that the results for the initial and replication samples were consistently the same. Therefore, only the set of results for the high ability group in Sample 2 is presented here, with the comparisons drawn with the results for the high ability group in the initial sample to highlight the stability of these results.

High Ability Group

As shown in Table 17, the means of transformed ability scores yielded by the five scoring methods were on average close to two points greater for CH-NTW-S2 than for the CH-ID1-S2. The means for the CH-NTW-S2 ranged from 61.36 to 61.80 (61.40 to 62.85 for the CH-NTW-S1), and from **59.68** to **60.22** for the CH-ID1-S1 (**59.54** to **60.02** for the CH-ID1-S1). The corresponding standard deviations varied from 4.43 to 5.43 for the CH-NTW-S2 (4.24 to 5.20 for the CH-NTW-S1) and from **4.46** to **5.30** for the CH-ID1-S2 (**4.55** to **5.38** for the CH-ID1-S1), indicating comparable variability of the scores yielded by the five models for the two sets of items.

As in the initial sample, the correlations between pairs of estimates yielded by the number right and one-parameter model, and between the scores yielded by the two- and three-parameter and nominal models were essentially the same for the CH-ID1-S2 and the CH-NTW-S2 subtest. These correlations were equal or greater than 0.97. The correlations between the remaining pairs of scores were found to be lower for the CH-ID1-S2 than for the CH-NTW-S2. These correlations varied from **0.90** to **0.91** for the CH-ID1-S2 (**0.86** to **0.89** for the CH-ID1-S1) and from 0.95 to 0.98 for the CH-NTW-S2 (0.94 to 0.97 for the CH-NTW-S1). These patterns and values are very similar to the patterns and values for the initial sample.

Taken together, although there were slight differences among the correlations obtained in the CH-NTW-S2 subtest, the students' ranked position will in effect be the same across the five scoring methods. In a case of the CH-ID1-S2, the high ability examinees' ranked position will be quite different when the number right or one-parameter scores are used in place of the two- and three-parameter and nominal model scores.

The pattern of agreements between pairs of estimates corresponded to the pattern observed for correlations. The closest agreements for both subtests were found between the scores produced by the number right and one-parameter models, and by the two- and three-parameter models. The values of the root mean square deviations for these pairs were, respectively, 1.14 and 0.87 for the CH-NTW-S2 and **0.94** and **0.86** for the CH-ID1-S2. Also for both subtests, there was slightly less agreement between the pairs of scores yielded by the two-parameter and nominal models, and by the three-parameter and nominal models. The root mean square deviations for these comparisons were found to be 1.15 and 1.40 for the CH-NTW-S2 and **1.10** and **1.46** for the CH-ID1-S2. In the initial study, the values of the root mean square deviations for the corresponding comparisons were 0.96, 1.09, 1.28, and 1.60 for the CH-NTW-S1 and **0.98**, **0.91**, **1.17**, and **1.55** for the CH-ID1-S1.

As for the initial sample, the lack of agreement between the number right and the two-, three-parameter, and nominal model scores and between the one-parameter and other IRT model scores was greater for the CH-ID1-S2 than for the CH-NTW-S2 (root mean square deviations **2.02** vs. 1.55, **2.34** vs. 1.76, **2.09** vs. 1.88, **2.03** vs. 1.17, **2.28** vs. 1.50, and **2.31** vs. 1.73). The corresponding root mean square deviations in the initial study ranged from **2.26** to **2.83** for the CH-ID1-S1 and from 1.18 to 1.91 for the CH-NTW-S1.

Given these findings, it appears that if a test contains items susceptible to the ID1 testwiseness strategy, for the high ability group, the number right or one-parameter models will likely produce somewhat different ability estimates when used in place of the two-, three-parameter, or nominal response models. In a case of a test without such items likely no significant differences in ability estimates will take place when ability scores yielded by the five scoring models are used in please of one another. This finding was again consistent for both samples.

Argument for not Combining Sample 1 and Sample 2

Following the argument for not combining samples in the social studies examination, and given that the Sample 1 and Sample 2 in the chemistry examination were both large, the initial and replication samples were not combined to provide the final set of results based on the larger size sample.

Discussion

As expected, the mean level of performance decreased with the decreasing ability in both samples for the CH-NTW and CH-ID1. The standard deviations were found to be from 1.25 to 2.25 T-scores higher for the CH-ID1 in the middle and low ability groups as compared to the corresponding values in the CH-NTW, while for the high ability group they were within 0.65 of a T - score point (cf. Table 13 with Table 17, Table 14 with Table 18, and Table 15 with Table 19).

The correlations between pairs of ability estimates yielded by the number right or one-parameter models and the two- and three-parameter and nominal models were consistently lower, particularly in the high and middle ability groups, for the CH-ID1 than for the CH-NTW (**0.85** to **0.93** vs. 0.90 to 0.98). Consequently, it appears that the greatest changes in the students' ranks will likely occur if the test contains items susceptible to the ID1 testwiseness strategy and when the number right or one-parameter model scores are used in place of the two-, three-parameter, or nominal response model scores. The differences in ranking will be slightly greater for the middle ability examines than for high and low ability examines (correlations of **0.85** to **0.89** vs. **0.86** and **0.91**). Some differences in ranking, particularly for middle and low ability examinees will also occur in a case of a test not containing items susceptible to testwiseness when the number right or one-parameter models scores are used in place of the three-parameter or nominal model scores. With two exceptions, these correlations ranged from 0.90 to 0.94. The exceptions were the correlations of 0.95 that were found for pairs of scores yielded by the number right or one-parameter models and the nominal model for the CH-NTW-S2 in the high ability group. The correlations between remaining pairs of scores were relatively high (0.95 to 1.00) for both subtests.

The patterns of root mean square deviations for the three ability groups also indicated differences between the two subtests. Congruent with the correlational findings, the root mean square deviations between ability estimates yielded by the number right or one-parameter models and the two-, three-parameter, or nominal response models were consistently greater for the CH-ID1 than for the CH-NTW (**2.02** to **3.54** vs. 1.17 to 2.39). These findings suggest that for all students, if the test contains items susceptible to the ID1 testwiseness strategy, the number right or one-parameter models will produce different ability estimates when used in place of the two- and three-parameter and nominal models. Some discrepancies in ability estimates will also likely occur for the CH-NTW when the number right scores are used in a place of the nominal models scores for the middle ability examinees (root mean square deviations of 2.00 and 2.11) or when number right or one-parameter model scores are used in place of nominal or three-parameter model scores in the low ability group (root mean square deviations from 2.12 to 2.39).

As expected, some differences were found between the items not susceptible to testwiseness and items susceptible to the ID1 testwiseness strategy. Since these differences were found only for the pairs of ability estimates obtained from the number right and the two- and three-parameter and nominal models and from the one-parameter model and the two- and three-parameter and nominal models, it appears that these differences are likely associated with the combination of two factors: item difficulty and scoring algorithm. First, the mean difficulty parameter for the CH-NTW was found to be higher than for the CH-ID1 (e.g., the mean b_i in the one-parameter model were -0.91 for the CH-NTW and **–1.71** for the CH-ID1; the summary of the mean item parameters across the five scoring models and two types of items for both samples is presented in Table 61 of Appendix E). Second, the presence of the discrimination parameter in the two-, three-parameter and nominal model algorithms seemed to increase the magnitude of differences between pairs of ability estimates obtained from these models and the estimates obtained from the number right and one parameter models that do not include a discrimination parameter. Taken together, it appears that that if the test contains relatively easy items and when the ability scores obtained from the models containing a discrimination parameter are used in place of scores yielded by the models not including an item discrimination parameter in their scoring algorithms, the differences in ability estimates will be greater than the differences that will occur only when one of these two factors is present. This conclusion is supported by the pattern and magnitude of root mean square deviations and correlation between the remaining pairs of ability estimates, which were similar for both the CH-NTW and CH-ID1. For example, no observed difference between the subtests in terms of differences between ability scores yielded by the number right and one-parameter models can be due to the fact that despite the differences in subtest difficulty, both models do not contain an item discrimination parameter. Similarly, no difference between subtests in ability scores yielded by the two-parameter and three-parameter and nominal models can be due to the similarity of these models containing an item discrimination parameter.

Further, contrary to what was anticipated, the use of information from incorrect responses appeared not to contribute to the differences between the CH-NTW and the CH-ID1. Despite the fact that differences in the incorrect item option trace lines were found for the CH-NTW as compared to the CH-ID1, it appears that the influence of information from incorrect responses on ability estimates obtained from these subtests is weak. The discrepancies between the pairs of scores yielded by nominal response model and the two- and three-parameter models were consistent for both subtests. Taken together, it appears that the information yielded by correct item option may suppress the information obtained from the incorrect item responses resulting, in the case of the chemistry examination, in no contribution of information from wrong responses to the differences between the CH-NTW and CH-ID1.

A question might be asked whether or not the observed differences between the CH-NTW and CH-ID1 might be attributable to the different number of items, 21 and 14, included in each subtest. However, if this was a factor, all the results should have been affected. Since they were not, the differences in ability estimates yielded by the number right or one-parameter model and the two- and three-parameter and nominal models appear not to be attributable to the number of items.

CHAPTER VI – COMPARISON ACROSS SUBJECT AREAS

This chapter provides comparison and discussion of the results obtained from the social studies and chemistry examination data analyses. As in the previous two chapters, the results for the non-susceptible to testwiseness items are reported in regular print, and the results for the items susceptible to the ID1 testwiseness strategy are reported in **bold** print. Given the similarity of the results obtained for the initial sample and the replication sample, the comparisons reported below are for the initial sample.

Comparison of Item Parameters

The results of item analyses conducted for the two types of items in the social studies and chemistry examinations for the total sample indicated that that there was less difference between the SS-NTW and the SS-ID1 in terms of their difficulty, than between the CH-NTW and the CH-ID1. For example, the mean b_i parameter estimates determined using the two-parameter model were –0.80 for the SS-NTW and **-1.06** for the SS-ID1, but for the chemistry test, the corresponding b_i parameter estimates were -0.82 for the CH-NTW and **–1.76** for the CH-ID1. The mean discriminations were similar across both subtests in both subject areas (e.g., a_i in the two-parameter model were 0.93 for the SS-NTW, **0.89** for the SS-ID1, 0.90 for the CH-NTW, and **0.88** for the CH-ID1). As expected, given the prior normal distribution of the pseudo-guessing parameter, the mean values of c_i were similar across all subtests (e.g., c_i were 0.23 for the SS-NTW and SS-ID1, 0.24 for the CH-NTW, and **0.26** for the CH-ID1).

Lastly, the distribution of wrong responses was more uniform across item foils for items not susceptible to testwiseness than across foils for items susceptible to the ID1 testwiseness strategy in both examinations. Across all items, 5.7% and 15.9% of item foils were selected by less than five percent of examinees in the SS-NTW and CH-NTW, respectively. The corresponding percentages were **37.7%** and **31.0%** for the SS-ID1 and CH-ID1, respectively. The analysis of item option trace lines supported these findings. In both the SS-NTW and the CH-NTW items, the distractors appeared to be fairly equally attractive for examinees at a given proficiency level. In the case of the SS-ID1 and the CH-ID1 items, the trace lines indicated either very low probability of selecting an absurd option by the majority of students or its attractiveness to low ability examinees only.

Comparison of Ability Estimates

The comparisons of the five sets of ability estimates were conducted for the examinees at, defined earlier, three ability levels: high, middle, and low. Overall, the patterns of means and standard deviations of the transformed ability scores, and the patterns of correlations and root mean square deviations between pairs of ability estimates were similar for both the social studies examination and the chemistry examination. However, differences between the two examinations were found in regard to the magnitude of some of the correlations and root mean square deviations for the two types of items and the three ability groups.

Comparison of Subtests

As expected, the mean level of performance decreased with the decreasing ability for both examinations. Further, for both examinations, the means were consistent within each ability group and the two subtests. For example, the mean ability T-scores obtained from the five scoring models in the high ability group ranged from 61.88 to 62.72 for the SS-NTW, from **61.10** to **61.72** for the SS-ID1, from 61.40 to 61.85 for the CH-NTW, and from **59.54** to **60.02** for the CH-ID1 subtest. The standard deviations of scores yielded by the five methods for both subtests were found to be comparable for the high ability groups in both examinations, but were on average one to two points lower for the social studies examination than for the chemistry examination in the middle and low ability groups. For example, in the high ability group, the standard deviations varied from 3.56 to 5.01 for the SS-NTW and from **3.96** to **5.15** for the SS-ID1 subtest, and from 4.24 to 5.20 for CH-NTW and from **4.55** to **5.38** for the CH-ID1 subtest. In contrast, for the middle ability group the standard deviations ranged form 3.59 to 3.77 for the SS-NTW and from **4.75** to **4.92** for the SS-ID1. The corresponding standard deviations for the chemistry examination varied from 4.42 to 4.66 for the CH-NTW and from **5.64** to **6.04** for CH-ID1.

With one exception, the correlations between pairs of scores yielded by the number right and one-parameter models, and between the scores yielded by the two-parameter, three-parameter and nominal models were equal to or greater than 0.95 for the two types of items in both examinations and for all ability groups. A lower correlation, 0.94, was found for the pairs of ability estimates yielded by the three-parameter and nominal models for the CH-NTW in the low ability group. The correlations between the remaining pairs of scores differed by subtest, ability group, and subject area.

The correlations between the pairs of scores yielded by the number right or one-parameter models and by the two-parameter, three-parameter, and nominal models were similar for both subtests in the social studies examination (0.91 to 0.98 for the SS-NTW vs. **0.92** to **0.97** for the SS-ID1). The corresponding correlations in the chemistry examination were found to be considerably higher for the subtest of non-susceptible items than for the subtest of items susceptible to the ID1 testwiseness strategy (0.90 to 0.97 for the CH-NTW vs. **0.85** to **0.93** for the CH-ID1). Given this finding it appears that greater differences in ranking will likely occur for the CH-ID1 than for the other subtests when the number right or one-parameter model scores are used in place of the other IRT model scores in each of the ability groups.

There will also be some differences in ranking for the SS-NTW, SS-ID1, and CH-NTW when the number right or one-parameter scores are used in place of the nominal scores in each of the three ability groups. The correlations for these pairs of scores were less than 0.95. Likewise, some differences in ranking will likely take place when the number right or one-parameter scores are used in place of the three-parameter model scores for the SS-NTW and SS-ID1 for in middle ability group (correlation of 0.94) and for the CH-NTW in the middle and low ability groups (correlations from 0.92 to 0.93).

The patterns of root mean square deviations reflected the patterns observed for the correlations. The agreements between the pairs of ability scores yielded by the number right and one-parameter models, and by the two- and three-parameter and nominal models were similar for both subtests on the social studies and chemistry examinations. For example, in the middle ability group, the root mean square deviations for the number

right and one-parameter model scores were 0.81 for the SS-NTW, **0.69** for the SS-ID1, 0.55 for the CH-NTW, and **0.59** for the CH-ID1. Similarly, in the middle ability group the root mean square deviations for the three-parameter and nominal model scores were 1.18 for the SS-NTW, **1.23** for the SS-ID1, 1.64 for the CH-NTW, and **1.69** for the CH-ID1-S1.

The agreements between the pairs of scores yielded by the number right or one-parameter models and the two- and three-parameter and nominal models were found to be similar for the two subtests in the social studies examination. The root mean square deviations for these comparisons varied from 0.99 to 2.18 for the SS-NTW and from **1.41** to **2.43** for the SS-ID1. In the chemistry examination, the lack of agreement between the corresponding pairs of scores was found to be greater for the subtest of items susceptible to the ID1 testwiseness strategy than for the subtest of items not susceptible to testwiseness. The root mean square deviations for these comparisons ranged from 1.18 to 2.36 for the CH-NTW and from **2.26** to **3.54** for the CH-ID1. Given these findings it appears that the greater absolute differences in ability for all students will likely occur for the CH-ID1 than for other subtests (CH-NTW, SS-NTW, and SS-ID1) when the number right or one-parameter model scores are used in place of the other IRT model scores.

Some discrepancies in ability estimates will also occur for the SS-NTW, the SS-ID1, and the CH-NTW when the number right scores are used in place of the nominal scores in each ability group, and for the CH-NTW when the number right or one-parameter scores are used in place of the three-parameter model scores, and when the three-parameter model scores are used in place of the nominal model scores for the low ability examinees. The root mean square deviations in each of these cases exceeded 2.00.

Comparison of Models

Based upon the results obtained in this study, the five scoring models can be grouped into two sets. The first set includes the number right and one-parameter scoring models. The second set contains the two-parameter, three-parameter, and nominal response scoring models. The distinguishing difference between the models in the two sets appears to be the presence of the discrimination or a_i parameter found in the models in the second set.

The correlations between scores yielded by the scoring models were higher within the two sets than across the two sets. For example, the correlations between pairs of scores yielded by the two- and three parameter models and the number right and three-parameter models in the high ability group were, respectively, 0.99 vs. 0.96 for the SS-NTW, **0.99** vs. **0.95** for the SS-ID1, 0.98 vs. 0.96 for the CH-NTW, and **0.99** vs. **0.86** for the CH-ID1.

Likewise, the root mean square deviations between ability estimates yielded by the scoring models were smaller within the two sets than across the sets. For example, the root mean square deviations for pairs of ability yielded by the two- and three-parameter models, and the number right and three-parameter models in the high ability group were, respectively, 0.77 vs. 1.92 for the SS-NTW, **0.55** vs. **1.94** for the SS-ID1, 0.96 vs. 1.65 for the CH-NTW, and **0.91** vs. **2.75** for the CH-ID1.

Also for both subtests on both diploma examinations, the differences in ability estimates yielded by the three-parameter and nominal response models were consistently greater in the low ability group than in the middle and high ability groups. The root mean

square deviations for these comparisons were 1.70 vs. 1.18 and 1.30 for the SS-NTW-S1, **1.69** vs. **1.23** and **1.10** for the SS-ID1-S1, 2.14 vs. 1.64 and 1.60 for the CH-NTW-S1, and **2.06** vs. **1.69** and **1.55** for the CH-ID1-S1. The differences between ability estimates yielded by the three-parameter and nominal response models may likely be due to the presence of the pseudo-guessing parameter in the three-parameter model and lack of such parameter in the nominal response model. The observation that these differences were greater in the low ability group may be associated with the combined influence of information from wrong responses and guessing behavior of low ability examinees on ability estimates.

Precision of measurement. The differences between pairs of ability estimates yielded by the nominal response model and the two- and three-parameter models for dichotomously scored items was less than expected. However, in agreement with research results obtained by Bock (1972, 1997), De Ayala (1989, 1992), Thissen (1976), and Thissen and Steinberg (1984), it was found that the mean standard error of ability was smaller for the low ability group when the nominal response model was used in place of the two- and three-parameter models. For example, in case of the SS-NTW, the mean standard error of θ for the ability scores produced by the nominal response model was 0.32 while the standard error of θ for the two-, and three-parameter models were respectively, 0.35, 0.43. The corresponding mean standard errors of θ for the SS-ID1 subtest were **0.40** vs. **0.43**, and **0.49**. Similar pattern of standard errors of θ was found for the low ability examinees in case of the chemistry examination (0.43 vs. 0.46, and 0.53 for the CH-NTW and **0.53** vs. **0.55** and **0.57** for the CH-ID1).

Less difference between the mean standard errors of ability scores yielded by the nominal and the two-parameter models was found for the middle and high ability group (e.g., 0.37, and 0.38 for the SS-NTW, middle ability group; and 0.57, and 0.58 for the CH-NTW, high ability group). The mean standard errors of ability scores yielded by the three-parameter models in the middle and high ability groups were smaller than the mean standard errors of scores yielded by the two-parameter and nominal models (e.g.; **0.44** vs. **0.47**, and **0.46** for the SS-ID1, middle ability group; and **0.48** vs. **0.58**, and **0.57** for the CH-NTW, high ability group). These findings are consistent with previous research and suggest that different models may yield different amounts of error in θ for examinees at different ability levels. It appears that while the nominal response model may be appropriate to use in order to reduce the amount of error in θ for low ability examinees, the three-parameter model may be more suitable to use for estimation of the high ability examinees' scores (Bock, 1972; De Ayala, 1989; Thissen and Steinberg, 1984).

Discussion

It was expected that items with one or more options that could be easily eliminated by the majority of examinees would be on average easier than items with the properly functioning foils. The expected differences between the subtest of items not susceptible to testwiseness and the subtest of items susceptible to the ID1 testwiseness strategy were not found for the social studies examination but were observed for the chemistry examination.

One possible explanation for this finding is associated with the quality of item distractors. It might have been that the properly functioning distractors in the social

studies ID1 susceptible items were quite attractive to examinees and even after the elimination of absurd options, the probability of selecting the correct response did not increase. On the other hand, the chemistry items susceptible to the ID1 testwiseness strategy could contain less attractive properly functioning options leading to the situation in which an examinee, after elimination of an absurd option, has a better chance of selecting the correct one. A plausible explanation for this is associated with the nature of the content area being tested. Examination of the item content on the social studies test suggests that interpretation of information and expression of one's opinions are important when responding to an item (e.g. questions calling for a "best answer"). Such items, even if they contain an absurd option, will likely have quite attractive distractors examinees are drawn towards. In contrast, the items on the chemistry tests measure mostly scientific facts and as such are less open to an examinee's interpretation. Such items often tend to have a more clear correct answer and therefore less attractive distractors. Consequently, elimination of one of the distractors as absurd would likely improve an examinee's chance to answer a chemistry item correctly.

Further, in the case of chemistry examination, only the discrepancies between the number right or one-parameter model scores and the two-, three-parameter, and nominal model scores were found to be greater for the subtest of items susceptible to the ID1 testwiseness strategy than for the items not susceptible to testwiseness. The differences between the remaining pairs of scores were comparable for both subtests. Consequently, it appears that these differences are associated with both subtest difficulty and scoring algorithm. It seems that if a test contains relatively easy items (i.e., CH-ID1) and if the ability estimates yielded by the model containing a discrimination parameter are used in place of scores yielded by the model not including an item discrimination parameter in its scoring algorithm, the differences in ability estimates will be greater than the differences that will occur when only one of these two elements is present.

Contrary to what was anticipated, the use of information from incorrect responses appeared not to contribute to the differences between the subtest of items not susceptible to testwiseness and the subtest of items susceptible to the ID1 testwiseness strategy. Despite the fact that differences in the incorrect option trace lines were found for the items susceptible to the ID1 testwiseness strategy as compared to the non-susceptible to testwiseness items, it appears that the influence of information from incorrect responses on ability estimates obtained from these subtests is weak. The discrepancies between the pairs of scores yielded by the nominal response model and the two- and three-parameter models were found to be consistent for both subtests.

CHAPTER VII – CONCLUSIONS AND IMPLICATIONS

A brief summary of the background of the study, the purpose of the study, and the procedures followed to address this purpose are presented in the beginning of this chapter. This summary is followed by a summary of the results. The limitations of the study are presented next, followed by the conclusions drawn in light of these limitations. The chapter concludes with recommendations for practice and future research in the areas of testwiseness and scoring multiple-choice items.

Summary of the Study

Background of the Study

Different scoring methods are used to score multiple-choice items and estimate examinees' ability. Among them, the number right model and the one-, two-, and three-parameter IRT models are used most often. These models, appropriate for dichotomously scored items, are often perceived as failing to incorporate examinees' partial knowledge in the process of ability estimation. Since many researchers (De Ayala, 1989, 1993; Lord, 1980; Tatsuoka, 1983) have found that examinees who do not possess the necessary knowledge to answer an item do not simply randomly select their answers but rather use partial knowledge about the item content to select their response, the use of the nominal response model that incorporates information from correct as well as incorrect responses in estimation of ability was proposed (Bock, 1972, 1997; Thissen et al., 1989). It has been shown that using this additional information yields more precise ability estimates, particularly for the examinees whose scores fall below the mean ability level, where incorrect responses occur more frequently (e.g., Bock, 1972, 1997; Levine & Drasgow, 1983; Thissen, 1976; Thissen et al, 1989).

Another factor that has been known to affect ability estimates is testwiseness. It has been found that examinees who do not know the correct answer to a test question but possess both partial knowledge and test-taking skills have a greater probability of selecting a correct response than their peers who possess partial knowledge but are not testwise or those who have knowledge of testwiseness principles but have low partial knowledge (e.g., Diamond & Evans, 1972; Millman, 1966; Rogers & Bateson, 1991b; Rogers & Yang, 1996; Towns & Robinson, 1993).

The present study was motivated by the need for research involving the application of the nominal response model in scoring testwise multiple-choice items to assess its utility for obtaining ability estimates.

Purpose of the Study

The purpose of this study was to investigate the comparability between ability estimates yielded by the number right model, the one-, two-, and three-parameter item response binary models, and the nominal response model using two types of items: items susceptible to testwiseness and items not susceptible to testwiseness. Since several researchers (Bock, 1972; Levine & Drasgow, 1983; Thissen, 1976; Thissen & Steinberg, 1984) found that examinees' ability level may likely be a factor associated with selection

of a particular item option, comparability of these estimates was investigated separately for the high, middle, and low ability examinees.

It was hypothesized that the differences between ability estimates yielded by the five scoring models would be greater for testwise susceptible items than for items not susceptible to testwiseness. It was expected that using the nominal response models would likely intensify these differences for the examinees at the lower and middle proficiency level. It was also anticipated that the differences among the ability estimates for the three ability groups would be consistent across the social studies examination (humanities) and the chemistry examination (science).

Method

Data Analyzed

Two data sets were used. Each set consisted of responses of high school students to the multiple-choice items contained in school-leaving examinations of social studies and chemistry. These "high stakes" tests count 50% of a student's final grade and are intended for students who are planning or wish to leave open the opportunity to pursue some form of tertiary education (Alberta Education, 1999c; 1999e).

Prior to conducting data analyses, two panels of experts, one for each subject area, were formed to analyze the items in each test for the presence of testwise cues. The raters used their content area expertise and the results of the classical item analysis to identify items susceptible to testwiseness. For both panels, consensus agreement was reached for each item. Based on item classification, two subtests were identified within each diploma examination for further analyses. For the Social Studies 30 Diploma Examination, one subtest consisted of 35 items not susceptible to testwiseness (SS-NTW) and the second contained 23 items susceptible to the ID1 testwiseness strategy – "eliminate option(s) that are known to be incorrect and choose form among the remaining alternatives" (SS-ID1). For the Chemistry 30 Diploma Examination, one subtest contained 21 items that were not susceptible to testwiseness (CH-NTW) and the other subtest contained 14 items sensitive to the presence of the ID1 testwise strategy (CH-ID1). The numbers of items sensitive to other testwiseness strategies in both diploma examinations were not sufficient to obtain stable ability estimates.

The total numbers of students that completed the social studies and chemistry examinations were 10,905 and 8,594, respectively. Two samples of 4,000 students each were randomly drawn without replacement from the total number of examinees for each test. The initial data analyses were performed using one sample for each test. In order to examine the stability of the results obtained from the initial analyses, replication studies were conducted using the second sample for each test. The examinees' total raw scores on the social studies and chemistry tests were used to identify members of the high, middle and low ability groups. Kelley's (1927) guidelines were used for this purpose.

Analyses Conducted

The psychometric characteristics of the items not susceptible to testwiseness and items susceptible to the ID1 testwiseness strategy in each diploma examination were examined using the five scoring models. The results of these analyses allowed for comparison of the two types of items and were expected to contribute to the explanation

of the ability estimation results. First, the Alberta Education standards were used to evaluate non-susceptible to testwiseness items and items susceptible to the ID1 testwiseness strategy using conventional item analysis. Next, following confirmation that the assumptions of unidimensionality, local independence, equal discrimination, non-speededness, and non-guessing underlying the IRT theory were tenable, item analyses using the one-, two-, three-parameter models and the nominal response model were conducted. In addition, item option trace lines for non-susceptible and susceptible to the ID1 testwiseness strategy items produced using the nominal response model were visually inspected to determine if there were differences between the options of the two types of items in each of the two subject areas. All item analyses were conducted using the full samples of 4000 examinees.

Ability estimates obtained from each scoring model were compared with each other for each subtest within each of the two subject areas. Prior to the comparative analyses, the ability estimates were converted to T-scores ($\mu = 50$, $\sigma = 10$). Correlations among the ability estimates were compared to examine if the students were ranked differently by the different scoring methods and the root mean square deviations among pairs of ability estimates were calculated to determine if the scores for examinees were equal in an absolute sense. These comparisons were conducted separately for the high, middle, and low ability examinees.

Results and Discussion

The results of item and ability estimation obtained in the initial and replication study were essentially the same for both examinations. Consequently, the results presented and discussed below are based on the analysis for the initial sample. The results for the items not susceptible to testwiseness are presented in regular print and the results for the items susceptible to the ID1 testwiseness strategy are reported in **bold** print.

Item Analysis

Classical item analysis. The results of the conventional item analyses conducted for the subtests of items not susceptible to testwiseness and items susceptible to the ID1 strategy of testwiseness indicated that that these two subtests differed less in terms of their difficulty in the case of the social studies examination than in the case of the chemistry examination. For example, the mean *p*-values were 0.65 for the SS-NTW and **0.68** for the SS-ID1, and 0.65 for the CH-NTW and **0.75** for the CH-ID1. For both examinations, the two subtests were comparable in terms of their items discrimination indices. Also for both examinations, the distribution of wrong responses was more uniform across item foils of items not susceptible to testwiseness than of items susceptible to the ID1 testwiseness strategy.

Item response item analysis. The results of the one-, two-, three-parameter and nominal model item analyses were consistent with the results of the classical item analysis. The subtest of items not susceptible to testwiseness and the subtest of items susceptible to the ID1 strategy of testwiseness were found to be less different in terms of their difficulty in the social studies examination than in the chemistry examination. For example, the mean b_i parameters in the two-parameter model were –0.80 for the SS-NTW and **-1.06** for the SS-ID1, and -0.82 for the CH-NTW and **–1.76** for the CH-ID1. The

mean discrimination parameters and the means of the pseudo-guessing parameters were similar across both subtests in both subject areas.

The examination of item option trace lines indicated that, for the non-susceptible to testwiseness items on both diploma examinations the distractors appeared to be fairly equally attractive for examinees at a given proficiency level. In the case of both the social studies and chemistry items susceptible to the ID1 testwiseness strategy the trace lines indicated either very low probability of selecting an absurd option by the majority of examinees or its attractiveness to low ability examinees only.

Comparison of Ability Estimates

As expected, the mean level of performance decreased with the decreasing ability in both examinations. Further, at each ability level, the means for each of the five sets of scores were consistent across two types of items for both the social studies examination and the chemistry examination.

Social Studies 30. The patterns and magnitude of correlations between the pairs of ability estimates yielded by the five models were similar for both non-susceptible to testwiseness and susceptible to the ID1 testwiseness strategy subtests in the social studies examination. In both subtests, the lowest correlations (0.91 to 0.94) in each ability group were for the pairs of scores yielded by the nominal response model and the number right or one-parameter models. Correlations for the pairs of scores yielded by the three-parameter model and the number right or one- parameter models were found to be lower in the middle ability group than in the high and low ability groups (0.94 vs. 0.95, 0.96, and 0.97). The remaining correlations were relatively high (0.95 to 1.00).

The largest (greater than 2.00) root mean square deviations were found for scores yielded by the number right and nominal response models for both subtests and in all ability groups. The root mean square deviations obtained for the remaining pairs of scores for both subtests were on average smaller. Although the pattern of differences between ability estimates yielded by the number right or one-parameter models and the remaining IRT models were slightly greater for the SS-ID1 than for the SS-NTW (**1.44** to **2.43** vs. 0.99 to 2.06) for the middle and low ability examinees, the differences between the two subtests were relatively small and can be considered non-significant.

Chemistry 30. In the chemistry examination, the pattern and magnitudes of the correlations were different for the subtest containing not susceptible to testwiseness items and the subtest containing items susceptible to the ID1 testwiseness strategy. For all three ability groups, the correlations between pairs of ability estimates yielded by the number right or one-parameter models and other IRT models were found to be considerably lower in the subtest of items susceptible to the ID1 testwiseness strategy than in the subtest of items not susceptible to testwiseness (**0.85** to **0.93** vs. 0.90 to 0.97). The correlations between remaining pairs of scores were greater than 0.95 for both subtests.

The patterns of root mean square deviations for the three ability groups also indicated differences between the two subtests. The root mean square deviations for the number right or one-parameter model scores and the other IRT model scores were consistently greater for the subtest of items susceptible to the ID1 testwiseness strategy than for the subtest of items not susceptible to testwiseness (**2.26** to **3.54** vs. 1.18 to 2.39). The agreements between the remaining pairs of scores were found to be similar for both subtests (**0.59** to **2.06** for the CH-ID1 and 0.55 to 2.14 for the CH-NTW).

Summary. The expected differences between the subtest of items not susceptible to testwiseness and the subtest of items susceptible to the ID1 testwiseness strategy were not found for the social studies examination but were observed for the chemistry examination. The results suggest that, for the chemistry examination, greater changes in examinees' ranks and their absolute ability scores will occur for the test containing items susceptible to the ID1 testwiseness strategy than for the test not containing such items when the number right or one-parameter model scores are used in place of the two-parameter, three-parameter, and nominal response model scores. The expected effect of information from incorrect responses on ability estimates obtained from the two types of item was not found. The differences between the pairs of scores yielded by nominal response model and by the two- and three-parameter models for dichotomously scored items were consistent for both subtests.

Further, for both subject areas and regardless of an item susceptibility to testwiseness, the differences between ability estimates yielded by the number right or one-parameter models and the two- and three-parameter and nominal models were greater than the differences between the remaining pairs of scores across all ability groups. Also, the differences between ability scores yielded by the three-parameter and nominal response model were found to be greater for low ability examinees than for middle and high ability students.

Limitations of the Study

A limitation of the present study is that only items susceptible to the ID1 testwiseness strategy were considered. The numbers of items susceptible to other testwiseness strategies were not sufficient to obtain stable ability estimates. For example, only 3 items on the social studies examination and no items on the chemistry examination were found to be susceptible to the ID2 strategy, and only 2 items on the social studies examination and 4 items on the chemistry examination were susceptible to the ID3 testwiseness strategy. These findings are similar to the previous research findings concerning the presence of testwise susceptible items on Alberta Education diploma examinations administered in June 1992. For instance, Rogers and Wilson (1993) found 2 items susceptible to the ID2 and 5 items susceptible to the ID3 strategy on the June 1992 social studies examination. In the case of the June 1992 chemistry examination, 3 items susceptible to the ID2 strategy and 1 item susceptible to the ID3 testwiseness strategy were identified.

Lack of sufficient number of items susceptible to other testwiseness strategies prevented the investigation of whether or not such items would produce different patterns of differences between ability estimates yielded by the five scoring models as compared to the pattern of differences between ability scores obtained from the subtests of items not susceptible to testwiseness and items susceptible to the ID1 testwiseness strategy.

Conclusions

The results of item analysis for the subtest of items susceptible to the ID1 testwiseness strategy and the subtest of items not susceptible to testwiseness were not consistent across the two subject areas. In the case of the chemistry examination, the

examinees earned a higher score on the subtest of items with one or more options that could be eliminated (testwiseness strategy ID1) than on the subtest of items not susceptible to testwiseness. In the case of the social studies examination the difference between the two subtests in terms of their difficulty was smaller.

In agreement with this finding, the differences among ability estimates yielded by the five scoring for the subtest of items susceptible to the ID1 testwiseness strategy and the subtest of items not susceptible to testwiseness were also not consistent across the two examinations. The discrepancies in ability estimates that were found for the pairs of scores yielded by the number right or one-parameter models and by the two- and three-parameter and nominal models were greater, particularly for the middle and low ability examinees, for the subtest of items susceptible to the ID1 strategy of testwiseness than for the subset of items not susceptible to testwiseness for the chemistry examination. In the case of the social studies examination differences between the two subtests were smaller.

One possible reason for the different behavior of items susceptible to the ID1 strategy of testwiseness on the social studies and chemistry examinations may be related to the quality of distractors on both examinations and the nature of the content being tested. It is possible that in the situation when interpretation of information and employing one's values is important when responding to an item (e.g., social studies) the ID1 strategy of testwiseness is not working as expected. In such case, the non-absurd options, including the correct option, appear to be more equally attractive to the examinees than when interpretation and values are less important. On the other hand, when the items measure scientific knowledge (i.e., chemistry), they are more likely to have a more unambiguous correct answer and therefore less attractive distractors. In such cases, employing the ID1 testwiseness strategy and elimination of the absurd option would likely improve an examinee's chance to answer an item correctly.

The magnitudes of differences between pairs of ability estimates yielded by five scoring models suggested that these models could be grouped in two sets. The first set contains the number right and one-parameter models and the second set includes the two-parameter, three-parameter, and nominal response models. The difference between the models in the two sets seems to be associated with the presence of the discrimination parameter found in the models in the second set. It appears that the models that do not contain the discrimination parameter will likely produce different ability scores as compared to the scores obtained from the models that include a discrimination parameter. In addition, low ability examinees will likely receive a different score when the three-parameter model is used in a place of the nominal response model. These discrepancies may be due to the mixed effect of information from incorrect responses and guessing behavior of low ability examinees (Hambleton, et al., 1989; Thissen & Steinberg, 1984, 1997).

To summarize, it appears that the quality of item distractors or the nature of the content area being tested will affect the test difficulty and determine whether or not item susceptibility to the ID1 testwiseness strategy operates as expected. Test difficulty in turn will influence agreements between ability scores yielded by different scoring methods. It seems that relatively easy tests will result in less agreement, especially for the middle and low ability examinees, for pairs of ability estimates yielded by different scoring models tests of higher difficulty. These conclusions have a few implications with respect to development and use of tests in educational setting.

Implications for Testing Practices

As pointed by Lord (1952), items of medium difficulty (*p*-value in the neighborhood of 0.50) tend to have a desirable higher possible maximum discrimination index than very easy or very difficult items. Such items are most appropriate for norm-reference testing. Given that the greatest discrepancies between pairs of ability estimates yielded by different scoring models were obtained for the subtest of relatively easy items (e.g., $\overline{X} = 75\%$ in case of CH-ID1) as compared to more difficult subtests (e.g., $\overline{X} =$ 65.29%, 67.57% and 64.67% in case of the SS-NTW, SS-ID1, and CH-NTW, respectively), it appears that by including fewer easy items in a test it may be possible to increase the agreement between different scoring models.

The problem from the test developer's point of view is that it is usually unknown in advance how examinees will respond to an item. Extensive field testing and employing several methods to analyze items would likely improve item quality. Although conventional item analysis based upon *p*-values and point-biserials for each item option may be used for these analyses, the item response models, particularly the nominal response model may provide some additional advantages. The IRT models permit items to vary in terms of their difficulty, discrimination, or guessing parameter and allow for calculation of probabilities of selecting the correct response across different ability levels. Such information about an item can be used in selection of best items, that is items with desirable properties, for the test. In addition, the nominal response model allows for calculation and graphical representation of the distribution of probabilities of selecting incorrect response options across ability levels. Although information from incorrect responses was found to be a minor factor affecting ability estimates, examination of the trace lines associated with each alternative on an item may be helpful in detection of specific flaws of the item (i.e. options that are easily eliminated by the majority of students). Consequently, the poor performance of the item options can be observed and such an item can either be modified or eliminated in the process of test development. Examination of trace lines and parameters associated with each alternative on a multiple-choice item may lead to the enhancement of a test developer's knowledge about and understanding of the properties of both the correct option and foils. Based on the nominal response model item analysis, distractors can be designed to be equivalent for an item that is to be scored dichotomously, which in turn would be expected to result in a more adequate meeting the specific needs and properties of a test (Bock, 1972; Thissen et al., 1989).

If item response theory is employed to obtain parameter estimates, a question arises which model may be most appropriate for a purpose of scoring multiple-choice items and individuals. Hambleton and Swaminathan (1985) recommended assessment of the appropriateness of different IRT models for various applications using three types of evidence: validity of the assumptions of the model for the data; extent to which the expected properties of the model are obtained; and the accuracy of model prediction using real data and simulated test data. Conducting several analyses designed to detect potential misfits of the models can be helpful in selecting a model that bests fits the data and the nature of the intended applications (Hambleton, et al., 1991).

It should be noted, however, that the item response theory requires large samples for parameter estimation purposes. While such samples are available in large testing

programs, they are not available at the classroom level. Thus, only the conventional method of item analysis and scoring students is appropriate for classroom testing (Hambleton et al., 1989; Lord, 1980).

Implications for Future Research

There are several issues for future research that can be suggested based on the findings of this study. First, interviewing grade 12 students to determine the strategies they used when responding to multiple-choice items contained in the diploma examinations is needed. Analysis of the "think aloud" protocols would provide some insight in examinees' thinking processes at the item level and help to understand the cognitive errors students are likely to commit (Rogers & Bateson, 1991b; Rogers and Wilson, 1993). Having such information would be helpful in understanding why some of the items susceptible to the ID1 strategy of testwiseness are similar in their difficulties to items not susceptible to testwiseness (e.g., social studies examination) while other ID1 testwise susceptible items are much easier than non-susceptible items (e.g., chemistry examination).

Further, given information on students' application of testwiseness strategies, it would be possible to identify students who are testwise and those who are not in order to examine comparability of ability estimates yielded by different scoring models for these two groups of test-takers. Examining differential functioning of items susceptible to testwiseness using testwise examinees as a focus group and non-testwise students as a reference group would help determine whether or not items susceptible to testwiseness display item bias.

Rogers & Harley (1999) found that the influence of testwiseness due to the presence of testwise susceptible items in high stake tests may be lessened if 3-option items are used instead of 4-option items. The psychometric properties of the 3-option and 4-option tests (e.g., reliability) were found to be fairly equivalent. Given these findings, additional studies in which item and test level analyses are conducted are needed in a variety of subject areas to clarify the changes that may be brought about by reduction of number of options in multiple-choice items from four to three (Rogers & Harley, 1999)

Further, only items susceptible to the ID1 strategy of testwiseness were considered in this study. As Ndalichako (1998) found, ability estimates obtained form a subtest of items susceptible to the ID2 and ID3 strategies of testwiseness contained in the Test of Testwiseness (Rogers & Wilson, 1993) yielded by the three-parameter and finite state theory models were different from the ability estimates produced by the number right and one- and two-parameter models. Given these findings, it appears that performance of different scoring models, especially the nominal response model, needs to be examined in the presence of items susceptible to different strategies of testwiseness.

As stated on page 34, in the absence of statistical guidelines, the criteria for interpretation of correlation coefficient and root mean square deviation magnitudes were arbitrarily chosen. Further studies are needed to establish criteria for interpretation of correlation coefficients hypothesized to be 1.00, but lower than 1.00 and for interpretation of root mean square deviations hypothesized to be 0.00, but greater than 0.00 in the presence of large sample size such as one used in this study.

In order to determine which model provides most accurate results, simulation studies are needed. Computer generated parameters of items with characteristics similar to those yielded by the real data and applied to a simulated population with known distribution of θ would provide information on the accuracy of the ability estimates yielded by different scoring models.

Lastly, only two subject areas, social studies and chemistry, were analyzed in this study. Further studies conducted across different subject areas would help to explain the findings of this study.

References

Alberta Education (1999a). Alberta Education Annual Report 1998-1999. Edmonton, AB: Author

Alberta Education (1999b). Chemistry 30 Diploma Examination Results Examiners' Report for June 1999. Edmonton, AB: Author.

Alberta Education (1999c). Chemistry 30 Grade 12 Diploma Examination, June 1999. Edmonton, AB: Author.

Alberta Education (1999d). Social Studies 30 Diploma Examination Results Examiners' Report for June 1999. Edmonton, AB: Author.

Alberta Education (1999e). Social Studies 30 Grade 12 Diploma Examination, June 1999. Edmonton, AB: Author.

Aiken, L. R. (1987). Testing with multiple-choice items. Journal of Research and Development in Education, 20, 44-58.

Allen, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. Language Testing, 9, 101-122.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. Review of Educational Research, 53, 571-585.

Bennet, R. E., & Ward, W. C. (1993). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrica, 37, 29-51.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (pp.33-49). New York, NY: Springer-Verlag Inc.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. Journal of Educational Measurement, 34, 197-211.

Brown, J. S., & Burton, R. R. (1978). Diagnostic model for procedural bugs in basic mathematical skills. Cognitive Science, 4, 379-426.

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

Crehan, K. D., Koehler, R. A., & Slakter, M. J. (1974). Longitudinal studies of test-wiseness. Journal of Educational Measurement, 11, 209-212

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. Educational and Psychological Measurement, 49, 789-805.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. Applied Psychological Measurement, 16, 327-343

De Ayala, R. J. (1993). Methods, plainly speaking: An introduction to polytomous item response theory models. Measurement and evaluation in Counseling and Development, 25, 172-189.

De Ayala, R. J. (1995). Item parameter recovery for the nominal response model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Diamond, J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 11, 209-212.

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, 1, 95-97.

Frary, R. B. (1989). Partial-credit scoring methods for multiple choice tests. Applied Measurement in Education, 2, 79-96.

Glass, G. V., & Stanley, J. C. (1970). Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gorsuch, R. L. (1983). Factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gulliksen, H. (1950). Theory of mental tests. New York, NY: Wiley.

Haladyna, T. M. (1999). Developing and validating multiple-choice test items. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. Applied Measurement in Education, 5, 1-16.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications, Inc.

Harman, H. H. (1976). Modern factor analysis. Chicago, IL: The University of Chicago Press.

Hughes, C. A., Salvia, J., & Bott, D. (1991). The nature and extent of test-wiseness cues in seventh- and tenth- grade classroom tests. Diagnostique, 16, 153-163.

Hughes, C. A., Schumaker, J. B., Deshler, D. D., & Mercer, C. D. (1988). The test-taking strategy. Lawrence, KS: Edge Enterprises.

Jacobs, P., & Vandeventer, M. (1970). Information in wrong responses. Psychological Reports, 36, 311-315

Lane, S., Stone, C. A., & Hsu, H. (1990). Diagnosing students' errors in solving algebra word problems. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect response option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7). Psychometric Society.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. Journal of Educational Measurement, 21, 239-243.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). Washington, DC: American Council on Education and Macmillian Publishing Co.

Millman, J. (1966). Test-wiseness in taking objective achievement and aptitude examination: Its nature and importance. Final report. New York, NY: College Entrance Examination Board.

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

Morse, D. T. (1994). The relative difficulty of Selected test-wiseness skills among college students. Paper presented at the Annual Meeting of the Mid-South educational Research Association, Nashville, TN, November 9-11, 1994 (ERIC ED 387 528)

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Nandakumar, R. (1994). Assessing dimensionality in set of item responses: Comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics, 18, 41-68.

Nedelsky, L. (1954). Ability to avoid gross ever as a measure of achievement. Educational and Psychological Measurement, 14, 459-472.

Nelson, L. (1983). Lertap3: A test, survey and general data analysis system for small computers. Dunedin, New Zealand: University of Otago, Department of Education.

Rogers, W. T., & Bateson, D. J. (1991a). The influence of test-wiseness upon performance of high school seniors on school leaving examination. Applied Measurement in Education, 4, 159-183.

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

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. Educational and Psychological Measurement, 59, 2, 234-247

Rogers, W. T., & Ndalichako, J. (2000). Number right, item response, and finite scoring: Robustness with respect to lack of equally classifiable options and item option independence. Educational and Psychological Measurement, 60, 5-19.

Rogers, W. T., & Wilson, C. (1993). The influence of test-wiseness upon performance on high school students on Alberta Education's diploma examinations. (Alberta Education, Student Evaluation Branch, Contract No. 91-0143), Edmonton, AB: University of Alberta

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. European Journal of Psychological Assessment, 12, 247-259.

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

Roznowski, M., & Bassett, J. (1992). Training test-wiseness and flawed item types. Applied Measurement in Education, 5, 1, 35-48.

Salvia, J., & Ysseldyke, J. E. (1995). Assessment (6th ed.). Boston, MA: Houghton Mifflin

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement No. 17.

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.

Sigel, I. E. (1963). How intelligence tests limit understanding of intelligence. Merrill-Palmer Quarterly, 9, 30-96.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970a). Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 7, 119-122

Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970b). Learning test-wiseness by programmed texts. Journal of Educational Measurement, 7, 247-254

Stout, W. F. (1987). A non parametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST manual. Urbana-Champaingn, IL: University of Illinois, Department of Statistics

Suen, H. K. (1990). Principles of test theories. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Swaminathan, H. (1983). Parameter estimation in item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 24-44). Vancouver, BC: Educational Research Institute of British Columbia.

Sympson, J. B., (1983). A new IRT model for calibration multiple-choice items. Paper presented at the meeting of the Psychometric Society, Los Angeles, CA.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 13, 201-214.

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 13, 201-214.

Thissen, D. (1991). MULTILOG User's Guide Version 6. Chicago, IL: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49, 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distracters are also part of the item. Journal of Educational Measurement, 26, 161-176.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (pp.51-65). New York, NY: Springer-Verlag Inc.

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, 7, 709-722.

Appendix A

Taxonomy of Testwiseness Principles

Taxonomy of Testwiseness Principles[1]

I. Elements independent of test constructor and 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 (see I.C. and II.B.) which resist a quick response.
4. Mark omitted items, or items, which 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 of response.
2. Pay careful attention to directions, determining clearly the nature of the question.
3. Ask 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 guessing.
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.

---

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

II. Elements dependent upon the test constructor or purpose.

A. Intent consideration strategy

1. Interpret and answer questions in view of the 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 the specific detail.

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) options.
    d. He places it in a certain physical position among the options (such as in the middle).
    e. He places it in a certain logical position among an ordered set of options (such as in the middle).
    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 stereotypical 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 resemblance 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

Social Studies 30 Diploma Examination: Testwiseness Susceptibility by Item

Code:

NTW - Not testwise susceptible

ID1 - Eliminate options known to be incorrect and choose from the remaining alternatives

ID2 - Choose neither or both of two options which imply the correctness of each other

ID3 - Choose neither or one of two statements, one of which, if correct, would imply the incorrectness of the other

IIB4 - Recognize and make use of the resemblance between the options and an aspect of the stem

Social Studies 30 Diploma Examination: Item Classification

Item	Raters				Agreement	Option(s)
	I	II	III	IV	TW principle	
1	ID1	ID1; IIB4	ID1	ID1	**ID1**	A & B
2	ID2	ID1; IIB4	ID1	ID2	**ID2**	A & C
3	IIB4	IIB4	ID1		**ID1; IIB4**	B; A
4		ID1	ID1		**NTW**	
5			ID1		**NTW**	
6		ID1; IIB4			**NTW**	
7	ID1	ID1	ID1; ID3	ID1	**ID1**	D
8	ID2	ID1; ID2	ID1		**ID2**	A & B
9	ID1		ID1	ID1	**ID1**	D
10		ID1; IIB4	ID1		**IIB4**	A
11	ID1	ID1	ID1	ID1	**ID1**	B
12	ID1		ID1	ID1	**ID1**	D
13	ID1	ID1	ID1	ID1	**ID1**	C
14	ID1	ID1	ID1	ID1	**ID1**	C & D
15		ID1		ID1	**NTW**	
16	ID2		ID1	ID1	**ID1**	A
17				ID1	**NTW**	
18		ID1		ID3	**NTW**	
19		ID1		ID1	**NTW**	
20				ID1	**NTW**	
21	ID2	ID1; IIB4		IIB4	**IIB4**	D
22		ID1; IIB4	IIB4	IIB4	**IIB4**	A
23	ID1	ID1	ID3	ID1	**ID1**	C
24	ID2	ID1		ID1	**ID1**	C
25				ID1	**NTW**	
26		ID1		ID1	**NTW**	
27		ID1		ID1	**NTW**	
28			ID1	ID3	**NTW**	

29	ID1		ID1	ID1	**ID1**	A & D
30				ID1	**NTW**	
31	ID3	ID3		ID1; ID3	**ID3**	A & B
32	ID2	ID1; ID2		ID2	**ID1; ID2**	B; A & C
33			ID1		**NTW**	
34	ID1	ID1			**ID1**	B
35	ID1	ID1	ID1		**ID1**	B & C
36		ID1	ID1	ID1	**ID1**	A
37		ID1	ID1	ID1	**ID1**	D
38	ID1		ID1	ID1	**ID1**	C
39					**NTW**	
40	ID2	ID1			**NTW**	
41		ID1			**NTW**	
42			ID1	ID1	**ID1**	C & D
43			ID1		**NTW**	
44	ID3	ID1		ID3	**ID3**	A & D
45				ID1	**NTW**	
46			ID1	ID3	**NTW**	
47			ID1		**NTW**	
48		ID1	ID1	ID1	**ID1**	C
49		ID1			**NTW**	
50	ID1	ID1; IIB4	ID1	ID1; IIB4	**ID1; IIB4**	C; A
51		ID1		ID1	**NTW**	
52		ID1	ID1		**NTW**	
53		IIB4	ID1		**NTW**	
54		ID1; IIB4			**NTW**	
55		ID1	ID1	ID1	**ID1**	A
56	ID1	ID1; IIB4	ID1; ID3		**ID1**	C
57			ID1		**NTW**	
58		ID1	ID1		**ID1**	D
59		ID1			**NTW**	
60	ID1		ID1	ID1	**ID1**	B

61				IIB4	**NTW**	
62				ID1	**NTW**	
63	ID1; IIB4	ID1; IIB4	ID1	ID1	**ID1; IIB4**	C; B
64				ID1	**NTW**	
65			ID1	ID1	**ID1**	D
66		ID1; ID2			**NTW**	
67		ID1; ID3			**NTW**	
68		ID1			**NTW**	
69				ID1	**NTW**	
70	ID2	ID2			**ID2**	B & C

Appendix C

Chemistry 30 Diploma Examination: Testwiseness Susceptibility by Item

Code:

NTW - Not testwise susceptible

ID1 - Eliminate options known to be incorrect and choose from the remaining alternatives

ID2 - Choose neither or both of two options which imply the correctness of each other

ID3 - Choose neither or one of two statements, one of which, if correct, would imply the incorrectness of the other

ID5 - Consider the subject matter and difficulty of neighboring items when interpreting and answering a given item

IIB1a - Recognize and make use of any consistent idiosyncrasies of the test constructor, which distinguish the correct answer from incorrect options: He makes it longer (shorter) than the incorrect options

IIB3 - Recognize and make use of specific determiners

IIB4 - Recognize and make use of the resemblance between the options and an aspect of the stem

Chemistry 30 Diploma Examination: Item Classification

Item	Rater I	Rater II	Rater III	Rater IV	Agreement TW Principle	Option(s)
1		ID3		ID1	**NTW**	
2		ID1			**NTW**	
3	ID3; IIB4	ID3; IIB4	ID3	ID3	**ID3; IIB4**	C & D; D
4	ID1	ID1			**NTW**	
5	ID1	ID1	ID1	ID1	**ID1**	D
6		ID1			**NTW**	
7	ID1	ID1	ID1	ID1	**ID1**	C
8	ID1	ID1		ID2	**ID1**	C & D
9		ID1			**NTW**	
10	ID3	ID3	ID1	ID3	**ID3**	A & B
11	ID3	ID3; IIB4	ID1; ID3	ID3	**ID3**	C & D
12	ID1	ID1	ID1; ID3	ID1	**ID1**	C & D
13	ID3	ID3			**NTW**	
14	IIB3	ID2; IIB3		IIB3	**IIB3**	C
15					**NTW**	
16			ID2		**NTW**	
17				ID1; IIB4	**NTW**	
18			ID1	ID1	**ID1**	D
19	ID1; ID2	ID1; ID2	ID1	ID1	**ID1**	C
20			ID1	ID1	**ID1**	D
21	ID1	ID1			**NTW**	
22			ID1	ID3	**NTW**	
23	ID3	ID3	ID3	ID1	**ID3**	A & C
24	ID3	ID3	ID1	ID3	**ID3**	A & B
25	ID1	ID1	ID3	ID1	**ID1**	B
26			ID2	ID3	**NTW**	
27			ID1		**NTW**	
28	ID1	ID1	ID1; ID3	ID1	**ID1**	A & C

29			ID1		**NTW**	
30	ID1	ID1	ID1	ID1	**ID1**	D
31		ID1	ID2	IIB4	**NTW**	
32					**NTW**	
33	ID1	ID1	ID1	ID1	**ID1**	A
34				ID1	**NTW**	
35			ID2	ID1	**NTW**	
36	ID1	ID1	ID1		**ID1**	D
37	IB1a	IB1a	IB1a; ID1	IB1a; ID1	**IB1a; ID1**	D; C
38			ID1	ID3	**NTW**	
39					**NTW**	
40	ID2	ID2	ID1	ID1; ID2	**ID1; ID2**	A & B; A & D
41	ID1	ID1	ID1	ID5	**ID1**	C
42	ID5	ID5	ID1; ID5	ID1	**ID5**	(item 44)
43			ID1		**NTW**	
44	ID1	ID1; 2B4	ID1		**ID1**	A

Appendix D

Social Studies 30 Diploma Examination: Item Analysis

Table 20
Results of Classical Item Analysis: SS-NTW-S1

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
4	1	354	0.09	-0.20	
	2	714	0.18	-0.22	
	3*	**2462**	**0.62**	**0.44**	**0.37**
	4	468	0.12	-0.22	
5	**1***	**1819**	**0.46**	**0.38**	**0.31**
	2	233	0.06	-0.13	
	3	849	0.21	-0.15	
	4	1093	0.27	-0.22	
6	1	346	0.09	-0.28	
	2*	**2901**	**0.73**	**0.43**	**0.37**
	3	451	0.11	-0.23	
	4	295	0.07	-0.14	
15	1	260	0.07	-0.23	
	2	145	0.04	-0.17	
	3	220	0.06	-0.21	
	4*	**3375**	**0.84**	**0.38**	**0.33**
17	1	404	0.10	-0.21	
	2	556	0.14	-0.22	
	3	406	0.10	-0.23	
	4*	**2631**	**0.66**	**0.44**	**0.38**
18	1	292	0.07	-0.23	
	2*	**3328**	**0.83**	**0.38**	**0.33**
	3	135	0.03	-0.16	
	4	245	0.06	-0.22	
19	1	227	0.06	-0.19	
	2	549	0.14	-0.21	
	3*	**2535**	**0.63**	**0.41**	**0.34**
	4	685	0.17	-0.21	
20	1	149	0.04	-0.17	
	2	212	0.05	-0.21	
	3*	**3376**	**0.84**	**0.37**	**0.32**
	4	260	0.07	-0.23	
25	1	510	0.13	-0.16	
	2	577	0.14	-0.16	
	3*	**2213**	**0.55**	**0.39**	**0.32**
	4	696	0.17	-0.22	
26	1	424	0.11	-0.32	
	2*	**2584**	**0.65**	**0.45**	**0.39**
	3	631	0.16	-0.19	
	4	361	0.09	-0.17	
27	1	412	0.10	-0.21	

	2*	**2207**	**0.55**	**0.51**	**0.45**
	3	958	0.24	-0.22	
	4	421	0.11	-0.31	
28	1	138	0.03	-0.20	
	2	249	0.06	-0.19	
	3	134	0.03	-0.18	
	4*	**3478**	**0.87**	**0.34**	**0.29**
30	**1***	**2513**	**0.63**	**0.40**	**0.33**
	2	495	0.12	-0.19	
	3	256	0.06	-0.22	
	4	734	0.18	-0.21	
33	**1***	**1433**	**0.36**	**0.30**	**0.23**
	2	624	0.16	-0.09	
	3	1339	0.34	-0.05	
	4	603	0.15	-0.23	
39	**1***	**2152**	**0.54**	**0.44**	**0.37**
	2	541	0.14	-0.19	
	3	535	0.13	-0.20	
	4	770	0.19	-0.23	
40	1	257	0.06	-0.12	
	2	405	0.10	-0.16	
	3	666	0.17	-0.22	
	4*	**2672**	**0.67**	**0.34**	**0.27**
41	1	451	0.11	-0.20	
	2	568	0.14	-0.13	
	3	196	0.05	-0.13	
	4*	**2783**	**0.70**	**0.30**	**0.23**
43	1	347	0.09	-0.24	
	2	536	0.13	-0.17	
	3	353	0.09	-0.14	
	4*	**2763**	**0.69**	**0.35**	**0.28**
45	1	585	0.15	-0.12	
	2*	**2643**	**0.66**	**0.40**	**0.33**
	3	322	0.08	-0.29	
	4	447	0.11	-0.21	
46	1	388	0.10	-0.23	
	2	219	0.06	-0.15	
	3*	**2831**	**0.71**	**0.43**	**0.37**
	4	561	0.14	-0.27	
47	1	361	0.09	-0.30	
	2	479	0.12	-0.22	
	3*	**2872**	**0.72**	**0.46**	**0.40**
	4	286	0.07	-0.19	
49	1	378	0.09	-0.12	
	2*	**1178**	**0.29**	**0.20**	**0.12**
	3	1456	0.36	-0.13	

	4	988	0.25	0.01	
51	1	257	0.06	-0.21	
	2	300	0.08	-0.11	
	3*	**2509**	**0.63**	**0.24**	**0.16**
	4	932	0.23	-0.08	
52	1	273	0.07	-0.26	
	2*	**2621**	**0.66**	**0.48**	**0.42**
	3	770	0.19	-0.25	
	4	335	0.08	-0.23	
53	**1***	**3192**	**0.80**	**0.37**	**0.31**
	2	240	0.06	-0.16	
	3	229	0.06	-0.18	
	4	338	0.08	-0.24	
54	1	486	0.12	-0.11	
	2*	**2377**	**0.59**	**0.48**	**0.42**
	3	626	0.16	-0.31	
	4	510	0.13	-0.26	
57	1	346	0.09	-0.13	
	2*	**3157**	**0.79**	**0.31**	**0.25**
	3	319	0.08	-0.21	
	4	178	0.05	-0.15	
59	1	406	0.10	-0.18	
	2*	**1968**	**0.49**	**0.37**	**0.30**
	3	1023	0.26	-0.10	
	4	600	0.15	-0.24	
61	1	481	0.12	-0.18	
	2	411	0.10	-0.23	
	3*	**2726**	**0.68**	**0.41**	**0.34**
	4	381	0.10	-0.22	
62	1	247	0.06	-0.20	
	2	504	0.13	-0.28	
	3*	**2822**	**0.71**	**0.47**	**0.41**
	4	427	0.11	-0.24	
64	1	379	0.10	-0.19	
	2	376	0.09	-0.23	
	3*	**2400**	**0.60**	**0.35**	**0.28**
	4	845	0.21	-0.12	
66	1	211	0.05	-0.21	
	2	271	0.07	-0.16	
	3	308	0.08	-0.24	
	4*	**3208**	**0.80**	**0.37**	**0.31**
67	**1***	**2107**	**0.53**	**0.41**	**0.34**
	2	987	0.25	-0.28	
	3	624	0.16	-0.11	
	4	282	0.07	-0.17	
68	1	569	0.14	-0.14	

	2	349	0.09	-0.18	
	3	411	0.10	-0.20	
	4*	**2666**	**0.67**	**0.34**	**0.27**
69	1	425	0.11	-0.18	
	2	357	0.09	-0.26	
	3*	**2902**	**0.73**	**0.39**	**0.33**
	4	315	0.08	-0.16	

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Table 21
Results of Classical Item Analysis: SS-ID1-S1

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
1	1	131	0.03	-0.16	
	2	109	0.03	-0.12	
	3	1107	0.28	-0.24	
	4*	**2649**	**0.66**	**0.33**	**0.22**
7	**1***	**2587**	**0.65**	**0.38**	**0.27**
	2	507	0.13	-0.20	
	3	747	0.19	-0.21	
	4	156	0.04	-0.16	
9	1	467	0.12	-0.32	
	2	874	0.22	-0.20	
	3*	**2477**	**0.62**	**0.45**	**0.35**
	4	181	0.05	-0.15	
11	1	160	0.04	-0.14	
	2	32	0.01	-0.07	
	3	466	0.12	-0.20	
	4*	**3342**	**0.84**	**0.27**	**0.18**
12	**1***	**1598**	**0.40**	**0.33**	**0.22**
	2	182	0.05	-0.15	
	3	2078	0.52	-0.21	
	4	139	0.04	-0.14	
13	1	1223	0.31	-0.32	
	2	272	0.07	-0.26	
	3	48	0.01	-0.06	
	4*	**2454**	**0.61**	**0.45**	**0.35**
14	1	517	0.13	-0.29	
	2*	**3106**	**0.78**	**0.41**	**0.32**
	3	228	0.06	-0.18	
	4	148	0.04	-0.17	
16	1	168	0.04	-0.16	
	2	920	0.23	-0.26	
	3*	**2263**	**0.57**	**0.41**	**0.30**
	4	644	0.16	-0.18	
23	1	578	0.14	-0.23	
	2	958	0.24	-0.32	
	3	166	0.04	-0.09	
	4*	**2297**	**0.57**	**0.48**	**0.38**
24	1	243	0.06	-0.14	
	2	232	0.06	-0.15	
	3	181	0.05	-0.22	
	4*	**3343**	**0.84**	**0.31**	**0.23**

29	1	144	0.04	-0.18	
	2	550	0.14	-0.31	
	3*	**3157**	**0.79**	**0.42**	**0.33**
	4	148	0.04	-0.17	
34	1	599	0.15	-0.18	
	2	263	0.07	-0.17	
	3	984	0.25	-0.18	
	4*	**2152**	**0.54**	**0.37**	**0.26**
35	**1***	**2576**	**0.64**	**0.39**	**0.28**
	2	611	0.15	-0.21	
	3	277	0.07	-0.14	
	4	536	0.13	-0.22	
36	1	162	0.04	-0.17	
	2*	**2556**	**0.64**	**0.37**	**0.26**
	3	286	0.07	-0.19	
	4	993	0.25	-0.22	
37	**1***	**3340**	**0.84**	**0.43**	**0.35**
	2	350	0.09	-0.29	
	3	163	0.04	-0.20	
	4	145	0.04	-0.21	
38	1	739	0.19	-0.09	
	2*	**2620**	**0.66**	**0.24**	**0.13**
	3	121	0.03	-0.13	
	4	514	0.13	-0.17	
42	1	809	0.20	-0.32	
	2*	**2894**	**0.72**	**0.45**	**0.36**
	3	164	0.04	-0.18	
	4	133	0.03	-0.19	
48	1	902	0.23	-0.20	
	2*	**2272**	**0.57**	**0.40**	**0.29**
	3	170	0.04	-0.19	
	4	655	0.16	-0.21	
55	1	129	0.03	-0.20	
	2	359	0.09	-0.30	
	3	207	0.05	-0.19	
	4*	**3302**	**0.83**	**0.43**	**0.35**
56	**1***	**2503**	**0.63**	**0.39**	**0.28**
	2	385	0.10	-0.20	
	3	244	0.06	-0.18	
	4	866	0.22	-0.21	
58	1	452	0.11	-0.24	
	2	248	0.06	-0.20	
	3*	**3198**	**0.80**	**0.37**	**0.28**
	4	101	0.03	-0.16	
60	**1***	**2943**	**0.74**	**0.45**	**0.36**
	2	139	0.04	-0.21	

	3	317	0.08	-0.23	
	4	601	0.15	-0.28	
65	1	913	0.23	-0.33	
	2	428	0.11	-0.32	
	3*	**2526**	**0.63**	**0.53**	**0.44**
	4	132	0.03	-0.12	

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Table 22
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: SS-NTW-S1

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
4	-0.63	1.04	-0.56	1.14	0.23	0.32
5	0.23	0.77	0.26	0.58	0.66	0.14
6	-1.28	1.10	-1.10	0.8	-0.52	0.26
15	-2.18	1.18	-1.77	0.73	-1.44	0.21
17	-0.86	1.09	-0.75	0.92	-0.10	0.27
18	-2.07	1.12	-1.75	0.76	-1.16	0.30
19	-0.73	0.91	-0.71	0.69	-0.11	0.23
20	-2.18	1.16	-1.79	0.81	-1.11	0.35
25	-0.29	0.82	-0.31	0.85	0.49	0.28
26	-0.80	1.12	-0.68	0.98	-0.02	0.28
27	-0.28	1.34	-0.23	1.05	0.10	0.15
28	-2.44	1.14	-2.03	0.69	-1.71	0.22
30	-0.70	0.89	-0.70	0.68	-0.07	0.23
33	0.76	0.56	1.11	0.5	1.51	0.15
39	-0.21	0.99	-0.20	0.72	0.18	0.15
40	-0.92	0.68	-1.13	0.47	-0.46	0.21
41	-1.09	0.58	-1.52	0.40	-0.72	0.22
43	-1.06	0.77	-1.18	0.61	-0.22	0.32
45	-0.88	0.87	-0.89	0.59	-0.43	0.18
46	-1.16	1.08	-1.01	0.71	-0.67	0.16
47	-1.23	1.25	-0.98	0.99	-0.36	0.29
49	1.14	0.35	2.53	0.99	2.07	0.24
51	-0.69	0.39	-1.38	0.30	0.00	0.25
52	-0.85	1.24	-0.68	0.83	-0.38	0.14
53	-1.79	0.98	-1.66	0.61	-1.29	0.19
54	-0.51	1.19	-0.42	0.94	0.00	0.19
57	-1.72	0.72	-2.01	0.49	-1.22	0.28
59	0.04	0.76	0.04	0.66	0.62	0.20
61	-1.00	0.95	-0.95	0.81	-0.16	0.31
62	-1.15	1.31	-0.89	1.02	-0.36	0.26
64	-0.54	0.68	-0.66	0.50	0.00	0.20
66	-1.82	1.02	-1.64	0.69	-1.03	0.28
67	-0.15	0.85	-0.15	0.63	0.29	0.16
68	-0.92	0.70	-1.10	0.56	-0.14	0.29
69	-1.28	0.91	-1.26	0.58	-0.84	0.17

Table 23
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: SS-ID1-S1

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
1	-0.91	0.57	-1.27	0.40	-0.47	0.21
7	-0.82	0.75	-0.90	0.52	-0.34	0.19
9	-0.66	0.95	-0.61	0.65	-0.24	0.15
11	-2.16	0.59	-2.92	0.38	-2.04	0.28
12	0.55	0.56	0.78	0.49	1.30	0.17
13	-0.63	1.00	-0.56	0.89	0.09	0.26
14	-1.67	1.04	-1.45	0.71	-0.92	0.25
16	-0.36	0.85	-0.36	0.69	0.24	0.22
23	-0.41	1.10	-0.34	0.75	-0.05	0.13
24	-2.16	0.73	-2.46	0.44	-2.01	0.20
29	-1.77	1.12	-1.45	0.76	-0.97	0.24
34	-0.21	0.69	-0.25	0.73	0.66	0.29
35	-0.80	0.77	-0.87	0.55	-0.27	0.20
36	-0.77	0.73	-0.87	0.71	0.23	0.35
37	-2.16	1.35	-1.57	0.85	-1.32	0.17
38	-0.87	0.31	-2.10	0.21	-0.75	0.21
42	-1.30	1.12	-1.07	0.74	-0.71	0.17
48	-0.37	0.79	-0.40	0.87	0.50	0.31
55	-2.07	1.27	-1.56	0.88	-1.08	0.27
56	-0.70	0.78	-0.75	0.58	-0.11	0.22
58	-1.85	0.92	-1.75	0.62	-1.15	0.26
60	-1.38	1.12	-1.14	0.88	-0.48	0.29
65	-0.73	1.47	-0.52	1.15	-0.15	0.19

Table 24
Results of IRT Item Analysis for NRM: SS-NTW-S1

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
4	1	-0.35	-0.81	45	1	0.25	0.05
	2	-0.12	0.01		**2***	**0.79**	**1.55**
	3*	**0.81**	**1.31**		3	-0.87	-1.22
	4	-0.34	-0.52		4	-0.16	-0.37
5	**1***	**0.62**	**0.84**	46	1	-0.34	-0.54
	2	-0.39	-1.29		2	-0.19	-1.03
	3	-0.07	0.12		**3***	**0.81**	**1.70**
	4	-0.16	0.34		4	-0.28	-0.13
6	1	-0.63	-0.86	47	1	-0.70	-0.89
	2*	**0.83**	**1.77**		2	-0.08	-0.16
	3	-0.24	-0.31		**3***	**0.98**	**1.80**
	4	0.04	-0.59		4	-0.21	-0.75
15	1	-0.30	-0.56	49	1	-0.33	-0.90
	2	-0.33	-1.16		**2***	**0.30**	**0.27**
	3	-0.27	-0.70		3	-0.11	0.50
	4*	**0.90**	**2.42**		4	0.13	0.13
17	1	-0.29	-0.60	51	1	-0.58	-1.16
	2	-0.15	-0.20		2	-0.07	-0.73
	3	-0.40	-0.66		**3***	**0.43**	**1.44**
	4*	**0.83**	**1.46**		4	0.22	0.45
18	1	-0.22	-0.42	52	1	-0.75	-1.25
	2	0.86	2.34		**2***	**1.08**	**1.65**
	3	-0.34	-1.27		3	0.06	0.35
	4	-0.30	-0.65		4	-0.39	-0.75
19	1	-0.52	-1.30	53	**1***	**0.75**	**2.06**
	2	-0.17	-0.21		2	-0.13	-0.71
	3*	**0.75**	**1.43**		3	-0.28	-0.85
	4	-0.05	0.07		4	-0.34	-0.50
20	1	-0.30	-1.11	54	1	0.13	-0.31
	2	-0.30	-0.76		**2***	**0.92**	**1.23**
	3*	**0.89**	**2.42**		3	-0.55	-0.38
	4	-0.29	-0.55		4	-0.49	-0.54
25	1	-0.17	-0.47	57	1	0.09	-0.28
	2	-0.16	-0.34		**2***	**0.58**	**1.98**
	3*	**0.61**	**1.02**		3	-0.35	-0.57
	4	-0.28	-0.20		4	-0.33	-1.14
26	1	-0.78	-0.89	59	1	-0.36	-0.78
	2*	**0.89**	**1.48**		**2***	**0.66**	**0.90**
	3	0.03	0.03		3	0.13	0.30
	4	-0.15	-0.61		4	-0.44	-0.43
27	1	-0.33	-0.62	61	1	-0.05	-0.28

	2*	**1.15**	**1.21**		2	-0.33	-0.58
	3	0.04	0.40		**3***	**0.74**	**1.53**
	4	-0.85	-0.99		4	-0.35	-0.67
28	1	-0.55	-1.28	62	1	-0.42	-1.07
	2	-0.03	-0.26		2	-0.30	-0.26
	3	-0.34	-1.11		**3***	**0.99**	**1.74**
	4*	**0.92**	**2.65**		4	-0.27	-0.41
30	**1***	**0.75**	**1.41**	64	1	-0.32	-0.72
	2	-0.13	-0.30		2	-0.45	-0.81
	3	-0.60	-1.25		**3***	**0.61**	**1.28**
	4	-0.03	0.14		4	0.17	0.25
33	**1***	**0.50**	**0.45**	66	1	-0.46	-1.05
	2	-0.05	-0.34		2	0.02	-0.49
	3	0.09	0.44		3	-0.34	-0.58
	4	**-0.53**	**-0.55**		**4***	**0.79**	**2.11**
39	**1***	**0.74**	**0.98**	67	**1***	**0.66**	**1.04**
	2	-0.25	-0.44		2	-0.28	0.22
	3	-0.28	-0.47		3	0.05	-0.14
	4	-0.21	-0.07		4	-0.44	-1.12
40	1	-0.16	-0.95	68	1	0.04	-0.12
	2	-0.17	-0.5		2	-0.30	-0.74
	3	-0.18	-0.01		3	-0.30	-0.58
	4*	**0.51**	**1.46**		**4***	**0.56**	**1.44**
41	1	-0.26	-0.36	69	1	-0.07	-0.31
	2	0.04	-0.02		2	-0.55	-0.78
	3	-0.25	-1.19		3	0.70	1.71
	4*	**0.47**	**1.58**		4*	-0.08	-0.62
43	1	-0.50	-0.83				
	2	-0.04	-0.15				
	3	-0.05	-0.57				
	4*	**0.60**	**1.54**				

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 25
Results of IRT Item Analysis for NRM: SS-ID1-S1

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
1	1	-0.52	-1.47	35	**1***	**0.59**	**1.37**
	2	-0.24	-1.46		2	-0.16	-0.13
	3	0.14	1.02		3	-0.20	-0.94
	4*	**0.62**	**1.91**		4	-0.23	-0.30
7	**1***	**0.65**	**1.54**	36	1	-0.53	-1.51
	2	-0.13	-0.18		**2***	**0.73**	**1.63**
	3	0.01	0.26		3	-0.32	-0.79
	4	-0.54	-1.61		4	0.12	0.67
9	1	-0.64	-0.61	37	**1***	**1.07**	**2.53**
	2	0.14	0.45		2	-0.28	-0.22
	3*	**0.82**	**1.49**		3	-0.32	-1.02
	4	-0.32	-1.33		4	-0.48	-1.28
11	1	-0.18	-0.66	38	1	0.22	0.30
	2	-0.35	-2.38		**2***	**0.35**	**1.56**
	3	-0.01	0.49		3	-0.46	-1.74
	4*	**0.54**	**2.55**		4	-0.11	-0.12
12	**1***	**0.69**	**1.20**	42	1	-0.03	0.66
	2	-0.44	-1.21		**2***	**1.00**	**2.11**
	3	0.23	1.51		3	-0.37	-1.18
	4	-0.48	-1.50		4	-0.60	-1.59
13	1	-0.02	1.13	48	1	0.07	0.40
	2	-0.74	-0.85		**2***	**0.73**	**1.32**
	3	-0.12	-2.15		3	-0.71	-1.73
	4*	**0.88**	**1.88**		4	-0.09	0.02
14	1	-0.23	0.03	55	1	-0.54	-1.49
	2*	**0.81**	**2.08**		2	-0.34	-0.28
	3	-0.17	-0.75		3	-0.13	-0.66
	4	-0.42	-1.36		**4***	**1.01**	**2.43**
16	1	-0.53	-1.60	56	**1***	**0.67**	**1.42**
	2	-0.13	0.33		2	-0.28	-0.61
	3*	**0.70**	**1.27**		3	-0.42	-1.15
	4	-0.05	0.00		4	0.03	0.34
23	1	-0.34	-0.21	58	1	-0.06	0.10
	2	-0.35	0.29		2	-0.20	-0.59
	3	-0.08	-1.35		**3***	**0.78**	**2.23**
	4*	**0.77**	**1.26**		4	-0.52	-1.74
24	1	0.01	-0.49	60	**1***	**0.98**	**2.03**
	2	-0.03	-0.56		2	-0.79	-1.83
	3	-0.57	-1.17		3	-0.17	-0.46
	4*	**0.59**	**2.22**		4	-0.02	0.27
29	1	-0.37	-1.26	65	1	-0.16	0.50

	2	-0.21	0.20	2	-0.72	-0.66
	3*	**0.88**	**2.24**	**3***	**1.13**	**1.65**
	4	-0.30	-1.18	4	-0.26	-1.49
34	1	-0.14	-0.22			
	2	-0.43	-1.18			
	3	0.00	0.32			
	4*	**0.57**	**1.09**			

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 26
Probabilities of Selecting Item Options in NRM: SS-NTW-S1

Item	Option	Probability of Selecting Item Option at θ level						
		-3	-2	-1	0	1	2	3
4	1	0.271	0.219	0.148	0.077	0.031	0.011	0.004
	2	0.308	0.314	0.268	0.175	0.090	0.040	0.016
	3*	**0.070**	**0.179**	**0.388**	**0.644**	**0.836**	**0.934**	**0.975**
	4	0.351	0.287	0.196	0.103	0.042	0.015	0.005
5	**1***	**0.073**	**0.149**	**0.276**	**0.452**	**0.639**	**0.790**	**0.888**
	2	0.181	0.133	0.090	0.054	0.028	0.012	0.005
	3	0.283	0.288	0.268	0.220	0.156	0.097	0.055
	4	0.463	0.430	0.366	0.274	0.178	0.101	0.052
6	1	0.530	0.347	0.165	0.056	0.015	0.004	0.001
	2*	**0.092**	**0.259**	**0.531**	**0.774**	**0.907**	**0.963**	**0.985**
	3	0.285	0.275	0.193	0.097	0.039	0.014	0.005
	4	0.093	0.119	0.111	0.073	0.039	0.019	0.009
15	1	0.341	0.239	0.120	0.045	0.015	0.005	0.001
	2	0.205	0.139	0.068	0.025	0.008	0.002	0.001
	3	0.271	0.196	0.101	0.039	0.013	0.004	0.001
	4*	**0.183**	**0.427**	**0.711**	**0.891**	**0.964**	**0.989**	**0.997**
17	1	0.281	0.242	0.169	0.089	0.036	0.013	0.004
	2	0.275	0.273	0.220	0.132	0.062	0.025	0.010
	3	0.368	0.284	0.178	0.083	0.031	0.010	0.003
	4*	**0.076**	**0.202**	**0.433**	**0.696**	**0.871**	**0.952**	**0.983**
18	1	0.309	0.233	0.130	0.055	0.021	0.007	0.002
	2*	**0.191**	**0.424**	**0.696**	**0.877**	**0.957**	**0.986**	**0.995**
	3	0.189	0.126	0.063	0.024	0.008	0.002	0.001
	4	0.312	0.217	0.112	0.044	0.015	0.005	0.002
19	1	0.299	0.191	0.101	0.043	0.015	0.005	0.001
	2	0.312	0.283	0.213	0.128	0.064	0.028	0.012
	3*	**0.102**	**0.232**	**0.437**	**0.660**	**0.826**	**0.919**	**0.964**
	4	0.288	0.294	0.249	0.169	0.095	0.048	0.022
20	1	0.197	0.137	0.069	0.026	0.009	0.003	0.001
	2	0.279	0.195	0.098	0.037	0.012	0.004	0.001
	3*	**0.189**	**0.433**	**0.714**	**0.891**	**0.964**	**0.989**	**0.997**
	4	0.335	0.235	0.119	0.046	0.015	0.005	0.001
25	1	0.230	0.214	0.178	0.127	0.077	0.041	0.020
	2	0.254	0.238	0.200	0.144	0.088	0.048	0.024
	3*	**0.098**	**0.199**	**0.362**	**0.563**	**0.744**	**0.868**	**0.937**
	4	0.418	0.349	0.260	0.166	0.090	0.043	0.019

26	1	0.670	0.444	0.207	0.064	0.015	0.003	0.001
	2*	**0.048**	**0.168**	**0.417**	**0.689**	**0.862**	**0.943**	**0.977**
	3	0.148	0.221	0.231	0.162	0.086	0.040	0.017
	4	0.134	0.167	0.146	0.085	0.038	0.015	0.005
27	1	0.190	0.217	0.182	0.093	0.030	0.008	0.002
	2*	**0.014**	**0.070**	**0.258**	**0.583**	**0.835**	**0.945**	**0.982**
	3	0.173	0.288	0.349	0.259	0.122	0.046	0.016
	4	0.623	0.425	0.211	0.065	0.013	0.002	0.000
28	1	0.353	0.183	0.065	0.018	0.004	0.001	0.000
	2	0.206	0.180	0.108	0.050	0.020	0.008	0.003
	3	0.223	0.143	0.063	0.021	0.006	0.002	0.001
	4*	**0.218**	**0.494**	**0.764**	**0.911**	**0.969**	**0.989**	**0.996**
30	**1***	**0.096**	**0.226**	**0.431**	**0.653**	**0.818**	**0.913**	**0.960**
	2	0.242	0.237	0.188	0.118	0.061	0.028	0.012
	3	0.384	0.235	0.116	0.046	0.015	0.004	0.001
	4	0.279	0.302	0.264	0.183	0.105	0.054	0.026
33	**1***	**0.067**	**0.133**	**0.232**	**0.356**	**0.488**	**0.612**	**0.718**
	2	0.159	0.182	0.183	0.161	0.128	0.092	0.063
	3	0.228	0.300	0.346	0.352	0.320	0.267	0.208
	4	0.545	0.385	0.239	0.131	0.064	0.029	0.012
39	**1***	**0.060**	**0.145**	**0.312**	**0.548**	**0.764**	**0.896**	**0.958**
	2	0.281	0.254	0.203	0.132	0.069	0.030	0.012
	3	0.298	0.262	0.203	0.128	0.065	0.027	0.011
	4	0.361	0.339	0.282	0.192	0.103	0.047	0.019
40	1	0.146	0.122	0.092	0.061	0.037	0.021	0.011
	2	0.237	0.196	0.145	0.096	0.058	0.032	0.017
	3	0.398	0.326	0.240	0.157	0.094	0.052	0.027
	4*	**0.219**	**0.356**	**0.523**	**0.685**	**0.811**	**0.895**	**0.944**
41	1	0.361	0.262	0.172	0.102	0.056	0.029	0.015
	2	0.206	0.202	0.179	0.143	0.107	0.075	0.052
	3	0.153	0.112	0.074	0.044	0.025	0.013	0.007
	4*	**0.281**	**0.424**	**0.576**	**0.710**	**0.812**	**0.882**	**0.927**
43	1	0.449	0.286	0.151	0.067	0.026	0.009	0.003
	2	0.223	0.225	0.188	0.132	0.082	0.047	0.026
	3	0.151	0.151	0.125	0.087	0.053	0.030	0.016
	4*	**0.177**	**0.339**	**0.537**	**0.715**	**0.839**	**0.913**	**0.954**
45	1	0.082	0.150	0.183	0.156	0.108	0.069	0.042
	2*	**0.073**	**0.229**	**0.478**	**0.698**	**0.834**	**0.909**	**0.950**
	3	0.662	0.397	0.158	0.044	0.010	0.002	0.000
	4	0.184	0.224	0.181	0.102	0.047	0.020	0.008
46	1	0.339	0.268	0.169	0.080	0.030	0.010	0.003

	2	0.133	0.122	0.089	0.049	0.022	0.009	0.003
	3*	**0.101**	**0.252**	**0.502**	**0.751**	**0.900**	**0.964**	**0.987**
	4	0.427	0.358	0.240	0.120	0.049	0.017	0.006
47	1	0.594	0.393	0.180	0.053	0.012	0.002	0.000
	2	0.192	0.236	0.201	0.109	0.045	0.016	0.006
	3*	**0.057**	**0.201**	**0.493**	**0.777**	**0.921**	**0.974**	**0.992**
	4	0.157	0.170	0.127	0.061	0.022	0.007	0.002
49	1	0.233	0.177	0.129	0.090	0.060	0.039	0.024
	2*	**0.114**	**0.162**	**0.222**	**0.291**	**0.366**	**0.442**	**0.516**
	3	0.489	0.463	0.421	0.366	0.306	0.245	0.190
	4	0.164	0.198	0.228	0.253	0.268	0.274	0.270
51	1	0.410	0.230	0.110	0.048	0.019	0.008	0.003
	2	0.137	0.127	0.102	0.073	0.050	0.032	0.021
	3*	**0.267**	**0.411**	**0.540**	**0.641**	**0.716**	**0.772**	**0.815**
	4	0.186	0.232	0.248	0.238	0.216	0.188	0.161
52	1	0.483	0.308	0.138	0.039	0.008	0.001	0.000
	2*	**0.036**	**0.144**	**0.401**	**0.705**	**0.887**	**0.960**	**0.986**
	3	0.211	0.302	0.303	0.192	0.087	0.034	0.013
	4	0.270	0.247	0.158	0.064	0.018	0.005	0.001
53	**1***	**0.196**	**0.404**	**0.652**	**0.837**	**0.933**	**0.974**	**0.990**
	2	0.172	0.147	0.099	0.052	0.024	0.011	0.004
	3	0.234	0.173	0.099	0.046	0.018	0.007	0.002
	4	0.398	0.276	0.150	0.065	0.024	0.009	0.003
54	1	0.073	0.120	0.155	0.135	0.082	0.041	0.020
	2*	**0.032**	**0.115**	**0.329**	**0.631**	**0.844**	**0.939**	**0.976**
	3	0.523	0.436	0.286	0.126	0.039	0.010	0.002
	4	0.372	0.329	0.230	0.107	0.035	0.010	0.002
57	1	0.133	0.135	0.115	0.085	0.057	0.037	0.023
	2*	**0.294**	**0.487**	**0.677**	**0.815**	**0.899**	**0.945**	**0.969**
	3	0.374	0.244	0.134	0.064	0.028	0.011	0.005
	4	0.199	0.133	0.074	0.036	0.016	0.007	0.003
59	1	0.268	0.224	0.159	0.093	0.045	0.019	0.008
	2*	**0.067**	**0.156**	**0.308**	**0.500**	**0.676**	**0.804**	**0.885**
	3	0.181	0.247	0.287	0.274	0.219	0.153	0.099
	4	0.483	0.373	0.245	0.132	0.060	0.024	0.009
61	1	0.202	0.209	0.176	0.117	0.064	0.032	0.015
	2	0.346	0.271	0.173	0.087	0.036	0.013	0.005
	3*	**0.115**	**0.263**	**0.489**	**0.716**	**0.867**	**0.943**	**0.976**
	4	0.336	0.258	0.161	0.079	0.032	0.012	0.004
62	1	0.247	0.193	0.115	0.046	0.014	0.004	0.001
	2	0.388	0.341	0.229	0.103	0.034	0.010	0.003

	3*	**0.060**	**0.191**	**0.465**	**0.762**	**0.922**	**0.977**	**0.994**
	4	0.305	0.276	0.191	0.089	0.030	0.009	0.003
64	1	0.293	0.231	0.152	0.084	0.040	0.018	0.007
	2	0.396	0.274	0.158	0.077	0.032	0.013	0.005
	3*	**0.133**	**0.266**	**0.444**	**0.619**	**0.754**	**0.845**	**0.902**
	4	0.178	0.229	0.246	0.221	0.173	0.125	0.086
66	1	0.324	0.206	0.098	0.036	0.011	0.003	0.001
	2	0.134	0.138	0.106	0.063	0.032	0.016	0.007
	3	0.362	0.259	0.138	0.057	0.021	0.007	0.002
	4*	**0.180**	**0.398**	**0.658**	**0.844**	**0.936**	**0.974**	**0.989**
67	**1***	**0.074**	**0.168**	**0.329**	**0.537**	**0.726**	**0.854**	**0.926**
	2	0.550	0.484	0.371	0.236	0.125	0.057	0.024
	3	0.143	0.174	0.186	0.165	0.121	0.077	0.046
	4	0.233	0.174	0.114	0.062	0.028	0.011	0.004
68	1	0.191	0.200	0.183	0.144	0.102	0.066	0.042
	2	0.285	0.213	0.138	0.078	0.039	0.018	0.008
	3	0.334	0.250	0.162	0.091	0.046	0.021	0.009
	4*	**0.191**	**0.337**	**0.517**	**0.687**	**0.814**	**0.894**	**0.941**
69	1	0.194	0.200	0.160	0.101	0.054	0.027	0.013
	2	0.513	0.327	0.162	0.063	0.021	0.006	0.002
	3	0.146	0.323	0.559	0.762	0.885	0.947	0.976
	4*	**0.147**	**0.150**	**0.119**	**0.074**	**0.039**	**0.019**	**0.009**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 27
Probabilities of Selecting Item Options in NRM: SS-ID1-S1

Item	Option	Probability of Selecting Item Option at θ level -3	-2	-1	0	1	2	3
1	1	0.246	0.128	0.058	0.023	0.009	0.003	0.001
	2	0.107	0.074	0.044	0.023	0.011	0.005	0.002
	3	0.410	0.413	0.358	0.278	0.199	0.135	0.088
	4*	**0.237**	**0.385**	**0.540**	**0.676**	**0.782**	**0.857**	**0.908**
7	**1***	**0.159**	**0.302**	**0.486**	**0.667**	**0.805**	**0.894**	**0.944**
	2	0.296	0.257	0.190	0.119	0.066	0.034	0.016
	3	0.302	0.302	0.256	0.185	0.118	0.069	0.038
	4	0.242	0.140	0.068	0.029	0.010	0.004	0.001
9	1	0.638	0.434	0.219	0.080	0.023	0.006	0.001
	2	0.177	0.263	0.289	0.230	0.146	0.082	0.044
	3*	**0.065**	**0.191**	**0.415**	**0.651**	**0.815**	**0.906**	**0.953**
	4	0.119	0.111	0.077	0.039	0.016	0.006	0.002
11	1	0.165	0.107	0.063	0.034	0.018	0.009	0.005
	2	0.049	0.027	0.013	0.006	0.003	0.001	0.000
	3	0.313	0.240	0.167	0.108	0.067	0.040	0.024
	4*	**0.472**	**0.627**	**0.757**	**0.851**	**0.912**	**0.949**	**0.971**
12	**1***	**0.088**	**0.167**	**0.274**	**0.397**	**0.523**	**0.641**	**0.742**
	2	0.235	0.144	0.076	0.036	0.015	0.006	0.002
	3	0.478	0.572	0.591	0.541	0.450	0.348	0.254
	4	0.198	0.117	0.059	0.027	0.011	0.004	0.001
13	1	0.418	0.505	0.457	0.304	0.159	0.072	0.031
	2	0.501	0.295	0.130	0.042	0.011	0.002	0.000
	3	0.021	0.023	0.019	0.011	0.005	0.002	0.001
	4*	**0.060**	**0.177**	**0.394**	**0.643**	**0.825**	**0.923**	**0.968**
14	1	0.462	0.365	0.223	0.106	0.042	0.016	0.006
	2*	**0.158**	**0.354**	**0.613**	**0.820**	**0.928**	**0.974**	**0.990**
	3	0.177	0.148	0.096	0.048	0.021	0.008	0.003
	4	0.203	0.133	0.067	0.026	0.009	0.003	0.001
16	1	0.213	0.133	0.072	0.033	0.013	0.004	0.001
	2	0.443	0.413	0.334	0.226	0.129	0.065	0.030
	3*	**0.094**	**0.201**	**0.373**	**0.579**	**0.758**	**0.876**	**0.940**
	4	0.250	0.253	0.221	0.163	0.101	0.055	0.028
23	1	0.333	0.299	0.230	0.137	0.062	0.023	0.008
	2	0.566	0.503	0.383	0.225	0.100	0.037	0.013
	3	0.049	0.057	0.057	0.044	0.026	0.012	0.006
	4*	**0.052**	**0.141**	**0.330**	**0.594**	**0.812**	**0.927**	**0.974**

24	1	0.132	0.120	0.089	0.057	0.034	0.020	0.011
	2	0.139	0.121	0.086	0.053	0.031	0.017	0.009
	3	0.381	0.194	0.080	0.029	0.010	0.003	0.001
	4*	**0.348**	**0.565**	**0.745**	**0.860**	**0.925**	**0.960**	**0.978**
29	1	0.188	0.128	0.066	0.025	0.008	0.002	0.001
	2	0.501	0.402	0.242	0.109	0.041	0.014	0.005
	3*	**0.146**	**0.349**	**0.626**	**0.838**	**0.941**	**0.980**	**0.993**
	4	0.165	0.121	0.067	0.027	0.009	0.003	0.001
34	1	0.287	0.258	0.207	0.147	0.093	0.053	0.029
	2	0.262	0.176	0.106	0.056	0.027	0.011	0.005
	3	0.324	0.335	0.309	0.252	0.183	0.121	0.075
	4*	**0.126**	**0.231**	**0.378**	**0.545**	**0.698**	**0.815**	**0.892**
35	**1***	**0.157**	**0.290**	**0.472**	**0.662**	**0.811**	**0.904**	**0.953**
	2	0.332	0.290	0.223	0.148	0.085	0.045	0.022
	3	0.166	0.140	0.103	0.066	0.037	0.018	0.009
	4	0.345	0.281	0.202	0.125	0.067	0.033	0.015
36	1	0.258	0.151	0.072	0.029	0.010	0.003	0.001
	2*	**0.136**	**0.281**	**0.474**	**0.660**	**0.799**	**0.887**	**0.938**
	3	0.282	0.204	0.120	0.059	0.025	0.010	0.004
	4	0.324	0.364	0.334	0.253	0.166	0.100	0.058
37	**1***	**0.113**	**0.344**	**0.682**	**0.897**	**0.972**	**0.993**	**0.998**
	2	0.415	0.327	0.168	0.057	0.016	0.004	0.001
	3	0.210	0.159	0.079	0.026	0.007	0.002	0.000
	4	0.262	0.169	0.071	0.020	0.005	0.001	0.000
38	1	0.162	0.182	0.190	0.188	0.180	0.168	0.155
	2*	**0.388**	**0.495**	**0.588**	**0.664**	**0.723**	**0.769**	**0.805**
	3	0.162	0.092	0.049	0.024	0.012	0.006	0.003
	4	0.287	0.231	0.174	0.124	0.085	0.057	0.038
42	1	0.451	0.457	0.341	0.181	0.076	0.029	0.011
	2*	**0.087**	**0.249**	**0.519**	**0.771**	**0.911**	**0.968**	**0.989**
	3	0.199	0.143	0.076	0.029	0.009	0.002	0.001
	4	0.263	0.151	0.064	0.019	0.005	0.001	0.000
48	1	0.271	0.315	0.298	0.232	0.154	0.092	0.051
	2*	**0.094**	**0.211**	**0.386**	**0.582**	**0.748**	**0.861**	**0.927**
	3	0.335	0.178	0.077	0.028	0.008	0.002	0.001
	4	0.300	0.296	0.239	0.159	0.090	0.045	0.022
55	1	0.250	0.153	0.063	0.018	0.004	0.001	0.000
	2	0.461	0.344	0.172	0.059	0.017	0.004	0.001
	3	0.168	0.155	0.095	0.040	0.014	0.005	0.001
	4*	**0.121**	**0.348**	**0.670**	**0.883**	**0.965**	**0.990**	**0.997**
56	**1***	**0.132**	**0.265**	**0.452**	**0.646**	**0.796**	**0.891**	**0.943**

	2	0.299	0.233	0.154	0.085	0.040	0.017	0.007
	3	0.265	0.179	0.103	0.049	0.020	0.008	0.003
	4	0.305	0.323	0.291	0.219	0.143	0.084	0.047
58	1	0.326	0.275	0.183	0.099	0.048	0.021	0.009
	2	0.249	0.183	0.106	0.050	0.021	0.008	0.003
	3*	**0.220**	**0.432**	**0.665**	**0.835**	**0.927**	**0.969**	**0.987**
	4	0.206	0.110	0.046	0.016	0.005	0.001	0.000
60	**1***	**0.088**	**0.261**	**0.540**	**0.784**	**0.915**	**0.969**	**0.989**
	2	0.376	0.190	0.067	0.017	0.003	0.001	0.000
	3	0.231	0.216	0.141	0.065	0.024	0.008	0.003
	4	0.305	0.332	0.252	0.135	0.058	0.023	0.008
65	1	0.341	0.422	0.389	0.217	0.078	0.023	0.007
	2	0.574	0.406	0.214	0.068	0.014	0.002	0.000
	3*	**0.022**	**0.101**	**0.338**	**0.685**	**0.898**	**0.972**	**0.992**
	4	0.063	0.071	0.059	0.030	0.010	0.003	0.001

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 28
Results of Classical Item Analysis: SS-NTW-S2

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
4	1	325	0.08	-0.18	
	2	721	0.18	-0.22	
	3*	**2427**	**0.61**	**0.45**	**0.39**
	4	523	0.13	-0.26	
5	**1***	**1899**	**0.48**	**0.38**	**0.30**
	2	247	0.06	-0.15	
	3	819	0.21	-0.14	
	4	1032	0.26	-0.23	
6	1	345	0.09	-0.27	
	2*	**2958**	**0.74**	**0.41**	**0.35**
	3	403	0.10	-0.23	
	4	289	0.07	-0.14	
15	1	268	0.07	-0.23	
	2	149	0.04	-0.17	
	3	199	0.05	-0.20	
	4*	**3382**	**0.85**	**0.37**	**0.31**
17	1	407	0.10	-0.21	
	2	574	0.14	-0.21	
	3	412	0.10	-0.27	
	4*	**2604**	**0.65**	**0.46**	**0.40**
18	1	313	0.08	-0.25	
	2*	**3334**	**0.83**	**0.41**	**0.36**
	3	113	0.03	-0.19	
	4	238	0.06	-0.24	
19	1	263	0.07	-0.19	
	2	570	0.14	-0.20	
	3*	**2445**	**0.61**	**0.40**	**0.33**
	4	715	0.18	-0.19	
20	1	150	0.04	-0.19	
	2	230	0.06	-0.22	
	3*	**3329**	**0.83**	**0.39**	**0.33**
	4	288	0.07	-0.22	
25	1	472	0.12	-0.18	
	2	585	0.15	-0.15	
	3*	**2207**	**0.55**	**0.41**	**0.34**
	4	735	0.18	-0.24	
26	1	387	0.10	-0.33	
	2*	**2504**	**0.63**	**0.44**	**0.37**
	3	707	0.18	-0.16	
	4	399	0.10	-0.18	
27	1	407	0.10	-0.17	

	2*	**2203**	**0.55**	**0.52**	**0.45**
	3	917	0.23	-0.23	
	4	472	0.12	-0.33	
28	1	166	0.04	-0.22	
	2	259	0.07	-0.19	
	3	121	0.03	-0.17	
	4*	**3454**	**0.86**	**0.35**	**0.30**
30	**1***	**2552**	**0.64**	**0.40**	**0.34**
	2	492	0.12	-0.20	
	3	224	0.06	-0.17	
	4	729	0.18	-0.23	
33	**1***	**1341**	**0.34**	**0.33**	**0.26**
	2	658	0.17	-0.10	
	3	1425	0.36	-0.08	
	4	575	0.14	-0.22	
39	**1***	**2136**	**0.53**	**0.46**	**0.39**
	2	547	0.14	-0.19	
	3	531	0.13	-0.20	
	4	785	0.20	-0.25	
40	1	278	0.07	-0.12	
	2	424	0.11	-0.15	
	3	671	0.17	-0.22	
	4*	**2625**	**0.66**	**0.33**	**0.26**
41	1	425	0.11	-0.21	
	2	560	0.14	-0.11	
	3	213	0.05	-0.14	
	4*	**2800**	**0.70**	**0.29**	**0.22**
43	1	295	0.07	-0.25	
	2	479	0.12	-0.15	
	3	355	0.09	-0.17	
	4*	**2869**	**0.72**	**0.36**	**0.29**
45	1	612	0.15	-0.10	
	2*	**2619**	**0.66**	**0.37**	**0.31**
	3	307	0.08	-0.29	
	4	459	0.12	-0.20	
46	1	360	0.09	-0.25	
	2	241	0.06	-0.11	
	3*	**2826**	**0.71**	**0.43**	**0.37**
	4	573	0.14	-0.28	
47	1	359	0.09	-0.31	
	2	477	0.12	-0.21	
	3*	**2931**	**0.73**	**0.43**	**0.37**
	4	230	0.06	-0.16	
49	1	411	0.10	-0.12	
	2*	**1200**	**0.30**	**0.21**	**0.14**
	3	1393	0.35	-0.12	

	4	990	0.25	-0.01	
51	1	254	0.06	-0.19	
	2	296	0.07	-0.10	
	3*	**2453**	**0.61**	**0.23**	**0.15**
	4	994	0.25	-0.09	
52	1	258	0.06	-0.23	
	2*	**2574**	**0.64**	**0.47**	**0.41**
	3	814	0.20	-0.26	
	4	352	0.09	-0.23	
53	**1***	**3169**	**0.79**	**0.39**	**0.33**
	2	233	0.06	-0.19	
	3	236	0.06	-0.19	
	4	359	0.09	-0.23	
54	1	506	0.13	-0.09	
	2*	**2320**	**0.58**	**0.49**	**0.42**
	3	592	0.15	-0.31	
	4	580	0.15	-0.28	
57	1	407	0.10	-0.14	
	2*	**3091**	**0.77**	**0.30**	**0.23**
	3	315	0.08	-0.20	
	4	185	0.05	-0.14	
59	1	423	0.11	-0.20	
	2*	**1937**	**0.48**	**0.38**	**0.31**
	3	1033	0.26	-0.09	
	4	605	0.15	-0.26	
61	1	509	0.13	-0.18	
	2	410	0.10	-0.22	
	3*	**2690**	**0.67**	**0.41**	**0.34**
	4	390	0.10	-0.22	
62	1	249	0.06	-0.21	
	2	476	0.12	-0.27	
	3*	**2878**	**0.72**	**0.46**	**0.40**
	4	395	0.10	-0.24	
64	1	353	0.09	-0.18	
	2	379	0.10	-0.25	
	3*	**2463**	**0.62**	**0.35**	**0.28**
	4	804	0.20	-0.12	
66	1	210	0.05	-0.20	
	2	285	0.07	-0.15	
	3	303	0.08	-0.25	
	4*	**3200**	**0.80**	**0.37**	**0.32**
67	**1***	**2082**	**0.52**	**0.39**	**0.32**
	2	968	0.24	-0.26	
	3	625	0.16	-0.12	
	4	323	0.08	-0.15	
68	1	625	0.16	-0.12	

	2	300	0.08	-0.17	
	3	402	0.10	-0.20	
	4*	**2665**	**0.67**	**0.32**	**0.25**
69	1	438	0.11	-0.18	
	2	382	0.10	-0.29	
	3*	**2883**	**0.72**	**0.41**	**0.35**
	4	296	0.07	-0.16	

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Table 29
Results of Classical Item Analysis: SS-ID1-S2

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
1	1	137	0.03	-0.17	
	2	110	0.03	-0.13	
	3	1069	0.27	-0.24	
	4*	**2681**	**0.67**	**0.35**	**0.24**
7	**1***	**2554**	**0.64**	**0.41**	**0.31**
	2	592	0.15	-0.21	
	3	697	0.17	-0.23	
	4	156	0.04	-0.17	
9	1	467	0.12	-0.34	
	2	869	0.22	-0.18	
	3*	**2495**	**0.62**	**0.44**	**0.34**
	4	167	0.04	-0.14	
11	1	147	0.04	-0.14	
	2	21	0.01	-0.06	
	3	479	0.12	-0.21	
	4*	**3353**	**0.84**	**0.27**	**0.19**
12	**1***	**1622**	**0.41**	**0.32**	**0.21**
	2	178	0.05	-0.15	
	3	2064	0.52	-0.21	
	4	133	0.03	-0.12	
13	1	1212	0.30	-0.33	
	2	245	0.06	-0.21	
	3	82	0.02	-0.14	
	4*	**2461**	**0.62**	**0.45**	**0.35**
14	1	491	0.12	-0.31	
	2*	**3144**	**0.79**	**0.41**	**0.32**
	3	212	0.05	-0.15	
	4	153	0.04	-0.18	
16	1	159	0.04	-0.14	
	2	866	0.22	-0.22	
	3*	**2281**	**0.57**	**0.42**	**0.32**
	4	690	0.17	-0.24	
23	1	610	0.15	-0.21	
	2	994	0.25	-0.35	
	3	142	0.04	-0.08	
	4*	**2250**	**0.56**	**0.49**	**0.39**
24	1	246	0.06	-0.10	
	2	261	0.07	-0.16	
	3	207	0.05	-0.27	
	4*	**3286**	**0.82**	**0.32**	**0.23**

29	1	145	0.04	-0.20	
	2	548	0.14	-0.28	
	3*	**3173**	**0.79**	**0.41**	**0.33**
	4	132	0.03	-0.18	
34	1	637	0.16	-0.17	
	2	254	0.06	-0.16	
	3	968	0.24	-0.20	
	4*	**2138**	**0.54**	**0.37**	**0.26**
35	**1***	**2607**	**0.65**	**0.37**	**0.27**
	2	584	0.15	-0.19	
	3	337	0.08	-0.16	
	4	471	0.12	-0.21	
36	1	137	0.03	-0.17	
	2*	**2560**	**0.64**	**0.41**	**0.31**
	3	316	0.08	-0.20	
	4	986	0.25	-0.26	
37	**1***	**3343**	**0.84**	**0.44**	**0.36**
	2	345	0.09	-0.27	
	3	153	0.04	-0.23	
	4	159	0.04	-0.22	
38	1	778	0.20	-0.08	
	2*	**2544**	**0.64**	**0.23**	**0.12**
	3	130	0.03	-0.12	
	4	545	0.14	-0.16	
42	1	815	0.20	-0.33	
	2*	**2855**	**0.71**	**0.45**	**0.36**
	3	169	0.04	-0.16	
	4	158	0.04	-0.19	
48	1	913	0.23	-0.23	
	2*	**2249**	**0.56**	**0.41**	**0.30**
	3	171	0.04	-0.16	
	4	665	0.17	-0.19	
55	1	130	0.03	-0.18	
	2	339	0.09	-0.28	
	3	228	0.06	-0.22	
	4*	**3295**	**0.82**	**0.42**	**0.34**
56	**1***	**2485**	**0.62**	**0.41**	**0.31**
	2	389	0.10	-0.20	
	3	249	0.06	-0.18	
	4	875	0.22	-0.23	
58	1	449	0.11	-0.21	
	2	250	0.06	-0.23	
	3*	**3194**	**0.80**	**0.37**	**0.28**
	4	104	0.03	-0.17	
60	**1***	**2935**	**0.73**	**0.43**	**0.34**
	2	144	0.04	-0.21	

	3	325	0.08	-0.23	
	4	594	0.15	-0.25	
65	1	950	0.24	-0.34	
	2	423	0.11	-0.29	
	3*	**2506**	**0.63**	**0.53**	**0.44**
	4	119	0.03	-0.14	

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Item Response Theory Assumption Tests for the SS-NTW-S2 and the SS-ID1-S2

Table 30
Unidimensionality Tests : SS-NTW-S2 and SS-ID1-S2

Principal Component Factor Analysis Eigenvalues	SS-NTW-S2	SS-ID1-S2
1	5.46	3.77
2	1.31	1.08
3	1.11	1.06
4	1.05	1.04
5	1.03	1.00
6	1.01	
DIMTEST		
Stout's T	1.49	0.52
p-value	0.08	0.30

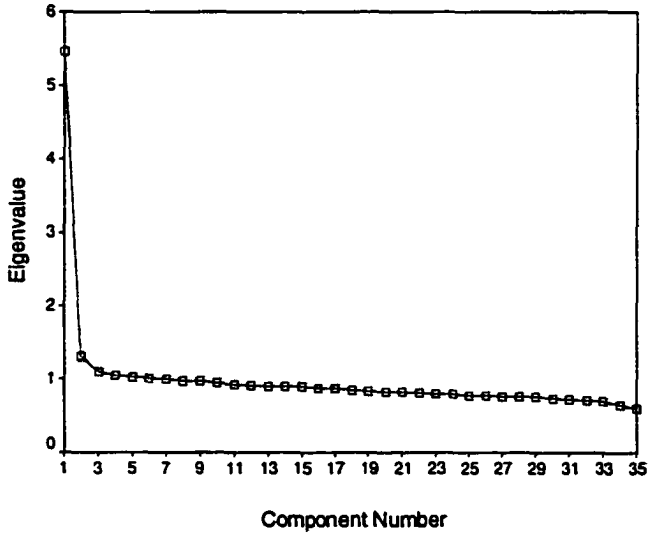


Figure 17. Scree Plot for the SS-NTW-S2

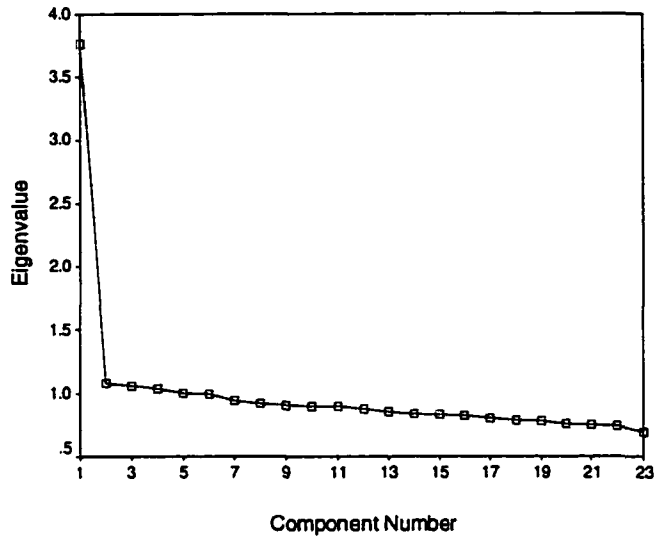


Figure 18. Scree Plot for the SS-ID1

The three methods used to assess unidimensionality of the SS-NTW-S2 and SS-ID1-S2 suggested that the hypothesis of the essential unidimensionality was tenable for both subtests.

The tenable assumption of unidimensionality indicates tenability of the assumption of local independence for the SS-NTW-S2 and SS-ID1-S2.

Table 31
Equal Discrimination Tests: SS-NTW-S2 and SS-ID1-S2

Point-Biserial Correlation	SS-NTW-S2	SS-ID1-S2
Minimum Value	0.21 (item 49)	0.23 (item 38)
Maximum Value	0.52 (item 27)	0.53 (item 65)

Large range of the point-biserial correlations for the SS-NTW-S2 and the SS-ID1-S2 indicates failure to meet the assumption of equal item discrimination. Therefore, the one-parameter model may not be appropriate for both subtests.

Table 32
Non-Speededness Tests: SS-NTW-S2 and SS-ID1-S2

Number of Examinees who did not Answer	SS-NTW-S2	SS-ID1-S2
Last Item	1	6
Last Two Items	0	0
Last Three Items	0	0

Given that, of 4000 examinees, one (less than 1.0%) did not complete one of the last three items on the SS-NTW-S2 subtest, and six (less than 1.0%) did not complete one of the last three items on the SS-ID1-S2 subtest, it was concluded that the assumption of non-speededness was tenable for both subtests.

Table 33
Non-Guessing Tests: SS-NTW-S2 and SS-ID1-S2

		SS-NTW-S2		SS-ID1-S2	
		Item	p-value	Item	p-value
Most Difficult Items	1	49	0.30	12	0.41
	2	33	0.34	34	0.53
	3	5	0.47	23	0.56
Number of Low Achievers Correctly Answering the Most Difficult Items	1 Item	54		64	
	2 Items	6		20	
	3 Items	0		2	

Examination of the performance of low achieving examinees on the three most difficult items on both subtests revealed that the effect of guessing was minimal. Additional evidence supporting this conclusion is provided by the fact that only 20 (less than 1.0%) out of 4000 examinees scored at or below the chance level (score of 17).

Table 34
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: SS-NTW-S2

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
4	-0.57	1.08	-0.51	1.13	0.22	0.30
5	0.13	0.77	0.14	0.55	0.52	0.13
6	-1.36	1.08	-1.19	0.81	-0.51	0.30
15	-2.18	1.16	-1.80	0.75	-1.34	0.26
17	-0.82	1.17	-0.69	0.91	-0.17	0.23
18	-2.07	1.34	-1.57	0.82	-1.31	0.19
19	-0.6	0.86	-0.62	0.59	-0.17	0.16
20	-2.06	1.20	-1.67	0.84	-1.05	0.33
25	-0.28	0.88	-0.28	0.90	0.45	0.27
26	-0.68	1.07	-0.60	1.26	0.23	0.34
27	-0.28	1.34	-0.22	1.21	0.19	0.19
28	-2.36	1.14	-1.97	0.70	-1.61	0.23
30	-0.75	0.91	-0.74	0.68	-0.16	0.22
33	0.89	0.67	1.12	0.59	1.41	0.13
39	-0.19	1.06	-0.17	0.76	0.14	0.13
40	-0.85	0.67	-1.07	0.45	-0.48	0.18
41	-1.11	0.57	-1.59	0.38	-0.84	0.21
43	-1.22	0.81	-1.32	0.66	-0.25	0.36
45	-0.84	0.79	-0.92	0.54	-0.38	0.19
46	-1.15	1.10	-1.00	0.79	-0.45	0.25
47	-1.32	1.14	-1.11	0.89	-0.43	0.31
49	1.1	0.37	2.34	1.28	1.89	0.25
51	-0.61	0.37	-1.29	0.56	1.25	0.47
52	-0.78	1.20	-0.64	0.98	-0.12	0.23
53	-1.73	1.06	-1.53	0.72	-0.95	0.27
54	-0.43	1.19	-0.36	0.83	-0.06	0.13
57	-1.59	0.67	-1.99	0.45	-1.22	0.26
59	0.08	0.79	0.08	0.71	0.64	0.20
61	-0.95	0.93	-0.92	0.66	-0.38	0.21
62	-1.23	1.28	-0.97	0.96	-0.46	0.25
64	-0.63	0.69	-0.76	0.45	-0.30	0.14
66	-1.79	1.02	-1.63	0.68	-1.08	0.26
67	-0.12	0.80	-0.12	0.69	0.51	0.22
68	-0.91	0.62	-1.21	0.77	0.48	0.46
69	-1.24	1.01	-1.13	0.64	-0.81	0.15

Table 35
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: SS-ID1-S2

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
1	-0.95	0.64	-1.22	0.49	-0.25	0.28
7	-0.77	0.86	-0.78	0.63	-0.17	0.22
9	-0.68	0.92	-0.66	0.64	-0.21	0.17
11	-2.17	0.6	-2.95	0.37	-2.20	0.25
12	0.51	0.52	0.78	0.46	1.37	0.18
13	-0.63	1.01	-0.57	0.92	0.13	0.28
14	-1.73	1.05	-1.50	0.72	-0.94	0.26
16	-0.39	0.84	-0.4	0.72	0.28	0.24
23	-0.34	1.16	-0.29	0.9	0.11	0.18
24	-2.01	0.75	-2.26	0.47	-1.73	0.21
29	-1.78	1.10	-1.51	0.72	-1.10	0.21
34	-0.19	0.69	-0.23	0.65	0.6	0.26
35	-0.84	0.75	-0.94	0.75	0.21	0.37
36	-0.78	0.88	-0.77	0.76	0.04	0.29
37	-2.14	1.38	-1.56	0.88	-1.26	0.19
38	-0.75	0.29	-1.93	0.21	-0.37	0.22
42	-1.22	1.11	-1.03	0.79	-0.54	0.22
48	-0.34	0.82	-0.36	1.13	0.61	0.35
55	-2.03	1.24	-1.58	0.80	-1.22	0.21
56	-0.67	0.85	-0.68	0.66	-0.03	0.23
58	-1.82	0.91	-1.75	0.58	-1.28	0.22
60	-1.35	1.08	-1.16	0.94	-0.31	0.36
65	-0.70	1.44	-0.51	1.09	-0.15	0.18

Table 36
Results of IRT Item Analysis for NRM: SS-NTW-S2

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
4	1	-0.31	-0.87	45	1	0.31	0.10
	2	-0.13	0.02		**2***	**0.75**	**1.54**
	3*	**0.84**	**1.29**		3	-0.91	-1.30
	4	-0.41	-0.45		4	-0.15	-0.34
5	**1***	**0.64**	**0.90**	46	1	-0.50	-0.72
	2	-0.45	-1.27		2	0.04	-0.82
	3	-0.02	0.09		**3***	**0.81**	**1.70**
	4	-0.17	0.28		4	-0.34	-0.15
6	1	-0.53	-0.77	47	1	-0.73	-0.88
	2*	**0.82**	**1.81**		2	-0.04	-0.10
	3	-0.26	-0.42		**3***	**0.92**	**1.86**
	4	-0.02	-0.62		4	-0.14	-0.88
15	1	-0.28	-0.50	49	1	-0.30	-0.83
	2	-0.28	-1.09		**2***	**0.31**	**0.27**
	3	-0.32	-0.84		3	-0.09	0.44
	4*	**0.88**	**2.43**		4	0.08	0.11
17	1	-0.24	-0.56	51	1	-0.54	-1.14
	2	-0.11	-0.15		2	-0.03	-0.75
	3	-0.56	-0.76		**3***	**0.40**	**1.40**
	4*	**0.91**	**1.47**		4	0.17	0.50
18	1	-0.13	-0.21	52	1	-0.58	-1.18
	2	1.08	2.54		**2***	**1.00**	**1.57**
	3	-0.65	-1.70		3	-0.03	0.32
	4	-0.30	-0.62		4	-0.39	-0.72
19	1	-0.49	-1.17	53	**1***	**0.81**	**2.06**
	2	-0.16	-0.21		2	-0.26	-0.83
	3*	**0.70**	**1.32**		3	-0.25	-0.82
	4	-0.05	0.06		4	-0.29	-0.42
20	1	-0.46	-1.28	54	1	0.21	-0.27
	2	-0.27	-0.69		**2***	**0.92**	**1.18**
	3*	**0.92**	**2.38**		3	-0.58	-0.46
	4	-0.19	-0.41		4	-0.54	-0.45
25	1	-0.29	-0.59	57	1	0.05	-0.18
	2	-0.10	-0.30		**2***	**0.54**	**1.89**
	3*	**0.67**	**1.03**		3	-0.34	-0.61
	4	-0.27	-0.14		4	-0.25	-1.10
26	1	-0.98	-1.17	59	1	-0.40	-0.76
	2*	**0.94**	**1.46**		**2***	**0.70**	**0.88**
	3	0.18	0.20		3	0.16	0.32
	4	-0.14	-0.50		4	-0.46	-0.44
27	1	-0.17	-0.56	61	1	-0.05	-0.24

	2*	**1.12**	**1.18**		2	-0.31	-0.58
	3	-0.02	0.32		**3***	**0.72**	**1.49**
	4	-0.93	-0.94		4	-0.36	-0.66
28	1	-0.52	-1.08	62	1	-0.43	-1.04
	2	0.01	-0.22		2	-0.32	-0.32
	3	-0.41	-1.30		**3***	**0.99**	**1.81**
	4*	**0.92**	**2.61**		4	-0.23	-0.44
30	**1***	**0.73**	**1.44**	64	1	-0.26	-0.74
	2	-0.19	-0.32		2	-0.55	-0.84
	3	-0.41	-1.23		**3***	**0.62**	**1.34**
	4	-0.13	0.11		4	0.19	0.24
33	**1***	**0.57**	**0.35**	66	1	-0.39	-1.00
	2	-0.11	-0.29		2	0.04	-0.43
	3	0.03	0.50		3	-0.43	-0.66
	4	-0.50	-0.57		**4***	**0.77**	**2.09**
39	**1***	**0.78**	**0.97**	67	**1***	**0.61**	**0.98**
	2	-0.23	-0.42		2	-0.26	0.17
	3	-0.29	-0.48		3	-0.04	-0.20
	4	-0.27	-0.08		4	-0.31	-0.95
40	1	-0.13	-0.89	68	1	0.09	0.01
	2	-0.15	-0.48		2	-0.32	-0.88
	3	-0.21	-0.04		3	-0.30	-0.58
	4*	**0.49**	**1.41**		**4***	**0.52**	**1.46**
41	1	-0.30	-0.44	69	1	-0.04	-0.26
	2	0.11	-0.02		2	-0.61	-0.76
	3	-0.29	-1.13		**3***	**0.78**	**1.72**
	4*	**0.48**	**1.60**		4	-0.13	-0.69
43	1	-0.61	-1.02				
	2	0.06	-0.15				
	3	-0.12	-0.52				
	4*	**0.66**	**1.68**				

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 37
Results of IRT Analysis for NRM: SS-ID1-S2

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
1	1	-0.52	-1.44	35	**1***	**0.57**	**1.38**
	2	-0.35	-1.53		2	-0.10	-0.16
	3	0.17	1.01		3	-0.22	-0.76
	4*	**0.70**	**1.96**		4	-0.26	-0.45
7	**1***	**0.73**	**1.52**	36	1	-0.61	-1.72
	2	-0.09	-0.02		**2***	**0.83**	**1.67**
	3	-0.07	0.15		3	-0.25	-0.62
	4	-0.57	-1.64		4	0.03	0.66
9	1	-0.76	-0.68	37	**1***	**1.15**	**2.59**
	2	0.20	0.50		2	-0.08	-0.07
	3*	**0.83**	**1.54**		3	-0.52	-1.27
	4	-0.27	-1.35		4	-0.55	-1.26
11	1	-0.20	-0.64	38	1	0.19	0.30
	2	-0.35	-2.68		**2***	**0.33**	**1.48**
	3	0.00	0.65		3	-0.43	-1.69
	4*	**0.55**	**2.68**		4	-0.09	-0.10
12	**1***	**0.65**	**1.21**	42	1	-0.08	0.57
	2	-0.40	-1.19		**2***	**0.95**	**2.00**
	3	0.18	1.49		3	-0.29	-1.14
	4	-0.43	-1.50		4	-0.58	-1.44
13	1	0.09	1.09	48	1	-0.06	0.34
	2	-0.42	-0.82		**2***	**0.71**	**1.27**
	3	-0.66	-2.12		3	-0.58	-1.63
	4*	**0.99**	**1.85**		4	-0.07	0.02
14	1	-0.33	-0.07	55	1	-0.41	-1.38
	2*	**0.80**	**2.11**		2	-0.31	-0.33
	3	-0.01	-0.71		3	-0.25	-0.68
	4	-0.47	-1.34		**4***	**0.97**	**2.38**
16	1	-0.43	-1.58	56	**1***	**0.72**	**1.40**
	2	-0.06	0.29		2	-0.27	-0.60
	3*	**0.69**	**1.28**		3	-0.42	-1.13
	4	-0.19	0.02		4	-0.03	0.32
23	1	-0.27	-0.09	58	1	0.10	0.18
	2	-0.42	0.32		2	-0.35	-0.69
	3	-0.15	-1.50		**3***	**0.82**	**2.25**
	4*	**0.84**	**1.26**		4	-0.56	-1.74
24	1	0.24	-0.42	60	**1***	**0.95**	**2.00**
	2	-0.03	-0.46		2	-0.76	-1.77
	3	-0.85	-1.31		3	-0.22	-0.49
	4*	**0.63**	**2.19**		4	0.02	0.26
29	1	-0.55	-1.38	65	1	-0.11	0.59

	2	-0.02	0.37	2	-0.55	-0.52
	3*	**0.96**	**2.35**	**3***	**1.19**	**1.69**
	4	-0.39	-1.33	4	-0.52	-1.76
34	1	-0.10	-0.15			
	2	-0.44	-1.22			
	3	-0.04	0.29			
	4*	**0.58**	**1.08**			

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 38
Probabilities of Selecting Item Options in NRM: SS-NTW-S2

Item	Option	Probability of Selecting Item Option at θ level -3	-2	-1	0	1	2	3
4	1	0.211	0.184	0.134	0.073	0.031	0.011	0.004
	2	0.299	0.313	0.273	0.179	0.089	0.038	0.015
	3*	**0.058**	**0.160**	**0.368**	**0.636**	**0.838**	**0.938**	**0.977**
	4	0.433	0.343	0.225	0.112	0.042	0.014	0.004
5	**1***	**0.075**	**0.156**	**0.293**	**0.477**	**0.663**	**0.808**	**0.898**
	2	0.225	0.158	0.100	0.054	0.025	0.010	0.004
	3	0.242	0.260	0.252	0.212	0.153	0.096	0.055
	4	0.458	0.425	0.355	0.257	0.159	0.086	0.043
6	1	0.473	0.319	0.161	0.060	0.018	0.005	0.001
	2*	**0.109**	**0.283**	**0.552**	**0.787**	**0.914**	**0.967**	**0.988**
	3	0.299	0.264	0.175	0.085	0.033	0.012	0.004
	4	0.119	0.134	0.112	0.069	0.035	0.016	0.007
15	1	0.341	0.240	0.122	0.048	0.016	0.005	0.002
	2	0.189	0.133	0.068	0.026	0.009	0.003	0.001
	3	0.274	0.185	0.091	0.034	0.011	0.003	0.001
	4*	**0.197**	**0.442**	**0.719**	**0.892**	**0.964**	**0.989**	**0.996**
17	1	0.227	0.223	0.171	0.091	0.037	0.013	0.004
	2	0.232	0.259	0.226	0.138	0.063	0.025	0.009
	3	0.486	0.347	0.192	0.075	0.022	0.005	0.001
	4*	**0.055**	**0.170**	**0.411**	**0.696**	**0.879**	**0.957**	**0.985**
18	1	0.278	0.252	0.146	0.057	0.018	0.006	0.002
	2*	**0.115**	**0.351**	**0.683**	**0.892**	**0.969**	**0.991**	**0.998**
	3	0.299	0.161	0.056	0.013	0.002	0.000	0.000
	4	0.308	0.235	0.115	0.038	0.010	0.003	0.001
19	1	0.310	0.205	0.114	0.052	0.020	0.007	0.002
	2	0.301	0.276	0.215	0.137	0.073	0.035	0.016
	3*	**0.105**	**0.229**	**0.419**	**0.632**	**0.799**	**0.901**	**0.954**
	4	0.283	0.290	0.252	0.179	0.107	0.057	0.029
20	1	0.270	0.164	0.071	0.023	0.006	0.002	0.000
	2	0.276	0.203	0.106	0.041	0.014	0.004	0.001
	3*	**0.167**	**0.404**	**0.694**	**0.882**	**0.961**	**0.988**	**0.996**
	4	0.287	0.229	0.129	0.054	0.019	0.007	0.002
25	1	0.284	0.240	0.179	0.112	0.057	0.026	0.011
	2	0.215	0.220	0.198	0.149	0.093	0.050	0.025
	3*	**0.081**	**0.178**	**0.347**	**0.564**	**0.758**	**0.883**	**0.947**
	4	0.420	0.362	0.276	0.175	0.092	0.042	0.018

26	1	0.756	0.488	0.196	0.048	0.009	0.001	0.000
	2*	**0.033**	**0.146**	**0.398**	**0.668**	**0.840**	**0.926**	**0.966**
	3	0.092	0.189	0.241	0.190	0.111	0.057	0.028
	4	0.119	0.178	0.165	0.094	0.040	0.015	0.005
27	1	0.107	0.158	0.164	0.102	0.040	0.013	0.004
	2*	**0.013**	**0.068**	**0.257**	**0.582**	**0.834**	**0.945**	**0.983**
	3	0.165	0.282	0.340	0.246	0.113	0.041	0.014
	4	0.716	0.493	0.239	0.070	0.013	0.002	0.000
28	1	0.386	0.212	0.079	0.023	0.006	0.001	0.000
	2	0.186	0.174	0.110	0.053	0.023	0.009	0.004
	3	0.223	0.137	0.057	0.018	0.005	0.001	0.000
	4*	**0.206**	**0.477**	**0.753**	**0.906**	**0.966**	**0.988**	**0.995**
30	**1***	**0.107**	**0.236**	**0.440**	**0.664**	**0.832**	**0.924**	**0.968**
	2	0.291	0.256	0.190	0.114	0.057	0.025	0.011
	3	0.227	0.160	0.095	0.046	0.018	0.007	0.002
	4	0.374	0.349	0.275	0.176	0.093	0.044	0.019
33	**1***	**0.048**	**0.101**	**0.192**	**0.324**	**0.481**	**0.634**	**0.761**
	2	0.195	0.208	0.200	0.171	0.128	0.086	0.052
	3	0.282	0.347	0.384	0.376	0.325	0.250	0.175
	4	0.475	0.343	0.224	0.129	0.066	0.030	0.012
39	**1***	**0.050**	**0.129**	**0.297**	**0.545**	**0.773**	**0.906**	**0.965**
	2	0.256	0.243	0.203	0.136	0.070	0.030	0.012
	3	0.289	0.258	0.203	0.128	0.062	0.025	0.009
	4	0.406	0.370	0.297	0.191	0.095	0.039	0.014
40	1	0.140	0.122	0.096	0.067	0.043	0.026	0.015
	2	0.224	0.191	0.147	0.102	0.064	0.037	0.021
	3	0.417	0.335	0.243	0.158	0.093	0.051	0.027
	4*	**0.218**	**0.352**	**0.514**	**0.673**	**0.800**	**0.886**	**0.938**
41	1	0.374	0.265	0.166	0.093	0.049	0.024	0.012
	2	0.166	0.177	0.167	0.142	0.111	0.083	0.060
	3	0.182	0.130	0.082	0.047	0.025	0.012	0.006
	4*	**0.277**	**0.428**	**0.584**	**0.718**	**0.815**	**0.880**	**0.922**
43	1	0.493	0.293	0.135	0.050	0.016	0.005	0.001
	2	0.158	0.183	0.165	0.120	0.076	0.045	0.026
	3	0.187	0.181	0.136	0.083	0.044	0.022	0.010
	4*	**0.162**	**0.343**	**0.564**	**0.747**	**0.864**	**0.929**	**0.963**
45	1	0.070	0.139	0.179	0.164	0.124	0.087	0.059
	2*	**0.079**	**0.243**	**0.488**	**0.691**	**0.816**	**0.888**	**0.931**
	3	0.672	0.393	0.150	0.040	0.009	0.002	0.000
	4	0.179	0.225	0.183	0.105	0.051	0.022	0.010
46	1	0.401	0.293	0.165	0.067	0.022	0.006	0.002

	2	0.072	0.090	0.087	0.061	0.034	0.017	0.008
	3*	**0.089**	**0.240**	**0.500**	**0.754**	**0.900**	**0.962**	**0.985**
	4	0.439	0.377	0.248	0.119	0.045	0.015	0.005
47	1	0.643	0.412	0.178	0.051	0.011	0.002	0.000
	2	0.177	0.226	0.195	0.111	0.050	0.020	0.008
	3*	**0.071**	**0.235**	**0.529**	**0.787**	**0.919**	**0.970**	**0.989**
	4	0.110	0.127	0.099	0.051	0.021	0.008	0.003
49	1	0.238	0.184	0.138	0.099	0.068	0.045	0.029
	2*	**0.115**	**0.164**	**0.225**	**0.297**	**0.377**	**0.461**	**0.544**
	3	0.452	0.431	0.397	0.352	0.300	0.246	0.194
	4	0.195	0.221	0.241	0.253	0.255	0.248	0.233
51	1	0.372	0.212	0.107	0.049	0.022	0.009	0.004
	2	0.119	0.113	0.095	0.073	0.053	0.037	0.026
	3*	**0.281**	**0.410**	**0.528**	**0.624**	**0.699**	**0.759**	**0.806**
	4	0.228	0.264	0.270	0.254	0.226	0.195	0.164
52	1	0.346	0.236	0.123	0.044	0.012	0.003	0.001
	2*	**0.047**	**0.157**	**0.397**	**0.689**	**0.877**	**0.956**	**0.985**
	3	0.297	0.352	0.319	0.197	0.090	0.035	0.013
	4	0.310	0.256	0.161	0.070	0.022	0.006	0.002
53	**1***	**0.167**	**0.371**	**0.635**	**0.837**	**0.938**	**0.978**	**0.992**
	2	0.230	0.175	0.103	0.046	0.018	0.006	0.002
	3	0.225	0.173	0.103	0.047	0.018	0.007	0.002
	4	0.379	0.280	0.160	0.070	0.026	0.009	0.003
54	1	0.055	0.102	0.150	0.144	0.096	0.053	0.027
	2*	**0.028**	**0.105**	**0.313**	**0.616**	**0.830**	**0.929**	**0.969**
	3	0.484	0.410	0.273	0.119	0.036	0.009	0.002
	4	0.434	0.382	0.264	0.121	0.038	0.010	0.002
57	1	0.170	0.163	0.136	0.100	0.068	0.044	0.028
	2*	**0.309**	**0.486**	**0.660**	**0.795**	**0.882**	**0.933**	**0.962**
	3	0.355	0.232	0.131	0.065	0.030	0.013	0.006
	4	0.166	0.119	0.073	0.040	0.020	0.010	0.005
59	1	0.295	0.245	0.171	0.095	0.044	0.017	0.006
	2*	**0.056**	**0.140**	**0.293**	**0.492**	**0.675**	**0.805**	**0.886**
	3	0.162	0.235	0.287	0.281	0.225	0.156	0.100
	4	0.487	0.380	0.250	0.131	0.057	0.021	0.007
61	1	0.209	0.216	0.184	0.125	0.070	0.036	0.017
	2	0.325	0.259	0.170	0.089	0.039	0.015	0.006
	3*	**0.117**	**0.261**	**0.481**	**0.704**	**0.857**	**0.937**	**0.973**
	4	0.349	0.264	0.165	0.082	0.034	0.013	0.004
62	1	0.269	0.205	0.117	0.045	0.013	0.003	0.001
	2	0.397	0.338	0.216	0.093	0.030	0.008	0.002

	3*	**0.066**	**0.207**	**0.491**	**0.780**	**0.928**	**0.979**	**0.994**
	4	0.269	0.250	0.175	0.082	0.029	0.009	0.003
64	1	0.226	0.197	0.138	0.080	0.040	0.018	0.008
	2	0.488	0.318	0.167	0.072	0.027	0.009	0.003
	3*	**0.129**	**0.271**	**0.459**	**0.637**	**0.767**	**0.852**	**0.906**
	4	0.156	0.213	0.235	0.212	0.166	0.120	0.083
66	1	0.267	0.184	0.095	0.038	0.013	0.004	0.001
	2	0.130	0.138	0.110	0.068	0.036	0.018	0.009
	3	0.423	0.280	0.139	0.054	0.018	0.006	0.002
	4*	**0.181**	**0.398**	**0.656**	**0.840**	**0.933**	**0.972**	**0.988**
67	**1***	**0.087**	**0.179**	**0.332**	**0.527**	**0.712**	**0.844**	**0.921**
	2	0.526	0.455	0.352	0.234	0.133	0.066	0.030
	3	0.188	0.202	0.195	0.162	0.114	0.071	0.040
	4	0.199	0.164	0.121	0.077	0.041	0.019	0.008
68	1	0.186	0.202	0.192	0.161	0.122	0.087	0.060
	2	0.262	0.189	0.119	0.066	0.033	0.016	0.007
	3	0.333	0.245	0.157	0.089	0.046	0.022	0.010
	4*	**0.219**	**0.365**	**0.532**	**0.684**	**0.799**	**0.875**	**0.923**
69	1	0.172	0.197	0.167	0.105	0.054	0.026	0.012
	2	0.576	0.373	0.180	0.064	0.019	0.005	0.001
	3*	**0.106**	**0.277**	**0.534**	**0.762**	**0.895**	**0.956**	**0.981**
	4	0.146	0.153	0.119	0.068	0.032	0.014	0.006

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 39
Probabilities of Selecting Item Option in NRM: SS-ID1-S2

Item	Option	Probability of Selecting Item Option at θ level -3	-2	-1	0	1	2	3
1	1	0.264	0.139	0.061	0.023	0.008	0.003	0.001
	2	0.145	0.091	0.047	0.021	0.009	0.003	0.001
	3	0.387	0.406	0.354	0.267	0.182	0.117	0.073
	4*	**0.204**	**0.364**	**0.538**	**0.689**	**0.801**	**0.877**	**0.925**
7	**1***	**0.119**	**0.254**	**0.453**	**0.662**	**0.820**	**0.912**	**0.959**
	2	0.299	0.281	0.220	0.142	0.077	0.038	0.018
	3	0.333	0.320	0.256	0.168	0.094	0.047	0.022
	4	0.249	0.145	0.071	0.028	0.009	0.003	0.001
9	1	0.725	0.487	0.225	0.072	0.018	0.004	0.001
	2	0.133	0.233	0.281	0.233	0.153	0.090	0.051
	3*	**0.057**	**0.187**	**0.423**	**0.659**	**0.814**	**0.900**	**0.947**
	4	0.085	0.094	0.071	0.037	0.015	0.006	0.002
11	1	0.164	0.102	0.058	0.031	0.016	0.008	0.004
	2	0.033	0.018	0.009	0.004	0.002	0.001	0.000
	3	0.326	0.249	0.173	0.112	0.069	0.042	0.025
	4*	**0.477**	**0.631**	**0.760**	**0.853**	**0.913**	**0.950**	**0.971**
12	**1***	**0.098**	**0.175**	**0.280**	**0.403**	**0.532**	**0.651**	**0.752**
	2	0.207	0.130	0.073	0.037	0.017	0.007	0.003
	3	0.529	0.594	0.593	0.533	0.440	0.337	0.243
	4	0.166	0.101	0.055	0.027	0.012	0.005	0.002
13	1	0.452	0.514	0.454	0.301	0.157	0.071	0.030
	2	0.309	0.211	0.112	0.045	0.014	0.004	0.001
	3	0.173	0.093	0.039	0.012	0.003	0.001	0.000
	4*	**0.065**	**0.182**	**0.395**	**0.643**	**0.826**	**0.924**	**0.968**
14	1	0.519	0.389	0.219	0.094	0.034	0.011	0.004
	2*	**0.155**	**0.359**	**0.626**	**0.830**	**0.933**	**0.975**	**0.990**
	3	0.105	0.108	0.084	0.049	0.025	0.011	0.005
	4	0.222	0.144	0.071	0.026	0.008	0.002	0.001
16	1	0.162	0.111	0.066	0.033	0.014	0.005	0.002
	2	0.347	0.343	0.297	0.217	0.134	0.073	0.037
	3*	**0.099**	**0.206**	**0.378**	**0.584**	**0.762**	**0.879**	**0.942**
	4	0.392	0.340	0.258	0.166	0.090	0.043	0.019
23	1	0.272	0.274	0.236	0.151	0.070	0.026	0.009
	2	0.644	0.558	0.413	0.228	0.091	0.029	0.009
	3	0.046	0.053	0.051	0.037	0.019	0.008	0.003
	4*	**0.038**	**0.115**	**0.300**	**0.584**	**0.820**	**0.936**	**0.979**

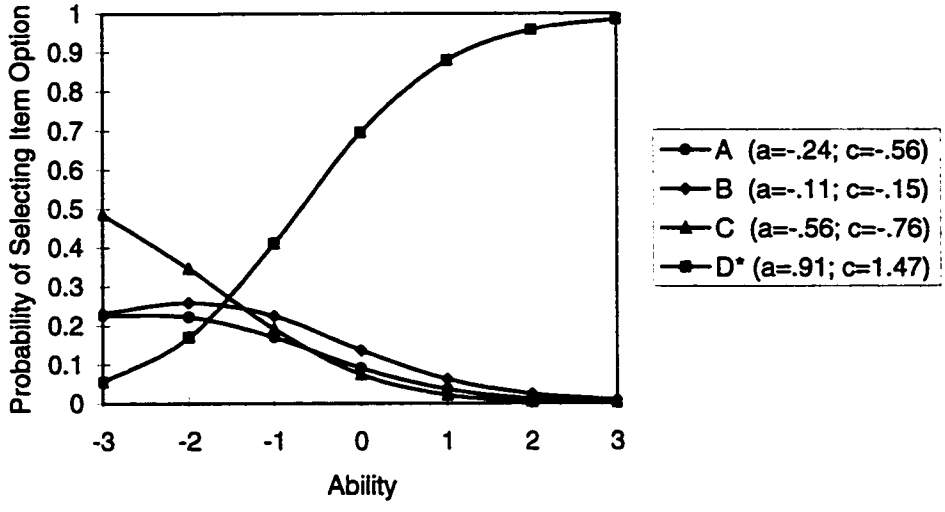
24	1	0.055	0.080	0.079	0.063	0.046	0.032	0.022
	2	0.119	0.132	0.099	0.060	0.033	0.018	0.009
	3	0.594	0.290	0.096	0.026	0.006	0.001	0.000
	4*	**0.232**	**0.498**	**0.726**	**0.852**	**0.915**	**0.949**	**0.968**
29	1	0.305	0.173	0.069	0.020	0.005	0.001	0.000
	2	0.359	0.344	0.234	0.116	0.049	0.019	0.007
	3*	**0.137**	**0.351**	**0.635**	**0.842**	**0.940**	**0.978**	**0.992**
	4	0.199	0.132	0.062	0.021	0.006	0.002	0.000
34	1	0.271	0.254	0.214	0.158	0.104	0.062	0.034
	2	0.258	0.172	0.103	0.054	0.025	0.011	0.004
	3	0.351	0.350	0.313	0.246	0.171	0.108	0.063
	4*	**0.120**	**0.223**	**0.371**	**0.542**	**0.700**	**0.820**	**0.898**
35	**1***	**0.173**	**0.309**	**0.489**	**0.670**	**0.811**	**0.900**	**0.950**
	2	0.276	0.253	0.205	0.144	0.089	0.051	0.027
	3	0.217	0.177	0.127	0.079	0.043	0.022	0.010
	4	0.334	0.261	0.180	0.107	0.057	0.027	0.013
36	1	0.250	0.140	0.063	0.022	0.007	0.002	0.000
	2*	**0.099**	**0.234**	**0.444**	**0.667**	**0.829**	**0.920**	**0.964**
	3	0.255	0.205	0.132	0.068	0.029	0.011	0.004
	4	0.396	0.421	0.360	0.243	0.136	0.068	0.032
37	**1***	**0.096**	**0.328**	**0.681**	**0.899**	**0.972**	**0.993**	**0.998**
	2	0.268	0.268	0.163	0.063	0.020	0.006	0.002
	3	0.302	0.195	0.076	0.019	0.004	0.001	0.000
	4	0.334	0.209	0.079	0.019	0.004	0.001	0.000
38	1	0.180	0.196	0.201	0.198	0.188	0.175	0.160
	2*	**0.384**	**0.482**	**0.569**	**0.643**	**0.703**	**0.752**	**0.791**
	3	0.158	0.093	0.051	0.027	0.014	0.007	0.003
	4	0.279	0.230	0.178	0.132	0.095	0.067	0.046
42	1	0.469	0.460	0.341	0.182	0.077	0.029	0.011
	2*	**0.089**	**0.245**	**0.508**	**0.761**	**0.905**	**0.966**	**0.988**
	3	0.159	0.127	0.076	0.033	0.011	0.003	0.001
	4	0.282	0.168	0.075	0.024	0.006	0.001	0.000
48	1	0.375	0.373	0.318	0.227	0.137	0.074	0.037
	2*	**0.094**	**0.203**	**0.374**	**0.576**	**0.752**	**0.870**	**0.936**
	3	0.249	0.147	0.075	0.032	0.011	0.004	0.001
	4	0.281	0.277	0.233	0.165	0.099	0.052	0.026
55	1	0.198	0.133	0.062	0.020	0.006	0.001	0.000
	2	0.419	0.311	0.161	0.059	0.018	0.005	0.001
	3	0.247	0.194	0.107	0.041	0.013	0.004	0.001
	4*	**0.136**	**0.361**	**0.671**	**0.880**	**0.963**	**0.989**	**0.997**
56	**1***	**0.108**	**0.234**	**0.429**	**0.643**	**0.809**	**0.907**	**0.957**

	2	0.284	0.229	0.156	0.087	0.041	0.017	0.007
	3	0.262	0.182	0.107	0.051	0.021	0.007	0.002
	4	0.347	0.356	0.308	0.218	0.130	0.069	0.034
58	1	0.218	0.224	0.172	0.105	0.057	0.029	0.014
	2	0.352	0.231	0.113	0.044	0.015	0.005	0.002
	3*	**0.199**	**0.421**	**0.665**	**0.835**	**0.924**	**0.965**	**0.984**
	4	0.231	0.123	0.049	0.015	0.004	0.001	0.000
60	**1***	**0.095**	**0.271**	**0.544**	**0.780**	**0.910**	**0.965**	**0.987**
	2	0.370	0.191	0.069	0.018	0.004	0.001	0.000
	3	0.263	0.233	0.145	0.065	0.023	0.008	0.002
	4	0.271	0.305	0.242	0.137	0.063	0.026	0.011
65	1	0.382	0.448	0.404	0.226	0.081	0.024	0.007
	2	0.471	0.356	0.207	0.074	0.017	0.003	0.001
	3*	**0.023**	**0.100**	**0.331**	**0.678**	**0.896**	**0.972**	**0.993**
	4	0.124	0.097	0.058	0.022	0.005	0.001	0.000

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Examples of Trace Lines for Two Non-Susceptible to Testwiseness Items

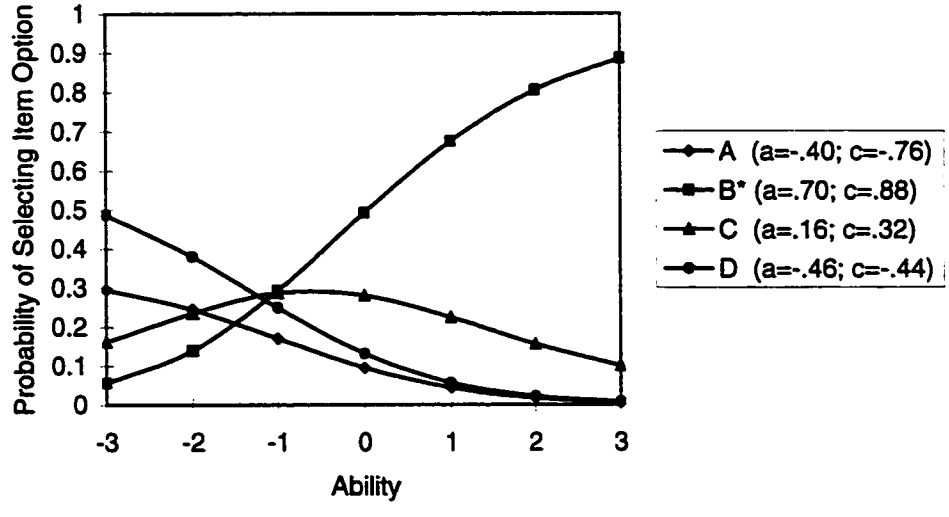
Item 17.



Note: D* indicates the correct option; A, B, and C are item foils

Figure 19. Non-Susceptible to Testwiseness Item 17, Social Studies 30, Sample 2

Item 59

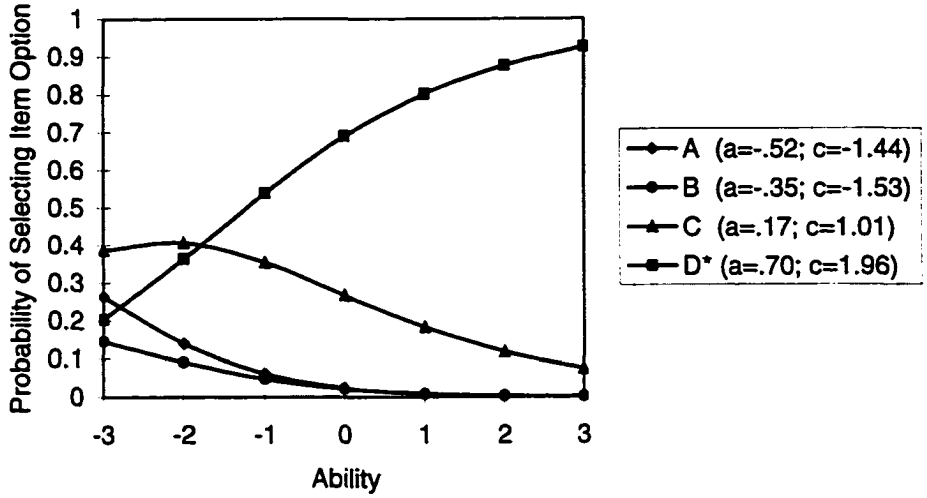


Note: B* indicates the correct option; A, C, and D are item foils

Figure 20. Non-Susceptible to Testwiseness Item 59, Social Studies 30, Sample 2

Examples of Trace Lines for Two Items Susceptible to the ID1 Testwiseness Strategy

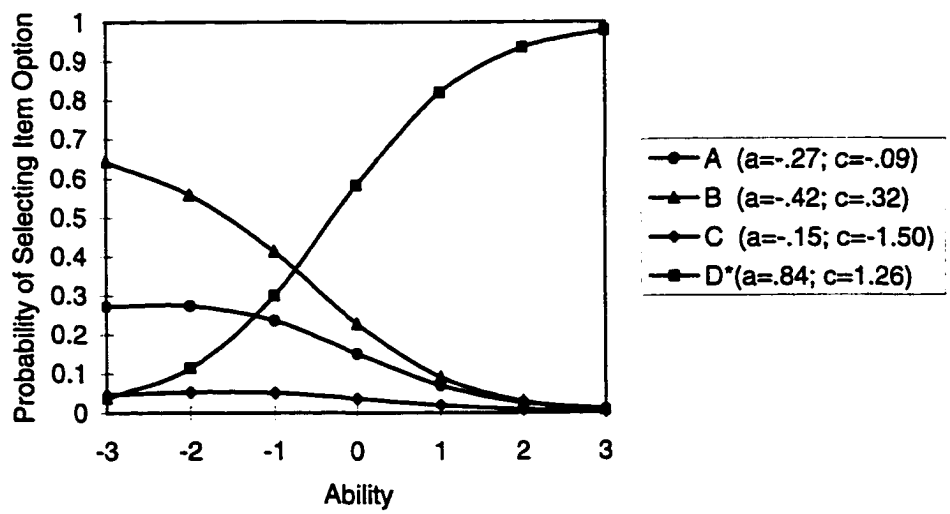
Item 1



Note: D* indicates the correct option; A and B are the absurd options; C is the properly functioning foil

Figure 21. Susceptible to ID1 Testwiseness Strategy Item 1, Social Studies 30, Sample 2

Item 23



Note: D* indicates the correct option; C is the absurd option; A and B are properly functioning foils.

Figure 22. Susceptible to ID1 Testwiseness Strategy Item 23, Social Studies 30, Sample 2

Table 40
Mean Item Parameter for the Social Studies Examination

Model	Mean Parameter	SS-NTW-S1	SS-NTW-S2	SS-ID1-S1	SS-ID1-S2
Number Right	p_i	0.65	0.65	0.68	0.67
		(0.13)	(0.13)	(0.11)	(0.11)
One-Parameter Model	b_i	-0.89	-0.87	-1.05	-1.03
		(0.79)	(0.79)	(0.73)	(0.72)
Two-Parameter Model	a_i	0.93	0.94	0.89	0.91
		(0.25)	(0.26)	(0.28)	(0.27)
	b_i	-0.80	-0.80	-1.06	-1.04
		(0.89)	(0.86)	(0.81)	(0.79)
Three-Parameter Model	a_i	0.72	0.76	0.67	0.71
		(0.20)	(0.22)	(0.21)	(0.22)
	b_i	-0.23	-0.19	-0.44	-0.37
		(0.78)	(0.80)	(0.80)	(0.83)
	c_i	0.23	0.24	0.23	0.24
		(0.06)	(0.08)	(0.06)	(0.06)
Nominal Response Model	a_{ik}	0.74	0.75	0.76	0.79
		(0.19)	(0.19)	(0.19)	(0.19)
	c_{ik}	1.52	1.52	1.78	1.79
		(0.52)	(0.53)	(0.46)	(0.47)

Note: The mean parameter values are reported in the upper part of each row and the standard deviations are shown in parentheses.

Appendix E

Chemistry 30 Diploma Examination: Item Analysis

Table 41
Results of Classical Item Analysis: CH-NTW-S1

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
1	1	188	0.05	-0.24	
	2	95	0.02	-0.08	
	3*	**3474**	**0.87**	**0.29**	**0.21**
	4	241	0.06	-0.15	
2	1	78	0.02	-0.13	
	2*	**3568**	**0.89**	**0.28**	**0.20**
	3	221	0.06	-0.19	
	4	133	0.03	-0.13	
4	1	167	0.04	-0.20	
	2*	**3306**	**0.83**	**0.44**	**0.35**
	3	123	0.03	-0.19	
	4	403	0.10	-0.31	
6	1	300	0.08	-0.18	
	2	167	0.04	-0.19	
	3*	**2512**	**0.63**	**0.47**	**0.36**
	4	1020	0.26	-0.33	
9	**1***	**1701**	**0.43**	**0.39**	**0.27**
	2	1016	0.25	-0.18	
	3	911	0.23	-0.21	
	4	322	0.08	-0.09	
13	**1***	**3161**	**0.79**	**0.40**	**0.30**
	2	357	0.09	-0.23	
	3	266	0.07	-0.25	
	4	212	0.05	-0.14	
15	**1***	**3193**	**0.80**	**0.43**	**0.33**
	2	203	0.05	-0.21	
	3	166	0.04	-0.20	
	4	435	0.11	-0.28	
16	1	381	0.10	-0.20	
	2	892	0.22	-0.12	
	3*	**2288**	**0.57**	**0.39**	**0.26**
	4	439	0.11	-0.26	
17	1	500	0.13	-0.07	
	2	1084	0.27	-0.34	
	3*	**1943**	**0.49**	**0.46**	**0.35**
	4	471	0.12	-0.17	
21	**1***	**3334**	**0.83**	**0.39**	**0.30**

	2	155	0.04	-0.18	
	3	127	0.03	-0.21	
	4	384	0.10	-0.24	
22	1	494	0.12	-0.27	
	2	360	0.09	-0.13	
	3*	**2703**	**0.68**	**0.45**	**0.34**
	4	439	0.11	-0.27	
26	1	550	0.14	-0.10	
	2*	**2319**	**0.58**	**0.42**	**0.29**
	3	523	0.13	-0.28	
	4	599	0.15	-0.21	
27	1	590	0.15	-0.19	
	2*	**2284**	**0.57**	**0.39**	**0.26**
	3	597	0.15	-0.18	
	4	526	0.13	-0.18	
29	1	575	0.14	-0.27	
	2	461	0.12	-0.20	
	3	488	0.12	-0.27	
	4*	**2473**	**0.62**	**0.51**	**0.40**
31	1	998	0.25	-0.23	
	2	440	0.11	-0.20	
	3*	**2297**	**0.57**	**0.41**	**0.28**
	4	256	0.06	-0.16	
32	1	225	0.06	-0.25	
	2	184	0.05	-0.15	
	3*	**3420**	**0.86**	**0.34**	**0.25**
	4	168	0.04	-0.15	
34	**1***	**1969**	**0.49**	**0.35**	**0.22**
	2	997	0.25	-0.12	
	3	399	0.10	-0.19	
	4	632	0.16	-0.18	
35	1	317	0.08	-0.17	
	2	734	0.18	-0.30	
	3	216	0.05	-0.14	
	4*	**2730**	**0.68**	**0.42**	**0.31**
38	1	576	0.14	-0.22	
	2	702	0.18	-0.20	
	3*	**2065**	**0.52**	**0.35**	**0.22**
	4	657	0.16	-0.05	
39	1	578	0.14	-0.04	
	2	251	0.06	-0.16	
	3	2564	0.64	-0.10	

	4*	601	0.15	0.28	0.18
43	1	248	0.06	-0.17	
	2*	**2980**	**0.75**	**0.37**	**0.26**
	3	489	0.12	-0.24	
	4	281	0.07	-0.15	

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Table 42
Results of Classical Item Analysis: CH-ID1-S1

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
5	1	465	0.12	-0.20	
	2	396	0.10	-0.24	
	3*	**3001**	**0.75**	**0.40**	**0.22**
	4	135	0.03	-0.19	
7	1	193	0.05	-0.15	
	2	290	0.07	-0.18	
	3*	**3053**	**0.76**	**0.34**	**0.16**
	4	463	0.12	-0.21	
8	**1***	**3451**	**0.86**	**0.41**	**0.27**
	2	349	0.09	-0.28	
	3	81	0.02	-0.18	
	4	116	0.03	-0.20	
12	1	304	0.08	-0.27	
	2*	**3470**	**0.87**	**0.41**	**0.28**
	3	128	0.03	-0.21	
	4	97	0.02	-0.19	
18	**1***	**2925**	**0.73**	**0.51**	**0.34**
	2	356	0.09	-0.25	
	3	532	0.13	-0.32	
	4	180	0.05	-0.21	
19	1	142	0.04	-0.18	
	2	35	0.01	-0.08	
	3	303	0.08	-0.36	
	4*	**3520**	**0.88**	**0.42**	**0.29**
20	1	1388	0.35	-0.25	
	2	477	0.12	-0.26	
	3*	**1884**	**0.47**	**0.45**	**0.25**
	4	219	0.06	-0.10	
25	**1***	**3059**	**0.77**	**0.47**	**0.30**
	2	172	0.04	-0.19	
	3	426	0.11	-0.32	
	4	343	0.09	-0.22	
28	1	135	0.03	-0.19	
	2*	**3592**	**0.90**	**0.35**	**0.23**
	3	80	0.02	-0.19	
	4	193	0.05	-0.21	
30	1	504	0.13	-0.16	

	2*	**2846**	**0.71**	**0.30**	**0.11**
	3	433	0.11	-0.16	
	4	215	0.05	-0.16	
33	1	35	0.01	-0.12	
	2	445	0.11	-0.14	
	3	905	0.23	-0.18	
	4*	**2615**	**0.65**	**0.28**	**0.07**
36	1	222	0.06	-0.13	
	2	617	0.15	-0.32	
	3*	**3130**	**0.78**	**0.38**	**0.21**
	4	31	0.01	-0.11	
41	**1***	**2502**	**0.63**	**0.38**	**0.18**
	2	450	0.11	-0.12	
	3	194	0.05	-0.15	
	4	853	0.21	-0.28	
44	1	145	0.04	-0.18	
	2	400	0.10	-0.27	
	3	330	0.08	-0.29	
	4*	**3071**	**0.77**	**0.47**	**0.31**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Table 43
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: CH-NTW-S1

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
1	-2.49	0.76	-2.75	0.47	-2.18	0.26
2	-2.77	0.81	-2.90	0.50	-2.36	0.26
4	-2.08	1.43	-1.47	0.92	-1.15	0.21
6	-0.71	1.09	-0.6	1.35	0.25	0.36
9	0.40	0.72	0.47	1.05	1.04	0.26
13	-1.77	1.04	-1.53	0.75	-0.88	0.29
15	-1.84	1.24	-1.42	0.85	-0.99	0.23
16	-0.40	0.69	-0.47	0.46	0.04	0.16
17	0.07	1.00	0.06	0.88	0.50	0.18
21	-2.14	1.12	-1.76	0.76	-1.20	0.28
22	-0.99	1.07	-0.85	0.74	-0.45	0.18
26	-0.44	0.81	-0.46	0.82	0.38	0.29
27	-0.39	0.71	-0.45	0.61	0.32	0.24
29	-0.66	1.27	-0.51	1.11	0.00	0.24
31	-0.41	0.77	-0.45	0.80	0.49	0.31
32	-2.35	0.94	-2.19	0.58	-1.78	0.22
34	0.04	0.55	0.06	0.58	1.01	0.27
35	-1.03	0.85	-1.04	0.59	-0.47	0.21
38	-0.09	0.54	-0.13	0.85	1.08	0.36
39	2.30	0.66	2.85	0.69	2.42	0.07
43	-1.44	0.74	-1.62	0.47	-1.09	0.19

Table 44
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: CH-ID1-S1

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
5	-1.59	0.75	-1.65	0.62	-0.47	0.37
7	-1.69	0.52	-2.39	0.34	-1.41	0.26
8	-2.62	1.21	-1.90	0.81	-1.43	0.26
12	-2.67	1.21	-1.94	0.79	-1.51	0.24
18	-1.45	1.25	-1.04	0.89	-0.60	0.22
19	-2.83	1.42	-1.85	0.98	-1.38	0.28
20	0.17	0.85	0.16	1.16	0.73	0.25
25	-1.70	1.22	-1.24	0.92	-0.69	0.26
28	-3.08	1.09	-2.38	0.71	-1.93	0.25
30	-1.31	0.31	-2.98	0.21	-1.32	0.26
33	-0.92	0.20	-3.13	0.15	-0.57	0.26
36	-1.84	0.74	-1.92	0.52	-1.10	0.28
41	-0.74	0.50	-1.08	0.37	-0.05	0.24
44	-1.72	1.09	-1.35	0.69	-1.00	0.18

Table 45
Results of IRT Item Analysis for NRM: CH-NTW-S1

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
1	1	-0.71	-1.03	26	1	0.14	-0.26
	2	0.13	-1.17		**2***	**0.63**	**1.14**
	3*	**0.56**	**2.48**		3	-0.56	-0.6
	4	0.02	-0.28		4	-0.22	-0.28
2	1	-0.45	-1.59	27	1	-0.18	-0.32
	2*	**0.64**	**2.68**		**2***	**0.54**	**1.07**
	3	-0.20	-0.35		3	-0.17	-0.31
	4	0.01	-0.74		4	-0.19	-0.44
4	1	-0.39	-1.05	29	1	-0.34	-0.32
	2*	**1.12**	**2.55**		2	-0.18	-0.45
	3	-0.49	-1.45		3	-0.44	-0.55
	4	-0.25	-0.04		**4***	**0.96**	**1.32**
6	1	-0.18	-0.65	31	1	-0.03	0.39
	2	-0.64	-1.57		2	-0.26	-0.52
	3*	**0.92**	**1.61**		**3***	**0.64**	**1.23**
	4	-0.10	0.61		4	-0.35	-1.11
9	**1***	**0.52**	**0.68**	32	1	-0.57	-0.88
	2	-0.14	0.20		2	-0.01	-0.68
	3	-0.27	0.06		**3***	**0.71**	**2.39**
	4	-0.11	-0.94		4	-0.13	-0.83
13	**1***	**0.79**	**2.07**	34	**1***	**0.48**	**0.88**
	2	-0.19	-0.35		2	0.09	0.23
	3	-0.61	-0.96		3	-0.37	-0.82
	4	0.01	-0.76		4	-0.20	-0.29
15	**1***	**0.96**	**2.24**	35	1	-0.16	-0.66
	2	-0.35	-0.95		2	-0.28	0.12
	3	-0.41	-1.21		3	-0.19	-1.05
	4	-0.20	-0.08		**4***	**0.63**	**1.59**
16	1	-0.36	-0.76	38	1	-0.37	-0.47
	2	0.26	0.30		2	-0.24	-0.22
	3*	**0.66**	**1.21**		**3***	**0.42**	**0.90**
	4	-0.56	-0.75		4	0.20	-0.21
17	1	0.10	-0.42	39	1	0.03	-0.06
	2	-0.48	0.17		2	-0.73	-1.19
	3*	**0.71**	**0.84**		3	0.05	1.43
	4	-0.33	-0.59		**4***	**0.66**	**-0.17**
21	**1***	**0.97**	**2.52**	43	1	-0.23	-0.88
	2	-0.26	-0.99		**2***	**0.57**	**1.76**
	3	-0.72	-1.64		3	-0.22	-0.19
	4	0.01	0.11		4	-0.11	-0.69
22	1	-0.39	-0.42				

2	0.09	-0.50
3*	**0.82**	**1.53**
4	-0.51	-0.62

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 46
Results of IRT Item Analysis for NRM: CH-ID1-S1

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
5	1	0.07	-0.01	25	**1***	**0.91**	**2.01**
	2	-0.15	-0.26		2	-0.37	-1.24
	3*	**0.65**	**1.91**		3	-0.45	-0.40
	4	-0.57	-1.64		4	-0.09	-0.36
7	1	-0.14	-1.03	28	1	-0.18	-0.79
	2	-0.20	-0.65		**2***	**0.89**	**2.89**
	3*	**0.40**	**1.81**		3	-0.61	-1.74
	4	-0.07	-0.13		4	-0.10	-0.37
8	**1***	**1.03**	**2.82**	30	1	0.04	-0.15
	2	-0.05	0.18		**2***	**0.28**	**1.58**
	3	-0.53	-1.72		3	0.01	-0.31
	4	-0.45	-1.28		4	-0.33	-1.12
12	1	-0.07	-0.01	33	1	-0.91	-3.09
	2*	**0.99**	**2.76**		2	0.20	0.20
	3	-0.41	-1.18		3	0.28	0.92
	4	-0.52	-1.57		**4***	**0.43**	**1.98**
18	**1***	**0.95**	**1.89**	36	1	0.20	-0.21
	2	-0.22	-0.47		2	-0.11	0.67
	3	-0.26	-0.09		**3***	**0.72**	**2.48**
	4	-0.47	-1.34		4	-0.81	-2.94
19	1	-0.13	-0.6	41	**1***	**0.39**	**1.38**
	2	-0.13	-2.01		2	0.11	-0.33
	3	-0.66	-0.33		3	-0.34	-1.31
	4*	**0.92**	**2.93**		4	-0.16	0.25
20	1	-0.04	0.73	44	1	-0.23	-1.26
	2	-0.56	-0.58		2	-0.22	-0.24
	3*	**0.71**	**0.98**		3	-0.39	-0.56
	4	-0.11	-1.13		**4***	**0.85**	**2.06**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 47
Probabilities of Selecting Item Options in NRM: CH-NTW-S1

Item	Option	Probability of Selecting Item Option at θ level						
		-3	-2	-1	0	1	2	3
1	1	0.488	0.233	0.085	0.027	0.008	0.002	0.001
	2	0.034	0.038	0.032	0.023	0.016	0.011	0.007
	3*	**0.362**	**0.615**	**0.797**	**0.894**	**0.941**	**0.966**	**0.980**
	4	0.116	0.115	0.087	0.057	0.035	0.021	0.012
2	1	0.168	0.083	0.034	0.013	0.005	0.002	0.001
	2*	**0.458**	**0.667**	**0.823**	**0.913**	**0.959**	**0.981**	**0.991**
	3	0.275	0.173	0.092	0.044	0.020	0.009	0.004
	4	0.099	0.077	0.051	0.030	0.017	0.009	0.005
4	1	0.244	0.176	0.082	0.024	0.006	0.001	0.000
	2*	**0.096**	**0.314**	**0.662**	**0.892**	**0.972**	**0.993**	**0.998**
	3	0.221	0.144	0.061	0.016	0.004	0.001	0.000
	4	0.440	0.365	0.195	0.067	0.019	0.005	0.001
6	1	0.175	0.165	0.124	0.069	0.030	0.011	0.004
	2	0.277	0.165	0.078	0.027	0.007	0.002	0.000
	3*	**0.062**	**0.175**	**0.395**	**0.661**	**0.850**	**0.942**	**0.979**
	4	0.486	0.495	0.403	0.243	0.113	0.045	0.017
9	**1***	**0.080**	**0.151**	**0.266**	**0.425**	**0.599**	**0.751**	**0.858**
	2	0.357	0.350	0.319	0.263	0.192	0.124	0.073
	3	0.459	0.394	0.316	0.228	0.146	0.083	0.043
	4	0.104	0.105	0.099	0.084	0.063	0.042	0.026
13	**1***	**0.153**	**0.369**	**0.640**	**0.836**	**0.933**	**0.973**	**0.989**
	2	0.258	0.233	0.152	0.074	0.031	0.012	0.005
	3	0.494	0.294	0.125	0.040	0.011	0.003	0.001
	4	0.094	0.104	0.082	0.049	0.025	0.012	0.006
15	**1***	**0.122**	**0.327**	**0.629**	**0.854**	**0.952**	**0.986**	**0.996**
	2	0.255	0.185	0.096	0.035	0.011	0.003	0.001
	3	0.235	0.161	0.079	0.027	0.008	0.002	0.001
	4	0.388	0.327	0.197	0.084	0.029	0.010	0.003
16	1	0.276	0.234	0.157	0.083	0.037	0.015	0.006
	2	0.124	0.195	0.244	0.239	0.198	0.149	0.107
	3*	**0.093**	**0.218**	**0.406**	**0.594**	**0.734**	**0.826**	**0.884**
	4	0.508	0.353	0.194	0.084	0.031	0.010	0.003
17	1	0.067	0.102	0.135	0.139	0.111	0.072	0.043
	2	0.689	0.588	0.433	0.251	0.112	0.041	0.013
	3*	**0.038**	**0.106**	**0.258**	**0.491**	**0.717**	**0.861**	**0.934**
	4	0.206	0.204	0.174	0.118	0.061	0.026	0.010

21	**1***	**0.159**	**0.413**	**0.704**	**0.881**	**0.956**	**0.984**	**0.994**
	2	0.191	0.145	0.072	0.026	0.008	0.003	0.001
	3	0.395	0.189	0.060	0.014	0.003	0.001	0.000
	4	0.255	0.253	0.165	0.079	0.033	0.013	0.005
22	1	0.388	0.331	0.218	0.102	0.037	0.012	0.004
	2	0.085	0.117	0.124	0.094	0.056	0.029	0.014
	3*	**0.072**	**0.207**	**0.457**	**0.719**	**0.880**	**0.951**	**0.980**
	4	0.455	0.345	0.201	0.084	0.027	0.008	0.002
26	1	0.094	0.135	0.158	0.148	0.116	0.080	0.052
	2*	**0.088**	**0.205**	**0.393**	**0.601**	**0.765**	**0.867**	**0.925**
	3	0.547	0.389	0.227	0.105	0.041	0.014	0.005
	4	0.271	0.271	0.222	0.145	0.079	0.038	0.017
27	1	0.298	0.260	0.206	0.145	0.090	0.050	0.027
	2*	**0.138**	**0.247**	**0.403**	**0.581**	**0.740**	**0.854**	**0.923**
	3	0.292	0.257	0.206	0.146	0.092	0.052	0.028
	4	0.272	0.235	0.185	0.128	0.079	0.044	0.023
29	1	0.368	0.334	0.248	0.128	0.046	0.014	0.004
	2	0.200	0.213	0.186	0.112	0.048	0.017	0.006
	3	0.394	0.324	0.218	0.102	0.033	0.009	0.002
	4*	**0.038**	**0.128**	**0.349**	**0.659**	**0.873**	**0.960**	**0.988**
31	1	0.371	0.375	0.333	0.254	0.166	0.098	0.054
	2	0.298	0.239	0.169	0.102	0.053	0.025	0.011
	3*	**0.115**	**0.227**	**0.395**	**0.588**	**0.753**	**0.866**	**0.931**
	4	0.216	0.159	0.102	0.057	0.027	0.012	0.005
32	1	0.482	0.258	0.103	0.034	0.010	0.003	0.001
	2	0.110	0.103	0.072	0.041	0.022	0.011	0.005
	3*	**0.273**	**0.526**	**0.755**	**0.889**	**0.952**	**0.979**	**0.991**
	4	0.135	0.113	0.070	0.036	0.016	0.007	0.003
34	**1***	**0.135**	**0.230**	**0.356**	**0.496**	**0.629**	**0.739**	**0.822**
	2	0.227	0.262	0.274	0.259	0.222	0.177	0.133
	3	0.316	0.230	0.152	0.091	0.049	0.025	0.012
	4	0.322	0.278	0.218	0.154	0.099	0.059	0.033
35	1	0.174	0.155	0.118	0.075	0.041	0.020	0.010
	2	0.543	0.430	0.291	0.163	0.079	0.035	0.015
	3	0.129	0.112	0.082	0.051	0.027	0.013	0.006
	4*	**0.154**	**0.303**	**0.509**	**0.711**	**0.853**	**0.932**	**0.970**
38	1	0.405	0.311	0.215	0.133	0.074	0.039	0.019
	2	0.352	0.308	0.243	0.171	0.109	0.064	0.036
	3*	**0.149**	**0.252**	**0.384**	**0.524**	**0.646**	**0.740**	**0.807**
	4	0.095	0.129	0.158	0.173	0.171	0.157	0.137
39	1	0.118	0.143	0.153	0.150	0.136	0.113	0.086

	2	0.373	0.211	0.106	0.049	0.021	0.008	0.003
	3	0.493	0.610	0.667	0.667	0.615	0.522	0.404
	4*	**0.016**	**0.036**	**0.073**	**0.135**	**0.229**	**0.357**	**0.508**
43	1	0.198	0.148	0.097	0.055	0.028	0.014	0.006
	2*	**0.252**	**0.420**	**0.609**	**0.769**	**0.877**	**0.938**	**0.970**
	3	0.383	0.290	0.191	0.109	0.057	0.027	0.013
	4	0.167	0.141	0.104	0.066	0.038	0.021	0.011

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 48
Probabilities of Selecting Item Options in NRM: CH-ID1-S1

Item	Option	Probability of Selecting Item Option at θ level						
		-3	-2	-1	0	1	2	3
5	1	0.198	0.198	0.162	0.114	0.072	0.043	0.025
	2	0.299	0.239	0.158	0.089	0.045	0.022	0.010
	3*	**0.238**	**0.423**	**0.620**	**0.775**	**0.876**	**0.933**	**0.964**
	4	0.265	0.139	0.060	0.022	0.007	0.002	0.001
7	1	0.123	0.094	0.067	0.045	0.029	0.018	0.011
	2	0.215	0.156	0.105	0.066	0.040	0.023	0.013
	3*	**0.417**	**0.548**	**0.673**	**0.777**	**0.854**	**0.908**	**0.943**
	4	0.245	0.202	0.155	0.112	0.077	0.051	0.033
8	**1***	**0.186**	**0.459**	**0.750**	**0.910**	**0.971**	**0.991**	**0.997**
	2	0.339	0.284	0.158	0.065	0.024	0.008	0.003
	3	0.214	0.111	0.038	0.010	0.002	0.000	0.000
	4	0.261	0.147	0.055	0.015	0.004	0.001	0.000
12	1	0.300	0.247	0.137	0.057	0.021	0.007	0.003
	2*	**0.199**	**0.474**	**0.758**	**0.913**	**0.971**	**0.991**	**0.997**
	3	0.258	0.151	0.060	0.018	0.005	0.001	0.000
	4	0.243	0.128	0.045	0.012	0.003	0.001	0.000
18	**1***	**0.082**	**0.238**	**0.518**	**0.786**	**0.926**	**0.977**	**0.993**
	2	0.260	0.233	0.158	0.074	0.027	0.009	0.003
	3	0.428	0.369	0.240	0.109	0.038	0.012	0.004
	4	0.230	0.161	0.085	0.031	0.009	0.002	0.001
19	1	0.110	0.109	0.065	0.027	0.010	0.004	0.001
	2	0.027	0.027	0.016	0.007	0.002	0.001	0.000
	3	0.704	0.411	0.144	0.036	0.008	0.002	0.000
	4*	**0.160**	**0.454**	**0.775**	**0.930**	**0.980**	**0.994**	**0.998**
20	1	0.383	0.449	0.449	0.369	0.249	0.143	0.075
	2	0.492	0.342	0.204	0.100	0.040	0.014	0.004
	3*	**0.052**	**0.129**	**0.272**	**0.474**	**0.676**	**0.824**	**0.911**
	4	0.074	0.080	0.075	0.057	0.036	0.019	0.009
25	**1***	**0.100**	**0.281**	**0.574**	**0.818**	**0.936**	**0.979**	**0.993**
	2	0.181	0.141	0.080	0.032	0.010	0.003	0.001
	3	0.532	0.383	0.201	0.073	0.022	0.006	0.002
	4	0.188	0.194	0.146	0.076	0.032	0.012	0.005
28	1	0.192	0.127	0.060	0.023	0.008	0.003	0.001
	2*	**0.308**	**0.592**	**0.819**	**0.932**	**0.976**	**0.991**	**0.997**
	3	0.270	0.116	0.036	0.009	0.002	0.000	0.000
	4	0.230	0.165	0.085	0.036	0.014	0.005	0.002

30	1	0.172	0.162	0.146	0.127	0.108	0.090	0.074
	2*	**0.471**	**0.564**	**0.646**	**0.717**	**0.774**	**0.821**	**0.859**
	3	0.160	0.146	0.128	0.108	0.089	0.072	0.058
	4	0.197	0.128	0.080	0.048	0.028	0.016	0.009
33	1	0.157	0.050	0.015	0.004	0.001	0.000	0.000
	2	0.151	0.146	0.130	0.111	0.093	0.078	0.065
	3	0.244	0.256	0.246	0.228	0.208	0.188	0.169
	4*	**0.449**	**0.548**	**0.610**	**0.657**	**0.697**	**0.733**	**0.766**
36	1	0.087	0.089	0.076	0.055	0.036	0.023	0.014
	2	0.529	0.401	0.249	0.132	0.064	0.029	0.013
	3*	**0.268**	**0.466**	**0.662**	**0.809**	**0.899**	**0.948**	**0.973**
	4	0.117	0.044	0.014	0.004	0.001	0.000	0.000
41	**1***	**0.270**	**0.388**	**0.515**	**0.636**	**0.738**	**0.815**	**0.871**
	2	0.113	0.123	0.123	0.115	0.101	0.084	0.068
	3	0.164	0.113	0.073	0.043	0.024	0.013	0.007
	4	0.454	0.376	0.289	0.206	0.137	0.088	0.054
44	1	0.125	0.103	0.065	0.030	0.011	0.004	0.001
	2	0.335	0.281	0.177	0.083	0.032	0.012	0.004
	3	0.405	0.286	0.152	0.060	0.020	0.006	0.002
	4*	**0.135**	**0.329**	**0.606**	**0.827**	**0.937**	**0.978**	**0.993**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 49
Results of Classical Item Analysis: CH-NTW-S2

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
1	1	206	0.05	-0.21	
	2	88	0.02	-0.10	
	3*	**3463**	**0.87**	**0.28**	**0.19**
	4	241	0.06	-0.15	
2	1	88	0.02	-0.18	
	2*	**3598**	**0.90**	**0.29**	**0.21**
	3	212	0.05	-0.21	
	4	101	0.03	-0.09	
4	1	207	0.05	-0.21	
	2*	**3299**	**0.83**	**0.41**	**0.31**
	3	114	0.03	-0.17	
	4	377	0.09	-0.27	
6	1	270	0.07	-0.15	
	2	158	0.04	-0.18	
	3*	**2531**	**0.63**	**0.46**	**0.35**
	4	1040	0.26	-0.34	
9	**1***	**1656**	**0.41**	**0.41**	**0.29**
	2	1035	0.26	-0.19	
	3	909	0.23	-0.19	
	4	349	0.09	-0.12	
13	**1***	**3175**	**0.79**	**0.40**	**0.30**
	2	353	0.09	-0.22	
	3	255	0.06	-0.27	
	4	216	0.05	-0.14	
15	**1***	**3236**	**0.81**	**0.43**	**0.33**
	2	203	0.05	-0.22	
	3	145	0.04	-0.19	
	4	415	0.10	-0.27	
16	1	382	0.10	-0.21	
	2	829	0.21	-0.16	
	3*	**2357**	**0.59**	**0.41**	**0.28**
	4	432	0.11	-0.24	
17	1	465	0.12	-0.10	
	2	1097	0.27	-0.34	
	3*	**1922**	**0.48**	**0.46**	**0.35**
	4	509	0.13	-0.14	
21	**1***	**3352**	**0.84**	**0.38**	**0.29**

	2	153	0.04	-0.18	
	3	119	0.03	-0.21	
	4	376	0.09	-0.24	
22	1	523	0.13	-0.29	
	2	346	0.09	-0.16	
	3*	**2743**	**0.69**	**0.47**	**0.36**
	4	385	0.10	-0.26	
26	1	605	0.15	-0.09	
	2*	**2269**	**0.57**	**0.39**	**0.27**
	3	496	0.12	-0.28	
	4	628	0.16	-0.19	
27	1	597	0.15	-0.19	
	2*	**2268**	**0.57**	**0.40**	**0.28**
	3	585	0.15	-0.21	
	4	547	0.14	-0.16	
29	1	544	0.14	-0.25	
	2	475	0.12	-0.24	
	3	474	0.12	-0.26	
	4*	**2507**	**0.63**	**0.51**	**0.41**
31	1	945	0.24	-0.22	
	2	436	0.11	-0.22	
	3*	**2339**	**0.59**	**0.44**	**0.32**
	4	273	0.07	-0.21	
32	1	212	0.05	-0.24	
	2	158	0.04	-0.13	
	3*	**3449**	**0.86**	**0.33**	**0.24**
	4	179	0.05	-0.17	
34	**1***	**1987**	**0.50**	**0.32**	**0.19**
	2	980	0.25	-0.09	
	3	395	0.10	-0.19	
	4	636	0.16	-0.17	
35	1	292	0.07	-0.17	
	2	737	0.18	-0.31	
	3	219	0.06	-0.13	
	4*	**2751**	**0.69**	**0.42**	**0.31**
38	1	528	0.13	-0.21	
	2	734	0.18	-0.20	
	3*	**2013**	**0.50**	**0.35**	**0.22**
	4	723	0.18	-0.06	
39	1	586	0.15	-0.05	
	2	260	0.07	-0.16	
	3	2583	0.65	-0.10	

	4*	**566**	**0.14**	**0.31**	**0.22**
43	1	240	0.06	-0.15	
	2*	**2958**	**0.74**	**0.36**	**0.25**
	3	503	0.13	-0.24	
	4	298	0.07	-0.16	

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Table 50
Results of Classical Item Analysis: CH-ID1-S2

Item	Option	N	p-value	r_{pb}	r_{cpb} for correct option
5	1	514	0.13	-0.22	
	2	355	0.09	-0.20	
	3*	**2997**	**0.75**	**0.37**	**0.19**
	4	129	0.03	-0.16	
7	1	202	0.05	-0.13	
	2	289	0.07	-0.18	
	3*	**3028**	**0.76**	**0.35**	**0.17**
	4	480	0.12	-0.24	
8	**1***	**3355**	**0.84**	**0.43**	**0.29**
	2	433	0.11	-0.30	
	3	94	0.02	-0.20	
	4	115	0.03	-0.22	
12	1	325	0.08	-0.27	
	2*	**3449**	**0.86**	**0.40**	**0.26**
	3	124	0.03	-0.17	
	4	101	0.03	-0.21	
18	**1***	**2944**	**0.74**	**0.51**	**0.35**
	2	358	0.09	-0.26	
	3	519	0.13	-0.32	
	4	179	0.05	-0.21	
19	1	121	0.03	-0.13	
	2	47	0.01	-0.11	
	3	300	0.08	-0.33	
	4*	**3531**	**0.88**	**0.38**	**0.25**
20	1	1424	0.36	-0.25	
	2	435	0.11	-0.24	
	3*	**1876**	**0.47**	**0.45**	**0.25**
	4	248	0.06	-0.11	
25	**1***	**3071**	**0.77**	**0.46**	**0.29**
	2	173	0.04	-0.17	
	3	432	0.11	-0.34	
	4	323	0.08	-0.19	
28	1	166	0.04	-0.17	
	2*	**3572**	**0.89**	**0.32**	**0.19**
	3	71	0.02	-0.16	
	4	190	0.05	-0.20	
30	1	434	0.11	-0.18	

	2*	**2915**	**0.73**	**0.34**	**0.16**
	3	444	0.11	-0.18	
	4	205	0.05	-0.18	
33	1	46	0.01	-0.13	
	2	403	0.10	-0.14	
	3	878	0.22	-0.20	
	4*	**2672**	**0.67**	**0.29**	**0.09**
36	1	201	0.05	-0.13	
	2	557	0.14	-0.31	
	3*	**3211**	**0.80**	**0.36**	**0.20**
	4	31	0.01	-0.08	
41	**1***	**2416**	**0.60**	**0.38**	**0.18**
	2	517	0.13	-0.13	
	3	188	0.05	-0.17	
	4	878	0.22	-0.26	
44	1	139	0.04	-0.17	
	2	420	0.11	-0.26	
	3	312	0.08	-0.28	
	4*	**3072**	**0.77**	**0.45**	**0.29**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters; N refers to the number of examinees who selected item option; r_{pb} is point-biserial correlation; r_{cpb} is corrected point-biserial correlation

Item Response Theory Assumption Tests for the CH-NTW-S2 and the CH-ID1-S2

Table 51
Unidimensionality Tests: CH-NTW-S2 and CH-ID1-S2

Principal Component Factor Analysis Eigenvalues	CH-NTW-S2	CH-ID1-S2
1	3.33	2.28
2	1.07	1.03
3	1.07	1.00
4	1.04	
5	1.00	
DIMTEST		
Stout's T	0.30	-
p-value	0.38	-

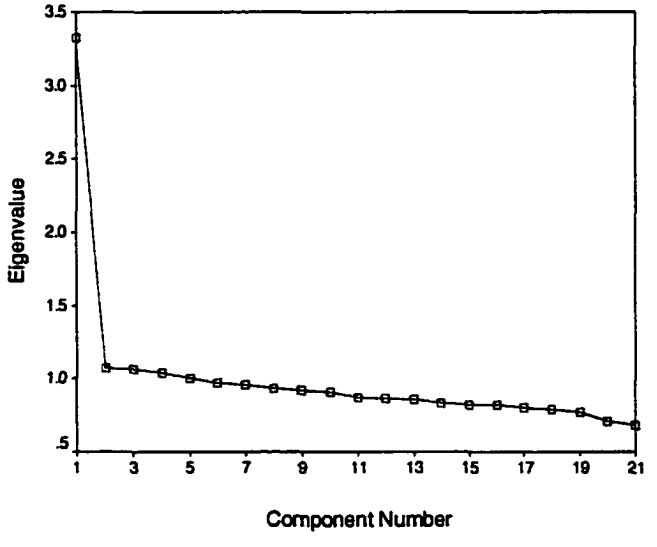


Figure 23. Scree Plot for the CH-NTW-S2

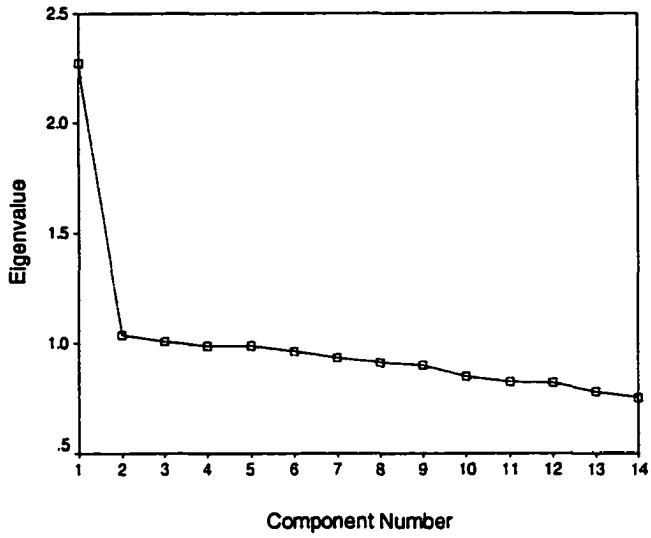


Figure 24. Scree Plot for the CH-ID1-S2

The results of the unidimensionality tests indicated that the assumption of the essential unidimensionality is met for both subtests.

The tenable assumption of unidimensionality indicates tenability of the assumption of local independence for the CH-NTW-S2 and CH-ID1-S2.

Table 52
Equal Discrimination Tests: CH-NTW-S2 and CH-ID1-S2

Point-Biserial Correlation	CH-NTW-S2	CH-ID1-S2
Minimum Value	0.28 (item 1)	0.29 (item 33)
Maximum Value	0.51 (item 29)	0.51 (item 18)

Large range of the point-biserial correlations for the CH-NTW-S2 and the CH-ID1-S2 indicates failure to meet the assumption of equal item discrimination. Therefore, the one-parameter model may not be appropriate for both subtests.

Table 53
Non-Speededness Tests: CH-NTW-S2 and CH-ID1-S2

Number of Examinees who did not Answer	CH-NTW-S2	CH-ID1-S2
Last Item	1	57
Last Two Items	0	0
Last Three Items	0	0

Given that, of 4000 examinees, one (less than 1.0%) did not complete the last item on the CH-NTW-S2 subtest, and 57 (1.4%) did not complete the last item on the CH-ID1-S2 subtest, it was concluded that the assumption of non-speededness was tenable for both subtests.

Table 54
Non-Guessing Tests: CH-NTW-S2 and CH-ID1-S2

		CH-NTW-S2		CH-ID1-S2	
		Item	p-value	Item	p-value
Most Difficult Items	1	39	0.14	20	0.47
	2	9	0.41	41	0.60
	3	17	0.48	33	0.67
Number of Low Achievers Correctly Answering the Most Difficult Items	1 Item	16		12	
	2 Items	1		11	
	3 Items	0		0	

Examination of the performance of low achieving examinees on the three most difficult items on both subtests revealed that the effect of guessing was minimal. Additional evidence supporting this conclusion is provided by the fact that only 10 (less than 1.0%) out of 4000 examinees scored at or below the chance level (score of 11).

Table 55
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: CH-NTW-S2

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
1	-2.43	0.70	-2.91	0.44	-2.14	0.29
2	-2.84	0.92	-2.71	0.55	-2.35	0.24
4	-2.04	1.22	-1.61	0.81	-1.15	0.25
6	-0.73	1.05	-0.64	1.10	0.18	0.33
9	0.46	0.80	0.49	1.09	0.98	0.23
13	-1.78	1.04	-1.57	0.71	-1.01	0.26
15	-1.90	1.25	-1.48	0.84	-1.11	0.20
16	-0.49	0.77	-0.54	0.58	0.08	0.20
17	0.10	0.99	0.09	1.11	0.63	0.24
21	-2.16	1.07	-1.85	0.72	-1.24	0.30
22	-1.04	1.16	-0.86	0.77	-0.53	0.15
26	-0.37	0.73	-0.42	0.81	0.59	0.33
27	-0.37	0.76	-0.41	0.66	0.37	0.26
29	-0.70	1.32	-0.54	1.35	0.08	0.29
31	-0.46	0.90	-0.46	0.94	0.38	0.31
32	-2.40	0.90	-2.34	0.56	-1.89	0.22
34	0.01	0.47	0.03	0.41	0.98	0.22
35	-1.06	0.89	-1.04	0.59	-0.60	0.17
38	-0.02	0.56	-0.03	0.58	0.99	0.28
39	2.36	0.83	2.46	0.91	2.10	0.07
43	-1.39	0.71	-1.62	0.46	-1.01	0.22

Table 56
Results of IRT Item Analysis for 1PL, 2PL, and 3PL: CH-ID1-S2

Item	One-Parameter	Two-Parameter		Three-Parameter		
	b_i	a_i	b_i	a_i	b_i	c_i
5	-1.59	0.63	-1.89	0.74	0.12	0.52
7	-1.65	0.56	-2.16	0.38	-1.19	0.27
8	-2.37	1.22	-1.71	0.79	-1.34	0.21
12	-2.62	1.19	-1.92	0.85	-1.30	0.31
18	-1.49	1.33	-1.02	1.09	-0.47	0.28
19	-2.88	1.20	-2.09	0.80	-1.57	0.29
20	0.18	0.82	0.17	1.00	0.75	0.24
25	-1.73	1.11	-1.33	0.75	-0.90	0.21
28	-3.02	0.86	-2.77	0.54	-2.24	0.25
30	-1.43	0.48	-2.17	0.34	-0.92	0.28
33	-1.02	0.25	-2.82	0.18	-0.76	0.25
36	-2.02	0.70	-2.20	0.48	-1.38	0.27
41	-0.62	0.52	-0.87	0.61	0.71	0.39
44	-1.73	1.09	-1.35	0.75	-0.87	0.23

Table 57
Results of IRT Item Analysis for NRM: CH-NTW-S2

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
1	1	-0.39	-0.70	26	1	0.17	-0.18
	2	-0.11	-1.38		**2***	**0.59**	**1.10**
	3*	**0.53**	**2.42**		3	-0.60	-0.69
	4	-0.02	-0.34		4	-0.16	-0.22
2	1	-0.72	-1.66	27	1	-0.21	-0.32
	2*	**0.73**	**2.82**		**2***	**0.56**	**1.06**
	3	-0.23	-0.34		3	-0.26	-0.37
	4	0.22	-0.83		4	-0.08	-0.37
4	1	-0.30	-0.79	29	1	-0.28	-0.33
	2*	**0.95**	**2.41**		2	-0.30	-0.48
	3	-0.46	-1.52		3	-0.41	-0.55
	4	-0.20	-0.11		**4***	**0.99**	**1.36**
6	1	-0.12	-0.70	31	1	0.06	0.40
	2	-0.63	-1.58		2	-0.31	-0.54
	3*	**0.88**	**1.64**		**3***	**0.79**	**1.30**
	4	-0.13	0.65		4	-0.54	-1.16
9	**1***	**0.62**	**0.63**	32	1	-0.54	-0.87
	2	-0.16	0.21		2	0.04	-0.78
	3	-0.20	0.07		**3***	**0.68**	**2.42**
	4	-0.26	-0.91		4	-0.17	-0.76
13	**1***	**0.82**	**2.10**	34	**1***	**0.42**	**0.89**
	2	-0.13	-0.31		2	0.13	0.21
	3	-0.71	-1.08		3	-0.37	-0.83
	4	0.02	-0.71		4	-0.19	-0.28
15	**1***	**0.98**	**2.32**	35	1	-0.23	-0.76
	2	-0.41	-0.97		2	-0.29	0.14
	3	-0.44	-1.33		3	-0.14	-1.00
	4	-0.14	-0.04		**4***	**0.65**	**1.61**
16	1	-0.38	-0.77	38	1	-0.38	-0.57
	2	0.14	0.22		2	-0.23	-0.18
	3*	**0.69**	**1.24**		**3***	**0.45**	**0.87**
	4	-0.45	-0.69		4	0.15	-0.12
17	1	-0.01	-0.52	39	1	-0.07	-0.04
	2	-0.47	0.18		2	-0.71	-1.10
	3*	**0.71**	**0.81**		3	0.00	1.45
	4	-0.22	-0.48		**4***	**0.78**	**-0.31**
21	**1***	**0.93**	**2.52**	43	1	-0.17	-0.89
	2	-0.21	-0.95		**2***	**0.53**	**1.73**
	3	-0.74	-1.69		3	-0.24	-0.18
	4	0.03	0.11		4	-0.13	-0.65
22	1	-0.37	-0.33				

2	-0.01	-0.54
3*	**0.88**	**1.59**
4	-0.50	-0.72

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 58
Results of Item Response Item Analysis for Nominal Response Model: CH-NTW-S2

Item	Option	a_{ik}	c_{ik}	Item	Option	a_{ik}	c_{ik}
5	1	0.01	0.05	25	**1***	**0.81**	**1.98**
	2	-0.14	-0.38		2	-0.25	-1.13
	3*	**0.54**	**1.87**		3	-0.59	-0.49
	4	-0.41	-1.55		4	0.02	-0.37
7	1	0.05	0.97	28	1	0.02	0.51
	2	-0.19	-0.66		**2***	**0.70**	**2.74**
	3*	**0.41**	**1.78**		3	-0.56	-1.79
	4	-0.18	-0.15		4	-0.11	-0.43
8	**1***	**1.12**	**2.77**	30	1	0.04	0.27
	2	0.06	0.42		**2***	**0.38**	**1.67**
	3	-0.62	-1.74		3	0.02	-0.23
	4	-0.56	-1.46		4	-0.37	-1.16
12	1	0.09	0.03	33	1	0.76	2.71
	2*	**1.01**	**2.75**		2	0.12	0.00
	3	-0.16	-0.99		3	0.23	0.79
	4	-0.75	-1.79		**4***	**0.42**	**1.91**
18	**1***	**1.01**	**1.93**	36	1	0.06	0.41
	2	-0.27	-0.48		2	-0.25	0.47
	3	-0.25	-0.10		**3***	**0.53**	**2.41**
	4	-0.49	-1.36		4	-0.34	-2.47
19	1	0.03	0.70	41	**1***	**0.45**	**1.34**
	2	-0.35	-1.93		2	0.15	-0.19
	3	-0.54	-0.25		3	-0.49	-1.44
	4*	**0.86**	**2.87**		4	-0.12	0.29
20	1	0.02	0.74	44	1	0.31	1.34
	2	-0.50	-0.66		2	-0.14	-0.12
	3*	**0.70**	**0.97**		3	-0.41	-0.62
	4	-0.19	-1.06		**4***	**0.86**	**2.09**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 59
Probabilities of Selecting Item Options in NRM: CH-NTW-S2

Item	Option	Probability of Selecting Item Option at θ level						
		-3	-2	-1	0	1	2	3
1	1	0.320	0.180	0.088	0.039	0.017	0.007	0.003
	2	0.070	0.052	0.034	0.020	0.011	0.006	0.003
	3*	**0.459**	**0.646**	**0.792**	**0.885**	**0.938**	**0.967**	**0.982**
	4	0.151	0.123	0.087	0.056	0.034	0.020	0.012
2	1	0.319	0.131	0.040	0.010	0.003	0.001	0.000
	2*	**0.363**	**0.638**	**0.832**	**0.926**	**0.967**	**0.984**	**0.992**
	3	0.274	0.185	0.092	0.039	0.016	0.006	0.002
	4	0.044	0.046	0.036	0.024	0.015	0.009	0.006
4	1	0.262	0.189	0.096	0.036	0.011	0.003	0.001
	2*	**0.151**	**0.380**	**0.677**	**0.877**	**0.960**	**0.988**	**0.996**
	3	0.204	0.125	0.054	0.017	0.005	0.001	0.000
	4	0.383	0.305	0.172	0.071	0.024	0.008	0.003
6	1	0.135	0.134	0.106	0.064	0.030	0.012	0.005
	2	0.259	0.154	0.073	0.026	0.007	0.002	0.000
	3*	**0.070**	**0.188**	**0.406**	**0.663**	**0.848**	**0.940**	**0.977**
	4	0.537	0.525	0.414	0.246	0.115	0.046	0.018
9	**1***	**0.057**	**0.120**	**0.235**	**0.409**	**0.609**	**0.778**	**0.887**
	2	0.390	0.376	0.337	0.269	0.183	0.107	0.056
	3	0.382	0.354	0.305	0.234	0.153	0.086	0.043
	4	0.172	0.150	0.122	0.088	0.054	0.029	0.014
13	**1***	**0.137**	**0.359**	**0.642**	**0.839**	**0.934**	**0.973**	**0.989**
	2	0.212	0.216	0.149	0.075	0.032	0.013	0.005
	3	0.560	0.318	0.123	0.035	0.008	0.002	0.000
	4	0.091	0.107	0.086	0.051	0.025	0.012	0.005
15	**1***	**0.125**	**0.341**	**0.647**	**0.864**	**0.956**	**0.986**	**0.996**
	2	0.303	0.205	0.097	0.032	0.009	0.002	0.001
	3	0.231	0.152	0.070	0.022	0.006	0.001	0.000
	4	0.341	0.302	0.187	0.082	0.029	0.010	0.003
16	1	0.312	0.245	0.158	0.082	0.035	0.014	0.005
	2	0.177	0.233	0.253	0.220	0.160	0.104	0.064
	3*	**0.094**	**0.215**	**0.405**	**0.610**	**0.769**	**0.869**	**0.926**
	4	0.417	0.306	0.184	0.089	0.036	0.013	0.004
17	1	0.088	0.117	0.137	0.128	0.092	0.054	0.029
	2	0.702	0.592	0.436	0.257	0.117	0.044	0.015
	3*	**0.038**	**0.105**	**0.252**	**0.483**	**0.714**	**0.865**	**0.941**
	4	0.172	0.186	0.176	0.133	0.078	0.037	0.016

21	**1***	**0.181**	**0.441**	**0.716**	**0.880**	**0.953**	**0.982**	**0.993**
	2	0.173	0.134	0.070	0.027	0.009	0.003	0.001
	3	0.404	0.185	0.056	0.013	0.003	0.001	0.000
	4	0.242	0.240	0.158	0.079	0.035	0.015	0.006
22	1	0.411	0.353	0.233	0.107	0.038	0.012	0.003
	2	0.113	0.139	0.132	0.087	0.044	0.019	0.008
	3*	**0.066**	**0.198**	**0.455**	**0.733**	**0.896**	**0.963**	**0.987**
	4	0.411	0.310	0.180	0.073	0.022	0.006	0.002
26	1	0.094	0.139	0.167	0.162	0.134	0.100	0.071
	2*	**0.096**	**0.215**	**0.394**	**0.584**	**0.735**	**0.837**	**0.900**
	3	0.568	0.388	0.216	0.098	0.037	0.013	0.004
	4	0.243	0.258	0.223	0.156	0.093	0.050	0.025
27	1	0.318	0.275	0.214	0.145	0.086	0.046	0.023
	2*	**0.126**	**0.234**	**0.394**	**0.578**	**0.742**	**0.857**	**0.925**
	3	0.352	0.289	0.214	0.138	0.078	0.040	0.019
	4	0.205	0.202	0.179	0.138	0.094	0.057	0.032
29	1	0.311	0.297	0.232	0.124	0.046	0.014	0.004
	2	0.284	0.266	0.204	0.106	0.039	0.012	0.003
	3	0.368	0.309	0.212	0.099	0.032	0.009	0.002
	4*	**0.037**	**0.127**	**0.353**	**0.671**	**0.883**	**0.966**	**0.990**
31	1	0.268	0.324	0.319	0.246	0.154	0.084	0.043
	2	0.318	0.265	0.180	0.096	0.042	0.016	0.006
	3*	**0.074**	**0.185**	**0.378**	**0.606**	**0.787**	**0.894**	**0.950**
	4	0.341	0.226	0.122	0.052	0.018	0.005	0.002
32	1	0.444	0.237	0.097	0.033	0.010	0.003	0.001
	2	0.085	0.081	0.059	0.036	0.020	0.011	0.006
	3*	**0.307**	**0.555**	**0.769**	**0.893**	**0.952**	**0.978**	**0.990**
	4	0.163	0.126	0.075	0.037	0.017	0.007	0.003
34	**1***	**0.165**	**0.261**	**0.378**	**0.501**	**0.614**	**0.708**	**0.782**
	2	0.200	0.237	0.256	0.254	0.233	0.201	0.166
	3	0.316	0.227	0.149	0.090	0.050	0.026	0.013
	4	0.319	0.275	0.216	0.155	0.104	0.065	0.039
35	1	0.188	0.159	0.114	0.067	0.033	0.015	0.006
	2	0.555	0.442	0.298	0.165	0.077	0.033	0.013
	3	0.113	0.105	0.082	0.053	0.029	0.014	0.007
	4*	**0.144**	**0.293**	**0.506**	**0.716**	**0.861**	**0.938**	**0.973**
38	1	0.383	0.291	0.199	0.121	0.066	0.034	0.016
	2	0.361	0.318	0.252	0.179	0.114	0.067	0.037
	3*	**0.134**	**0.233**	**0.366**	**0.511**	**0.643**	**0.747**	**0.822**
	4	0.122	0.158	0.183	0.190	0.177	0.152	0.124
39	1	0.142	0.160	0.163	0.153	0.129	0.096	0.061

	2	0.337	0.200	0.107	0.053	0.024	0.009	0.003
	3	0.512	0.618	0.676	0.678	0.616	0.492	0.336
	4*	**0.008**	**0.022**	**0.053**	**0.117**	**0.231**	**0.403**	**0.600**
43	1	0.158	0.127	0.089	0.055	0.031	0.017	0.009
	2*	**0.266**	**0.429**	**0.608**	**0.761**	**0.868**	**0.931**	**0.965**
	3	0.397	0.296	0.194	0.113	0.059	0.030	0.014
	4	0.178	0.149	0.109	0.070	0.042	0.023	0.012

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Table 60
Probabilities of Selecting Item Options in NRM: CH-ID1-S2

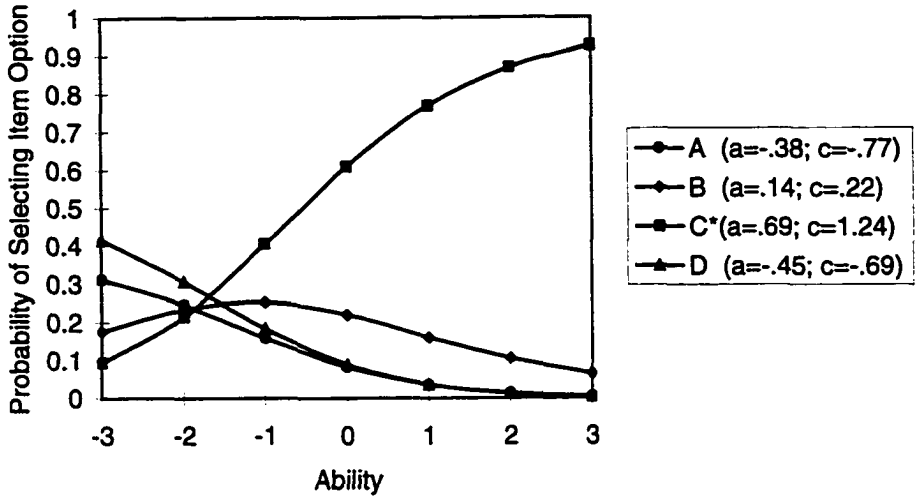
Item	Option	Probability of Selecting Item Option at θ level -3	-2	-1	0	1	2	3
5	1	0.251	0.223	0.176	0.125	0.082	0.052	0.032
	2	0.256	0.196	0.133	0.081	0.046	0.025	0.013
	3*	**0.315**	**0.477**	**0.638**	**0.769**	**0.861**	**0.919**	**0.954**
	4	0.178	0.104	0.054	0.025	0.011	0.004	0.002
7	1	0.355	0.342	0.310	0.265	0.216	0.168	0.127
	2	0.143	0.108	0.077	0.052	0.033	0.020	0.012
	3*	**0.271**	**0.374**	**0.486**	**0.596**	**0.695**	**0.777**	**0.840**
	4	0.231	0.177	0.127	0.087	0.056	0.035	0.021
8	**1***	**0.132**	**0.389**	**0.706**	**0.892**	**0.964**	**0.988**	**0.996**
	2	0.303	0.309	0.194	0.085	0.032	0.011	0.004
	3	0.269	0.139	0.044	0.010	0.002	0.000	0.000
	4	0.297	0.163	0.055	0.013	0.003	0.000	0.000
12	1	0.211	0.205	0.127	0.060	0.025	0.010	0.004
	2*	**0.203**	**0.495**	**0.767**	**0.909**	**0.966**	**0.987**	**0.995**
	3	0.161	0.122	0.059	0.022	0.007	0.002	0.001
	4	0.425	0.178	0.048	0.010	0.002	0.000	0.000
18	**1***	**0.070**	**0.220**	**0.512**	**0.795**	**0.934**	**0.981**	**0.995**
	2	0.292	0.256	0.165	0.071	0.023	0.007	0.002
	3	0.403	0.359	0.237	0.104	0.035	0.010	0.003
	4	0.235	0.165	0.085	0.030	0.008	0.002	0.000
19	1	0.245	0.248	0.178	0.098	0.047	0.021	0.009
	2	0.055	0.038	0.019	0.007	0.002	0.001	0.000
	3	0.523	0.300	0.122	0.038	0.010	0.003	0.001
	4*	**0.178**	**0.413**	**0.681**	**0.857**	**0.941**	**0.975**	**0.990**
20	1	0.378	0.440	0.443	0.374	0.266	0.164	0.092
	2	0.443	0.307	0.184	0.092	0.039	0.014	0.005
	3	**0.062**	**0.142**	**0.283**	**0.471**	**0.660**	**0.804**	**0.895**
	4	0.117	0.111	0.090	0.062	0.036	0.018	0.008
25	**1***	**0.115**	**0.310**	**0.595**	**0.817**	**0.926**	**0.971**	**0.988**
	2	0.123	0.115	0.077	0.036	0.014	0.005	0.002
	3	0.646	0.431	0.204	0.069	0.019	0.005	0.001
	4	0.117	0.144	0.125	0.078	0.040	0.019	0.009
28	1	0.298	0.237	0.158	0.093	0.051	0.027	0.014
	2*	**0.360**	**0.567**	**0.744**	**0.862**	**0.929**	**0.965**	**0.982**
	3	0.170	0.076	0.028	0.009	0.003	0.001	0.000
	4	0.172	0.120	0.070	0.036	0.017	0.008	0.004

30	1	0.255	0.236	0.206	0.169	0.134	0.103	0.078
	2*	**0.373**	**0.486**	**0.593**	**0.687**	**0.765**	**0.826**	**0.873**
	3	0.164	0.149	0.127	0.103	0.080	0.060	0.044
	4	0.209	0.128	0.074	0.041	0.021	0.011	0.005
33	1	0.149	0.056	0.020	0.007	0.002	0.001	0.000
	2	0.187	0.154	0.125	0.100	0.079	0.061	0.047
	3	0.297	0.272	0.246	0.220	0.194	0.169	0.145
	4*	**0.514**	**0.571**	**0.624**	**0.674**	**0.718**	**0.755**	**0.786**
36	1	0.176	0.167	0.140	0.105	0.073	0.049	0.032
	2	0.474	0.330	0.202	0.112	0.057	0.028	0.013
	3*	**0.318**	**0.482**	**0.646**	**0.777**	**0.867**	**0.922**	**0.955**
	4	0.033	0.021	0.012	0.006	0.003	0.001	0.001
41	**1***	**0.222**	**0.345**	**0.483**	**0.614**	**0.723**	**0.806**	**0.866**
	2	0.118	0.136	0.141	0.133	0.116	0.096	0.076
	3	0.231	0.140	0.077	0.038	0.018	0.008	0.003
	4	0.429	0.378	0.299	0.215	0.143	0.090	0.055
44	1	0.284	0.348	0.348	0.287	0.205	0.134	0.083
	2	0.254	0.199	0.127	0.067	0.030	0.013	0.005
	3	0.347	0.207	0.101	0.040	0.014	0.004	0.001
	4*	**0.115**	**0.246**	**0.425**	**0.607**	**0.751**	**0.849**	**0.911**

Note: * and values in **bold** print refer to the correct item option parameters; numbers in regular print refer to the incorrect alternative parameters.

Examples of Trace Lines for Non-Susceptible to Testwiseness Items

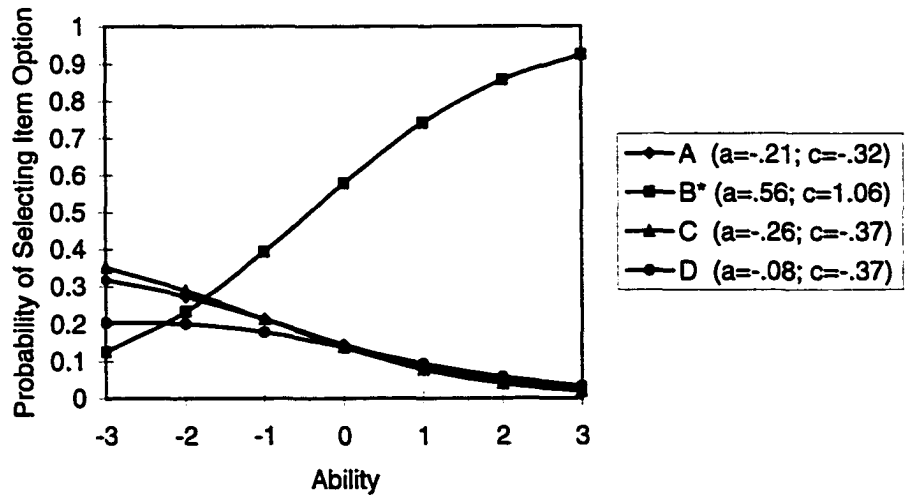
Item 16



Note: C* is the correct option; A, B, and D are item foils.

Figure 25. Non-Susceptible to Testwiseness Item 16, Chemistry 30, Sample 2

Item 27

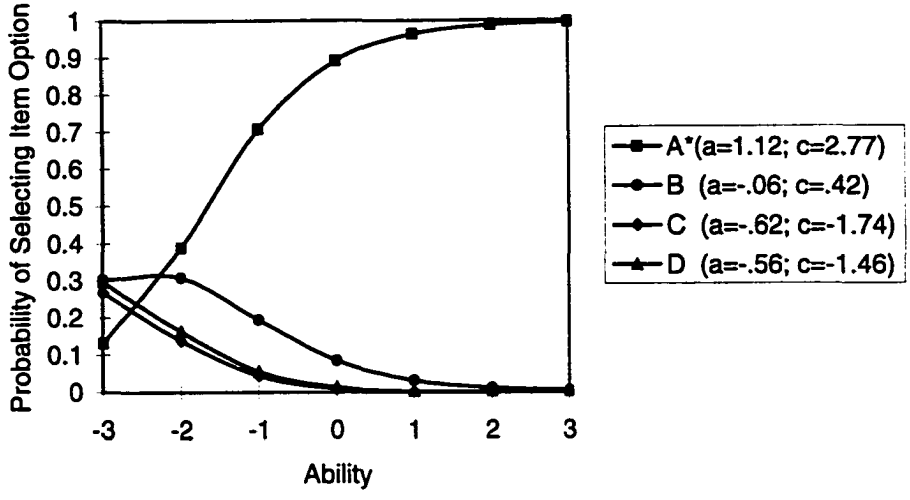


Note: B* is the correct option; A, C, and D are item foils.

Figure 26. Non-Susceptible to Testwiseness Item 27, Chemistry 30, Sample 2

Examples of Trace Lines for Items Susceptible to the ID1 Testwiseness Strategy

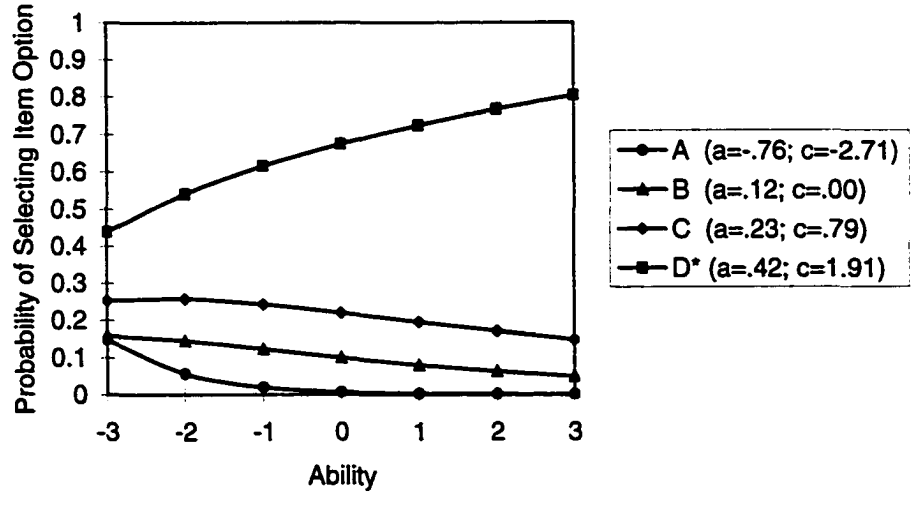
Item 8



Note: A* is the correct option; C and D are the absurd options; B is a properly functioning item foil.

Figure 27. Susceptible to ID1 Testwiseness Strategy Item 8, Chemistry 30, Sample 2

Item 33



Note: D* is the correct option; A is the absurd option; B and C are properly functioning item foils.

Figure 28. Susceptible to ID1 Testwiseness Strategy Item 33, Chemistry 30, Sample 2

Table 61
Mean Item Parameters for the Chemistry Examination

Model	Mean Parameter	CH-NTW-S1	CH-NTW-S2	CH-ID1-S1	CH-ID1-S2
Number Right	p_i	0.65	0.65	0.75	0.75
		(0.18)	(0.18)	(0.11)	(0.12)
One-Parameter Model	b_i	-0.91	-0.92	-1.71	-1.71
		(1.19)	(1.21)	(0.89)	(0.87)
Two-Parameter Model	a_i	0.90	0.91	0.88	0.85
		(0.24)	(0.22)	(0.39)	(0.34)
	b_i	-0.82	-0.86	-1.76	-1.72
		(1.24)	(1.20)	(0.85)	(0.80)
Three-Parameter Model	a_i	0.75	0.76	0.65	0.66
		(0.23)	(0.25)	(0.30)	(0.26)
	b_i	-0.24	-0.27	-0.91	-0.81
		(1.19)	(1.17)	(0.69)	(0.85)
	c_i	0.24	0.24	0.26	0.29
		(0.06)	(0.06)	(0.04)	(0.08)
Nominal Response Model	a_{ik}	0.71	0.72	0.72	0.70
		(0.19)	(0.17)	(0.25)	(0.23)
	c_{ik}	1.55	1.55	2.11	2.08
		(0.75)	(0.74)	(0.60)	(0.58)

Note: The means of item parameters are in the upper row of each cell; the standard deviations are in parentheses.